Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Pure and Applied Mathematics (38^{th}cycle)

Decision making in clinical development

By

Luca Rondano

******

Supervisor(s):
Prof. Mauro Gasparini, Supervisor
Stefano Vezzoli, Co-Supervisor

Doctoral Examination Committee:
Dr. Sofia Villar, Referee, University of Cambridge
Prof. Thomas Jaki, Referee, University of Regensburg

Politecnico di Torino
2026

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Luca Rondano
2026

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Project partners

Politecnico di Torino
Prof. Mauro Gasparini

Chiesi Farmaceutici S.p.A.
Stefano Vezzoli

Other significant contributions

Gaëlle Saint-Hilary, Pavel Mozgunov, Gianmarco Caruso.

Acknowledgements

I would first like to express my sincere gratitude to Chiesi Farmaceutici S.p.A. for their continuous support, collaboration, and financial contribution throughout my doctoral research. Their openness to innovation and research has been fundamental to the completion of this work.

My deepest thanks go to Stefano for his constant guidance, scientific insight, and encouragement over these years. Without him, this thesis would not exist. Our dinners in Parma have a special place in my heart.

I am very grateful to Gaëlle for her supervision and for her kindness.

I would like to thank Pavel for welcoming me into the MRC Biostatistics Unit, and for making that experience both enriching and enjoyable. I am equally thankful to Gianmarco for his crystal clear explanations at the office desk and for his company at the game table.

Finally, my gratitude goes to Mauro, sometimes distant, always caring. I thank him for giving me the opportunity to undertake this PhD and for looking after me from the heights of DISMA.

Abstract

Drug development requires a series of complex decisions, many of which depend on rigorous statistical methodology. This thesis, carried out in collaboration with Chiesi Farmaceutici S.p.A., contributes three new methodological advances addressing challenges in trial planning, randomisation, and bioequivalence assessment. The first contribution examines how prior knowledge can be incorporated into trial design through the *probability of success* (PoS). We introduce the notion of PoS *post interim* — defined conditionally on the Data Monitoring Committee's recommendation to continue the trial following an interim analysis — and characterise its formal relationship with the classical PoS. By analysing how efficacy and futility boundaries influence these probabilities, we provide insights that support boundary selection and enhance the interpretation of interim decisions. The second contribution focuses on response-adaptive randomisation in multi-arm phase II trials. Building on weighted information-theoretic principles, we develop a Bayesian randomisation method for normally distributed outcomes with unknown means and variances. The design favours treatments aligned with clinically desirable profiles while preserving sufficient allocation to control, and it balances exploration and exploitation through tunable parameters. Extensive simulations compare the method against established alternatives. The third contribution concerns power calculation in bioequivalence studies — an application that relies on the multivariate non-central t distribution. We clarify inconsistencies in the literature on the definition of the non-central t and and provide the correct expression for power. We then apply this result to improve blinded sample-size reassessment, incorporating (blinded) interim information on both variance and means. Together, these contributions offer theoretical and practical advances that support more efficient and informative drug development.

Contents

List of Figures **ix**

List of Tables **x**

Introduction **1**

1 Investigating the impact of Data Monitoring Committee recommendations on the probability of trial success **8**

1.1 Chapter introduction 8

1.2 PoS and PoS post interim 10

1.3 Special case for normally distributed data 14

1.4 Examples 16

1.4.1 Example 1: No early stop for efficacy 16

1.4.2 Example 2: O'Brien-Fleming efficacy boundary 19

1.4.3 Example 3: Pocock efficacy boundary 21

1.4.4 Further comments on the examples 23

1.4.5 Impact of the information fraction 25

1.5 Discussion 28

1.6 Proofs and supplementary material 29

1.6.1 Proof that PoS_{post} is increasing in θ_{eff} 29

1.6.2 Proof of (1.5) 32

1.6.3 Probabilities in (1.3) for normally distributed data 33

2 A Bayesian entropy-based response-adaptive design jointly targeting mean and variance in multi-arm trials with continuous outcomes **35**

2.1 Chapter introduction 35

2.2 Methodology . . . 38
2.2.1 Context-dependent information measure . . . 38
2.2.2 Background methodology . . . 40
2.2.3 Uniform-variance weight function . . . 42
2.2.4 Targeted-variance weight function . . . 44
2.2.5 Design based on the context-dependent measure . . . 47
2.3 Procedure for best treatment arm's identification . . . 49
2.3.1 Definition of best treatment arm and hypothesis testing procedure 49
2.3.2 Control of the FWER . . . 51
2.4 Robust strategies for selecting κ and ω . . . 51
2.4.1 Selection strategy . . . 51
2.4.2 Objective function . . . 52
2.5 Simulation study of a multi-arm trial . . . 53
2.5.1 Study setting . . . 53
2.5.2 Competing designs . . . 54
2.5.3 Calibration of η, κ and ω . . . 56
2.5.4 Average performance in randomly generated scenarios . . . 57
2.5.5 Designs optimality . . . 60
2.5.6 Characterisation of the scenarios based on the optimal design for patient benefit . . . 62
2.5.7 Individual scenarios performance . . . 63
2.6 Discussion . . . 66
2.7 Proofs and supplementary material . . . 67
2.7.1 Proof of Theorem 2.1 . . . 67
2.7.2 Proof of Theorem 2.2 . . . 68
2.7.3 Proof of Theorem 2.3 . . . 69
2.7.4 Alternative definition of best treatment arm . . . 72
2.7.5 Selection of the null scenarios . . . 73
2.7.6 Randomisation with UWE . . . 75
2.7.7 Randomisation with RGI . . . 75

3 Considerations on the multivariate non-central t distributions with applications to a sample size reassessment design for bioequivalence trials. 77
3.1 Chapter introduction . . . 77
3.2 The two versions of the non-central t distribution . . . 80
3.2.1 In one dimension . . . 80

3.2.2 In more than one dimension 81
3.3 Applications to bioequivalence trials 82
3.3.1 Bioequivalence trials 82
3.3.2 Power and its approximation for $n_T = n_R$ 86
3.3.3 Linear dependence of $(T_1^{sup}, T_2^{sup}, T_1^{eq}, T_2^{eq})$ 87
3.3.4 Sample size reassessment based on the interim estimation of mean and variance 88
3.4 Simulation studies 90
3.4.1 Study setting 90
3.4.2 Results of the simulation study 91
3.5 Discussion 92

Conclusions **95**

References **97**

List of Figures

1.1 PoS and other relevant probabilities. Example with no efficacy stopping rule. 18
1.2 PoS and other relevant probabilities. Example with O'Brien-Fleming type of boundary. 20
1.3 PoS and other relevant probabilities. Example with Pocock type of boundary. 22
1.4 PoS post interim for varying efficacy and futility boundaries. 24
1.5 PoS and PoS_{post} for varying information fractions. 26
1.6 PoS and other relevant probabilities for small and large information fractions. Example with O'Brien-Fleming type of boundary. 27

2.1 Uniform-variance and targeted-variance information gains. 44
2.2 Cut-off values η_S that control the FWER$= 0.05$. 56
2.3 Evaluation of the operating characteristics for varying configurations of the tuning parameters κ and ω. 58
2.4 FWER of the considered designs. 58
2.5 Operating characteristics of the considered designs. Results of the simulation study. 59
2.6 Operating characteristics of the considered designs. Results of the simulation study when using the alternative definition of best arm. 74

3.1 Comparison of univariate $t_\nu^{(1)}(\mu)$ and $t_\nu^{(2)}(\mu)$. 82

List of Tables

1.1 PoS and PoS_{post} tradeoff. 23
1.2 PoS and other relevant probabilities for small and large information fractions. Example with O'Brien-Fleming type of boundary. 27

2.1 Optimal designs for patient benefit, percentage of correct selection and power. 61
2.2 Characterisation of optimal designs for patient benefit. 63
2.3 Six additional hand-picked scenarios. 65

3.1 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 0.9$ 93
3.2 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.0$ 93
3.3 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.1$ 94
3.4 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 0.8$ 94
3.5 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.25$ 94

Introduction

Drug development is a long process that involves numerous complex and high-stakes decisions [Bacchieri and Cioppa, 2007; Piantadosi, 2017]. Modern clinical research increasingly relies on sophisticated statistical methodologies to ensure rigorous inference, efficient trial conduct, and ethically sound decision-making. As clinical development programs grow more complex (often involving interim analyses, multiple treatment arms, adaptive allocation strategies, and other modern design features), the statistical tools used by investigators are becoming more refined and specific to the individual trial challenges. Strategic choices include whether to continue, terminate, or expand the development of a candidate drug [Jennison and Turnbull, 1999; Grieve, 2022], which randomisation approach to use [Rosenberger and Lachin, 2016, Hu and Rosenberger, 2006], and how to assess the required sample size as precisely as possible [Chow et al., 2017]. New data and evidence generated throughout development can be integrated with existing knowledge to support these decisions. Across different trial settings, subtle modelling of the design choices can have meaningful consequences for trial efficiency, interpretability and operating characteristics. This thesis brings together three methodological contributions addressing these challenges from three different perspectives.

1. The first contribution concerns the incorporation of prior information into clinical trial design through the concept of probability of success (PoS), also known as assurance. Under the standard frequentist approach, "success" is typically defined as rejecting a null hypothesis (usually stating that the experimental treatment is ineffective). The likelihood of success is evaluated at a fixed value of the parameter of interest (usually at a value corresponding to a clinically relevant effect) by computing the statistical power of a test. This approach, however, is incapable of incorporating previous or subjective information into the study, when this is available. For example, a prior distribution for the treatment effect may be obtained from historical data

[Ibrahim et al., 2015] or by elicitation from experts [Crisp et al., 2018]. The investigators may want to take advantage of such a prior to evaluate the average power over all possible values of the treatment effect in the parametric space, which is, by definition, the PoS [Chuang-Stein, 2006; Rufibach, Burger, and Abt, 2016; Spiegelhalter et al., 1986]. In a hybrid Bayesian/frequentist setting, prior beliefs can also be incorporated into the PoS exclusively at the design stage, without affecting the frequentist analysis of the final data [Grieve, 2022].

The idea of PoS/assurance traces back to O'Hagan et al., 2005, and has been further developed by many authors in subsequent years. Some early works on the properties of PoS include Gasparini et al., 2013 and Carrol, 2013, while, more recently, Grieve, 2023 gives a detailed mathematical break-down on the evaluation of PoS in group sequential designs. Temple and Robertson, 2021 and Rufibach, Jordan, and Abt, 2016 show how the PoS can be updated to incorporate new evidence, as more information is collected on the treatment effects, for example from concurrent trials or from an interim analysis. Our attention goes in particular to the information disclosed by the Data Monitoring Committee (DMC) in an interim analysis, usually consisting in a simple statement that advises to continue or to stop the trial (either for efficacy or futility), following the pre-specified rules defined in the protocol. The first part of this work investigates the formal relationship between the initial PoS of a trial and the conditional probability of success given the DMC recommendation to continue the trial after the interim, which we refer to as "PoS post interim". By analysing how interim efficacy and futility boundaries shape these quantities, we provide insights that can guide their choice, and improve the interpretation of DMC recommendations.

2. The second contribution focuses on response-adaptive randomisation (RAR) in multi-arm trials. Randomising patients to treatments is a defining element of a well-conducted clinical study. It ensures comparability of treatment groups, mitigates bias, and provides a foundation for valid statistical inference [Rosenberger and Lachin, 2016].

   In clinical trials, fixed and equal randomisation schemes, in which allocation probabilities remain unchanged throughout the study, are still the most commonly used procedures due to their simplicity and favourable statistical properties. Fixed randomisation protects effectively against confirmation and indication biases and, on

average, balances all prognostic factors across treatment groups [Piantadosi, 2017, Chapter 13]. Moreover, key operating characteristics, such as power and type I error, are usually straightforward to compute analytically.

However, in trial with heterogeneous variances or with more than two arms, allocating the same number of patients to each arm can be inefficient, sometimes performing worse than alternative allocation strategies on key operating characteristics [Lu and Yeh-Fong, 2022; Woods et al., 1998]. More generally, fixed randomisation strategies may expose a substantial proportion of participants to less effective or less safe treatments, as they do not allow the allocation procedure to shift dynamically in favour of superior treatments. A further challenge arises when the number of available patients is small relative to the number of treatment options (or doses). In such settings, pre-specifying appropriate sample sizes for each group is difficult, and dividing patients evenly across all arms can lead to insufficient power. More broadly, the rigidity of fixed designs may limit their efficiency or the benefit of trial participants [Freedman, 1987] when true treatment effects differ across arms.

RAR attempts to address these limitations by sequentially updating allocation probabilities based on accrued outcomes. These probabilities can be adapted, for example, to favour treatments with more desirable performance or to optimise specific operating characteristics, such as the proportion of patients receiving the most effective treatment or the overall statistical power [Rosenberger et al., 2001; Tymofyeyev et al., 2007; Zhang and Rosenberger, 2006]. The flexibility of RAR, however, also comes with some downsides. It requires periodic unblinding of the data, which may raise concerns about trial integrity and complicate logistics. Statistically, the assumption that samplings are independent generally no longer holds, requiring the use of more nuanced testing procedures [Smith and Villar, 2018]. In addition, the complexity of RAR designs often necessitates extensive simulations to evaluate operating characteristics, as analytical expressions for quantities such as power and type I error are typically unavailable. All of these considerations were thoroughly addressed in Hu and Rosenberger, 2006. More recently, Robertson et al., 2023 provides a comprehensive overview of the arguments for and against the use of RAR.

Although RAR has been extensively studied in the biostatistical literature since the seminal work by Thompson, 1933, its practical uptake in clinical trials remains

limited. The first real implementation — the (in)famous ECMO trial [Bartlett et al., 1985] — generated substantial controversy due to concerns about its statistical methodology. Since then, a number of statistically rigorous trials have incorporated some form of RAR across various research areas, with the most frequent applications in oncology and multi-arm phase II studies [Wilson et al., 2025].

RAR in multi-arm phase II studies is precisely the focus of the second chapter of this work. The main challenge in this setting is balancing the gathering of information, which requires spreading the allocations across all arms, with the goal of giving the best possible treatment to as many patients as possible, which means maximising the number of allocations to the most effective one.

The challenge of balancing these competing goals is typically described as the "exploration vs exploitation" (or "learning vs earning") tradeoff [Azriel et al., 2011]. Multi-Arm Bandit (MAB) designs represent a wide class of (not necessarily randomised) response-adaptive approaches specifically developed to address this problem (e.g., Villar et al., 2015; Aziz et al., 2021 for categorical responses; Smith and Villar, 2018; Williamson and Villar, 2020 for continuous responses). The theory of weighted (or context-dependent) information measures [Belis and Guiasu, 1968; Kelbert and Mozgunov, 2015; Suhov et al., 2016] offers an alternative to MAB methods when investigators are interested not only in identifying the best treatment, but in selecting a treatment that meets specific clinical profiles. By modelling the treatment effect with a Bayesian prior, this class of approaches derives a decision-making framework based on the difference between the Shannon differential entropy and the weighted Shannon differential entropy of the posterior distribution of the treatment effects. This quantity, called *information gain*, quantifies the value of learning about an arm when estimates lying in certain regions of the parameter space are of particular clinical interest. This is achieved by specifying an arbitrary parametric weight function that assigns higher importance to clinically relevant areas. Recent examples of context-dependent designs are Kasianova et al., 2021, Mozgunov and Jaki, 2020a and Kasianova et al., 2023 for binary responses, Mozgunov and Jaki, 2020b for multinomial responses and Caruso and Mozgunov, 2024 for continuous responses

Building on these developments, we propose a Bayesian RAR design based on a weighted information gain, tailored for normally distributed endpoints in which both the mean and variance of the treatment arms are unknown. The design favours

treatments whose characteristics align with pre-specified clinical targets for the mean and variance while protecting the allocation to the control arm. Through the use of an appropriate bivariate weight function and carefully calibrated tuning parameters, the method balances exploration and exploitation and aims to improve both patient benefit and estimation precision. This randomisation method broadens the class of RAR designs available for multi-arm phase II trials, particularly when investigators are interested not only in identifying the single best treatment but in selecting treatments that meet specific clinical profiles.

3. The third contribution regards the exact formula for the power of bioequivalence (BE) trials. In conventional comparative clinical trials, statistical inference focuses on detecting differences between treatments. However, failing to reject the hypothesis of no difference does not establish equivalence, and a statistically detectable difference may be so small as to be clinically irrelevant [Blackwelder, 1982; Blackwelder and Chang, 1984]. Absolute equivalence can never be proven; rather, one can only assert with high probability that the true difference falls inside a clinically acceptable interval.

   BE trials aim at establishing equivalence at the level of exposure, when it is believed that it serves as a good surrogate for equivalence in efficacy and safety [Senn, 2007, Chapter 22]. Their goal is to demonstrate that a test formulation and a reference formulation yield equivalent bioavailabilities (i.e., concentrations) of the active ingredient in the body, hence the prefix "bio". BE studies rest on the principle that comparable systemic exposure implies comparable therapeutic response, thereby reducing or eliminating the need for full-scale clinical development of the new product. The justification is pharmacological: once absorbed, a drug follows well-characterised pharmacokinetic pathways governing its time-concentration profile; from there, effect-site dynamics and pharmacodynamic processes produce a clinical response which is based on the concentration of the drug in the body.

   The standard approach for assessing BE is the two one-sided test procedure, that was first described in Schuirmann, 1987, consisting of two (correlated) tests of non-inferiority. When the response follows a normal distribution, equivalence testing gives rise to multiple correlated statistics following non-central t distributions [Phillips, 1990; Hauschke et al., 1999; Chang et al., 2014]. Yang and Sun, 2019 gives an exact approach to compute power and sample size for an overall BE assessment

involving two superiority tests (test drug vs placebo and reference drug vs placebo) and one equivalence test (test drug vs reference drug), using a multivariate non-central t distribution. Proving the superiority of both treatments to placebo guarantees that the sample size of the study is large enough to detect effect differences, validating an eventual equivalence verdict on the test and reference [ICH, 2000]. Building on the formula for power mentioned above, Zhu and Sun, 2019 presents four designs for blinded and unblinded sample-size reassessment, while Hinds and Sun, 2025 presents an additional unblinded approach that introduces the re-estimation of the mean during the interim analysis. However, there are two coexisting definitions of non-central t in the statistical literature and the power formula given in Yang and Sun, 2019 is based on the incorrect one. An explanation for why the error may have gone unnoticed in three consecutive papers is that the numerical differences that come from the two definitions are almost unnoticeable in the specific context of their research. Nevertheless, we provide the correct expression and clarify this confusion. We then proceed to give some rudimentary results on the multivariate t distribution applied to the power calculations of a BE trial.

In conclusion, we apply the correct power formula in a simple design for blinded sample size reassessment and we introduce an original modification of it. Sample size reassessment typically relies on an interim estimation of the variance, enabling investigators to recalculate the required sample size mid-trial to ensure adequate power. We introduce a (blinded) design that, additionally, factors in the estimates of the treatment means. Including mean estimation can offer a meaningful improvement because, when equivalence is assessed through the mean ratio, even small deviations in the treatment effects can lead to substantial changes in statistical power and therefore in the required sample size [Hauschke et al., 1999].

Collectively, the above contributions provide a combination of methodological developments, theoretical insights, and practical considerations that ultimately support more efficient and informative drug development. Summarising, the outline of the thesis is as follows.

In Chapter 1, we begin by introducing the general definitions of probability of success and probability of success post interim, and by outlining how these quantities relate to one another, while also examining how different choices of futility and efficacy boundaries influence these probabilities. Next, we show their explicit expressions in a two-arm trial

where the treatment effect is defined as the mean treatment difference and is assumed to follow a normal distribution. The chapter concludes with three illustrative examples based on a fictional study, which offer practical guidance on selecting the interim boundaries. The impact of the information fraction of the interim analysis is also investigated.

In Chapter 2, we derive an information-theoretic criterion for a family of weight functions centred around two target parameters γ and ξ, corresponding to desirable values for the treatment mean and variance. We then introduce a decision rule for identifying the best arm at the end of the study and we describe an hypothesis-testing procedure for establishing the superiority of the selected arm over control. The exploration-exploitation tradeoff inherent in the information-theoretic approach can be controlled through the appropriate choice of the two tuning parameters, κ and ω. We therefore propose a simulation-based strategy for selecting these parameters that identifies optimal values for a desired operating characteristic. The chapter closes with a comparative evaluation of our method against several widely used alternatives — including fixed randomisation and Thompson sampling — using results from an extensive simulation study.

In Chapter 3, we give an overview of the two coexisting definitions of the non-central t distribution. We then apply the multivariate non-central t to the context of bioequivalence trials and present the correct expression for computing power in this setting. Furthermore, we present a refinement of a simple design for blinded sample size reassessment that improves the accuracy of the re-estimation when the test and the reference formulations are expected to produce very similar effect.

Chapter 1

Investigating the impact of Data Monitoring Committee recommendations on the probability of trial success

1.1 Chapter introduction[1]

In a clinical study analysed according to standard frequentist principles, the sample size and the threshold for declaring statistical significance can be determined as a function of the type I and type II error, the treatment effect size of interest (e.g., expressed as mean difference, hazard ratio, etc.) and other relevant parameters (e.g., standard deviation of the response variable, hazard in the control group, etc.). A prior distribution of the treatment effect is not strictly necessary to design a clinical trial, but when such prior information is available, the sponsor may want to use it to make better informed decisions on how to conduct the study. In a frequentist trial, this prior distribution may play a role in the design, but not in the analysis of the trial. In particular, this prior distribution can be used to determine the *probability of success* (PoS) of a study, also known as *assurance*. The

[1]This chapter is based on the article:
Rondano, L., Saint-Hilary, G., Gasparini, M., Vezzoli, S. "Investigating the impact of Data Monitoring Committee recommendations on the probability of trial success." *Journal of Biopharmaceutical Statistics*, online ahead of print.

concept of PoS/*assurance* traces back to O'Hagan et al., 2005, and has been studied by many other authors in the following years. Chuang-Stein, 2006 introduces the very similar concept of *average power*, and some early works on the properties of PoS, to name a few, are Gasparini et al., 2013 and Carrol, 2013. Rufibach, Burger, and Abt, 2016 gives useful reccomendations on which prior to use when computing PoS. Crisp et al., 2018 reports the practical experience of GSK with the use of PoS. All these concepts surrounding *PoS* are summarised in Chuang-Stein and Kirby, 2017 and a review of the different terminologies can also be found in Kunzmann et al., 2021.

As the prior distribution plays no role in the frequentist analysis of the study, this approach is defined as hybrid Bayesian/frequentist. A thorough review of hybrid Bayesian/frequentist designs can be found in Grieve, 2022. If an interim analysis is planned, the PoS calculated at the design stage can be updated to incorporate the information disclosed by the Data Monitoring Committee (DMC), usually consisting in a simple statement that advises to continue or to stop the trial (either for efficacy or futility), following the pre-specified rules defined in the protocol.

Our current research is driven by the observation that, in several instances, the study team exhibited either excessive optimism or pessimism regarding the final results following the DMC recommendations to continue the trial. In practice, the impact of the DMC recommendation on the PoS for the trial was often found to be relatively minor, but further research was needed to understand their relationship.

In this chapter we focus on the relationship between PoS and its updated version after the DMC recommendation to continue the trial, extending the work of Temple and Robertson, 2021 and Rufibach, Jordan, and Abt, 2016. Temple and Robertson, 2021 extends the use of PoS of a single study to the conditional PoS of a subsequent study given a success in a previous study with the same endpoint (for example a Phase II study with the same treatment). Rufibach, Jordan, and Abt, 2016 discusses statistical approaches that can be used to sequentially update PoS of a Phase III study, using either external (e.g., coming from another concurrent study) or internal (e.g., coming from the results of an interim analysis) information, in a time-to-event setting. Furthermore, a very recent paper by Grieve, 2023 develops the use of PoS in group sequential designs for an arbitrary number of interim analyses.

Compared to these previous work, our research explores more specifically the relationship between PoS and its updated version, hereafter defined as PoS *post interim*. Moreover, we illustrate in detail the influence of interim boundaries on these probabilities. We believe that the assessment of PoS and PoS *post interim* helps inform the choice of the boundaries for efficacy and for futility.

In Section 1.2, we define PoS and PoS *post interim* in the general case and show their relationship, together with other operating characteristics of interest. In addition, we demonstrate the impact of the choice of the boundaries for futility and efficacy on these probabilites.

In Section 1.3, we show how to compute PoS and PoS *post interim* in the case of a two-arm trial where the treatment effect is the mean difference between the two groups, assumed to be normally distributed. Under these conditions, the computations become straightforward.

In Section 1.4, we provide three examples of a fictional study. In these examples we explore the impact of different futility and efficacy boundaries on PoS and PoS *post interim*, as well as providing some recommendations on selecting such boundaries. Furthermore, we analyse the impact of changing the information fraction in the same three examples.

Conclusive remarks are presented in Section 1.5.

1.2 PoS and PoS post interim

Consider a double-blind clinical trial with one interim analysis. Denote the estimators of the true treatment effect θ at the interim and final analysis as $\hat{\theta}_{int}$ and $\hat{\theta}_{fin}$, respectively. The threshold to reach statistical significance at the final analysis is denoted by θ_{suc}. The trial may also be stopped early at the interim analysis, for futility or efficacy, if $\hat{\theta}_{int}$ is lower than a futility boundary θ_{fut} or greater than an efficacy boundary θ_{eff}. We define the (overall) success of the trial as reaching statistical significance at the interim analysis or at the final analysis, i.e., $\hat{\theta}_{int} > \theta_{eff}$ or $\hat{\theta}_{fin} > \theta_{suc}$. For the sake of simplicity, nuisance parameters are assumed known.

Let us assume that a prior distribution for the treatment effect θ is available. For example, it was obtained from historical data or by elicitation from experts Crisp et al., 2018. Let $q_0(\theta)$ be the probability density function of the prior distribution of θ. Given this prior

distribution, we consider the classical definition of *probability of success* (PoS), with succes defined as reaching statistical significance at the interim or final analysis:

$$\begin{aligned}\text{PoS} &= P(\textit{no early stop and success at final analysis}) + P(\textit{early stop for efficacy}) \\ &= \int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}, \hat{\theta}_{fin} > \theta_{suc}|\theta)\, q_0(\theta) d\theta + \int P(\hat{\theta}_{int} > \theta_{eff}|\theta)\, q_0(\theta) d\theta.\end{aligned}$$

Since the trial is double-blind, the DMC recommendation to continue the study at the interim analysis only informs the sponsor that $\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}$, but the exact value of the estimated treatment effect obtained at the interim analysis remains unknown. How does this information affect the PoS? Let us define $q_1(\theta)$ as the posterior density of θ when the trial is continued after the interim:

$$q_1(\theta) = q_1(\theta|\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}) = \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta)\, q_0(\theta)}{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta')\, q_0(\theta') d\theta'}.$$

Then, we define the *probability of success post interim* (PoS_{post}) as the probability of reaching statistical significance at the end of the trial, given that the trial was continued after the interim:

$$\text{PoS}_{post} = \int P(\hat{\theta}_{fin} > \theta_{suc}|\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}, \theta)\, q_1(\theta) d\theta. \tag{1.1}$$

PoS_{post} is useful to evaluate the confidence we would have in the success of the trial if it continues after the interim. Prior to the trial start, we can effectively fine-tune such confidence by adjusting the boundaries for futility and efficacy. However, there is a cost in aiming for higher PoS_{post}, as discussed below.

In order to assess the impact of boundary selection, it is first helpful to clarify the relationship between PoS_{post} and PoS. By expanding the conditional probability and the posterior density in equation (1.1), we get

$$\begin{aligned}\text{PoS}_{post} &= \int \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}, \hat{\theta}_{fin} > \theta_{suc}|\theta)}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta)} \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta)\, q_0(\theta)}{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta')\, q_0(\theta')d\theta'} d\theta \\ &= \frac{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}, \hat{\theta}_{fin} > \theta_{suc}|\theta)\, q_0(\theta)d\theta}{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}|\theta')\, q_0(\theta')d\theta'}.\end{aligned} \tag{1.2}$$

The numerator is the probability of not stopping early and reaching statistical significance at final analysis. The denominator is the probability of not stopping the trial early, respectively. Hence

$$\begin{aligned}\text{PoS}_{post} &= \frac{P(\textit{no early stop and success at final analysis})}{P(\textit{no early stop})} \\ &= \frac{\text{PoS} - P(\textit{early stop for efficacy})}{P(\textit{no early stop})}.\end{aligned} \tag{1.3}$$

The values of PoS_{post}, PoS, $P(\textit{no early stop and success at final analysis})$, $P(\textit{early stop for efficacy})$ and $P(\textit{no early stop})$ depend on the choice of the boundaries θ_{fut} and θ_{eff}.

It should be noted that, in our calculations, we assume that the DMC recommendation after the interim analysis will follow exactly the pre-specified stopping rules. In reality, the futility rule is likely to be non-binding and this might not happen all the times. For example, if the threshold for futility is crossed by a small margin while a number of secondary endpoints show positive results, the DMC may decide to suggest the continuation of the trial, instead of stopping it for futility. For this reason, even though the assumption is that the DMC always follows the futility boundary, θ_{suc} is calculated in a non-binding fashion and thus the type I error is not affected by the futility boundaries. However, we should expect that the DMC will behave at least somewhat consistently with the pre-specified stopping rules. We discuss the potential implications of the non-binding nature of the futility rule in Section 1.5.

If the interim estimate of a trial does not exceed the efficacy boundary, our confidence in the final success might be reduced. Such loss of confidence is higher for lower (easier

to reach) values of θ_{eff}, because failing to reach them suggests that the treatment may not be as good as expected. At the same time, if the interim estimate is above the futility boundary, our confidence in the final success should increase. Such increase is higher for larger values of θ_{fut}, because exceeding them suggests that the treatment has at least a small effect. When computing PoS_{post} we are conditioning on $\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}$, therefore both conditions are fulfilled. This means that PoS_{post} is increasing in both θ_{eff} and θ_{fut} (all proofs are in Section 1.6).

Therefore PoS_{post} is lower if the futility and efficacy boundaries are small, and higher if the futility and efficacy boundaries are large.

Consider the simpler case where early stopping for efficacy is not planned at all in the protocol, i.e., $\theta_{eff} = +\infty$. Then the expression in equation (1.3) simplifies to

$$\text{PoS}_{post} = \frac{\text{PoS}}{P(\textit{no early stop for futility})}.$$

If the trial can only be stopped early for futility, it is intuitive that our confidence in a final success can only increase with respect to PoS if the trial is not stopped at the interim analysis; indeed, $\text{PoS}_{post} \geq \text{PoS}$ for any choice of θ_{fut} and $\text{PoS}_{post} = \text{PoS} \Leftrightarrow P(\textit{no early stop for futility}) = 1 \Leftrightarrow \theta_{fut} = -\infty$.

Of note, by inverting the probability of not stopping early, we obtain the relative gain in probability of success after the interim:

$$P(\textit{no early stop for futility})^{-1} = \frac{\text{PoS}_{post}}{\text{PoS}}.$$

These probabilities are functions of the futility boundary. However, a change in the boundary has opposite effects on the different probabilities:

$$\theta_{fut} \nearrow \Longrightarrow \begin{array}{l} P(\textit{no early stop for futility}) \searrow \\ \text{PoS} \searrow \\ \text{PoS}_{post} \nearrow \end{array}$$

Since PoS and PoS_{post} have opposite trends, there is a tradeoff in setting a more (or less) aggressive futility boundary. Computing both probabilities of success for different values of θ_{fut} allows an assessment of this tradeoff and facilitates an informed decision on the choice of a futility boundary.

1.3 Special case for normally distributed data

We consider a two-arm trial with a sample size of n subjects per group in which θ is the mean treatment difference between a new drug and a control. The sample size is set according to the choice of the type I error rate α, the power $1-\beta$ and a mean treatment effect of interest Δ. The objective of the trial is to demonstrate that the new drug is superior to the control, which is translated in statistical terms as rejecting the null hypothesis H_0: $\theta \leq 0$ in favour of the alternative H_1: $\theta > 0$. The standard deviation of the response variable σ is assumed to be known. An interim analysis is scheduled after n_{int} subjects in each group have completed the study. We set a futility boundary θ_{fut} and an efficacy boundary θ_{eff} so that we may stop either for futility or for efficacy at the interim analysis. The threshold to reach statistical significance at the final analysis is again denoted by θ_{suc}.

In this particular case, we assume that the mean treatment effect θ is normally distributed. We will show that under this assumption PoS and PoS_{post} can be computed using a bivariate normal distribution. Given the mean treatment difference θ, the estimator $\hat{\theta}_{fin}|\theta$ at the final analysis is

$$\hat{\theta}_{fin}|\theta \sim \mathcal{N}\left(\theta, \frac{2\sigma^2}{n}\right).$$

The estimator $\hat{\theta}_{int}|\theta$ at the interim analysis also follows a normal distribution:

$$\hat{\theta}_{int}|\theta \sim \mathcal{N}\left(\theta, \frac{2\sigma^2}{n_{int}}\right)$$

and the conditional bivariate normal distribution of both is

$$\begin{pmatrix} \hat{\theta}_{int} \\ \hat{\theta}_{fin} \end{pmatrix} | \theta \sim N \left(\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} \frac{2\sigma^2}{n_{int}} & \frac{2\sigma^2}{n} \\ \frac{2\sigma^2}{n} & \frac{2\sigma^2}{n} \end{pmatrix} \right).$$

Let us also assume that the prior for theta is a normal distribution:

$$\theta \sim \mathcal{N} \left(\theta_0, \frac{2\sigma^2}{n_0} \right), \tag{1.4}$$

where n_0 is a fixed positive real number. For example, n_0 might be set to be equal (or smaller, in case of discounting) to the sample size of a previous study providing an estimate of the mean treatment effect. In general n_0 is not necessarily an integer and it may be chosen freely to obtain any sensible value for the variance of the prior Grieve, 2022, Chapter 2. Expressing the variance of the prior as $\frac{2\sigma^2}{n_0}$ is convenient when deriving the unconditional distributions of the estimators $\hat{\theta}_{int}$ and $\hat{\theta}_{fin}$. Since both $\hat{\theta}_{fin}|\theta$ and θ are normally distributed, it follows that the unconditional distribution of $\hat{\theta}_{fin}$ is normal too. The same is true for the for unconditional distribution of $\hat{\theta}_{int}$. Their unconditional bivariate distribution is

$$\begin{pmatrix} \hat{\theta}_{int} \\ \hat{\theta}_{fin} \end{pmatrix} \sim N \left(\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}, \begin{pmatrix} 2\sigma^2 \left(\frac{n_{int}+n_0}{n_{int} n_0} \right) & 2\sigma^2 \left(\frac{n+n_0}{nn_0} \right) \\ 2\sigma^2 \left(\frac{n+n_0}{nn_0} \right) & 2\sigma^2 \left(\frac{n+n_0}{nn_0} \right) \end{pmatrix} \right). \tag{1.5}$$

For the reader's convenience, we provide a proof of (1.5) in Section 1.6.2.

A very similar result can be obtained in a time-to-event setting if the treatment effect is measured as hazard ratio, by making use of the normal approximation of the log hazard ratio, as shown in Rufibach, Burger, and Abt, 2016.

When the distribution in (1.5) holds, the probabilities in equation (1.3) become easy to compute by making use of the bivariate normal cumulative distribution function of $(\hat{\theta}_{int}, \hat{\theta}_{fin})$. Numerically, this corresponds to evaluating two nested integrals. The explicit formulae of these probabilities are given in the Section 1.6.3.

More generally, the unconditional multivariate distribution for an arbitrary number K of interim analyses follows a K-variate normal, as explained in detail in Grieve, 2023. If we extend the above framework to include more than one interim analysis, the computations become more complex because all interim estimators must be incorporated into the calculations. In a K-stage design, the unconditional PoS and the other PoS-related metrics are derived from a K-variate normal distribution, which requires the evaluation of K nested integrals. However, this hardly poses a challenge for modern computers provided that K is reasonably small.

In order to provide a straightforward characterization of the probabilities found in equation (1.3), in what follows we will present three examples from a fictive two-stage trial.

1.4 Examples

1.4.1 Example 1: No early stop for efficacy

Consider a parallel group trial for a new treatment where the treatment effect is assumed normally distributed and the standard deviation is known ($\sigma = 1$). Assume $\Delta = 0.3$ as the treatment effect size of interest, expressed in terms of mean difference between treatments. This choice of Δ and σ reflects a median standardised effect size of 0.3 found in pivotal trials Rothwell et al., 2018. The power and the type I error rate are set to $1-\beta = 0.9$ and one-sided $\alpha = 0.025$ respectively. Based on these assumptions, a total of 468 subjects ($n = 234$ per group) are needed to reach the target power. The threshold for statistical significance θ_{suc} is computed as:

$$\theta_{suc} = z_{1-\alpha}\sqrt{\frac{2}{n}}\sigma = 0.181.$$

An interim analysis for futility is scheduled when half of the subjects have completed the study ($n_{int} = 117$ in each group). In this first example, the trial is not allowed to stop early for efficacy, i.e., the efficacy boundary is fixed at $\theta_{eff} = +\infty$.

We consider three normal prior distributions to evaluate PoS and PoS_{post} of this trial. These are defined as in (1.4) with $n_0 = 10$ and different choices of θ_0. We call the priors *pessimistic*, *realistic* and *optimistic*, as an evocative reflection of their characteristics. The

pessimistic prior is centered in a value below the effect size of interest (i.e., $\theta_0 = \Delta - 0.2$), the realistic prior is exactly centered in Δ (i.e., $\theta_0 = \Delta$) and the optimistic prior is centered in a value above Δ (i.e., $\theta_0 = \Delta + 0.2$). The same variance is assumed in all scenarios. These three prior distributions may be seen as plausible posterior distibutions obtained from a small study conducted on a total of $2n_0$ subjects before the start of the new trial. The pessimistic prior reflects a scenario where the treatment effect was below expectation in the small trial. In such a scenario, we may still consider advancing the development program, for example in a disease with a high unmet need. The realistic and optimistic priors correspond to cases with an observed effect in the small trial in line with expectations or above expectations, respectively.

As already noted, if a trial has a stopping rule for futility only, then $\text{PoS}_{post} \geq \text{PoS}$. Therefore, if the trial is not stopped at the interim look, the probability of reaching significance at the final analysis increases, regardless of the choice of the futility threshold. Relevant gains in PoS_{post} with negligible losses in PoS can ususally be achieved through an appropriate choice of θ_{fut}.

Figure 1.1 compares PoS_{post} and PoS as functions of θ_{fut}. In all three scenarios, PoS is almost constant for $\theta_{fut} < 0$, as the probability of success is practically unaffected by the definition of a small futility boundary. But even for those extremely cautious futility boundaries, there is an evident gain in terms of PoS_{post}.

To evaluate the gain in PoS_{post} against the loss of PoS, let us define the maximum possible PoS (MPPoS in short) as the upper bound for PoS. MPPoS is PoS in a trial without a stopping rule for futility, i.e., when $\theta_{fut} = -\infty$.

In the investigated scenarios, PoS decreases slowly while PoS_{post} increases relatively fast with increasing θ_{fut}. For this reason, a quite large futility boundary could be set to increase PoS_{post} for a small cost in terms of PoS reduction. For example, let us consider the trial design with the realistic prior, in which $\text{MPPoS} = 0.60$. Examining PoS_{post} as a function of θ_{fut}, we can identify the futility boundary that gives $\text{PoS} = 0.59$; that is $\theta_{fut} = 0.11$. This choice of θ_{fut} leads to $\text{PoS}_{post} = 0.90$. In a similar fashion, we can identify the smallest futility boundary that gives $\text{PoS} = 0.58$; that is $\theta_{fut} = 0.15$. This choice of θ_{fut} leads to $\text{PoS}_{post} = 0.93$. Limited losses of 1% or 2% compared to MPPoS lead to a significantly higher PoS_{post}, which corresponds to a stronger confidence in a final success in case of

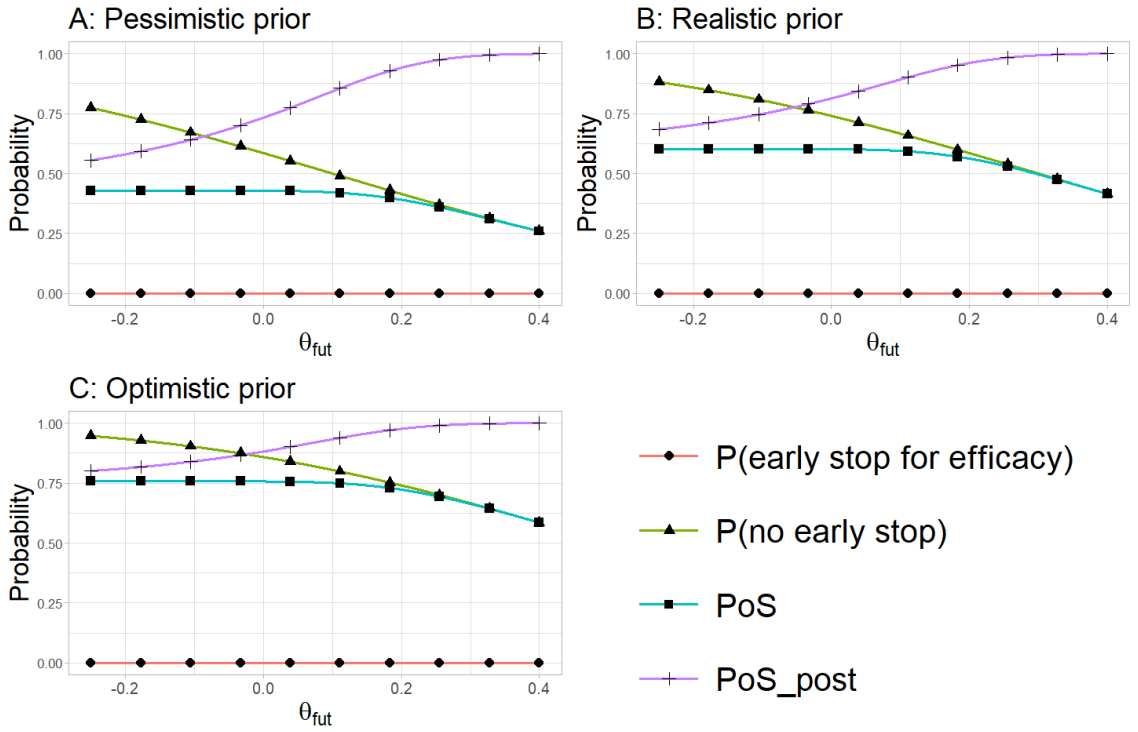

Fig. 1.1 PoS_{post}, PoS, $P(\textit{early stop for efficacy})$ and $P(\textit{no early stop})$ as functions of θ_{fut} for three different priors when there is no stopping rule for efficacy. $P(\textit{early stop for efficacy}) = 0$ because early stop for efficacy is not allowed. PoS is always smaller than $P(\textit{no early stop})$ because a success in this case requires that the trial is not stopped for futility at the interim. $\text{PoS}_{post} \geq \text{PoS}$ because the confidence in a success increases if the trial is not stopped at the interim analysis.

no early stop for futility. These findings are summarised in Table 1.1, including also the results for the following other examples.

The difference between the three priors, instead, is shown in Figure 1.1. As one would expect, the pessimistic prior gives the lowest PoS and PoS_{post}, while the optimistic prior gives the highest ones (θ_{fut} being equal). On the other hand, the largest increase in PoS_{post} compared to PoS is associated with the pessimistic prior, the lowest with the optimistic prior. This is because the optimistic prior assumes that the treatment effect is likely to be above Δ. Therefore, limited evidence of a positive treatment effect does not substantially alter the expected outcome. In contrast, the same evidence has a greater impact when the pessimistic prior is initially assumed.

1.4.2 Example 2: O'Brien-Fleming efficacy boundary

In the same scenarios previously described, we consider adding the possibility of an early stop for efficacy if the interim estimated effect is above a chosen efficacy boundary θ_{eff}, obtained using an O'Brien-Fleming alpha spending function DeMets and Lan, 1994

$$\alpha_{int}(I) = 2 - 2\phi\left(\frac{z_{1-\alpha/2}}{\sqrt{I}}\right).$$

Assuming the information fraction to be $I = 1/2$, i.e., half the subjects are considered for the interim analysis, $\alpha_{int}(1/2) = 0.0015$. The corresponding efficacy boundary at the interim analysis is

$$\theta_{eff} = z_{1-\alpha_{int}(1/2)}\sqrt{\frac{2}{n_{int}}}\sigma = 0.387. \tag{1.6}$$

Since we are introducing the option to stop early for efficacy, there are now two opportunities to reject the null hypothesis. Therefore, the threshold to reach statistical significance at the final analysis θ_{suc} has to be increased in order to preserve the overall type I error rate. In this setting, $\theta_{suc} = 0.182$ should be set.

As we can see in Figure 1.2, when adding a stopping rule for efficacy ($\theta_{eff} = 0.387$), $\text{PoS}_{post} \geq \text{PoS}$ does not hold for every possible choice of θ_{fut}, as was the case in Example 1 ($\theta_{eff} = +\infty$), but only when θ_{fut} is large enough. This is due to the negative impact of a failed interim for efficacy on PoS_{post} (PoS_{post} is increasing in θ_{eff}).

As it was the case in the previous example, PoS and PoS_{post} are highest with the optimistic prior but the difference $\text{PoS}_{post} - \text{PoS}$ is largest with the pessimistic prior. Indeed, we can achieve $\text{PoS}_{post} \geq \text{PoS}$ with a much smaller θ_{fut} when using the pessimistic prior, compared to the other scenarios.

Regardless of the prior used, there is always the possibility to select a futility boundary that gives PoS slightly below the MPPoS in order to increase PoS_{post} at a minimal cost (see Table 1.1 for the detailed results).

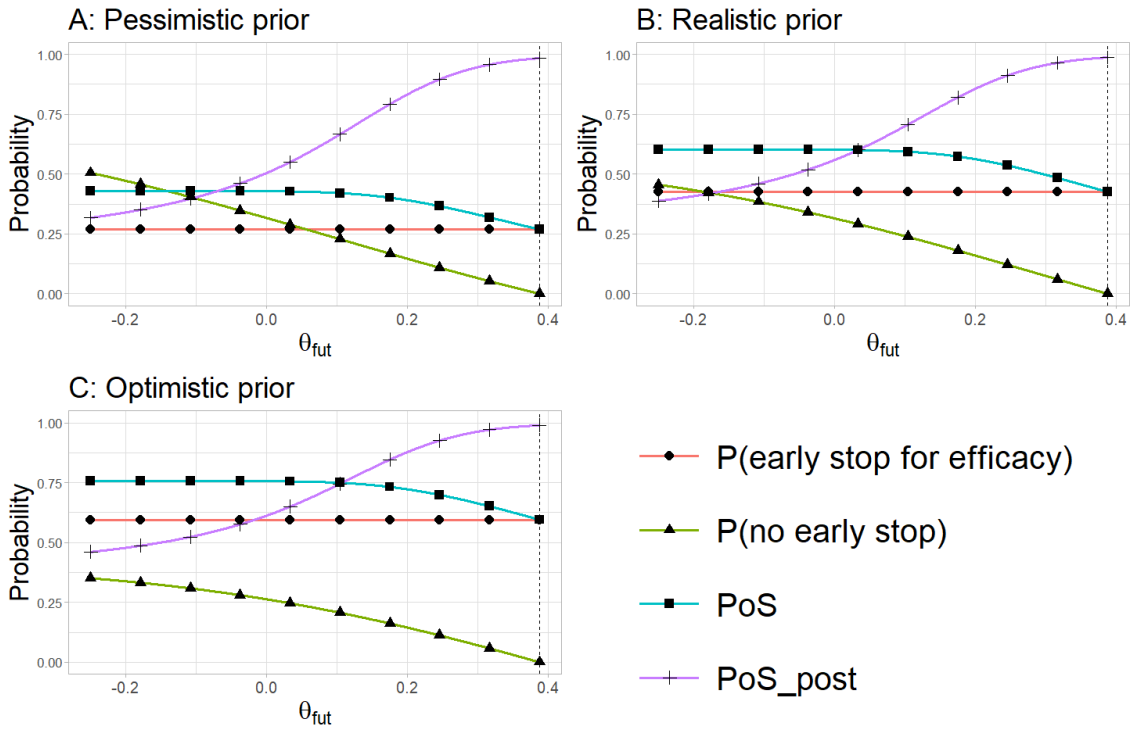

Fig. 1.2 PoS_{post}, PoS, $P(\textit{early stop for efficacy})$ and $P(\textit{no early stop})$ as functions of θ_{fut} for three different priors when an O'Brien-Fleming efficacy boundary (vertical dashed line) is used.

1.4.3 Example 3: Pocock efficacy boundary

This example is based on the same setting of Example 2, with one interim analysis with stopping rules for futility or efficacy, but using a different efficacy boundary, based on the Pocock alpha spending function DeMets and Lan, 1994

$$\alpha_{int}(I) = \alpha \ln\left[1 + (e-1)I\right].$$

Using the formula in (1.6), we obtain $\theta_{eff} = 0.282$. In order to preserve the overall type I error rate, the threshold for the final analysis is now $\theta_{suc} = 0.203$

Figure 1.3 presents the behaviour of PoS and PoS_{post} in this example. Since the Pocock efficacy boundary is smaller than the prior mean of both the realistic and optimistic prior, in both scenarios a very large futility boundary is needed to obtain $\text{PoS}_{post} \geq \text{PoS}$. Intuitively this is reasonable, because continuing after an interim analysis implies that the observed effect was smaller than the target effect ($\hat{\theta}_{int} \leq 0.282 < \Delta = 0.3$). On the other hand, with the pessimistic prior even a moderate futility boundary can result in a PoS_{post} more favorable than the initial PoS. Indeed, in the pessimistic scenario $\text{PoS}_{post} \geq \text{PoS}$ for the choice of θ_{fut} that reduces PoS by only 0.01, as shown in Table 1.1.

In any case, by choosing a futility boundary that gives PoS slightly below the MPPoS, we can always increase PoS_{post} (see Table 1.1 for detailed results), as in the previous examples. Although PoS_{post} will not necessarily exceed PoS, it may still be a valid option to choose a large futility boundary, in order to increase the confidence in a final success of the trial in case it continues after the interim analysis.

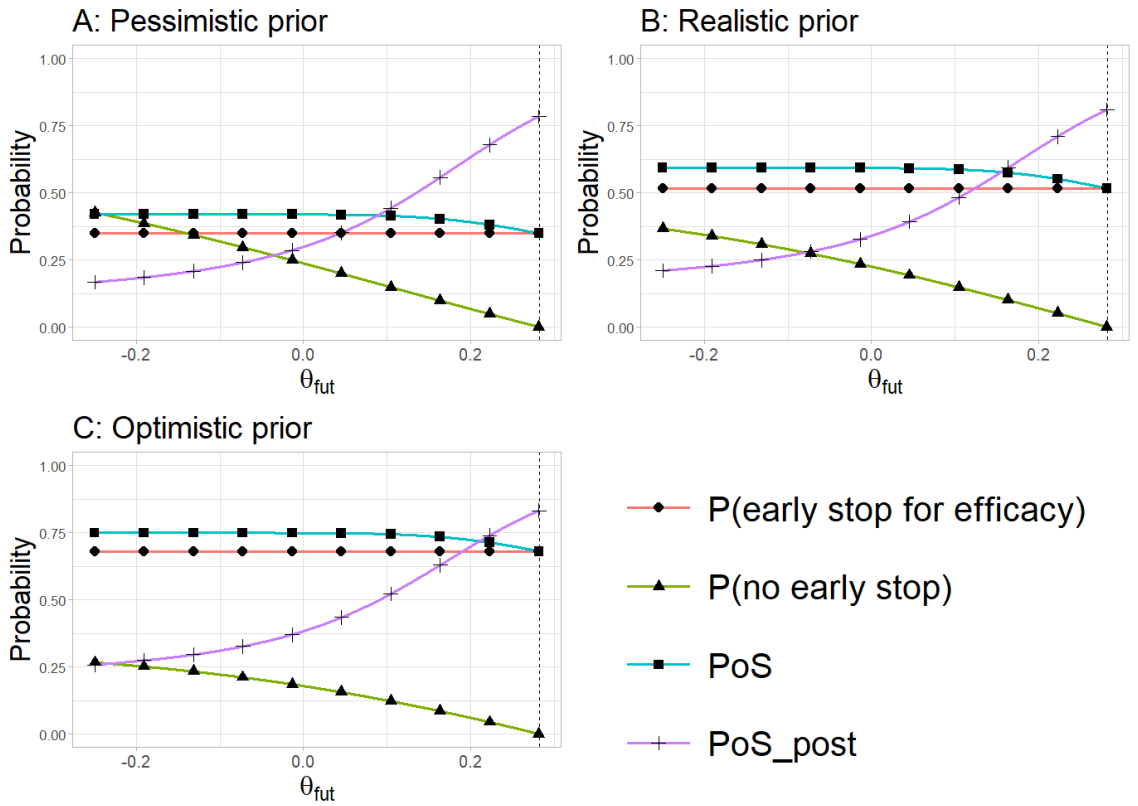

Fig. 1.3 PoS_{post}, PoS, $P(\textit{early stop for efficacy})$ and $P(\textit{no early stop})$ as functions of θ_{fut} for three different priors when a Pocock efficacy boundary (vertical dashed line) is used.

1.4.4 Further comments on the examples

Example	Prior	PoS for $\theta_{fut} = -\infty$ (MPPoS)	PoS_{post} for $\theta_{fut} = -\infty$	PoS_{post} when PoS is reduced by 0.01	PoS_{post} when PoS is reduced by 0.02
No early stop for efficacy	Pessimistic Realistic Optimistic	0.43 0.60 0.76	0.43 0.60 0.76	0.85 0.90 0.94	0.90 0.93 0.96
O'Brien-Fleming boundary	Pessimistic Realistic Optimistic	0.43 0.60 0.76	0.22 0.31 0.40	0.68 0.72 0.77	0.75 0.78 0.83
Pocock boundary	Pessimistic Realistic Optimistic	0.42 0.60 0.75	0.11 0.16 0.21	0.50 0.53 0.58	0.58 0.61 0.66

Table 1.1 PoS and PoS_{post} tradeoff according to futility boundaries. PoS is at its maximum (MPPoS) and PoS_{post} at its minimum when there is no stopping rule for futility ($\theta_{fut} = -\infty$). The two columns on the right-hand side show the value of PoS_{post} for the choice of futility boundaries that reduce PoS by 0.01 and 0.02 compared to MPPoS, respectively.

The effect of the chosen efficacy boundary on PoS_{post} can be seen in Figure 1.4, where the different PoS_{post} functions from the three examples are plotted together. As we have already mentioned, PoS_{post} is lower for smaller efficacy boundaries. Indeed PoS_{post} is lowest when a Pocock efficacy boundary is used, and highest when there is no stopping rule for efficacy.

Table 1.1 shows that even a quite small futility boundary, leading to a reduction in PoS of only 0.01, yields a PoS_{post} significantly greater than the one obtained without futility stopping rules. However, the benefit of choosing a larger futility boundary may not always outweigh the loss of PoS. For example, it looks reasonable to reduce PoS by 0.02 if an efficacy boundary is set (either the O'Brien-Fleming or the Pocock type), because PoS_{post} would increase further by 0.06 or more. On the other hand, in Example 1 ("No early stop for efficacy"), the gain in PoS_{post} is only 0.05 if we are considering the pessimistic prior, or less otherwise. This phenomenon is graphically represented in Figure 1.4, where the PoS_{post} function in the "No early stop for efficacy" example is less steep than in the "Pocock" example, meaning that in the latter situation PoS_{post} increases more quickly for larger values of θ_{fut}.

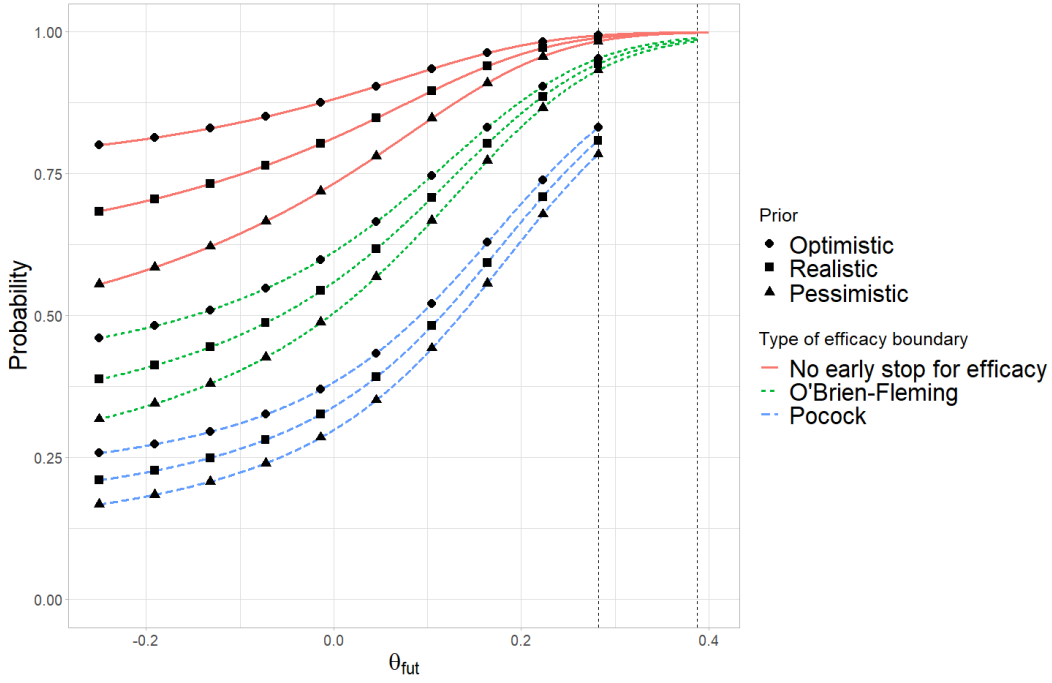

Fig. 1.4 Comparison of PoS_{post} as a function of θ_{fut} in the three discussed examples. From left to right, the vertical dashed lines correspond to the Pocock and the O'Brien-Fleming efficacy boundaries from example 3 and 2 respectively.

In Figure 1.4 we can also observe the impact of the three priors (the pessimistic, realistic and optimistic ones) on PoS_{post}. For low values of θ_{eff}, i.e., when a Pocock alpha-spending function is used, the difference of PoS_{post} in the pessimistic and realistic scenarios, as well as the difference between the realistic and the optimistic scenarios, is relatively constant for all values of θ_{fut}. For higher values of θ_{eff} (the extreme case being the "no early stop for efficacy" example), PoS_{post} is quite different between the three scenarios for low values of θ_{fut}, but then quickly converges to the same value with increasing θ_{fut}.

Moreover, focusing on the efficacy boundary, PoS_{post} decreases significantly for smaller θ_{eff}. We can observe a relevant drop in PoS_{post} when moving from the example with no early stop for efficacy to the example with the O'Brien-Fleming boundary and from the O'Brien-Fleming to the Pocock boundary. This does not necessarily mean that a lower efficacy boundary is detrimental to the study in general, but it means that, if the study continues after the interim, the confidence in its success may be reduced.

1.4.5 Impact of the information fraction

Our previous examples focused on trials with a single interim analysis conducted when half the subjects completed the study, corresponding to an information fraction $I = 0.5$. The theoretical results from Sections 1.2 and 1.3 remain valid regardless of the information fraction. Nevertheless, despite the general validity of their described relationship, variations in information fraction may influence PoS and PoS_{post} values. In this section, we investigate this impact within the specific context of the designs previously examined.

Figure 1.5 compares PoS and PoS_{post} for information fractions from 0.2 to 0.8, assuming the realistic prior. Note that the range of θ_{fut} depends on I, as θ_{fut} must be less than θ_{eff}, which decreases as I increases. Figure 1.6 and Table 1.2 provide more details for $I = 0.2$ and $I = 0.8$ using the O'Brien-Fleming alpha spending function under the same prior.

In the trial design with no early stop for efficacy, the information fraction has a limited impact on both PoS and PoS_{post}. PoS is almost not affected by the information fraction. Regarding PoS_{post}, the curve describing its relationship with θ_{fut} becomes flatter as I decreases, because the results informing the interim decision become less predictive of the final outcome.

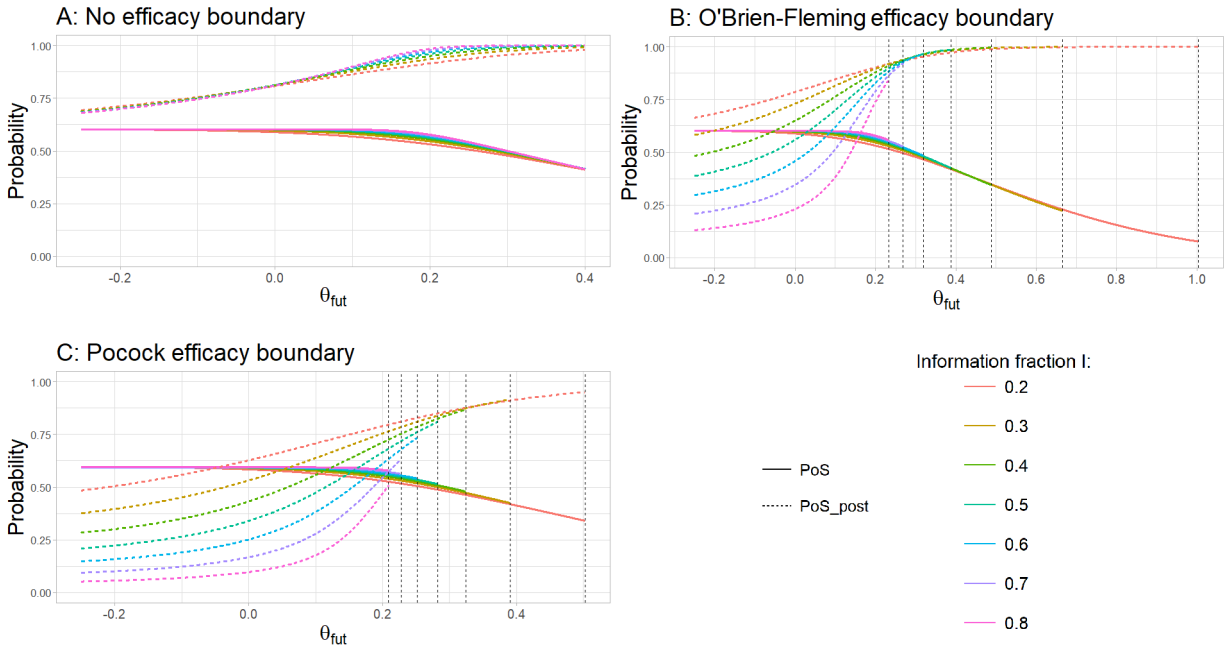

Fig. 1.5 PoS and PoS$_{post}$ under the realistic prior as functions of θ_{fut} for information fractions between $I = 0.2$ and $I = 0.8$. The vertical dashed lines correspond to the efficacy boundaries.

In the two designs allowing early efficacy stop, the information fraction minimally affects PoS overall. Despite its negligible impact on overall PoS, the information fraction significantly influences its components (PoS $= P(\textit{early stop for efficacy}) + P(\textit{no early stop}) \cdot$ PoS$_{post}$), as shown in Table 1.2: higher I values increase the probability of early efficacy stop, while reducing the relative contribution of final analysis success to PoS. This occurs because larger information fractions lead to greater alpha spending at the interim analysis, resulting in smaller, more easily crossed efficacy boundaries θ_{eff}. Contrary to PoS, PoS$_{post}$ is greatly affected by the information fraction. As in the design without early efficacy stop, the PoS$_{post}$ vs. θ_{fut} curve flatten with lower I, but this trend is significantly more pronounced when an interim efficacy analysis is planned. This is induced by gradually decreasing values of θ_{eff} as information accumulates. Failing to cross a smaller efficacy boundary later in the study has a more substantial negative impact on PoS$_{post}$ compared to not reaching a larger, more challenging boundary earlier.

For all three trial designs, consistent trends are observed when considering pessimistic and optimistic prior distributions.

	$I = 0.2$					$I = 0.8$				
θ_{fut}	PoS	PoS_{post}	$P(no\ early\ stop)$	$P(stop\ for\ eff.)$	$P(stop\ for\ fut.)$	PoS	PoS_{post}	$P(no\ early\ stop)$	$P(stop\ for\ eff.)$	$P(stop\ for\ fut.)$
-0.20	0.60	0.68	0.77	0.08	0.15	0.60	0.14	0.30	0.56	0.14
-0.15	0.60	0.70	0.74	0.08	0.18	0.60	0.15	0.28	0.56	0.16
-0.10	0.60	0.73	0.72	0.08	0.21	0.60	0.17	0.25	0.56	0.19
-0.05	0.59	0.76	0.68	0.08	0.24	0.60	0.19	0.22	0.56	0.22
0.00	0.59	0.79	0.65	0.08	0.27	0.60	0.23	0.19	0.56	0.26
0.05	0.58	0.82	0.62	0.08	0.31	0.60	0.29	0.15	0.56	0.29
0.10	0.57	0.85	0.58	0.08	0.34	0.60	0.38	0.11	0.56	0.33
0.15	0.55	0.88	0.54	0.08	0.38	0.60	0.54	0.07	0.56	0.37
0.20	0.53	0.90	0.50	0.08	0.42	0.58	0.74	0.03	0.56	0.41

Table 1.2 PoS, PoS_{post}, $P(no\ early\ stop)$, $P(early\ stop\ for\ efficacy)$ and $P(early\ stop\ for\ futility)$ as functions of θ_{fut} for information fractions $I = 0.2$ and $I = 0.8$ when an O'Brien-Fleming efficacy boundary is used and the realistic prior is assumed.

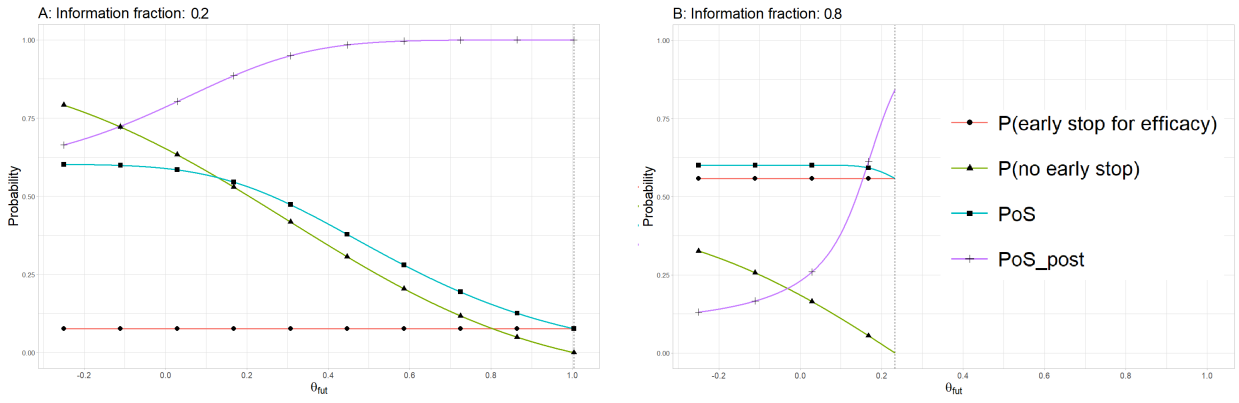

Fig. 1.6 PoS_{post}, PoS, $P(early\ stop\ for\ efficacy)$ and $P(no\ early\ stop)$ as functions of θ_{fut} for information fractions $I = 0.2$ and $I = 0.8$ when an O'Brien-Fleming efficacy boundary (vertical dashed lines) is used and the realistic prior is assumed.

1.5 Discussion

PoS is the unconditional probability of success of the study, based on a prior distribution on the treatment effect θ. PoS_{post}, on the other hand, is the PoS conditioned on the fact that the trial continues after the interim analysis, i.e., $\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}$.

Our experience has demonstrated that, on many occasions, the study team tends to express excessive optimism regarding the future outcome of the trial upon receiving the news that the trial will continue following the interim phase. Our research objectively examines the relationship between PoS and PoS_{post} to determine if this behavior is warranted, and to provide recommendations for adjusting the interim boundaries to align with the expectations. Moreover, we believe that PoS_{post} offers a straightforward interpretation of the impact of the DMC recommendations to continue the trial after the interim analysis, making it a valuable tool for communication with non-statisticians.

We considered the case where the treatment effect is measured as a mean difference, and provided some examples of a fictive trial with a fixed maximum sample size. In these exmples we explored the use of three different prior distributions and efficacy boundaries, while the futility boundary was allowed to vary on the entire continuous scale.

For a given choice of θ_{eff}, we have shown that PoS_{post} is increasing in θ_{fut}, while PoS is decreasing. PoS_{post} is also increasing in θ_{eff}, hence a higher efficacy boundary will yield a higher PoS_{post}, but at the cost of a smaller chance of an early stop for efficacy.

Our results show that, in certain situations, a somewhat large futility boundary may be chosen to obtain a substantial increase of PoS_{post} at the cost of a slight reduction of PoS. On the other hand, if the futility rule is not aggressive enough, the fact of passing an interim analysis should not be received with too much optimism, since the chances of success may remain quite low. This is particularly true for studies with efficacy boundaries that are quite easy to reach. When instead there are no efficacy stopping rules, the chances of success always increase after passing the futility interim analysis.

However, it is hard to recommend a general strategy for selecting boundaries based on the metrics discussed in this chapter because of the conceptual difference between PoS and PoS_{post}. A reduction in PoS implies directly that the trial is less likely to succeed, while an increment in PoS_{post} is only advantageous if the trial is continued after the interim. Their

tradeoff should be carefully evaluated case by case, depending on the characteristics of the trial and on financial considerations from the sponsor. Sometimes it may be worthwhile to sacrifice some of the PoS to increase PoS_{post}, but providing general guidelines for how much can be reasonably sacrificed is above the scope of this research.

The main reason for understanding the impact of the interim boundaries on PoS and PoS_{post} is to provide the sponsor and study team with clear and realistic expectations about the trial's trajectory, both before and after the interim analysis. Fine-tuning the interim boundaries to adjust these probabilities is possible, but it should rarely be the sole reason for their selection.

To the best of our knowledge, our research is the first to study the relationship between PoS and PoS_{post} and to connect it with the choice of the interim boundaries. We based our definitions of PoS and PoS_{post} on the assumption that the DMC would always recommend to continue a trial if the futility rule is fulfilled. Although in reality the futility rule is likely to be non-binding, it is reasonable to assume that the DMC recommendation will be at least somewhat consistent with the pre-specified stopping rules. Incorporating the probability that the DMC follows the futility rule into our calculations would be theoretically possible. However, the parameters of such modeling would certainly be quite subjective and context dependent: this is an area for future research.

1.6 Proofs and supplementary material

1.6.1 Proof that PoS_{post} is increasing in θ_{eff}

Given a futility boundary θ_{fut}, let us explicitly write the dependence of PoS_{post} on the efficacy boundary θ_{eff} as $\text{PoS}_{post}(\theta_{eff})$. $\text{PoS}_{post}(\theta_{eff})$ is an increasing function of θ_{eff} if

$$\text{PoS}_{post}(\theta_{eff1}) < \text{PoS}_{post}(\theta_{eff2})$$

for any choice of θ_{eff1}, θ_{eff2} such that $\theta_{fut} < \theta_{eff1} < \theta_{eff2}$.

Let us choose any pair $(\theta_{eff1}, \theta_{eff2})$ satisfying $\theta_{fut} < \theta_{eff1} < \theta_{eff2}$. In a study with an interim analysis for efficacy, the threshold at the final analysis have to be determined according to the choice of the efficacy boundary in order to preserve the overall type I error rate. Let us call θ_{suc1} and θ_{suc2} the thresholds for significance at the final analysis

corresponding to the choice of θ_{eff1} and θ_{eff2} as efficacy boundaries, respectively. Since $\theta_{eff1} < \theta_{eff2}$, it follows that

$$\theta_{suc1} > \theta_{suc2}. \tag{1.7}$$

Let $q_0(\theta)$ be the probability density function of the prior distribution of the treatment effect θ. For the law of total probability, the unconditional joint distribution of $\hat{\theta}_{int}$ and $\hat{\theta}_{fin}$ is

$$P(\hat{\theta}_{int} \in A, \hat{\theta}_{fin} \in B) = \int P(\hat{\theta}_{int} \in A, \hat{\theta}_{fin} \in B|\theta) q_0(\theta) d\theta \quad \forall A, B \subset \mathbb{R}.$$

For equation (1.2),

$$\begin{aligned} \text{PoS}_{post}(\theta_{eff1}) &= \frac{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1}|\theta)\, q_0(\theta) d\theta}{\int P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}|\theta')\, q_0(\theta') d\theta'} \\ &= \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1})}. \end{aligned}$$

Let us prove the following inequality first:

$$\frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1})} < \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2})}. \tag{1.8}$$

The right-hand side of (1.8) is equal to

$$\frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1}) + P(\theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}) + P(\theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2})}.$$

Let us call

$$\begin{aligned}
a &= P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1}),\\
b &= P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}),\\
c &= P(\theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2}, \hat{\theta}_{fin} > \theta_{suc1}),\\
d &= P(\theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2}).
\end{aligned}$$

Since $a, b, c, d > 0$,

$$\frac{a}{b} < \frac{a+c}{b+d} \Longleftrightarrow a(b+d) < (a+c)b \Longleftrightarrow ad < cb \Longleftrightarrow \frac{a}{b} < \frac{c}{d}.$$

By definition of conditional probability,

$$\begin{aligned}
\frac{a}{b} &= P(\hat{\theta}_{fin} > \theta_{suc1} | \theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}),\\
\frac{c}{d} &= P(\hat{\theta}_{fin} > \theta_{suc1} | \theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2}).
\end{aligned}$$

Hence (1.8) follows from the following inequality:

$$P(\hat{\theta}_{fin} > \theta_{suc1} | \theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}) < P(\hat{\theta}_{fin} > \theta_{suc1} | \theta_{eff1} < \hat{\theta}_{int} \leq \theta_{eff2}).$$

The above inequality is true because $\hat{\theta}_{fin}$ depends linearly on $\hat{\theta}_{int}$. In particular

$$\hat{\theta}_{fin} = \left(\frac{n_{int}}{n}\right) \hat{\theta}_{int} + \left(1 - \frac{n_{int}}{n}\right) \hat{\theta}_{post},$$

where $\hat{\theta}_{post}$ is the estimator of θ in the second part of the trial Grieve, 2022, Chapter 3.

Using (1.8), we can complete the proof as follows

$$\begin{aligned}
\text{PoS}_{post}(\theta_{eff1}) &\overset{(1.2)}{=} \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff1})} \\
&\overset{(1.8)}{<} \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2}, \hat{\theta}_{fin} > \theta_{suc1})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2})} \\
&\overset{(1.7)}{<} \frac{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2}, \hat{\theta}_{fin} > \theta_{suc2})}{P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff2})} \\
&\overset{(1.2)}{=} \text{PoS}_{post}(\theta_{eff2}).
\end{aligned}$$

The proof that PoS_{post} is increasing in θ_{fut} is analogous to this one.

1.6.2 Proof of (1.5)

From the law of total expectation

$$\mathbb{E}[\hat{\theta}_{int}] = \mathbb{E}\left[\mathbb{E}[\hat{\theta}_{int}|\theta]\right] = \mathbb{E}[\theta] = \theta_0.$$

From the law of total variance

$$var[\hat{\theta}_{int}] = \mathbb{E}\left[var[\hat{\theta}_{int}|\theta]\right] + var\left[\mathbb{E}[\hat{\theta}_{int}|\theta]\right] = \mathbb{E}\left[\frac{2\sigma^2}{n_{int}}\right] + var\left[\theta\right] = \frac{2\sigma^2}{n_{int}} + \frac{2\sigma^2}{n_0} = 2\sigma^2\left(\frac{n_{int}+n_0}{n_{int}n_0}\right).$$

Similarly, $\mathbb{E}[\hat{\theta}_{fin}] = \theta_0$ and $var[\hat{\theta}_{fin}] = 2\sigma^2\left(\frac{n+n_0}{nn_0}\right)$.

Let us define $\hat{\theta}_{post}$ as the estimator of the mean treatment effect in the second part of the trial. Notice that, conditionally on θ, $\hat{\theta}_{int}|\theta$ and $\hat{\theta}_{post}|\theta$ are independent and $\hat{\theta}_{fin}|\theta = \left[\left(\frac{n_{int}}{n}\right)\hat{\theta}_{int} + \left(1-\frac{n_{int}}{n}\right)\hat{\theta}_{post}\right]|\theta$. Moreover, from the law of total covariance

$$cov(\hat{\theta}_{int}, \hat{\theta}_{fin}) = \mathbb{E}\left(cov(\hat{\theta}_{int}|\theta, \hat{\theta}_{fin}|\theta)\right) + cov\left(\mathbb{E}(\hat{\theta}_{int}|\theta), \mathbb{E}(\hat{\theta}_{fin}|\theta)\right).$$

Let us compute the two terms in the right-hand side:

$$\begin{aligned}\mathbb{E}\left(cov(\hat{\theta}_{int}|\theta,\hat{\theta}_{fin}|\theta)\right) &= \mathbb{E}\left(cov(\hat{\theta}_{int}|\theta,\left[\left(\frac{n_{int}}{n}\right)\hat{\theta}_{int}+\left(1-\frac{n_{int}}{n}\right)\hat{\theta}_{post}\right]|\theta)\right)\\ &= \mathbb{E}\left(\frac{n_{int}}{n}var(\hat{\theta}_{int}|\theta)+0\right)\\ &= \mathbb{E}\left(\frac{n_{int}}{n}\frac{2\sigma^2}{n_{int}}\right)=\frac{2\sigma^2}{n},\end{aligned}$$

$$cov\left(\mathbb{E}(\hat{\theta}_{int}|\theta),\mathbb{E}(\hat{\theta}_{fin}|\theta)\right)=var(\theta)=\frac{2\sigma^2}{n_0}.$$

It immediately follows

$$cov(\hat{\theta}_{int},\hat{\theta}_{fin})=\frac{2\sigma^2}{n}+\frac{2\sigma^2}{n_0}=2\sigma^2\left(\frac{n+n_0}{nn_0}\right).$$

1.6.3 Probabilities in (1.3) for normally distributed data

$$\begin{aligned}&P(\textit{early stop for efficacy})=P(\hat{\theta}_{int}>\theta_{eff})\\ &=1-\phi\left(\frac{\theta_{eff}-\theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right)\end{aligned}$$

$$\begin{aligned}&P(\textit{no early stop})=P(\theta_{fut}\leq\hat{\theta}_{int}\leq\theta_{eff})\\ &=\phi\left(\frac{\theta_{eff}-\theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right)-\phi\left(\frac{\theta_{fut}-\theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right)\end{aligned}$$

Let us define the bivariate normal cumulative distribution function of $(\hat{\theta}_{int},\hat{\theta}_{fin})$ as $\phi_{biv}(\cdot,\cdot)$.

$$\begin{aligned}\text{PoS} &= P(\hat{\theta}_{int} > \theta_{eff}) + P(\theta_{fut} \leq \hat{\theta}_{int} \leq \theta_{eff}, \hat{\theta}_{fin} > \theta_{suc}) \\ &= 1 - \left[\phi\left(\frac{\theta_{fut} - \theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right) + \phi_{biv}(\theta_{eff}, \theta_{suc}) - \phi_{biv}(\theta_{fut}, \theta_{suc})\right]\end{aligned}$$

$$\begin{aligned}\text{PoS}_{post} &= \frac{\text{PoS} - P(\textit{early stop for efficacy})}{P(\textit{no early stop})} \\ &= 1 - \frac{\phi_{biv}(\theta_{eff}, \theta_{suc}) - \phi_{biv}(\theta_{fut}, \theta_{suc})}{\phi\left(\frac{\theta_{eff} - \theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right) - \phi\left(\frac{\theta_{fut} - \theta_0}{\sqrt{2}\sigma}\sqrt{\frac{n_{int}n_0}{n_{int}+n_0}}\right)}\end{aligned}$$

Chapter 2

A Bayesian entropy-based response-adaptive design jointly targeting mean and variance in multi-arm trials with continuous outcomes

2.1 Chapter introduction

A common practice in clinical trials is to assign participants evenly to each treatment arm, using fixed and equal probabilities during randomisation. This approach is attractive for its straightforward implementation and because, when the variability across treatment groups is assumed to be the same, it yields good statistical performance, providing reliable power, accurate estimates, and appropriate control of type I error. However, when the assumption of homogeneous variance does not hold, assigning the same number of patients to each treatment can become inefficient, sometimes performing worse than alternative allocation strategies on these key operating characteristics [Lu and Yeh-Fong, 2022]. Furthermore, a balanced allocation design may expose many participants to less effective treatments, and if the total number of patients is small relative to the number of treatment options or doses, dividing them evenly across all groups can result in a poor use of limited trial

resources [Woods et al., 1998]. More generally, fixed randomisation designs specify allocation probabilities in advance and maintain them throughout the trial, regardless of accumulating outcome information. This rigidity can limit their efficiency or the benefit of trial participants when true treatment effects differ. Response-adaptive designs (RAR) try to remedy to those potential issues, by sequentially updating the randomisation probabilities based on the previously observed patient outcomes. These probabilities can be adjusted to favor treatment arms with desirable properties or to optimise specific operating characteristics, such as the proportion of patients allocated to the most effective treatment or statistical power [Biswas et al., 2011]. A rigorous introduction to non-Bayesian adaptive designs is the book by Hu and Rosenberger, 2006. For a more recent review of RAR methodologies refer instead to Robertson et al., 2023.

A number of existing clinical trials have adopted some form of RAR in different research settings, but by far the most common use is during phase II studies with more than two treatment arms [Wilson et al., 2025]. Multi-Arm Bandit (MAB) approaches form a prominent category within response-adaptive designs, offering a flexible framework for assigning participants in multi-arm trials. The core idea behind these methods is to dynamically balance two competing goals: exploring the less-visited treatments, and focusing on the treatments that appear to be the most promising. Over the past several years, a variety of MAB designs have been introduced to accommodate different types of trial outcomes. Some designs have been specifically developed for discrete endpoints (e.g., Villar et al., 2015), while others have extended the methodology to handle continuous outcomes (e.g., Smith and Villar, 2018).

However, while they generally focus on the traditional goal of identifying the treatment with highest (or lowest) response, in some contexts the interest may lie instead in an optimal range or in achieving a specific value of the treatment effect. Similar considerations were discussed in the recent work by Caruso and Mozgunov, 2024, which examined the "exploration versus exploitation" trade-off [Azriel et al., 2011] in the setting of multi-arm trials when information on the desirable average outcome is available. Caruso's approach allocates each newly enrolled patient to the arm which maximises a certain information value, based upon a context-dependent criterion that accounts for both the precision of the estimation and the interest around the specific value of the estimate. This allocation procedure inherently favours the treatments with greater variability, as more evidence is required to estimate their effect. Moreover, Caruso's approach assumes known treatment variances and, while it contains a tuning parameter to regulate their relevance in the

allocation procedure, it lacks a handle to adjust more nuanced variance considerations. We extend their work by proposing a RAR design for multi-arm trials that adjusts the randomisation probabilities based on both the treatment means and variances. This addition allows the design to prevent excessive allocations to treatments with very low variance, since less information is needed to estimate their effect, as well as to treatments with unexpectedly large variance, where outcome unpredictability may pose safety risks.

Our methodology is based on the theory of weighted (or context-dependent) information measures [Kelbert and Mozgunov, 2015, Suhov et al., 2016] and the criterion used for the randomisation rule is derived from evaluating the information gain (defined as a difference of the Shannon differential entropy and the weighted Shannon differential entropy). By making use of a Bayesian framework and a suitable parametric weight function, the weighted information approach is able to incorporate the investigators' interest in a particular subset of the parameter space into the randomisation process. The idea of taking into account the information given by the "context" of the experiment directly into the decision-making has already been applied in clinical trial designs with binary [Mozgunov and Jaki, 2020a], multinomial [Mozgunov and Jaki, 2020b] and, most recently, continuous endpoints [Caruso and Mozgunov, 2024]. We propose a new application for normally distributed endpoints, where the treatment variances are assumed to be neither known nor homogeneous. Both the mean and the variance of each treatment response are modeled with a joint conjugate prior, and the corresponding information measure is derived from a combination of their posterior distribution with the weight function. By defining the randomisation probabilities in terms of the information measure, the weighted information approach is able to randomise more patients to treatment arms whose characteristics are more closely aligned with the clinical targets, while still achieving high level of power, compared to fixed randomisation.

Two tuning parameters within the weight function tailor the importance given to different aspects of the investigation: the first parameter controls the trade-off between exploring less-visited arms vs. maximising the allocations to the most promising one, while the other regulates the importance of the treatment variances. A systematic approach for calibrating these parameters according to the trial's objectives is proposed, and the resulting design performance under different parameter settings is evaluated through Monte Carlo simulations.

Finally, and distinctively from Caruso's design, we consider trials that include a control group. This allows us to construct a hypothesis-testing procedure to assess the superiority of an alternative treatment over the control when the goal is to identify a response close to a pre-specified target value set in advance by clinicians. In particular, we define the testing procedure within a hybrid Bayesian-frequentist framework, in which rejection of the null hypothesis is based on the posterior distributions of the treatment means, while frequentist power and type I error are quantified through a simulation-based approach.

In Section 2.2, we derive the information-theoretic randomisation criterion for a class of weight functions focused around the target values γ and ξ, by expanding on the work of Caruso and Mozgunov, 2024. In Section 2.3, we give the definition of "best" treatment arm and we present a decision rule to identify it at the end of the study. Moreover, we illustrate a hypothesis testing procedure to claim the superiority of the best arm over control and we discuss how to select the cut-off value for the rejection of the null hypothesis in a way that controls the type I error. In Section 2.4, we describe a robust strategy for selecting the tuning parameters κ and ω, based on the optimisation of appropriate objective functions targeting specific operating characteristics. In Section 2.5, we illustrate the performance of our approach compared to other popular alternatives, including fixed and response-adaptive randomisation methods, by analyzing the results from a large simulative study. Conclusive remarks are presented in Section 2.6.

2.2 Methodology

2.2.1 Context-dependent information measure

Consider a multi-arm clinical trial with K treatment arms and a continuous endpoint. We denote the first arm, $j = 1$, as the control arm. Let us assume the n_j responses from an arm j to be independent and identically distributed (iid) as follows: $X_{1,j},\dots,X_{n_j,j} \overset{iid}{\sim} N(M_j, V_j)$, with $M_j \in \mathbb{R}$, $V_J \in \mathbb{R}^+$ $(j = 1,\dots,K)$, $n_j \geq 1$. We assume a Bayesian framework in which both the mean and the variance of the response are modeled via a joint Normal inverse-Gamma prior distribution $(\mu_j, \sigma_j^2) \sim NIG(\mu_j^{(0)}, \nu_j, \alpha_j^{(0)}, \beta_j^{(0)})$. This is equivalent to having an inverse-Gamma prior distribution $IG(\alpha_j^{(0)}, \beta_j^{(0)})$ for the variance σ_j^2 and a Normal prior distribution $N(\mu_j^{(0)}, \sigma_j^2/\nu_j)$ for the mean μ_j (conditionally on σ_j^2). The joint posterior distribution of μ_j and σ_j^2 given $n_j \geq 2$ observations $\underline{x}_j = (x_1,\dots,x_{n_j})$ is a conjugate Normal inverse-Gamma distribution:

$$(\mu_j, \sigma_j^2)|\underline{x}_j \sim NIG\left(\overline{\mu}_j,\, m_j,\, \alpha_j,\, \beta_j\right),$$

where $\overline{\mu}_j = \frac{n_j\bar{x}_j + \nu_j\mu_j^{(0)}}{n_j+\nu_j}$, $m_j = n_j + \nu_j$, $\alpha_j = \alpha_j^{(0)} + \frac{n_j}{2}$, $\beta_j = \beta_j^{(0)} + \frac{n_j-1}{2}\bar{s}_j^2 + \frac{n_j\nu_j}{n_j+\nu_j}\frac{(\mu_j^{(0)}-\bar{x}_j)^2}{2}$, $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i$ and $\bar{s}_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(x_i-\bar{x}_j)^2$. Equivalently, $\mu_j|\sigma_j^2, \underline{x}_j \sim N\left(\overline{\mu}_j \frac{\sigma_j^2}{m_j}\right)$ and $\sigma_j^2|\underline{x}_j \sim IG\left(\alpha_j,\, \beta_j\right)$.

The posterior density of $(\mu_j, \sigma_j^2|\underline{x}_j)$ is

$$p(\mu_j, \sigma_j^2|\underline{x}_j) = \frac{\sqrt{m_j}}{\sqrt{2\pi\sigma_j^2}}\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}(\sigma_j^2)^{-\alpha_j-1}\exp\left(-\frac{2\beta_j + m_j\left(\mu_j-\overline{\mu}_j\right)^2}{2\sigma_j^2}\right), \tag{2.1}$$

which concentrates around the true mean and variance of the treatment, (M_j, V_j), as the sample size increases. A measure of the amount of information needed to estimate (M_j, V_j) can be given by the Shannon differential entropy of $p(\mu_j, \sigma_j^2|\underline{x}_j)$, i.e.,

$$h(\mu_j, \sigma_j^2) = -\int_0^\infty \int_{\mathbb{R}} p(\mu_j, \sigma_j^2|\underline{x}_j)\ln\left[p(\mu_j, \sigma_j^2|\underline{x}_j)\right] d\mu_j\, d\sigma_j^2. \tag{2.2}$$

However, this quantity does not depend on (M_j, V_j) directly. In fact, it does not indicate at all what is being estimated, but it only reflects the amount of uncertainty surrounding the estimation.

In a multi-arm phase II trial, the primary goal is often not only that of estimating all the treatment effects, but also that of identifying a promising treatment with characteristics aligned to clinical targets. For example, obtaining a rough estimate of the mean of a beneficial treatment may be more valuable than obtaining a precise estimate of the mean of an ineffective one. The Shannon entropy, on its own, ignores the fact that some parameter values are intrinsically more interesting than others. As previously pointed out in Mozgunov and Jaki, 2020b and Caruso and Mozgunov, 2024, the investigators' interest around certain clinically desirable values can be incorporated into the estimation measure by combining the Shannon entropy in (2.2) with an appropriate weight function $\phi(\mu_j, \sigma_j^2)$ that emphasises a specific region of the parametric space. This results in the definition of

the weighted Shannon differential entropy of $p(\mu_j, \sigma_j^2|\underline{x}_j)$, i.e.,

$$h^{\phi}(\mu_j, \sigma_j^2) = -\int_0^{\infty}\int_{\mathbb{R}} \phi(\mu_j, \sigma_j^2) p(\mu_j, \sigma_j^2|\underline{x}_j) \ln\left[p(\mu_j, \sigma_j^2|\underline{x}_j)\right] d\mu_j \, d\sigma_j^2.$$

Notice that when $\phi(\mu_j, \sigma_j^2)$ is uniformly 1 (i.e., the same degree of interest is applied to all values of the parameter space), we obtain the traditional Shannon differential entropy.

We can now give the definition of information gain

$$\Delta(\mu_j, \sigma_j^2) = h(\mu_j, \sigma_j^2) - h^{\phi}(\mu_j, \sigma_j^2), \tag{2.3}$$

which can be interpreted as the average amount of additional statistical information that is required when considering the context-dependent instead of the traditional estimation problem. In other words, $\Delta(\mu_j, \sigma_j^2)$ conveys the value of learning about arm j when considering the fact that finding estimates that are within a specific subset of the parametric space is more desirable.

2.2.2 Background methodology

The foundation of our research lies in the univariate weight function defined in Caruso and Mozgunov, 2024, which assings more importance to values of μ_j that are close to a target value $\gamma \in \mathbb{R}$, assumed to be set in advance by the clinicians as the most desirable value for the treatment mean of a continuous endpoint. We recall Caruso's weight function below:

$$\phi_{\gamma}(\mu_j) = C_{\gamma}(\overline{\mu}_j, \sigma_j^2, \gamma, m_j, \kappa) \exp\left\{-\frac{1}{2}\frac{(\mu_j - \gamma)^2}{\sigma_j^p / m_j^{\kappa}}\right\}, \tag{2.4}$$

where $C_{\gamma}(\overline{\mu}_j, \sigma_j^2, \gamma, m_j, \kappa)$ is a normalisation factor, and $\kappa, p \in \mathbb{R}$ are two tuning parameters. $\phi_{\gamma}(\mu_j)$ has a Gaussian kernel centered in γ, while $\sigma_j^p / m_j^{\kappa}$ quantifies how the curve is dispersed around this target mean. The parameter p can reduce or amplify the impact of the response variability, while κ influences the rates at which $\phi_{\gamma}(\mu_j)$ concentrates near γ as the number of observations increases.

Differently from the present work, Caruso and Mozgunov, 2024 assumes an improper uniform prior on μ_j, resulting in the following information gain:

$$\Delta_j^{\gamma} = \frac{1}{2}\frac{\sigma_j^{2-p}n_j^{\kappa}}{\sigma_j^{2-p}n_j^{\kappa}+n_j} - \frac{1}{2}\frac{\left(\gamma-\bar{x}_j\right)^2}{\sigma_j^2}n_j\left(\frac{\sigma_j^{2-p}n_j^{\kappa}}{\sigma_j^{2-p}n_j^{\kappa}+n_j}\right)^2. \tag{2.5}$$

In the next section we will discuss the properties of a similar information gain, but based on a bivariate weight function. Indeed, while in (2.4) the variance is assumed as a nuisance parameter that can be fixed, in our setting μ_j and σ_j^2 are jointly modeled with a Bayesian prior. For this reason, we need to introduce a bivariate weight function $\phi(\mu_j,\sigma_j^2)$, which additionally allows us to handle more precisely the interest around specific values of σ_j^2.

In particular, we will consider a class of bivariate weight functions that can be decomposed in the product of two weight components: one relative to the conditional mean, $\phi_1(\mu_j,\sigma_j^2)$, and one relative to the variance, $\phi_2(\sigma_j^2)$. Using this hierarchical structure, the preferences regarding both the mean and the variance can be incorporated directly into the weight function. Moreover, the following theorem shows that by defining a weight function in such a way, we can simplify the expression of the associated weighted Shannon entropy.

Theorem 2.2.1. *Let $\phi_1(\mu_j,\sigma_j^2)$ and $\phi_2(\sigma_j^2)$ be weight functions satisfying the normalisation condition for $p(\mu_j|\sigma_j^2,\underline{x}_j)$ and $p(\sigma_j^2|\underline{x}_j)$, given by*

$$\int_{\mathbb{R}} \phi_1(\mu_j,\sigma_j^2)p(\mu_j|\sigma_j^2,\underline{x}_j)d\mu_j = 1 \;\forall\sigma_j^2>0 \tag{2.6}$$

and

$$\int_0^{\infty} \phi_2(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)d\sigma_j^2 = 1, \tag{2.7}$$

respectively.

Then $\phi(\mu_j,\sigma_j^2)=\phi_1(\mu_j,\sigma_j^2)\phi_2(\sigma_j^2)$ is a weight function that satisfies the normalisation condition for $p(\mu_j,\sigma_j^2|\underline{x}_j)$:

$$\int_0^{\infty}\int_{\mathbb{R}} \phi(\mu_j,\sigma_j^2)p(\mu_j,\sigma_j^2|\underline{x}_j)d\mu_j\,d\sigma_j^2 = 1.$$

Moreover,

$$h^{\phi}(\mu_j, \sigma_j^2) = \int_0^{\infty} h^{\phi_1}(\mu_j|\sigma_j^2)\phi_2(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)d\sigma_j^2 + h^{\phi_2}(\sigma_j^2). \tag{2.8}$$

Proofs for this and subsequent theorems are given in Section 2.7.

In Sections 2.2.3 and 2.2.4, we describe two distinct classes of weight functions based on the decomposition in Theorem 1: one that only gives indication on the target mean value and the other jointly targeting both mean and variance. For each class, we present the corresponding expressions for the information gain.

2.2.3 Uniform-variance weight function

We consider the same weight function as in (2.4), with $p = 2$, for the conditional mean component of the weight, although with the critical difference that σ_j^2 here is unknown and provided with its own prior distribution, rather than being an assumed constant:

$$\phi_{\gamma}(\mu_j, \sigma_j^2) = C_{\gamma}(\overline{\mu}_j, \sigma_j^2, \gamma, m_j, \kappa) \exp\left\{ -\frac{1}{2} \frac{(\mu_j - \gamma)^2}{\sigma_j^2/m_j^{\kappa}} \right\}. \tag{2.9}$$

If the variance is not of primary interest in the context of the experiment, we can assume a uniform weight for it, i.e.,

$$\phi(\sigma_j^2) = 1 \; \forall \sigma_j^2 > 0, \tag{2.10}$$

which treats values of σ^2 equally, therefore maintaining the full focus on the target mean γ. The following theorem shows the explicit expression of the ensuing information gain, as well as its asymptotic behaviour.

Theorem 2.2.2. *Given a target value $\gamma \in \mathbb{R}$ and a tuning parameter $\kappa \in \mathbb{R}$, consider the weight functions $\phi_{\gamma}(\mu_j, \sigma_j^2)$ and $\phi(\sigma_j^2)$, as given in* (2.9) *and* (2.10) *respectively. Then, the information gain of the posterior Normal inverse-Gamma distribution in* (2.1) *with the weight function $\phi(\mu_j, \sigma_j^2) = \phi_{\gamma}(\mu_j, \sigma_j^2)\phi(\sigma_j^2)$ is*

$$\Delta_j^{\gamma} = \frac{1}{2} \frac{m_j^{\kappa}}{m_j^{\kappa} + m_j} - \frac{1}{2} \left(\gamma - \overline{\mu}_j \right)^2 m_j \left(\frac{m_j^{\kappa}}{m_j^{\kappa} + m_j} \right)^2 \frac{\alpha_j}{\beta_j}. \tag{2.11}$$

Its asymptotic expression is

$$\Delta_j^{\gamma} \overset{n_j \to \infty}{\sim} -\frac{1}{2} \frac{\left(\gamma - \bar{x}_j\right)^2}{\bar{s}_j^2} \frac{n_j}{(n_j^{1-\kappa}+1)^2} + c_\kappa, \tag{2.12}$$

where

$$c_\kappa = \begin{cases} 0\,, \textit{ if } \kappa < 1 \\ 1/4\,, \textit{ if } \kappa = 1 \\ 1/2\,, \textit{ if } \kappa > 1 \end{cases}.$$

The asymptotic expression easily follows from (2.11). Then, the information gain Δ_j^{γ}, asymptotically, can be interpreted as the negative standardised distance between $\bar{x}_j$ and γ, multiplied by a power of n_j. This means that learning from arms with sample mean closer to the target and that received a smaller number of allocations is more valuable. Panel A of Figure 2.1 gives a visual representation of Δ_j^{γ} as a function of the posterior mean $\overline{\mu}_j$ (left plot), the sample variance $\bar{s}_j^2$ (central plot) and the sample size n_j (right plot). With respect to $\overline{\mu}_j$, the information gain is symmetrical around its maximum γ, meaning that sampling from the arm with mean effect closest to the target is more valuable, while values that are equally distant above or below the target are treated identically. It is monotonically increasing in $\bar{s}_j^2$, meaning that the learning value of sampling from an arm increases with the uncertainty around its effect. It is monotonically decreasing in n_j, therefore exploring less visited arms is more valuable, everything else being equal. Moreover, larger values of κ correspond to a higher concentration of the information gain around γ, with respect to $\bar{x}_j$, and to a quicker decline of the information gain, with respect to n_j.

The information gain in (2.11) is similar to the one in (2.5) for $p = 2$, since they are based on the same weight function. The difference lies in the way that μ_j and σ_j are modeled in the design: Caruso and Mozgunov, 2024 assumes an improper uniform prior for the mean and a fixed variance, while we model them jointly with a normal inverse-Gamma. Interestingly, since the prior contribution tends to be negligible as the sample size increases, the two information gains are asymptotically equivalent.

While the information gain described in this section has several desirable properties, it may overemphasize exploration of treatments with exceedingly large variances. In the

following section, we introduce a new weight function that allows one to regulate the extent to which very large or very small variances are prioritised.

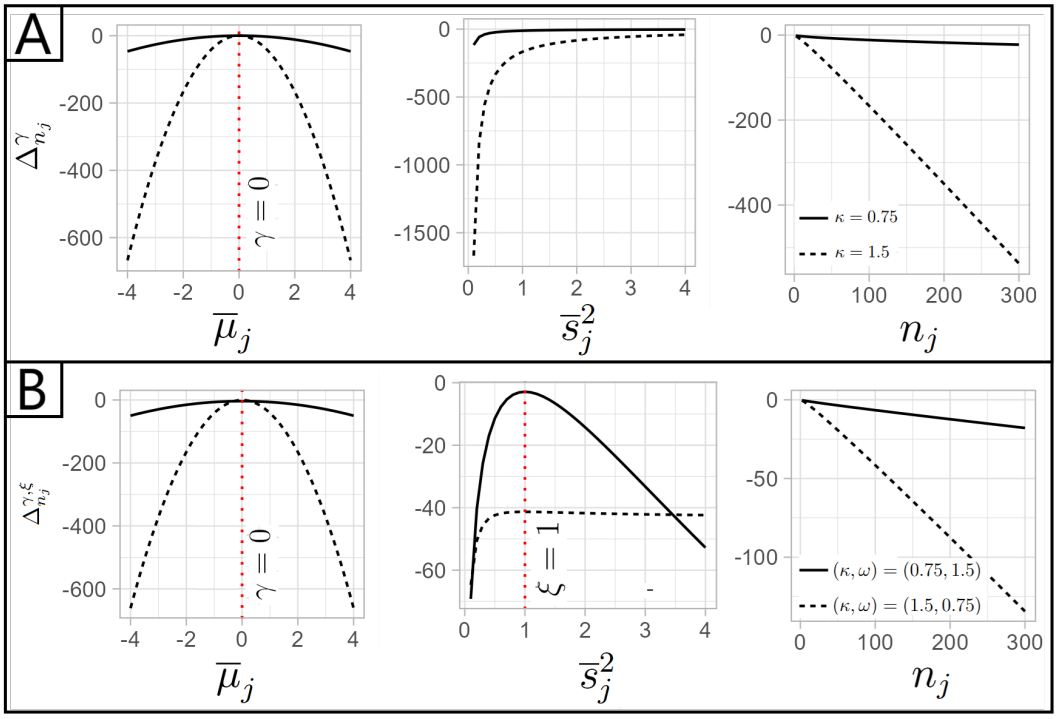

Fig. 2.1 Behaviour of the information gain Δ^{γ}_j in (2.11) (panel A) and of the information gain $\Delta^{\gamma,\xi}_j$ in (2.14) (panel B) as functions of the posterior mean $\overline{\mu}_j$, the estimated variance $\overline{s}^2_j$ and the current sample size n_{j_k} of the j-th arm. The vertical lines corresponds to fixed values of γ and ξ.

2.2.4 Targeted-variance weight function

In certain contexts, it may be preferable to focus the search for the most effective treatment on those with variance that is not excessively large. For instance, this is relevant when only a small number of patients are available, making it practically impossible to estimate the effect of treatments with high variance. It is also important when large deviations from typical values are associated with serious adverse events.

In this section, we consider a target value ξ for the treatment variance and define a weight function $\phi_\xi(\sigma^2_j)$ that is near zero for very small variances, reaches a maximum at ξ, and decreases as σ^2_j exceeds ξ. This formulation captures the interest in learning from arms with larger variance, provided it does not exceed the specified threshold ξ. A mathematically convenient choice is a weight function with the kernel of an inverse-Gamma

with mode equal to ξ, i.e.,

$$\phi_\xi(\sigma_j^2) = \frac{C_\xi(\alpha_j,\beta_j,\xi,m_j,\omega)}{(\sigma_j^2)^{m_j^\omega+1}} \exp\left(-\frac{\xi(m_j^\omega+1)}{\sigma_j^2}\right), \omega > 0, \tag{2.13}$$

where $C_\xi(\alpha_j,\beta_j,\xi,m_j,\omega)$ is a constant satisfying the normalisation condition (2.7). $\phi_\xi(\sigma_j^2)$ concentrates around ξ as the number of observations increases and the rate of this convergence can be tuned with the parameter ω. Combining (2.13) with (2.9) results in a closed-form expression of the information gain.

Theorem 2.2.3. *Given two target values $\gamma \in \mathbb{R}$, $\xi \in \mathbb{R}^+$ and two tuning parameters $\kappa, \omega \in \mathbb{R}$, consider the weight functions $\phi_\gamma(\mu_j,\sigma_j^2)$ and $\phi_\xi(\sigma_j^2)$, as given in* (2.9) *and* (2.13) *respectively. Then, the information gain of the posterior Normal inverse-Gamma distribution in* (2.1) *with the weight function $\phi(\mu_j,\sigma_j^2) = \phi_\gamma(\mu_j,\sigma_j^2)\phi_\xi(\sigma_j^2)$ is*

$$\begin{aligned}\Delta_j^{\gamma,\xi} = \alpha_j + \frac{2\alpha_j+3}{2}\left[\left(\psi\left(m_j^\omega+\alpha_j+1\right)-\psi(\alpha_j)\right)-\left(\ln\left(\xi(m_j^\omega+1)+\beta_j\right)-\ln(\beta_j)\right)\right] \\ + \frac{1}{2}\frac{m_j^\kappa}{m_j^\kappa+m_j} - \frac{m_j^\omega+\alpha_j+1}{\xi(m_j^\omega+1)+\beta_j}\left[\beta_j + \frac{1}{2}\left(\gamma-\overline{\mu}_j\right)^2 m_j\left(\frac{m_j^\kappa}{m_j^\kappa+m_j}\right)^2\right],\end{aligned} \tag{2.14}$$

where $\psi(\cdot) = \frac{d}{dz}\ln\Gamma(\cdot)$ is the digamma function (i.e., the logarithmic derivative of the gamma function).

Its asymptotic expression is

$$\Delta_j^{\gamma,\xi} \overset{n_j\to\infty}{\sim} -\frac{1}{2}\left[\frac{\overline{s}_j^2}{\lambda_j} - \ln\left(\frac{\overline{s}_j^2}{\lambda_j}\right) - 1\right] n_j - \frac{1}{2}\frac{\left(\gamma-\overline{x}_j\right)^2}{\lambda_j}\frac{n_j}{(n_j^{1-\kappa}+1)^2} + C_\kappa, \tag{2.15}$$

where

$$\lambda_j = \frac{\xi n_j^\omega + \overline{s}_j^2 n_j}{n_j^\omega + n_j}$$

and

$$C_\kappa = \begin{cases} 1/2\,, \textit{ if } \kappa < 1 \\ 3/4\,, \textit{ if } \kappa = 1 \\ 1\,, \textit{ if } \kappa > 1. \end{cases}$$

The asymptotic information gain is maximum, and equal to C_κ, when $\bar{x}_j = \gamma$ and $\bar{s}_j^2 = \xi$. By comparing the solid lines with the dotted lines in Figure 2.1, we can observe that larger values of κ and ω correspond to a higher concentration of the information gain (hence to a greater interest) around the target values. $\Delta_j^{\gamma,\xi}$ goes asymptotically to $-\infty$ when $\kappa > 0.5$ or $\omega \geq 1$, meaning that the context of the study becomes more relevant as more information is collected on a specific arm. Conversely, when $\omega < 1$ the first term of (2.15) converges to a constant, and when $\kappa \leq 0.5$ the same happens to the second term. In those cases, the information which is gained by including the context would be relevant only for small sample sizes.

The first and second terms of (2.15) are non-positive values multiplied by a power of n_j. In particular, $\left[\frac{\bar{s}_j^2}{\lambda_j} - \ln(\frac{\bar{s}_j^2}{\lambda_j}) - 1\right]$ is a measure of the distance between the sample variance $\bar{s}_j^2$ and λ_j, expressed as a ratio (which has been argued to be an appropriate distance definition on the positive real line by Mozgunov et al., 2019). λ_j is a weighted average of $\bar{s}_j^2$ and ξ depending on ω. For $\omega < 1$, λ_j converges to $\bar{s}_j^2$, and therefore $\left[\frac{\bar{s}_j^2}{\lambda_j} - \ln(\frac{\bar{s}_j^2}{\lambda_j}) - 1\right]$ tends to zero. For $\omega > 1$, λ_j converges to ξ. For $\omega = 1$, we obtain exactly the distance between $\bar{s}_j^2$ and $\frac{\bar{s}_j^2+\xi}{2}$. Therefore, we can tune the gain in estimating a treatment variance close to ξ by calibrating ω.

Similarly, $\frac{(\gamma-\bar{x}_j)^2}{\lambda_j}$ in the second term is the quadratic distance between the sample mean and γ, standardised by the same weighted average λ_j. This standardised squared distance is multiplied by a power of n_j depending on κ, which can be calibrated in order to tune the gain in estimating a treatment mean close to γ.

The information gain $\Delta_j^{\gamma,\xi}$ is the learning value associated with sampling from arm j when taking into account the context of the study, but, asymptotically, it can also be interpreted as the negative distance between $(\bar{x}_j, \bar{s}_j^2)$ and (γ, ξ), multiplied by a power of n_j. It follows that learning is most valuable from arms that are both closely aligned with the clinical targets and less frequently allocated. The trade-off between exploring under-sampled arms and prioritising arms near the target values can be controlled through the choice of the tuning parameters κ and ω.

Panel B of Figure 2.1 gives a visual representation of $\Delta_j^{\gamma,\xi}$ as a function of $\bar{\mu}_j$ (left plot), $\bar{s}_j^2$ (central plot) and n_j (right plot) for two different combinations of (κ, ω). For

$(\kappa, \omega) = (1.5, 0.75)$ the information gain varies substantially as μ_j gets further from γ, but only modestly when σ_j^2 increases over ξ. On the other hand, for $(\kappa, \omega) = (0.75, 1.5)$ we observe the opposite behaviour. Thus the relative values of κ and ω can determine the impact that different values of the mean and variance have on the information gain.

In the same figure we can also observe that $\Delta_j^{\gamma,\xi}$ (Panel B), as a function of $\overline{\mu}_j$ (left plot), is symmetrical around its maximum γ, like Δ_j^{γ} (Panel A). Furthermore, both $\Delta_j^{\gamma,\xi}$ and Δ_j^{γ} are decreasing in n_j (right plots). However, relative to $\overline{s}_j^2$ (central plots), the trends of the two information gains differ: Δ_j^{γ} is monotonically increasing, while $\Delta_j^{\gamma,\xi}$ increases rapidly only for values below ξ and slowly decreases for values above ξ, reflecting a preference for exploring arms with greater uncertainty, as long as their variance is not excessively above ξ. This difference is crucial, as it prevents situations in which an undesirable treatment with very large variance could provide the greatest information gain, a risk that arises when using the information gain defined in (2.11).

2.2.5 Design based on the context-dependent measure

Because of its easier interpretation and desirable properties, we base our randomisation methodology on the asymptotic expression of the information gain found in (2.15). In Section 2.7.6 we also provide a criterion based on the asymptotic information gain (2.12), derived from the uniform-variance weight function.

The constant part C_k is dropped from (2.15) as it would be the same for all treatment arms, making it irrelevant in the decision making. We obtain the following criterion:

$$\widehat{\Delta}_j^{\gamma,\xi} = -\frac{1}{2}\left[\frac{\overline{s}_j^2}{\lambda_j} - \ln\left(\frac{\overline{s}_j^2}{\lambda_j}\right) - 1\right] n_j - \frac{1}{2}\frac{\left(\gamma - \overline{x}_j\right)^2}{\lambda_j}\frac{n_j}{(n_j^{1-\kappa}+1)^2}.$$

$\widehat{\Delta}_j^{\gamma,\xi}$ attains its maximum value of zero at $(\overline{x}_j, \overline{s}_j^2) = (\gamma, \xi)$ and is strictly negative elsewhere.

To have a first estimate of the sample means and variances in all of the K arms, it is necessary to begin with a burn-in phase: the first $B \cdot K$ patients are allocated in equal batches of $B \geq 2$ individuals to all of the K arms. After the burn-in phase, each newly recruited patient is randomised according to the following rule.

Given $n = \sum_{j=1}^{K} n_j$ responses from the previously treated patients, patient $n+1$ is randomised to arm $j \in \{1, ..., K\}$ with probability

$$p_j = \begin{cases} 1/K & , \text{if } j = 1 \\ \dfrac{(\widehat{\Delta}_j^{\gamma,\xi})^{-1}}{\sum_{j=1}^{K}(\widehat{\Delta}_j^{\gamma,\xi})^{-1}} \dfrac{K-1}{K} & , \text{if } j > 1. \end{cases} \tag{2.16}$$

Compared to Caruso and Mozgunov, 2024's criterion, our methodology is fully randomised, instead of deterministic, and is intended for trials that include a control arm. Compared to an equal randomisation strategy, which would allocate, on average, $1/K$ patients to each arm, we only fix that probability for the control arm, $j = 1$, in order to guarantee a sensible comparison of the $K-1$ treatments with the control. The idea of protecting the control arm in RAR is well-established in the scientific literature. In particular, Trippa et al., 2012 demonstrated that protecting the control group can improve the power of adaptive trials with multiple treatment arms.

The remaining $(K-1)/K$ of the probability is distributed between the treatment arms, proportionally to the inverse of $\widehat{\Delta}_j^{\gamma,\xi}$. This means that the treatment arm with the information gain closest to zero has the highest randomisation probability. This approach aligns with our focus on finding arms whose mean and variance are closest to the target values.

It is worth mentioning that since the tuning parameters κ and ω affect the information gain, they also have a direct impact on the probabilities defined in (2.16). In particular, larger values of ω tend to amplify the effect that differences in variance have on the information gain, making the randomisation probabilities more sensitive to variability in the responses. The parameter κ, instead, decreases an arm's randomisation probability as additional observations are collected for it, thereby increasing the opportunities for other arms to be explored. In Section 2.4, we outline strategies for selecting κ and ω in accordance with the trial objectives, while in Section 2.5 we provide additional insights into how these parameters influence the operating characteristics of the proposed design.

2.3 Procedure for best treatment arm's identification

2.3.1 Definition of best treatment arm and hypothesis testing procedure

The primary objective of a multi-arm trial is typically that of identifying the best arm among many, according to some definition of "best". A key advantage of our approach is precisely its flexibility in allowing different definitions of "best" to be incorporated within the same framework. This enables the design to align more closely with diverse clinical or decision-making objectives, such as optimizing mean outcomes, minimizing risk, or targeting other relevant criteria.

In this section we justify the use of a natural definition of "best" treatment arm, based on the proximity of the treatment means to the target γ. Theoretical and empirical results for an alternative definition of the best arm, which also accounts for treatment variance, are provided in Section 2.7.4.

Let us consider the vector of the true mean treatment effects $(M_1,...,M_K)$ and the vector of true variances $(V_1,...,V_K)$ of K arm, where arm 1 is the control group. Among the $K-1$ other alternatives, let

$$j^\star = \underset{j=2,...,K}{\operatorname{argmin}} \left|M_j - \gamma\right| \tag{2.17}$$

be the best treatment arm. We can estimate the best arm $\widehat{j^\star}$ at the end of the trial by comparing the posterior distributions of the treatments means $\mu_j|\underline{x}_j$, $j=2,...,K$, with the posterior distribution of the control mean $\mu_1|\underline{x}_1$, as follows:

$$\widehat{j^\star} = \underset{j=2,...,K}{\operatorname{argmax}} P\left(\left|\mu_j - \gamma\right| < \left|\mu_1 - \gamma\right| \,\middle|\, \underline{x}_1, \underline{x}_j\right). \tag{2.18}$$

The marginal posterior distribution of $\mu_j|\underline{x}_j$ follows a location-scale version of the t distribution with $2\alpha_j$ degrees of freedom, i.e.,

$$\mu_j|\underline{x}_j \sim t_{2\alpha_j}\left(\overline{\mu}_j, \frac{\beta_j}{\alpha_j m_j}\right).$$

Notice that the prior specification affects the Bayesian parameter inference described here, while it does not influence the randomisation rule (2.16), as previously mentioned at the beginning of Section 2.5. It is possible to compute the probability in (2.18) by numerical

integration or by Monte Carlo simulations. We correctly identify the best arm if we claim that $\widehat{j^\star} = j^\star$ at the end of trial.

This definition is consistent with a typical hypothesis testing framework for assessing if any of the treatment means is better than the control mean. In particular, we propose a testing procedure that assesses whether M_j is significantly closer to the target γ than M_1. The rigorous definition of the pairwise hypothesis testing is

$$H_{0,j} : \left|M_j - \gamma\right| = |M_1 - \gamma| \quad \text{VS} \quad H_{1,j} : \left|M_j - \gamma\right| < |M_1 - \gamma| \tag{2.19}$$

for $j = 2, ..., K$.

We reject $H_{0,j}$ and claim the superiority of arm j against control if

$$P\left(\left|\mu_j - \gamma\right| < |\mu_1 - \gamma| \;\middle|\; \underline{x}_1, \underline{x}_j\right) > \eta, \tag{2.20}$$

for a fixed cut-off probability $\eta \in (0,1)$. Under the global null hypothesis, all the treatment means are at the same distance from the target as the control mean, i.e.,

$$H_0^G : \left|M_j - \gamma\right| = |M_1 - \gamma| \;\forall j. \tag{2.21}$$

We make a type I error if we incorrectly reject at least one $H_{0,j}$, for j in $2, ..., K$. Since η is the same for all pairwise comparisons, and by definition of the estimated best arm in (2.18), this is equivalent to incorrectly rejecting $H_{0,\widehat{j^\star}}$. We define the family-wise error rate (FWER) as the probability of rejecting $H_{0,\widehat{j^\star}}$ under the global null hypothesis, i.e.,

$$\text{FWER} = P\left[P\left(\left|\mu_{\widehat{j^\star}} - \gamma\right| < |\mu_1 - \gamma| \;\middle|\; \underline{X}_1, \underline{X}_{\widehat{j^\star}}\right) > \eta \;\middle|\; H_0^G\right], \tag{2.22}$$

where $\underline{X}_j = (X_{1,j}, \ldots, X_{n_j,j})$ is the random vector of the observations in arm j. In the next subsection, we describe how to select η to control the FWER under a pre-specified threshold.

As for the power of the trial, we define it as the probability of rejecting $H_{0,j^\star}$, under the alternative $H_{1,j^\star}$, i.e.,

$$1 - \beta = P\left[P\left(\left|\mu_{j^\star} - \gamma\right| < |\mu_1 - \gamma| \;\middle|\; \underline{X}_1, \underline{X}_{j^\star}\right) > \eta \;\middle|\; H_{1,j^\star}\right].$$

2.3.2 Control of the FWER

The global null hypothesis consists of an infinite set of scenarios satisfying the condition in (2.21) and with possibly heterogeneous variances, therefore it is not possible to explore every single one. We recommend focusing on a subset of plausible null scenarios, by identifying the cut-off value η that controls the FWER for them, as defined in (2.22). Which scenarios are plausible should be determined by the investigators, based upon what is known about the endpoint and the treatments.

Let us consider a subset of plausible null scenarios S_{H_0}. Each scenario $S \in S_{H_0}$ is characterised by a vector of means $(M_1, ..., M_K)$, such that (2.21) is satisfied, and by a vector of positive variances $(V_1, ..., V_K)$. The FWER is controlled for all of the scenarios in S_{H_0} if

$$P\left[P\left(\left|\mu_{\widehat{j}^\star} - \gamma\right| < |\mu_1 - \gamma| \;\middle|\; \underline{X}_{n_1}, \underline{X}_{n_{\widehat{j}^\star}}\right) > \eta \;\middle|\; S\right] < \text{FWER} \quad \forall S \in S_{H_0}.$$

Operatively, we need to find the cut-off value η_S that gives the desired FWER for each $S \in S_{H_0}$ and then take the maximum

$$\eta = \max_{S \in S_{H_0}} \eta_S.$$

Since the cut-off value can only be obtained through simulations, we can estimate η in a design with equal randomisation, which is less computationally expensive, and then apply that same η to compute the FWER (and other relevant operating characteristics) of other randomisation methods. This allows us to make a quick comparison between the FWER obtained under fixed randomisation with that achieved by other methods. In Section 2.5 we present a simulation study, which includes an application of the discussed strategy for calibrating η.

2.4 Robust strategies for selecting κ and ω

2.4.1 Selection strategy

In this section, we propose an approach to select the robust optimal values of κ and ω that emerge from evaluating a large variety of plausible alternative scenarios. By "robust optimal", we mean a parameter configuration that, on average, maximises a specific operating characteristic of interest (e.g., average proportion of patients allocated to the

best arm, statistical power, percentage of times the best arm is selected in repeated trials). Other than the treatment means and variances, that we assume to be unknown, there are more variables that influence the optimal choice of κ and ω, like the total sample size, the burn-in size, the number of arms and the target values. For simplicity, we make the assumption that those can be fixed in advance by the investigators and the clinicians.

Let us consider a set S_{H_1} of plausible alternative scenarios and a grid of values for κ and ω. Notice that an alternative scenario has to satisfy $H_{1,j}$ as in (2.19) for at least one $j \in \{2,...,K\}$. Then, the procedure is as follows:

1. Select an operating characteristic of interest. We denote by $u^S(\kappa,\omega)$ its value under a scenario $S \in S_{H_1}$ for fixed κ and ω.

2. Define an objective function $g\left(u^S(\kappa,\omega)\right)$ to evaluate the selected operating characteristic.

3. Compute $u^S(\kappa,\omega)$ for each scenario $S \in S_{H_1}$ and for all values of κ and ω in the considered grid.

4. Find the robust optimal pair $(\kappa^\star,\omega^\star) = \text{argmin}_{(\kappa,\omega)} \sum_{S\in S_{H_1}} g\left(u^S(\kappa,\omega)\right)$.

The selected pair is the optimal choice for the given operating characteristic, on average. In this work we shall evaluate the robust optimal pair $(\kappa^\star,\omega^\star)$ for three different operating characteristics.

2.4.2 Objective function

Following Caruso and Mozgunov, 2024, we evaluate the following operating characteristics (for any alternative scenario $S \in S_{H_1}$ and a large grid of values of κ and ω) by simulation over M pseudo-trials with a fixed total sample size N. We will omit the dependence on S, κ and ω for readability.

Let us define the patient benefit (PB), that is, the average proportion of patients allocated to the best arm, as

$$PB = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\text{patient } i \text{ of pseudo-trial } m \text{ is assigned to arm } j^\star)\right\}.$$

Let us define the percentage of correct selections (PCS), that is, the percentage of times the best arms is selected at the end of the trial, as

$$PCS = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(\widehat{j^\star} = j^\star).$$

Lastly, let us evaluate the power of rejecting the null hypothesis for the best treatment as

$$1-\beta = \frac{1}{M}\sum_{m=1}^{M}\left[P\left(\left|\mu_{j^\star}-\gamma\right| < \left|\mu_1-\gamma\right| \,\middle|\, \underline{x}_{n_1}, \underline{x}_{n_{j^\star}}\right) > \eta\right].$$

For PB, we propose to adopt the following squared difference as the objective function:

$$g_1\left(u^S(\kappa,\omega)\right) = \left(u^S(\kappa,\omega) - \max_{(\kappa,\omega)} u^S(\kappa,\omega)\right)^2,$$

quantifying how far each value $u^S(\kappa,\omega)$ is from the highest achievable value in the given scenario S. The idea is to select the robust optimal pair $(\kappa^\star,\omega^\star)$ that minimises the average (squared) loss of PB among the considered alternative scenarios.

For PCS and $1-\beta$, we propose a binary objective function that returns one if the power achieved under the information theoretic approach is smaller than the power achieved under fixed randomisation:

$$g_2\left(u^S(\kappa,\omega)\right) = \mathbb{1}\left(u^S(\kappa,\omega) < u^S_{FR}(\kappa,\omega)\right),$$

where $u^S_{FR}(\kappa,\omega)$ is the power attained using fixed randomisation. In this case, the robust optimal pair $(\kappa^\star,\omega^\star)$ is the one which underperforms fixed randomisation, in terms of PCS or $1-\beta$, in the smallest number of simulated scenarios.

2.5 Simulation study of a multi-arm trial

2.5.1 Study setting

Consider a multi-arm Phase II trial with $K = 4$ arms, one for the control and three for alternative treatments (or doses), and with a continuous endpoint. We assume a total

sample size of $N = 100$. The objective is to detect the superiority of the treatment with effect closest to γ, as explained in Section 2.3. For simplicity we assume $\gamma = 0$, meaning that the best treatment is the one with the smallest mean response in absolute value. We assume a target variance of $\xi = 2$.

Let us assume we are dealing with new treatments of which we have very limited knowledge, by considering weakly informative Normal inverse-gamma priors for the mean and the variance of each arm, as explained at the beginning of Section 2.2. Specifically, we set $\nu = \alpha_j^{(0)} = \beta_j^{(0)} = 0.0001$ and $\mu_j^{(0)} = 0$, $j = 1,...,4$.

In order to assess the performance of our method, we investigate 500 alternative scenarios. Each scenario is randomly generated as follows:

1. a vector of means (M_1, M_2, M_3, M_4) is generated by sampling four times from a $Unif[\gamma - 3\sqrt{\xi}, \gamma + 3\sqrt{\xi}]$ and assigning the furthest value from γ to M_1, with the first arm acting as the control.

2. a vector of variances (V_1, V_2, V_3, V_4) is generated by sampling four times from a $Unif[\xi/2, 2\xi]$.

Operating characteristics of interest are $1 - \beta$, PB and PCS. For each given scenario, the operating characteristics are assessed over 10^4 pseudo-trials via Monte Carlo simulations. We also evaluate the FWER in correspondence of specific null scenarios.

2.5.2 Competing designs

In the following, we will denote TWE(κ, ω) the class of designs based on the targeted-variance weighted entropy which follow the randomisation rule in (2.16). We will denote UWE(κ) the class of designs which make use of a modification of that randomisation rule, based on the uniform-variance weight function (2.10) (more details can be found in Section 2.7.6). For TWE(κ, ω) we will consider three optimal pairs of κ and ω obtained through the robust strategies explained in Section 2.4. For UWE(κ), we will use two of the same κ values identified in the optimal pairs, omitting the unnecessary parameter ω.

We compare the performance of TWE(κ, ω) against UWE(κ) and three other alternative randomisation procedures which have good properties either in terms of statistical power or patient benefit:

1. Fixed and equal randomisation (FR), where each patient can be treated in a specific arm with fixed probability $1/K$.

2. Randomised Thompson sampling (RTS), where the randomisation probabilities are proportional to the adjusted posterior probability of being the best, i.e., patient n is randomised to arm $j \in \{2,...,K\}$ with probability

$$p_j = P(j = j^\star|\text{data})^c = P\left(j = \underset{j}{\text{argmin}}\,|\mu_j - \gamma| \,\middle|\, \underline{x}_2,...,\underline{x}_K\right)^c,$$

   where the exponent $c = n/(2N)$ helps stabilise the resulting allocations. This method is a randomised version of Thompson sampling (Thompson, 1933).

3. Randomised targeted Gittins index (RGI), which is a modification of the deterministic allocation procedure outlined in Smith and Villar, 2018 that sequentially allocates patients to the arm with highest Gittins index. The Gittins index is a measure of the effectiveness of a treatment combined with its discounted variance, based on the number of previous allocations. Caruso and Mozgunov, 2024 adapts Smith and Villar's proposal to the case where the best treatment is not defined as the one with highest (or lowest) effect size, but the one with mean closest to a target value γ. It is important to mention that their modification is based on some heuristic justifications, rather than on analytical or statistical considerations. RGI is simply a randomised version of such modification in which the allocation probabilities are proportional to a modified Gittins index. More details on this methodology can be found in Section 2.7.7.

   It is important to note that the more recent study by Williamson and Villar, 2020 introduces a randomised procedure based on the Gittins index and that formally incorporates the same form of control protection considered in this work. However, we chose to exclude this method from our comparative simulation study because it employs block randomisation with block sizes $b \geq 2$, whereas our focus is on procedures that randomise patients individually, one at a time.

In order to present a fair comparison with TWE(κ, ω), the probability of allocation to control was fixed to $1/K = 0.25$ in all of the designs. In practice, the patients are first assigned to either the control arm (with probability 0.25) or any of the treatment arms

(with probability 0.75). The patients assigned to the treatment arms are then randomised to a specific one following the rules of the selected method.

Following the considerations in Caruso and Mozgunov, 2024, we fix the burn-in size for all RAR designs to $B = 5$.

2.5.3 Calibration of η, κ and ω

The cut-off value η to be used in the rejection rule (2.20) is obtained using the method described in Section 2.3.2. In particular, we considered a set of 49 null scenarios S_{H_0} and, for each $S \in S_{H_0}$, we estimated the cut-off value η_S that controls the FWER at the level 0.05 when applying FR. The null scenarios considered and the corresponding cut-off values are represented in Figure 2.2 (more information on how we selected the null scenarios is given in Section 2.7.5).

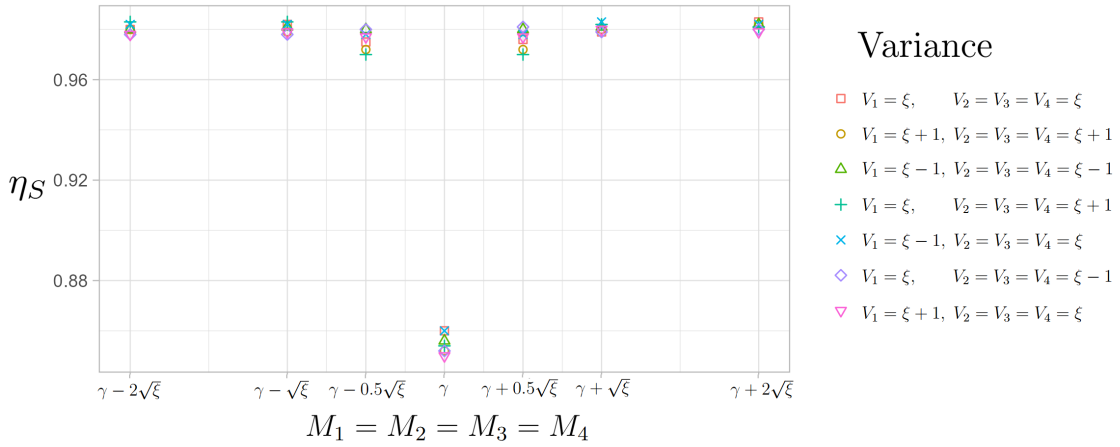

Fig. 2.2 Cut-off values η_S that control the FWER$= 0.05$ in 49 different null scenarios, under FR. Each null scenario is characterised by a common mean (on the x-axis) and a vector of variances (different shapes). The target values are fixed to $\gamma = 0$ and $\xi = 2$. The cut-off value is smallest when the means are equal to γ, and then rapidly increases to a plateau for means farther from γ. The variance, on the other hand, has only a marginal impact on η_S. The cut-off value that is selected for the actual study is $\eta = 0.983$.

We obtained $\max_{S \in S_{H_0}} \eta_S = 0.984$ from simulating 10^4 times each scenario, but we decided to consider the cut-off value $\eta = 0.983$ instead, since it resulted in only two scenarios having a negligible inflation of the FWER up to 0.051, while at the same time increasing the overall power of the study.

Following the procedure outlined in Section 2.4, we also selected the robust optimal pairs to be employed in the TWE(κ, ω) design by conducting simulations over a comprehensive grid of κ and ω values across the 500 randomly generated alternative scenarios. Since this is a time-consuming grid search, we restricted the research of the optimal tuning parameters in the range $[0.1, 1.9]$ for κ and $[0.1, 1.7]$ for ω. Higher values could be considered, but, in our investigation, they resulted in similar operating characteristics that could be achieved at the extremes of the given intervals, making them redundant.

The results of the robust selection strategy are illustrated in Figure 2.3. The optimal pair for patient benefit is $(0.3, 0.1)$, for the percentage of correct selection it is $(1.7, 0.8)$ and for power it is $(1.3, 1.3)$. Therefore those are the pairs considered for the TWE methods in the simulation study. For the UWE methods, we selected the κ values 0.3 and 1.3, taken from the optimal pairs above for patient benefit and power.

PCS and power are generally higher for larger κ, however we did not observe a strictly monotonic behaviour. Moreover, the optimal pairs selected for the two operating characteristics showed only a marginal advantage over other similar pairs. Therefore, applying the selection strategy to slightly different alternative scenarios might lead to slightly different optimal pairs.

PB, instead, is highest when κ and ω are close to zero, and it is generally decreasing in ω. Moreover, among the pairs considered, the optimal one shows a distinctly superior performance in this case.

2.5.4 Average performance in randomly generated scenarios

We applied the cut-off value η that was obtained with FR to all the other considered designs in order to evaluate their FWER under the 49 null scenarios, as shown in Figure 2.4. Almost all the scenario/method combinations present a FWER below 0.05, the only exceptions being the aforementioned two scenarios under FR and three other scenarios under TWE$(0.3, 0.1)$, which showed a small amount of inflation below the 0.055 level. This is likely due to the bias introduced by the optimal parameters for patient benefit, as treatments that appear to perform well are more likely to be explored further, even when their apparent success is due to random chance and they are, in fact, suboptimal. The outliers below 0.01 consist of the null scenarios where all treatments are sharing the common mean γ.

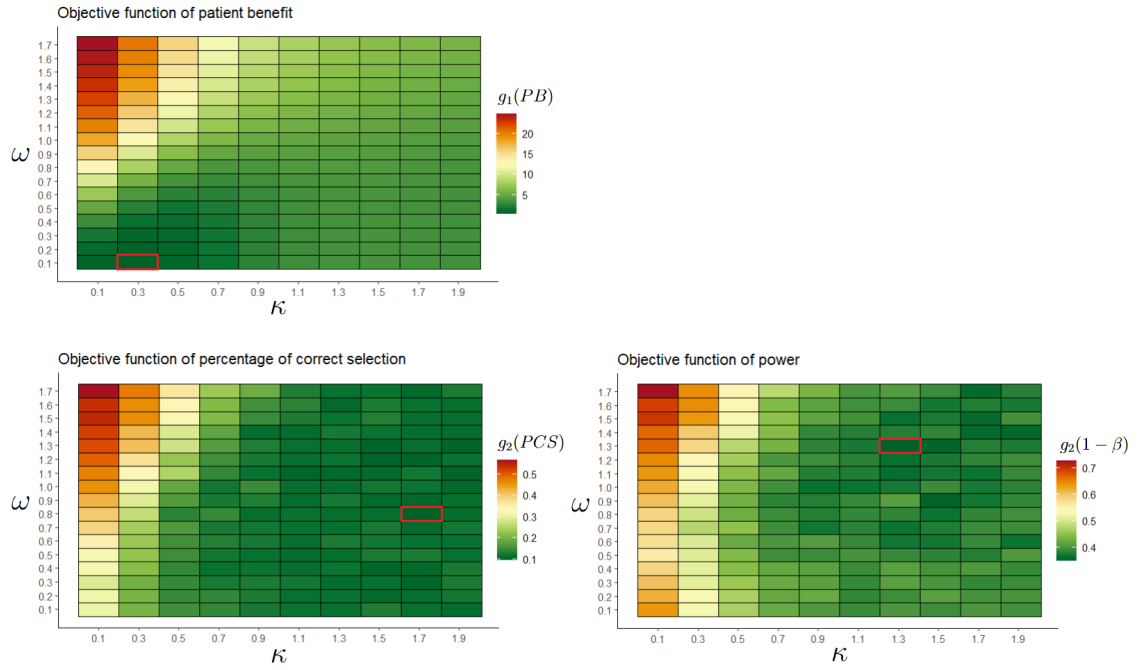

Fig. 2.3 Evaluation of PB, PCS and $1-\beta$ with the objective functions defined in Section 2.4, under different parameter configurations of TWE(κ,ω). The optimal pair of $(\kappa^\star,\omega^\star)$, corresponding to the smallest value, is highlighted with a red box.

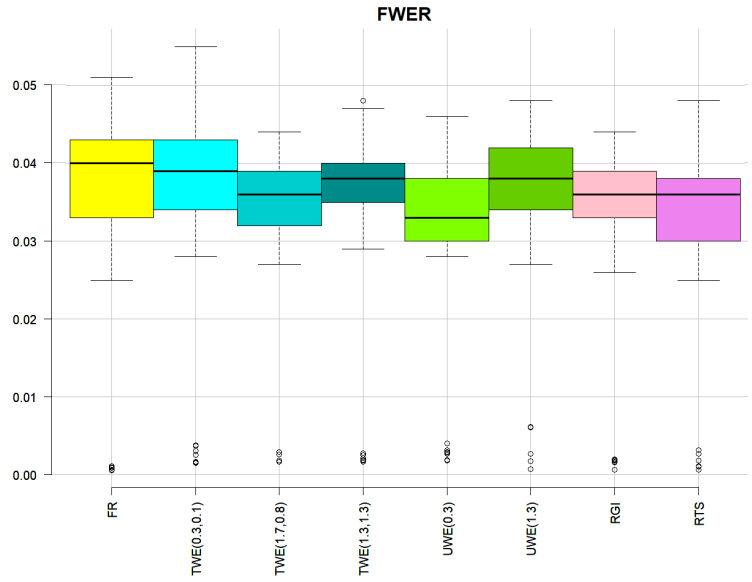

Fig. 2.4 FWER for the considered set of designs across the 49 null scenarios, when the cut-off value is $\eta = 0.983$. Each null scenario is characterised by a common mean and a vector of variances. The target values are fixed to $\gamma = 0$ and $\xi = 2$.

Figure 2.5 presents a comprehensive summary of the operating characteristics of the evaluated designs, aggregating the results across all 500 alternative scenarios randomly generated as described in Section 2.5.1. In many of the scenarios considered in our simulation, both the statistical power and the percentage of correct selection were very close to, or exactly, 100%. The cause of this lies in the way we generated the scenarios, which resulted in many instances where the control arm was clearly inferior to the best-performing treatment arm. The outliers correspond to the fewer cases where the advantage of the alternative treatments over the control was minimal; in these scenarios, both power and percentage of correct selection declined.

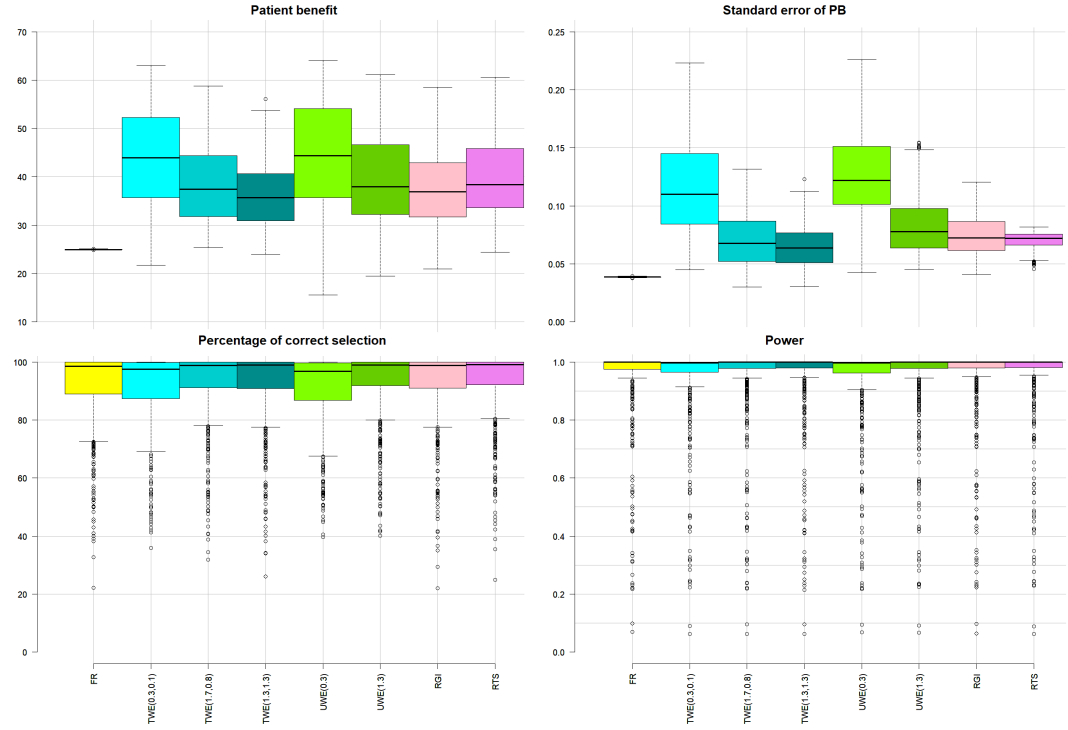

Fig. 2.5 Operating characteristics across $S = 500$ randomly generated scenarios for the considered set of designs, when the first definition of best arm (2.17) is used. For all scenarios, we assume $K = 4$, $N = 100$. The target values are fixed to $\gamma = 0$ and $\xi = 2$. Burn-in size is fixed to $B = 5$.

All the designs showed very similar percentage of correct selection and power. In terms of patient benefit, TWE$(0.3, 0.1)$ and UWE(0.3) are by far the best designs, on average. TWE$(1.7, 0.8)$, TWE$(1.3, 1.3)$, UWE(1.3) RTS and RGI displayed comparable patient benefit, usually superior to FR.

The UWE designs performed slightly better than the TWE designs, as the variances of the randomly generated scenarios we considered were drawn from a narrow range, i.e.

$[\xi/2, 2\xi]$, which limited the effectiveness of the variance-targeting component of the TWE approach. The UWE designs fail when an inferior treatment is associated with large variance, as we will discuss later while analysing some additional scenarios. Interestingly, a high level of patient benefit does not necessarily imply a reduction in statistical power. This is largely due to the fixed allocation probability to the control arm, which typically ensures that a sufficient number of patients are randomised to it. As a result, designs that increase patient benefit often lead to trials where the majority of patients contribute to the comparison between the best-performing arm and the control, ultimately yielding high power.

We also reported the standard error associated to the patient benefit, which quantifies the precision of our estimate of the average number of patients allocated to the best arm. The standard error is highest for TWE$(0.3, 0.1)$ and UWE(0.3), reflecting the high variability in the final number of patients assigned to the best arm under these two designs. This variability arises from their patient benefit-oriented parameter configuration, which heavily skews the randomisation probabilities toward any arm that produced desirable outcomes. As a result, these designs can sometimes allocate a large number of patients to an inferior arm that appeared promising early in the trial.

In contrast, FR usually exhibits the smallest standard error due to its fixed randomisation probabilities. TWE$(1.3, 1.3)$ and TWE$(1.7, 0.8)$ also exhibit relatively small standard errors (at times even smaller than that of FR) due to their emphasis on learning, which leads to more consistent patient allocations. RGI and RTS have similar standard errors, on average, compared to TWE$(1.3, 1.3)$ and TWE$(1.7, 0.8)$, but never less than that of FR.

2.5.5 Designs optimality

In Table 2.1 we reported for each design the percentage of the 500 evaluated scenarios in which it outperformed the others or was tied with the best design in terms of each operating characteristic, along with a direct comparison with FR. TWE$(0.3, 0.1)$ is optimal for patient benefit in 27.2% of the scenarios and underperforms FR only once, in a case where all treatments are clearly superior to control and quite similar to each other, but the best arm has the furthest variance from ξ (see Scenario VI in Table 2.3). UWE(0.3), which is the other patient benefit-oriented design, is optimal in 56.4% of the scenarios, but it loses to FR more often than TWE$(0.3, 0.1)$, because of its tendency of favouring high-variance treatments. When the treatment variances are in a narrow range or when the

Table 2.1 Percentage of scenarios in which each design outperformed or matched all the others in terms of patient benefit (Column 1), percentage of correct selection (Column 3), and power (Column 5). The sum of those percentages can go over 100% because of ties. Even-numbered columns show the percentage of scenarios where each design outperformed FR.

	PB		PCS		$1-\beta$	
Design	optimal or tied for optimal	$\geq \text{PB}_{FR}$	optimal or tied for optimal	$\geq \text{PCS}_{FR}$	optimal or tied for optimal	$\geq 1-\beta_{FR}$
FR	0.0		14.8		51.0	
TWE(0.3,0.1)	28.2	99.8	11.4	61.4	25.4	48.8
TWE(1.7,0.8)	0.0	100	26.8	82.8	39.2	71.6
TWE(1.3,1.3)	0.8	99.6	30.4	84.2	39.4	72.2
UWE(0.3)	56.6	96.2	18.2	45.0	22.4	38.6
UWE(1.3)	0.0	97.0	37.2	73.4	44.0	66.6
RGI	1.6	98.4	18.2	75.8	36.2	69.0
RTS	12.8	99.6	39.2	88.0	46.6	70.0

best treatment has large variance, UWE(0.3) allocates more patients than TWE$(0.3,0.1)$ to the best arm. However, it performs poorly when an inferior arm has the largest variance (see Scenario V in Table 2.3, for example). RTS is optimal for patient benefit in 12.2% of the scenarios and RGI in 1.6%. TWE$(1.7,0.8)$, which is never optimal, is the only design that never underperforms FR. The other designs that are never optimal are FR and UWE(1.3).

RTS emerged as the best design in terms of the percentage of correct selection, being optimal (or tied with other optimal designs) for this operating characteristic in 39.2% of the scenarios. Additionally, it outperforms FR in 88% of the cases, more frequently than any other design. UWE(1.3) is the second best design for percentage of correct selection, followed by TWE$(1.3,1.3)$ and TWE$(1.7,0.8)$. However UWE(1.3) loses to FR more often than the TWE designs. UWE(1.3) is the design that performs worst against FR, achieving a greater or equal percentage of correct selections in only 45% of the scenarios.

FR is optimal (or tied with other optimal designs) for power in 51% of the scenarios, followed by RTS, at 46.6%, and UWE(1.3), at 44.0%. The total exceeds 100% because ties are counted for each method involved. Although TWE$(1.7,0.8)$ and TWE$(1.3,1.3)$ are slightly less frequently optimal for power, they outperform FR more often than any

other design. TWE$(0.3, 0.1)$ and UWE(0.3), which prioritise patient benefit, are optimal for power in only 25.4% and 22.4% of scenarios, respectively.

Overall, TWE$(0.3, 0.1)$ and UWE(0.3) appear as the best designs for patient benefit, though lacking in percentage of correct selection and power. RTS, instead, performs particularly well on those two operating characteristics, followed by TWE$(1.7, 0.8)$, TWE$(1.3, 1.3)$ and UWE(1.3). RGI often outperforms several other designs on individual operating characteristics, but it is rarely the unequivocally optimal design. Notably, it is less frequently optimal than RTS across all categories. FR usually results in the highest power, albeit by a narrow margin, but at the cost of a substantially worse patient benefit.

2.5.6 Characterisation of the scenarios based on the optimal design for patient benefit

Since no ties for the optimal design with respect to patient benefit were observed across the 500 scenarios, we are able to characterise, without ambiguity, the conditions under which each design yields the greatest patient benefit.

In Table 2.2 we reported the number of times each design is optimal in such operating characteristic, together with summary measures for the treatment effect of the best arm (denoted by $j^\star$) and the second-best arm (denoted by $j^{\star\star}$). TWE$(0.3, 0.1)$ and UWE(0.3), as the more patient benefit-oriented designs, tend to be optimal in scenarios where $\mu_{j^\star}$ is clearly closer to γ than $\mu_{j^{\star\star}}$. Between these two, TWE$(0.3, 0.1)$ "wins" when $\sigma^2_{j^\star}$ is closer to the target variance $\xi = 2$, whereas UWE(0.3) performs better when the best arm is also the one with the greatest variance. TWE$(1.3, 1.3)$ is optimal only in 4 scenarios, which are characterised by both the best and the second-best arms having means very close to γ while $\sigma^2_{j^\star}$ is the variance closest to $\xi = 2$. An explanation is that TWE$(1.3, 1.3)$, as the TWE design with largest ω, favours treatments with variances close to ξ more strongly than any other design, when the means are similar to one another. On the other hand, when the variance of the best arm is notably smaller than ξ, all the weighted entropy designs tend to reduce the number of allocations to it. In those scenarios RGI and RTS result in higher patient benefit. RTS is optimal in scenarios where all treatment means, including $\mu_{j^\star}$, lie relatively far from γ. This behaviour arises because RTS does not account for the overall context of the experiment; it simply prioritises arms whose characteristics are closest to the target values, even when those are not clinically relevant. RGI, instead, is

optimal when all the other designs struggle, that is when the means have similar values and $\sigma^2_{j^\star}$ is the smallest variance (like in Scenario VI in Table 2.3).

Table 2.2 Summary measures describing the scenarios in which each design is optimal for patient benefit. $j^\star$ and $j^{\star\star}$ are the indices of the best and second best arm, respectively.

Optimal design for PB	number of scenarios	average $\lvert\mu_{j^\star} - \gamma\rvert$	average $\lvert\mu_{j^{\star\star}} - \gamma\rvert$	average $\sigma^2_{j^\star}$	average $\sigma^2_{j^{\star\star}}$
TWE(0.3,0.1)	141	0.79	1.63	2.22	3.00
TWE(1.3,1.3)	4	0.14	0.24	2.16	3.30
UWE(0.3)	283	0.67	1.59	2.87	2.21
RGI	8	0.34	0.42	1.40	2.86
RTS	64	1.83	2.49	1.69	2.70
Total	500	0.84	1.68	2.51	2.51

2.5.7 Individual scenarios performance

To better assess the performance of our proposal, we analysed six additional hand-picked scenarios and reported their operating characteristics under the considered designs in Table 2.3. In Scenario I there is a clear winner in terms of both proximity to the target mean and to the target variance. All of the TWE and the UWE designs achieve high levels of power, more than 12% above FR, while allocating almost half of the patients to the best arm at the same time. The more patient benefit-oriented designs, namely TWE$(0.3, 0.1)$ and UWE(0.3), are particularly effective in this case because they undervalue exploring the inferior arms and focus on the clearly superior treatment. RTS and RGI are comparable to them in correctly identifying the best arm, but they are notably inferior in terms of power and patient benefit.

Scenario II is similar to the previous one, except that the control and inferior treatments have lower variance. As a result, both the percentage of correct selection and the statistical power are increased. All designs, except for FR, allocate more patients to the best arm on average compared to the previous scenario, as the reduced variance makes it less likely for an inferior treatment to appear as the most effective by chance. This is also reflected in the lower standard errors. The patient benefit of UWE(0.3) is the highest, due to its tendency to overvalue treatments with higher variance.

In Scenario III, all treatments outperform the control and are more similar to each other, which reduces the advantage of using TWE. Nevertheless, TWE$(0.3, 0.1)$ and UWE(0.3)

again yield the highest patient benefit, which is about 5% better than both RTS and RGI, and 19% better than FR. No large differences in power or percentage of correct selection are observed across the designs.

Scenario IV, like Scenarios I and II, features a clear best treatment. However, both the treatment means and variances deviate more substantially from the target values. In this setting, RTS performs optimally in terms of patient benefit — achieving 3% more than TWE$(0.3, 0.1)$ and 8% more than UWE(0.3) — as well as power. We observe that when the variance of the best arm is lower than that of the competing arms, UWE designs tend to allocate fewer patients to it, as is the case here.

Under Scenario V, the best arm exhibits a much lower variance compared to the other treatments. In this case, the UWE designs allocate even fewer patients to the best arm than FR, due to their inherent tendency to favor exploration of treatments with higher variance. This pitfall in scenarios where the best arm has low variability illustrates a limitation of UWE, which TWE seeks to overcome through the bivariate information gain (2.15) introduced in this work. RTS achieves the highest power and patient benefit, likely because the best treatment mean and variance lie far from the target values, limiting the effectiveness of our proposed methodology. Nevertheless, TWE still offers a clear advantage over FR, both in terms of power and patient benefit.

Scenario VI, taken from the set of 500 simulation scenarios, is noteworthy as the patient benefit of most of the designs drops below 25%. In this scenario, the means of the two best arms are very similar, and closely followed by the mean of the remaining treatment arm. We have already mentioned that the UWE designs favour larger variances, resulting in this case in less than 25% of the patients allocated to the best arm. In the TWE methods this does not happen, however, given the same distance of the variance from ξ, the arms with variance smaller than ξ are penalised more than the arms with variance larger than ξ. Since TWE$(0.3, 0.1)$ allocates more aggressively to the arm with mean and variance closest to the target values, it ends us assigning most of the patients to the second best arm. The other TWE designs are somewhat protected by this behaviour as they tend to spread the allocations more across all the treatments. This is one of the few scenarios in which RGI is optimal in terms of patient benefit, closely followed by RTS.

Overall, all the considered RAR methods tend to result in a substantial increase in the number of patients allocated to the best treatment compared to FR. Furthermore, due to the

fixed allocation probability to the control group, all the considered RAR designs show high power and PCS, particularly when one of the three active treatments clearly outperforms the others. TWE designs exhibit the best performance when one treatment closely aligns with the target characteristics. In scenarios where this is not the case, RTS and RGI appear to be the more appropriate choices. TWE and UWE designs perform similarly when no inferior arm with high variance is present. However, in the presence of such an arm, UWE can perform poorly, occasionally allocating even fewer patients to the best treatment than FR. TWE$(0.3, 0.1)$, on the other hand, can occasionally fail when the best arm is not well distinguishable and has very low variance. However, we argue that in such cases, allocating patients to the second-best treatment may still yield meaningful benefits to the patients. Furthermore, a lower-than-expected variance should result in a higher chance of rejecting the null hypothesis, even if slightly fewer patient were allocated to that treatment.

Table 2.3 Operating characteristics of the considered designs under six hand-picked scenarios. The target values are fixed to $\gamma = 0$ and $\xi = 2$. Results are based on 10^4 replicated trials. In each scenario, the best arm according to definition (2.17) is in bold.

	Scenario I $M_{1,2,3,4} = (1.0, \mathbf{0.1}, 1.0, 1.0)$ $V_{1,2,3,4} = (3.0, \mathbf{2.1}, 3.0, 3.0)$			Scenario II $M_{1,2,3,4} = (1.0, \mathbf{0.1}, 1.0, 1.0)$ $V_{1,2,3,4} = (1.5, \mathbf{2.1}, 1.5, 1.5)$			Scenario III $M_{1,2,3,4} = (2.0, \mathbf{0.8}, 1.2, 1.2)$ $V_{1,2,3,4} = (2.5, \mathbf{2.5}, 2.5, 2.5)$		
Design	PB(s.e.)	PCS	1-β	PB(s.e.)	PCS	1-β	PB(s.e.)	PCS	1-β
FR	25(0.04)	94.9	0.25	25(0.04)	96.8	0.38	25(0.04)	69.7	0.68
TWE(0.3,0.1)	48(0.16)	95.4	0.40	51(0.15)	97.1	0.62	35(0.14)	70.2	0.67
TWE(1.7,0.8)	44(0.10)	96.7	0.38	46(0.09)	98.1	0.58	31(0.08)	71.4	0.68
TWE(1.3,1.3)	42(0.10)	96.8	0.37	42(0.09)	97.9	0.56	30(0.07)	70.7	0.68
UWE(0.3)	47(0.17)	93.5	0.38	54(0.12)	97.6	0.64	35(0.16)	70.1	0.66
UWE(1.3)	43(0.12)	95.9	0.37	49(0.10)	98.4	0.61	35(0.16)	70.1	0.66
RGI	37(0.10)	96.3	0.34	40(0.09)	98.0	0.54	31(0.08)	71.7	0.68
RTS	35(0.07)	96.3	0.33	37(0.07)	97.9	0.52	30(0.07)	71.3	0.69
	Scenario IV $M_{1,2,3,4} = (4.0, \mathbf{3.0}, 4.0, 4.0)$ $V_{1,2,3,4} = (4.0, \mathbf{3.0}, 4.0, 4.0)$			Scenario V $M_{1,2,3,4} = (4.0, \mathbf{2.0}, 3.0, 3.0)$ $V_{1,2,3,4} = (4.0, \mathbf{1.0}, 4.0, 4.0)$			VI $M_{1,2,3,4} = (3.8, 1.0, \mathbf{0.5}, 0.6)$ $V_{1,2,3,4} = (2.3, 2.6, \mathbf{1.0}, 3.3)$		
Design	PB(s.e.)	PCS	1-β	PB(s.e.)	PCS	1-β	PB(s.e.)	PCS	1-β
FR	25(0.04)	95.1	0.38	25(0.04)	99.7	0.98	25(0.04)	99.9	1
TWE(0.3,0.1)	34(0.07)	95.7	0.42	34(0.08)	99.7	0.98	21(0.14)	96.3	1
TWE(1.7,0.8)	30(0.04)	95.6	0.40	31(0.04)	99.8	0.98	27(0.08)	99.8	1
TWE(1.3,1.3)	29(0.04)	95.6	0.40	30(0.04)	99.7	0.98	25(0.07)	99.9	1
UWE(0.3)	29(0.11)	93.5	0.39	17(0.08)	97.3	0.96	22(0.15)	91.6	1
UWE(1.3)	27(0.05)	95.0	0.39	20(0.05)	99.3	0.98	23(0.09)	98.9	1
RGI	29(0.05)	95.5	0.40	29(0.05)	99.7	0.98	30(0.09)	99.8	1
RTS	37(0.08)	96.1	0.43	38(0.07)	99.8	0.98	29(0.07)	99.9	1

2.6 Discussion

We discussed a novel class of RAR designs for patient allocation based on the theory of weighted information measure and applied to a multi-armed trial with continuous outcomes, extending the methodology proposed by Caruso and Mozgunov, 2024 by incorporating some novelties: 1) unknown response variances fully modelled in a Bayesian way, 2) allocation rule based on randomised probabilities instead of derministic and 3) the possibility to joint targeting desiderable values of response mean and variance.

By making use of an appropriate bivariate weight function centered around a pre-specified pair of clinical targets (γ, ξ), we derived an adaptive information gain criterion which incorporates the information from the previous patient's outcomes and skews the randomisation probabilities in favour of the arms with characteristics more aligned to the clinical targets. At the same time, we protected the control arm by fixing the probability of allocation to it at $1/K$. Our methodology is guided by the configuration of appropriate tuning parameters κ and ω, for which we described a robust selection strategy, allowing flexibility in tackling the exploration vs. exploitation trade-off.

We adopted a hybrid Bayesian-frequentist approach, in which the unknown mean and variance parameters of the response are modelled by a joint Normal inverse-Gamma prior distribution, while the hypothesis testing is based on frequentist principles. This practice is common, as regulatory bodies often require explicit evidence that frequentist error rates are appropriately controlled [Shi and Yin, 2019]. We included a simulation-based strategy for selecting the cut-off value η used in the rejection rule such that the family-wise error rate is controlled under a given threshold, in consideration of the multiple comparisons between the control and the $K-1$ alternative treatments.

The performance of our proposed methodology was compared to several allocation procedures which are known to have good properties either in terms of statistical power or patient benefit, including a fully randomised version of Caruso and Mozgunov's criterion. For a fair comparison, we only considered designs with random allocations and with randomisation probability to the control arm fixed to $1/K$. The operating characteristics of the considered designs were evaluated by simulation under 500 randomly generated scenarios. Our proposed design exhibited noticeable gains in both patient benefit and statistical power compared to fixed randomisation when the characteristics of one treatment were closely aligned with the target values. Moreover, it notably improves on Caruso and

Mozgunov's criterion when an inferior treatment displays very large variance. However, it can underperform when the variance of the best arm is the furthest from ξ or when the means of the treatment arms are in a region far from γ. Our results indicate that the randomisation approache based of Thompson sampling tends to perform better in such scenarios, as it prioritises identifying the treatment most similar to the target characteristics, rather than one that exactly meets them. The strength of our methodology is best demonstrated when the context of the study is central and when identifying a promising treatment aligned with specific clinical targets is more important than simply determining the overall best option. In those circumstances, our methodology offers an efficient way to identify the best treatment, while also allocating most of the patients to it.

Our work focuses on a natural definition of the "best" arm, based on the proximity of treatment means to a target mean. However, alternative definitions — such as those incorporating both the mean and variance — can also be accommodated within the same framework. One such alternative has been explored, and its description and related considerations are provided in the supplementary material.

2.7 Proofs and supplementary material

2.7.1 Proof of Theorem 2.1

Proof. Remember that $p(\mu_j, \sigma_j^2|\underline{x}_j) = p(\mu_j|\sigma_j^2, \underline{x}_j)p(\sigma_j^2|\underline{x}_j)$ and $\phi(\mu_j, \sigma_j^2) = \phi_1(\mu_j|\sigma_j^2)\phi_2(\sigma_j^2)$.

The normalisation condition for $\phi(\mu_j, \sigma_j^2)$ follows from the normalisation conditions (2.6) and (2.7):

$$\begin{aligned}&\int_0^\infty \int_{\mathbb{R}} \phi(\mu_j, \sigma_j^2) p(\mu_j, \sigma_j^2|\underline{x}_j) d\mu_j \, d\sigma_j^2 \\ &= \int_0^\infty \phi_2(\sigma_j^2) p(\sigma_j^2|\underline{x}_j) \int_{\mathbb{R}} \phi_1(\mu_j, \sigma_j^2) p(\mu_j|\sigma_j^2, \underline{x}_j) d\mu_j \, d\sigma_j^2 = 1.\end{aligned}$$

Then, (2.8) is derived with some algebraic manipulation:

$$\begin{aligned}
h^{\phi}(\mu_j,\sigma_j^2) &= -\int_0^{\infty}\int_{\mathbb{R}} \phi(\mu_j,\sigma_j^2)p(\mu_j,\sigma_j^2|\underline{x}_j)\ln\left[p(\mu_j,\sigma_j^2|\underline{x}_j)\right]d\mu_j\,d\sigma_j^2 \\
&= -\int_0^{\infty}\int_{\mathbb{R}} \phi_1(\mu_j,\sigma_j^2)\phi_2(\sigma_j^2)p(\mu_j|\sigma_j^2,\underline{x}_j)p(\sigma_j^2|\underline{x}_j)\ln\left[p(\mu_j|\sigma_j^2,\underline{x}_j)p(\sigma_j^2|\underline{x}_j)\right]d\mu_j\,d\sigma_j^2 \\
&= \int_0^{\infty}\phi_2(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)\int_{\mathbb{R}} -\phi_1(\mu_j,\sigma_j^2)p(\mu_j|\sigma_j^2,\underline{x}_j)\ln\left[p(\mu_j|\sigma_j^2,\underline{x}_j)\right]d\mu_j\,d\sigma_j^2 \\
&\quad -\int_0^{\infty}\phi_2(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)\ln\left[p(\sigma_j^2|\underline{x}_j)\right]\int_{\mathbb{R}}\phi_1(\mu_j,\sigma_j^2)p(\mu_j|\sigma_j^2,\underline{x}_j)d\mu_j\,d\sigma_j^2 \\
&= \int_0^{\infty}\phi_2(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)h^{\phi_1}(\mu_j|\sigma_j^2)d\sigma_j^2 + h^{\phi_2}(\sigma_j^2).
\end{aligned}$$

□

2.7.2 Proof of Theorem 2.2

Proof. The Shannon entropy corresponds to the weighted Shannon entropy with a uniform weight function. Hence we can apply Theorem 2.1 to find the Shannon entropy of (μ_j,σ_j^2) by assuming uniform weight functions:

$$\begin{aligned}
h(\mu_j,\sigma_j^2) &= \int_0^{\infty} p(\sigma_j^2|\underline{x}_j)h(\mu_j|\sigma_j^2)d\sigma_j^2 + h(\sigma_j^2) \\
&= \mathbb{E}_{\sigma_j^2\sim p(\sigma_j^2|\underline{x}_j)}\left[h(\mu_j|\sigma_j^2)\right] + h(\sigma_j^2).
\end{aligned}$$

The Shannon entropy of the normal distributions is well known in the literature:

$$h\left(\mu_j|\sigma_j^2\right) = \frac{1}{2} + \frac{1}{2}\ln\left(\frac{2\pi\sigma_j^2}{m_j}\right)$$

The expected value of the logarithm of an inverse-Gamma random variable $\sigma^2 \sim IG(\alpha,\beta)$ is known to be $\mathbb{E}[\ln\sigma^2] = \ln(\beta) - \psi(\alpha)$. Therefore

$$\mathbb{E}_{\sigma_j^2\sim p(\sigma_j^2|\underline{x}_j)}\left[h(\mu|\sigma_j^2)\right] = \frac{1}{2} + \frac{1}{2}\left[\ln\left(\frac{2\pi}{m_j}\right) + \ln(\beta_j) - \psi(\alpha_j)\right]. \tag{2.23}$$

From Theorem 2.1, by assuming the weight functions (2.9) and (2.10),

$$h^{\phi}(\mu_j, \sigma_j^2) = \mathbb{E}_{\sigma_j^2 \sim p(\sigma_j^2|\underline{x}_j)}\left[h^{\phi_\gamma}(\mu_j|\sigma_j^2)\right] + h(\sigma_j^2).$$

From the proof of Theorem 1 in Caruso and Mozgunov, 2024,

$$\begin{aligned} h^{\phi_\gamma}\left(\mu_j|\sigma_j^2\right) &= \frac{1}{2}\ln\left(\frac{2\pi\sigma_j^2}{m_j}\right) + \frac{\frac{\sigma_j^2/m_j\ \sigma_j^2/m_j^k}{\sigma_j^2/m_j + \sigma_j^2/m_j^k} + \left(\frac{\gamma\sigma_j^2/m_j + \overline{\mu}_j\sigma_j^2/m_j^k}{\sigma_j^2/m_j + \sigma_j^2/m_j^k} - \overline{\mu}_j\right)^2}{2\sigma_j^2/m_j} \\ &= \frac{1}{2}\ln\left(\frac{2\pi\sigma_j^2}{m_j}\right) + \frac{1}{2}\left[1 - \frac{m_j^\kappa}{m_j^\kappa + m_j} + \frac{(\gamma - \overline{\mu}_j)^2}{\sigma_j^2} m_j \left(\frac{m_j^\kappa}{m_j^\kappa + m_j}\right)^2\right]. \end{aligned}$$

The expected value of the inverse of an inverse-Gamma random variable $\sigma^2 \sim IG(\alpha, \beta)$ is known to be $\mathbb{E}[1/\sigma^2] = \alpha/\beta$. Therefore

$$\mathbb{E}_{\sigma_j^2 \sim p(\sigma_j^2|\underline{x}_j)}\left[h^{\phi_\gamma}\left(\mu_j|\sigma_j^2\right)\right] = \frac{1}{2}\ln\left(\frac{2\pi\sigma_j^2}{m_j}\right) + \frac{1}{2}\left[1 - \frac{m_j^\kappa}{m_j^\kappa + m_j} + (\gamma - \overline{\mu}_j)^2 m_j \left(\frac{m_j^\kappa}{m_j^\kappa + m_j}\right)^2 \frac{\alpha}{\beta}\right]. \tag{2.24}$$

We obtain the information gain by subtracting (2.24) from (2.23).

$$\begin{aligned} \Delta_j^\gamma &= h\left(\mu_j, \sigma_j^2\right) - h^{\phi}\left(\mu_j, \sigma_j^2\right) \\ &= \mathbb{E}_{\sigma_j^2 \sim p(\sigma_j^2)}\left[h\left(\mu_j|\sigma_j^2\right)\right] - \mathbb{E}_{\sigma_j^2 \sim p(\sigma_j^2)}\left[h^{\phi_\gamma}\left(\mu_j|\sigma_j^2\right)\right] \\ &= \frac{1}{2}\frac{m_j^\kappa}{m_j^\kappa + m_j} - \frac{1}{2}\left(\gamma - \overline{\mu}_j\right)^2 m_j \left(\frac{m_j^\kappa}{m_j^\kappa + m_j}\right)^2 \frac{\alpha_j}{\beta_j}. \end{aligned}$$

The asymptotic expression (2.12) is easily obtained by observing that $\frac{\beta_j}{\alpha_j} \overset{n_j \to \infty}{\sim} \overline{s}_j^2$. □

2.7.3 Proof of Theorem 2.3

Proof. Let us define $z = m_j^\omega$. Observe that

$$\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j) \propto (\sigma_j^2)^{-z-1}\exp\left(-\frac{\xi(z+1)}{\sigma_j^2}\right)(\sigma_j^2)^{-\alpha_j-1}\exp\left(-\frac{\beta_j}{\sigma_j^2}\right)$$
$$= (\sigma_j^2)^{-(z+\alpha_j+1)-1}\exp\left(-\frac{\xi(z+1)+\beta_j}{\sigma_j^2}\right).$$

Notably, because of the normalisation condition (2.7), $\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)$ is the density function of an inverse-Gamma with parameters $\alpha^\star = z+\alpha_j+1$ and $\beta^\star = \xi(z+1)+\beta_j$.

From Theorem 2.1, by assuming the weight functions (2.9) and (2.13),

$$h^\phi(\mu_j,\sigma_j^2) = \mathbb{E}_{\sigma_j^2\sim\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)}\left[h^{\phi_\gamma}(\mu_j|\sigma_j^2)\right] + h^{\phi_\xi}(\sigma_j^2).$$

Similarly to the proof of Theorem 2.2, we can use the known expression of the Shannon entropy of a Normal distribution and the known expression of the expected value of the logarithm of an inverse-Gamma in order to write the first term explicitly:

$$\mathbb{E}_{\sigma_j^2\sim\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)}\left[h^{\phi_\gamma}(\mu_j|\sigma_j^2)\right] = \frac{1}{2}\left[1+\ln(2\pi)-\ln(m_j)+\ln(\beta^\star)-\psi(\alpha^\star)\right]$$
$$-\frac{1}{2}\frac{m_j^\kappa}{m_j^\kappa+m_j}+\frac{1}{2}\left(\gamma_j-\overline{\mu}_j\right)^2 m_j\left(\frac{m_j^\kappa}{m_j^\kappa+m_j}\right)^2\frac{\alpha^\star}{\beta^\star}, \tag{2.25}$$

We can proceed in a similar way for the second term:

$$\begin{aligned} h^{\phi_\xi}(\sigma_j^2) &= -\int_0^\infty \phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)\ln\left[p(\sigma_j^2|\underline{x}_j)\right]d\sigma_j^2 \\ &= -\mathbb{E}_{\sigma_j^2\sim\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)}\left[p(\sigma_j^2|\underline{x}_j)\right] \\ &= -\mathbb{E}_{\sigma_j^2\sim\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)}\left[\ln\left(\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}\right)-(\alpha_j+1)\ln(\sigma_j^2)-\frac{\beta_j}{\sigma_j^2}\right] \\ &= -\ln\left(\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}\right)+(\alpha_j+1)\left[\ln(\beta^\star)-\psi(\alpha^\star)\right]+\beta_j\frac{\alpha^\star}{\beta^\star}. \end{aligned} \tag{2.26}$$

The entropy of an inverse-Gamma is well known as well. In particular the entropy of σ_j^2 is

$$h\left(\sigma_j^2\right) = \alpha_j + \ln(\beta_j) + \ln\left(\Gamma(\alpha_j)\right) - (\alpha_j + 1)\psi(\alpha_j). \tag{2.27}$$

We can derive the final expression (2.14) with some algebraic manipulation, after substituting (2.23), (2.27), (2.25) and (2.26) in the definition of information gain (2.3):

$$\begin{aligned}
\Delta_j^{\gamma\cdot\xi} &= h(\mu_j, \sigma_j^2) - h^{\phi_{\gamma_j,\xi}}(\mu_j, \sigma_j^2) \\
&= \left\{\mathbb{E}_{\sigma_j^2 \sim p(\sigma_j^2|\underline{x}_j)}\left[h(\mu|\sigma_j^2)\right] + h(\sigma_j^2)\right\} - \left\{\mathbb{E}_{\sigma_j^2 \sim \phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)}\left[h^{\phi_{\gamma_j}}(\mu_j|\sigma_j^2)\right] + h^{\phi_\xi}(\sigma_j^2)\right\} \\
&= \alpha_j + \frac{2\alpha_j + 3}{2}\left[\left(\psi(\alpha^\star) - \psi(\alpha_j)\right) - \left(\ln(\beta^\star) - \ln(\beta_j)\right)\right] \\
&+ \frac{1}{2}\frac{m_j^\kappa}{m_j^\kappa + m_j} - \frac{\alpha^\star}{\beta^\star}\left[\beta_j + \frac{1}{2}\left(\gamma_j - \overline{\mu}_j\right)^2 m_j \left(\frac{m_j^\kappa}{m_j^\kappa + m_j}\right)^2\right].
\end{aligned} \tag{2.28}$$

The asymptotic expression (2.15) can be derived using the fact that $\psi(\alpha^\star) \overset{n_j \to \infty}{\sim} \ln(\alpha^\star) - \frac{1}{2\alpha^\star}$.

We were able to deduce the information gain without writing explicitly the expression of $\phi_\xi(\sigma_j^2)$. Indeed, all we needed to complete the proof was the kernel associated with the weight function and the normalisation condition. For the sake of completeness, we will now derive the explicit expression of $\phi_\xi(\sigma_j^2)$.

Previously, we stated that $\phi_\xi(\sigma_j^2)p(\sigma_j^2|\underline{x}_j)$ is the density of an inverse-Gamma with parameters $\alpha^\star$ and $\beta^\star$. It follows that its normalisation factor is equal to the product of the normalisation factors of $\phi_\xi(\sigma_j^2)$ and $p(\sigma_j^2|\underline{x}_j)$, that is,

$$\frac{\beta^{\star\alpha^\star}}{\Gamma(\alpha^\star)} = C_\xi \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)},$$

where C_ξ is the factor of $\phi_\xi(\sigma_j^2)$.

Rearranging, we find C_ξ and, as a direct consequence, the expression of $\phi_\xi(\sigma_j^2)$:

$$\phi_\xi(\sigma_j^2) = \frac{\beta^{\star\alpha^\star}}{\Gamma(\alpha^\star)} \frac{\Gamma(\alpha_j)}{\beta_j^{\alpha_j}} (\sigma_j^2)^{-z-1} \exp\left(-\frac{\xi(z+1)}{\sigma_j^2}\right).$$

□

2.7.4 Alternative definition of best treatment arm

A second definition of best arm follows from the observation that, under the assumption of normally distributed endpoint, targeting (γ, ξ) is analogous to claiming that the most clinical desirable distribution for a treatment is $Y \sim N(\gamma, \xi)$. Therefore, we can define the arm whose distribution $X_j \sim N(M_j, V_j)$ has the smallest Kullback-Leibler divergence from Y as the best one, i.e.,

$$j_\xi^\star = \underset{j=2,\dots,K}{\operatorname{argmin}} KL(f_Y \,||\, f_{X_j}) = \underset{j=2,\dots,K}{\operatorname{argmin}} \frac{1}{2}\left[\frac{(M_j-\gamma)^2}{V_j} + \frac{\xi}{V_j} - \ln\frac{\xi}{V_j} - 1\right], \tag{2.29}$$

where f_Y and f_{X_j} are the density functions of Y and X_j. The closed form expression of the Kullback-Leibler divergence for two normally distributed random variables is well known in the literature [for example, see Soch, 2024].

This definition is consistent with the idea of targeting γ and ξ at the same time, however it is not suited for a traditional hypothesis testing approach, as it is based on the estimation of the whole distribution of the endpoint. While methods for testing which distribution is significantly closer to a given target distribution have been developed in the field of applied statistics, they have never been used in the context of a clinical trial, to the best of our knowledge. For this reason, we only considered the testing approach described in (2.20).

In line with this definition, we estimate the best arm $\widehat{j_\xi^\star}$ at the end of the study as follows:

$$\widehat{j_\xi^\star} = \underset{j=2,\dots,K}{\operatorname{argmin}} KL(f_Y \,||\, f_{\widetilde{X}_j|\underline{x}_j}), \tag{2.30}$$

where $\widetilde{X}_j|\underline{x}_j$ is the posterior predictive distribution of X_j and $f_{\widetilde{X}_j|\underline{x}_j}$ is its density function. The posterior distribution of $\widetilde{X}_j|\underline{x}_j$ follows a location-scale version of the t distribution

with $2\alpha_j$ degrees of freedom, i.e.,

$$\widetilde{X}_j|\underline{x}_j \sim t_{2\alpha_j}\left(\overline{\mu}_j, \frac{\beta_j(m_j+1)}{\alpha_j m_j}\right).$$

We correctly identify the best arm according to definition (2.29) if we claim that $\widehat{j}^{\star}_{\xi} = j^{\star}_{\xi}$ at the end of the trial.

By applying the robust strategy from Section 2.4 for selecting κ and ω, while using this definition of best arm, we find the optimal pairs $(0.3, 0.1)$ for patient benefit, $(1.5, 1.5)$ for the percentage of correct selection and $(1.3, 1.3)$ for power. Two pairs out of three coincide with the ones obtained while using the definition of "best" given in (2.17). The only difference is in the optimal pair for the percentage of correct selection, because we use different approaches for identifying the best arm, depending on the definition of "best" that we are using, while the randomisation rule and the hypothesis testing procedure remain the same.

Figures 2.6 presents a comprehensive synthesis of the operating characteristics associated with the evaluated designs, aggregating the outcomes across all the 500 randomly generated alternative scenarios when the best arm is defined as in definition (2.29).

Figures (2.5) and (2.6), which are based on two different definitions of best arm, exhibit a high degree of similarity. A big part of this similarity can be explained by the fact that in 443 out of 500 of the randomly generated scenarios, the two definitions of "best" actually identify the same best treatment arm.

2.7.5 Selection of the null scenarios

For the calibration of η and the evaluation of the FWER, we selected 49 null scenarios satisfying the global null (2.21). Each scenario is characterised by homogeneous means close to or equal to γ and by varying variances. Specifically, we considered seven mean values ranging from $\gamma - 2\sqrt{\xi}$ to $\gamma + 2\sqrt{\xi}$, which corresponds to a range of treatment effect of interest, given that the target mean and variance are γ and ξ. The FWER is smallest when all means are equal to γ, and it increases rapidly to a plateau as the means deviate from this value. For each mean value, we examined seven variance configurations — three homogeneous and four heterogeneous — with variances close to or equal to ξ. The homogeneous cases correspond to $V_1 = V_2 = V_3 = V_4$, with V_1 equal to $\xi - 1$, ξ or

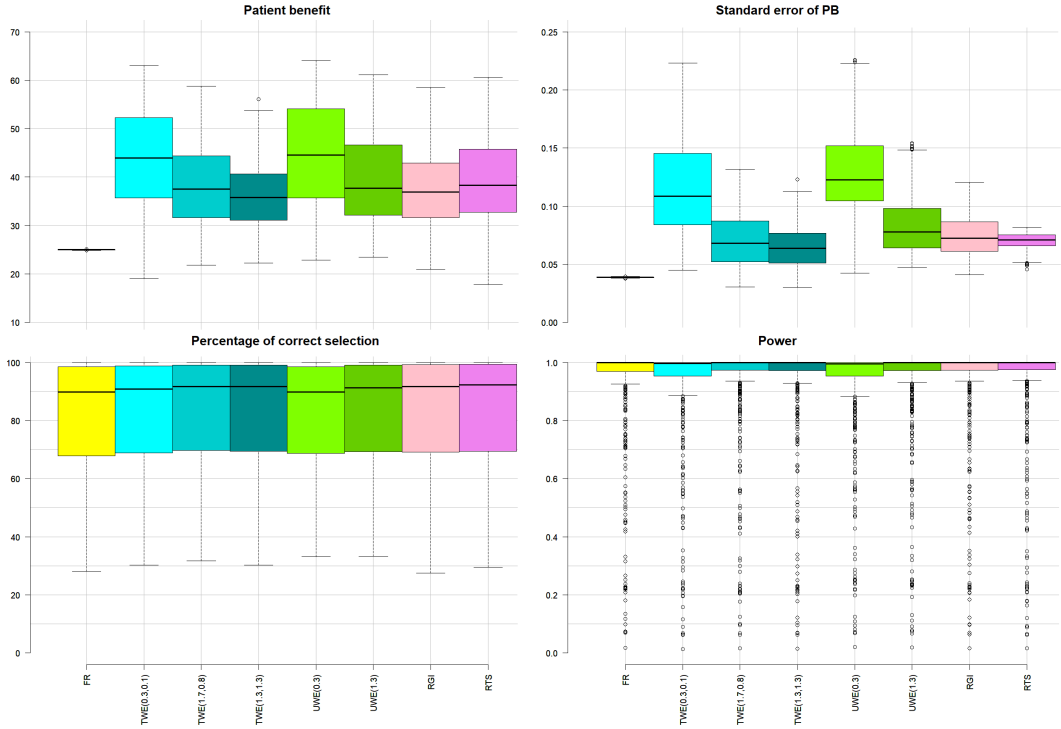

Fig. 2.6 Operating characteristics across $S = 500$ randomly generated scenarios for the considered set of designs, when the second definition of best arm (2.29) is used. For all scenarios, we assume $K = 4$, $N = 100$. The target values are fixed to $\gamma = 0$ and $\xi = 2$. Burn-in size is fixed to $B = 5$.

$\xi + 1$. Among the heterogeneous configurations, we considered two cases where the control variance was smaller than the others ($V_1 = \xi$, $V_2 = V_3 = V_4 = \xi + 1$; $V_1 = \xi - 1$, $V_2 = V_3 = V_4 = \xi$), and two where it was larger ($V_1 = \xi$, $V_2 = V_3 = V_4 = \xi - 1$; $V_1 = \xi + 1$, $V_2 = V_3 = V_4 = \xi$). Overall, differences in variance configurations only slightly affect the FWER, especially when the means are closer to γ.

2.7.6 Randomisation with UWE

In Section 2.2.2 we derived the information gain based on the uniform-variance weight function defined in Caruso and Mozgunov, 2024. In a similar fashion to what we did in Section 2.2.4, we can construct a randomisation rule based on its asymptotic expression (2.12). Let us define the following criterion by dropping the constant part c_κ of (2.12):

$$\widehat{\Delta}_j^\gamma = -\frac{1}{2}\frac{\left(\gamma - \bar{x}_j\right)^2}{\bar{s}_j^2}\frac{n_j}{(n_j^{1-\kappa}+1)^2}$$

Then, UWE(κ) randomises patient $n+1$ to arm $j \in \{1,...,K\}$ with probability

$$p_j = \begin{cases} 1/K & \text{, if } j = 1 \\ \dfrac{(\widehat{\Delta}_j^\gamma)^{-1}}{\sum_{j=1}^K (\widehat{\Delta}_j^\gamma)^{-1}}\dfrac{K-1}{K} & \text{, if } j > 1. \end{cases}$$

2.7.7 Randomisation with RGI

Smith and Villar, 2018 proposed a deterministic response-adaptive allocation method based on the Gittins index, which is typically associated with high patient benefit when the trial's objective is to identify the treatments with the highest (or lowest) effect size. However, this approach is not well-suited when the goal is to target a specific effect size. Caruso and Mozgunov, 2024 formulated two heuristic modifications of Smith and Villar, 2018's proposal to favour outcomes closer to a target value, rather than simply the highest (or lowest) ones. In particular, we expand on the targeted Gittins index criterion. In order to make it a fair competitor to the other RAR designs, we change it into a randomised method instead of a deterministic one, and we fix the probability of allocation to control to $1/K$. What we obtain is the randomised targeted Gittins index (RGI).

RGI randomises patient $n+1$ to arm $j \in \{1,...,K\}$ with probability

$$p_j = \begin{cases} 1/K & \text{, if } j = 1 \\ 1/\delta_{\mathscr{G}_j,\mathscr{G}_\gamma} & \text{, if } j > 1, \end{cases}$$

where $\delta_{\mathscr{G}_j,\mathscr{G}_\gamma}$ is the criterion found in Caruso and Mozgunov, 2024, supplementary material, Section D, equation (24).

Chapter 3

Considerations on the multivariate non-central t distributions with applications to a sample size reassessment design for bioequivalence trials.

3.1 Chapter introduction

Two common definitions of the non-central t distribution coexist in statistics, and the general forms of their cumulative distribution functions can be found in Genz and Bretz, 2009. However, confusion may arise when one encounters only a single definition without the broader context.

While researching bioequivalence (BE) trials for a collaboration with Chiesi Farmaceutici, I came across three recent papers addressing the exact calculation of power in BE studies under a normality assumption — an application that relies on the multivariate non-central t distribution. In all three papers, the authors use the wrong form of the non-central t distribution. Under the conditions considered in those works, as well as in many practical applications, this mistake is effectively invisible because the two definitions yield nearly

identical numerical results. Nevertheless, using the incorrect expression can, in general, lead to inaccuracies, albeit small. For this reason, the first part of this chapter provides clarification on the two definitions. We present the correct expression that should be used in BE trials, along with additional results and initial considerations for potential future extensions. The second part of the chapter applies these considerations to sample size reassessment in BE trials.

The goal of a BE trial is to demonstrate that a test drug and a reference drug achieve equivalent concentrations of the active ingredient. BE studies are based on the principle that establishing bioequivalence should also imply equivalence of the treatment responses, hence eliminating the need for full clinical development of the new product. The justification is given by the following reasoning: once absorbed, a drug follows a pharmacokinetic process that determines its concentration in the bloodstream. From there, the drug moves to and from the site of action according to effect-site dynamics. Through the pharmacodynamic response mechanism, the resulting concentration ultimately produces a clinical response. A detailed analysis of BE trials and related concepts is the book by Hauschke et al., 2007.

BE is typically assessed by comparing the areas under the curve (AUC) of the concentration of drug in the blood over time for the two treatments of interest [Senn, 2007, chapter 22]. If these profiles are essentially indistinguishable, it is inferred that the two products are therapeutically equivalent. In particular, BE inferences are traditionally made on the ratio of the mean AUC values, which usually follow a log-normal distribution. In those cases, the regulatory agencies require that equivalence should be assessed on the logarithmic scale [FDA, 1992]. Thus, by taking logarithmic transforms, the ratio of means becomes a difference of normally distributed outcomes. However, many clinical trial settings justify assuming normality on the original (untransformed) scale. For example, when assessing therapeutic equivalence between two inhalers for the relief of asthma symptoms, the morning peak expiratory flow rate is often used as a measure of airflow obstruction, and this variable is typically well approximated by a normal distribution [Jones et al., 1996]. We will consider this case, in which the untransformed variable is normally distributed with unknown variance. Moreover, for simplicity, we will consider the case of double-blind, randomised, three-arm parallel clinical trial, although many BE trials are cross-over. The three arms are for the test treatment, the reference treatment and placebo. The common practice for demonstrating bioequivalence between test and reference requires evidence that both are superior to placebo in terms of mean treatment difference, and that the test is

equivalent to the reference in terms of their mean ratio [Davit et al., 2009]. This procedure is also aimed at preventing that a hypothetical trail of BE trials leads to the registration of drugs inferior to placebo. Consequently, the overall BE assessment comprises two superiority tests (test vs. placebo and reference vs. placebo) and one equivalence test (test vs. reference).

Since equivalence tests are known to be conservative, particularly for highly variable drugs, it is relevant to mention that several corrections have recently been proposed to improve the power of equivalence trials (e.g., Boulaguiem et al., 2024 and Boulaguiem et al., 2025). However, to the best of our knowledge, these approaches do not account for the additional complexity introduced by the two superiority tests, and we therefore do not discuss them further here.

Yang and Sun, 2019 propose an exact approach to compute the power of a BE trial, for normally distributed treatments, by working directly with the distribution of the treatment differences, which is shown to be a multivariate non-central t distributions [Hauschke et al., 1999 and Chang et al., 2014]. However, their method relies on the incorrect definition of the non-central t. Zhu and Sun, 2019 later introduces four designs for blinded and unblinded sample size reassessment, based on the wrong power formula from Yang and Sun, 2019. More recently, Hinds and Sun, 2025 presents an unblinded approach that introduces the re-estimation of the mean during the interim analysis, which, again, makes use of the wrong fromula. Although the above papers use the incorrect definition of the non-central t, their results remain practically valid because the numerical differences are minimal. Nevertheless, we provide the correct expression.

In Section 3.4, partly to give an example of the correct use of the multivariate non-central t, partly to give a small contribution to the research on adaptive BE trials, we introduce a modification of a simple design for blinded sample size reassessment. Sample size reassessment typically relies on an interim estimation of the variance, enabling investigators to recalculate the required sample size mid-trial to ensure adequate power. It is worth noting that, even when blinding is formally maintained, sample size reassessment may still reveal information about the treatment effect. For example, a higher estimated variance can indirectly indicate larger differences between treatment groups (reflecting the correlation between the treatment effect and the variance estimator). This concern, however, is generally more pronounced in trials with binary or time-to-event endpoints (Hróbjartsson et al., 2014), whereas this chapter focuses on continuous endpoints.

Similarly to Hinds and Sun, 2025, we introduce a design that, additionally, factors in the estimates of the treatment means. However, differently from them, we propose a blinded approach. Including mean estimation can offer a meaningful improvement because, when equivalence is assessed through the mean ratio, even small deviations in the treatment effects can lead to substantial changes in statistical power and therefore in the required sample size [Hauschke et al., 1999].

Finally, our proposed design is compared by simulation to the design described in Zhu and Sun, 2019, our source of motivation.

3.2 The two versions of the non-central t distribution

3.2.1 In one dimension

There are two common definitions of non-central t distribution often used in Statistics. One refers simply to a location shifted version of the central t distribution, i.e.,

$$t_\nu^{(1)}(\mu) = \frac{Z}{\sqrt{Y/\nu}} + \mu, \tag{3.1}$$

where $Z \sim N(0,1)$ is a standard normal, $Y \sim \chi^2_\nu$ is a chi-squared with ν degrees of freedom and μ is the non-centrality parameter.

The other definition refers to a normal random variable centered in μ, and divided by the same chi-squared as before, i.e,

$$t_\nu^{(2)}(\mu) = \frac{Z+\mu}{\sqrt{Y/\nu}}. \tag{3.2}$$

The two definitions are identical when $\mu = 0$. In addition, as ν goes to infinity, the term Y/ν at the denominator converges to 1 and both definitions converge to a normal with mean equal to μ. It follows that their distributions become quite similar when μ is close to zero or ν is large. Figure 3.1 contains the densities of $t_\nu^{(1)}(\mu)$ and $t_\nu^{(2)}(\mu)$ for $\mu = 10$ and $\nu = 40$.

The cumulative distribution function of a $T \sim t_\nu^{(1)}(\mu)$ random variable is

$$P(T < z) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\sqrt{\nu\pi}} \int_{-\infty}^{z} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx, \tag{3.3}$$

while the cumulative distribution function of a $T \sim t_\nu^{(2)}(\mu)$ random variable is

$$P(T < z) = \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \int_{-\infty}^{\frac{sz}{\sqrt{\nu}}-\mu} \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{x^2}{2}\right) dx\, s^{\nu-1} e^{-s^2/2}\, ds. \tag{3.4}$$

3.2.2 In more than one dimension

Definitions (3.1) and (3.2) can be generalised to the case where at the numerator we have a zero-mean multivariate normal $\boldsymbol{Z} \sim N(\mathbf{0}, \rho)$ in k dimensions and a corresponding non-centrality vector $\boldsymbol{\mu} \in \mathbb{R}^n$, instead of the univariate Z and μ, while the chi-squared distribution at the denominator remains univariate. Notice that the variance-covariance matrix ρ is also the correlation matrix, since $\boldsymbol{Z}$ is chosen to have all unit variances. This generalisation is different from the classical works leading to Hotelling on T^2 random vectors, which are based on a multidimensional extension of the chi-squared distribution at the denominator, rather than on a multivariate extension of the numerator, as done here.

We denote the two versions of multivariate non-central t by $t_\nu^{(1)}(\boldsymbol{\mu}, \rho)$ and $t_\nu^{(2)}(\boldsymbol{\mu}, \rho)$, respectively.

The distribution functions for the multivariate generalisations of both versions of the non-central t is given in Genz and Bretz, 2009. We recall that the definition of distribution function of a random vector $\boldsymbol{T}$ of dimension $k > 1$ is

$$F_{\boldsymbol{T}}(z_1, \ldots, z_k) = P(T_1 < z_1,\ \ldots,\ T_k < z_k).$$

For any non-centrality parameter vector $\boldsymbol{\mu} = (\mu_1, ...\mu_k)$, $k \in \mathbb{N}$, and any correlation matrix ρ, the joint distribution function of a random vector $(T_1, ..., T_k) \sim t_\nu^{(1)}(\boldsymbol{\mu}, \rho)$ is

$$\begin{aligned} &P(T_1 < z_1, \ ..., \ T_k < z_k) \\ &= \frac{\Gamma(\frac{\nu+k}{2})}{\Gamma(\nu/2)\sqrt{\det(\rho)(\nu\pi)^k}} \int_{-\infty}^{z_1} ... \int_{-\infty}^{z_k} \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \rho^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+k}{2}} d\boldsymbol{x} \end{aligned} \tag{3.5}$$

and the joint distribution function of a random vector $(T_1, ..., T_k) \sim t_\nu^{(2)}(\boldsymbol{\mu}, \rho)$ is

$$\begin{aligned} &P(T_1 < z_1, \ ..., \ T_k < z_k) \\ &= \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} \int_0^\infty \int_{-\infty}^{\frac{s z_1}{\sqrt{\nu}} - \mu_1} ... \int_{-\infty}^{\frac{s z_k}{\sqrt{\nu}} - \mu_k} \frac{\exp\left(\boldsymbol{x}^T \rho^{-1} \boldsymbol{x}\right)}{\sqrt{(2\pi)^k \det(\rho)}} \, dx \, s^{\nu-1} e^{-s^2/2} \, ds. \end{aligned} \tag{3.6}$$

Kotz and Nadarajah, 2004 extensively cover both $t_\nu^{(1)}(\boldsymbol{\mu}, \rho)$ and $t_\nu^{(2)}(\boldsymbol{\mu}, \rho)$, but consider $t_\nu^{(1)}(\boldsymbol{\mu}, \rho)$ to be somehow more fundamental. We had the opposite experience in our encounter with the multivariate t in clinical trials. Both formulae (3.5) and (3.6) can be computed efficiently in R by using the function "pmvt" from the package "mvtnorm" [Genz et al., 2025].

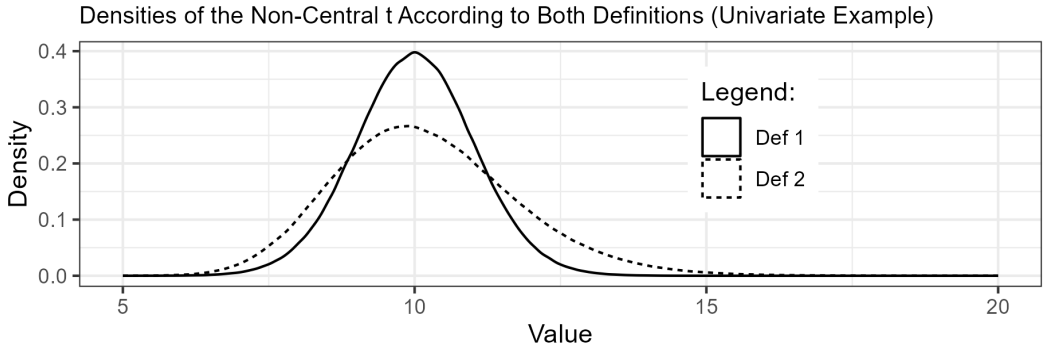

Fig. 3.1 Plot of the densities of $t_\nu^{(1)}(\mu)$ (solid line) and $t_\nu^{(2)}(\mu)$ (dotted line) for $\mu = 10$ and $\nu = 40$.

3.3 Applications to bioequivalence trials

3.3.1 Bioequivalence trials

A bioequivalence clinical trial usually aims to establish BE between a generic drug (called a test) and an innovator drug (reference). A placebo is often included, as explained in

the Chapter introduction, to demonstrate the sensitivity of the study. We consider a BE trial with test, reference, and placebo arms, and we assume that the clinical outcomes in each arm are normally distributed with a common unknown variance, i.e.$X_T \sim N(\mu_T, \sigma^2)$, $X_R \sim N(\mu_R, \sigma^2)$ and $X_P \sim N(\mu_P, \sigma^2)$.

The testing hypotheses for superiority are

$$H_{0,T}^{sup} : \mu_T - \mu_P \leq 0 \quad vs. \quad H_{1,T}^{sup} : \mu_T - \mu_P > 0$$

for the superiority of the test treatment over placebo and

$$H_{0,R}^{sup} : \mu_R - \mu_P \leq 0 \quad vs. \quad H_{1,R}^{sup} : \mu_R - \mu_P > 0$$

for the superiority of the reference treatment over placebo.

Ratio of means is often used to test equivalence in a BE study. Therefore, the testing hypotheses for the equivalence of test and reference are

$$H_0^{eq} : \frac{\mu_T}{\mu_R} \leq \theta_1 \text{ or } \frac{\mu_T}{\mu_R} \geq \theta_2 \quad vs. \quad H_1^{eq} : \theta_1 < \frac{\mu_T}{\mu_R} < \theta_2,$$

where $0 < \theta_1 < 1 < \theta_2$ are the boundaries of the prespecified equivalence range.

Sasabuchi, 1988 demonstrated that each of these hypotheses can be tested separately using simple t-tests. The test statistics used in the superiority assessments are

$$T_1^{sup} = \frac{\overline{X}_T - \overline{X}_P}{s\sqrt{\frac{1}{n_T} + \frac{1}{n_P}}}, \quad T_2^{sup} = \frac{\overline{X}_R - \overline{X}_P}{s\sqrt{\frac{1}{n_R} + \frac{1}{n_P}}} \tag{3.7}$$

while the test statistics used in the equivalence assessment are

$$T_1^{eq} = \frac{\overline{X}_T - \theta_1\overline{X}_R}{s\sqrt{\frac{1}{n_T} + \frac{\theta_1^2}{n_R}}}, \quad T_2^{eq} = \frac{\overline{X}_T - \theta_2\overline{X}_R}{s\sqrt{\frac{1}{n_T} + \frac{\theta_2^2}{n_R}}}, \tag{3.8}$$

where n_T, n_R and n_P are the sample sizes of of the test, reference and placebo group, respectively; $\overline{X}_T$, $\overline{X}_R$ and $\overline{X}_P$ are their sample means; s^2 is the pooled variance.

Notice that by dividing both the numerator and denominator of T_1^{sup} by $\sigma\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}$, we would obtain

$$N = \frac{\overline{X}_T - \overline{X}_P}{\sigma\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}} \quad \text{and} \quad D = \frac{s\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}}{\sigma\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}} = \frac{s}{\sigma}, \tag{3.9}$$

respectively. Here, N is normally distributed with mean $\mu = (\mu_T - \mu_P)/(\sigma\sqrt{\frac{1}{n_T}+\frac{1}{n_P}})$ and variance 1. Moreover, it can be shown that $D^2 \times (n_T + n_P - 2)$ has the same distribution of a chi-squared with $\nu = n_T + n_P - 2$ degrees of freedom. By defining the standard normal random variable $Z = N - \mu$ and the chi-squared random variable $Y = D^2 \times \nu$, we can rewrite T_1^{sup} as $\frac{Z+\mu}{\sqrt{Y/\nu}}$. This decomposition makes it clear that T_1^{sup} follows a t-distribution as defined in (3.2). A similar argument can be made for T_2^{sup}, T_1^{eq} and T_2^{eq}.

Superiority of both test and reference over placebo is demonstrated if $T_1^{sup} \geq t_{\alpha^{sup}}$ and $T_2^{sup} \geq t_{\alpha^{sup}}$, while equivalence of test and reference is demonstrated if $T_1^{eq} \geq t_{\alpha^{eq}}$ and $T_2^{eq} \leq -t_{\alpha^{eq}}$, given two pre-specified type I error rates α^{sup} and α^{eq}.

By applying the general expression of the multivariate non-central t given in Genz and Bretz, 2009 to the context of BE trials, Hauschke et al., 1999 showed that (T_1^{eq}, T_2^{eq}) follow a bivariate distribution of the form (3.6). More recently, Yang and Sun, 2019 extended the result to the entire distribution of $(T_1^{sup}, T_2^{sup}, T_1^{eq}, T_2^{eq})$, by showing that they also follow a multivariate non-central t distribution with $\nu = n_T + n_R + n_P - 3$ degrees of freedom, non-centrality parameter vector

$$\left(\Theta_1^{sup}, \Theta_2^{sup}, \Theta_1^{eq}, \Theta_2^{eq}\right) = \left(\frac{\mu_T - \mu_P}{\sigma\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}}, \frac{\mu_R - \mu_P}{\sigma\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}}, \frac{\mu_T - \theta_1\mu_R}{\sigma\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}}, \frac{\mu_T - \theta_2\mu_R}{\sigma\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} \right),$$

and correlation matrix

$$\rho = \begin{pmatrix} 1 & \frac{\frac{1}{n_P}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}} & \frac{\frac{1}{n_T}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}} & \frac{\frac{1}{n_T}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} \\ \frac{\frac{1}{n_P}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}} & 1 & \frac{-\frac{\theta_1}{n_R}}{\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}} & \frac{-\frac{\theta_2}{n_R}}{\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} \\ \frac{\frac{1}{n_T}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}} & \frac{-\frac{\theta_1}{n_R}}{\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}} & 1 & \frac{\frac{1}{n_T}+\frac{\theta_1\theta_2}{n_R}}{\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} \\ \frac{\frac{1}{n_T}}{\sqrt{\frac{1}{n_T}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} & \frac{-\frac{\theta_2}{n_R}}{\sqrt{\frac{1}{n_R}+\frac{1}{n_P}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} & \frac{\frac{1}{n_T}+\frac{\theta_1\theta_2}{n_R}}{\sqrt{\frac{1}{n_T}+\frac{\theta_1^2}{n_R}}\sqrt{\frac{1}{n_T}+\frac{\theta_2^2}{n_R}}} & 1 \end{pmatrix}. \tag{3.10}$$

Yang and Sun are correct in claiming that the multivariate distribution of the test statistics is non-central t, however they mistakenly present the wrong expression for the power of a BE trial $P(T_1^{sup} \geq t_{\alpha^{sup}}, T_2^{sup} \geq t_{\alpha^{sup}}, T_1^{eq} \geq t_{\alpha^{eq}}, T_2^{eq} \leq -t_{\alpha^{eq}})$, by relying on the wrong definition of non-central t. They derive the expression for power from (3.5), while the correct expression comes from (3.6), as we show next. The same mistake is present in a subsequent paper from one of the two authors Zhu and Sun, 2019, which compares different sample size reassessment approaches for BE trials.

It is important to notice that for the sample sizes commonly considered in BE trials, the discrepancy between the two expressions is minimal. As mentioned in Section 3.2, the two definitions of non-central t converge as the degrees of freedom, i.e., the sample size, increases; as a result, the plots and tables presented in those papers are, in practice, accurate.

For the sake of precision, we are going to show the correct expression of the power $P(T_1^{sup} \geq t_{\alpha^{sup}}, T_2^{sup} \geq t_{\alpha^{sup}}, T_1^{eq} \geq t_{\alpha^{eq}}, T_2^{eq} \leq -t_{\alpha^{eq}})$, while providing some additional rudimentary results.

3.3.2 Power and its approximation for $n_T = n_R$

In general, the power of a BE trial is obtained by computing the probability, derived from the distribution function (3.6), of $(T_1^{sup}, T_2^{sup}, T_1^{eq}, T_2^{eq})$ being in the rejection region $(t_{\alpha^{sup}}, \infty) \times (t_{\alpha^{sup}}, \infty) \times (t_{\alpha^{eq}}, \infty) \times (-\infty, t_{\alpha^{eq}})$, that is,

$$\begin{aligned} & P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}},\, T_2^{eq} \leq -t_{\alpha^{eq}}) \\ & = \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} \int_0^{\infty} \int_{\frac{s t_{\alpha^{sup}}}{\sqrt{\nu}} - \Theta_1^{sup}}^{\infty} \int_{\frac{s t_{\alpha^{sup}}}{\sqrt{\nu}} - \Theta_2^{sup}}^{\infty} \int_{\frac{s t_{\alpha^{eq}}}{\sqrt{\nu}} - \Theta_1^{eq}}^{\infty} \int_{-\infty}^{\frac{s t_{\alpha^{eq}}}{\sqrt{\nu}} - \Theta_2^{eq}} \\ & \quad \frac{1}{(2\pi)^2 \sqrt{|\det(\rho)|}} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \rho^{-1} \boldsymbol{x}\right) d\boldsymbol{x}\, s^{\nu-1} e^{-s^2/2}\, ds, \end{aligned} \tag{3.11}$$

where $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$.

If $\mu_T = \mu_R$ and $\theta_1 = 1/\theta_2$, we can rewrite Θ_1^{eq} and Θ_2^{eq} as

$$\Theta_1^{eq} = \frac{\mu_T}{\sigma} \sqrt{\frac{\theta_1 + \theta_2 - 2}{\frac{\theta_1}{n_T} + \frac{\theta_2}{n_R}}} \quad \text{and} \quad \Theta_2^{eq} = -\frac{\mu_T}{\sigma} \sqrt{\frac{\theta_1 + \theta_2 - 2}{\frac{\theta_2}{n_T} + \frac{\theta_1}{n_R}}}. \tag{3.12}$$

If in addition $n_T = n_R$, then the centrality parameters of T_1^{eq} and T_2^{eq} satisfy $\Theta_1^{eq} = -\Theta_2^{eq}$ and their correlation becomes $\rho_{3,4} = 2/\sqrt{\theta_1^2 + \theta_2^2 + 2}$. For the values of $\theta_1 = 0.8$ and $\theta_2 = 1.25$, usually adopted in real trials, we would obtain a correlation of $\rho_{3,4} = 0.976$ between T_1^{eq} and T_2^{eq}. In that case, the first statistic would almost coincide with a translation of the latter: $T_1^{eq} \approx T_2^{eq} + 2\Theta_1^{eq}$. In Kieser and Hauschke, 1999 they use this fact to justify the approximation

$$P(T_1^{eq} \geq t_\alpha, T_2^{eq} \geq -t_\alpha) \approx P(T_2^{eq} \geq -t_\alpha) \tag{3.13}$$

and they show that it is very accurate even when $\mu_T/\mu_R \in (\theta_1, \theta_2)$ for the usual sample sizes considered in BE trials. Similarly, we can approximate

$$\begin{aligned} & P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}},\, T_2^{eq} \geq -t_{\alpha^{eq}}) \\ \approx & P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_2^{eq} \geq -t_{\alpha^{eq}}). \end{aligned} \tag{3.14}$$

It follows that the power of a BE trial — the result of a four-dimensional operation — can be approximated by the difference of two three-dimensional probabilities:

$$P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}},\, T_2^{eq} \leq -t_{\alpha^{eq}}) \approx$$
$$P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}}) -$$
$$P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_2^{eq} \geq -t_{\alpha^{eq}}). \quad (3.15)$$

3.3.3 Linear dependence of $(T_1^{sup}, T_2^{sup}, T_1^{eq}, T_2^{eq})$

Not only are T_1^{sup}, T_2^{sup}, T_1^{eq} and T_2^{eq} correlated, but each one of them is a linear combination of the other three. In particular, we notice that

$$T_2^{eq} = a\,T_1^{sup} - a\,T_2^{sup} - b\,T_1^{eq}, \quad (3.16)$$

where

$$a = \frac{(\theta_2 - \theta_1)}{(1-\theta_1)} \sqrt{\frac{\frac{1}{n_T} + \frac{1}{n_P}}{\frac{1}{n_T} + \frac{\theta_2^2}{n_R}}} \text{ and } b = \frac{(\theta_2 - 1)}{(1-\theta_1)} \sqrt{\frac{\frac{1}{n_T} + \frac{\theta_1^2}{n_R}}{\frac{1}{n_T} + \frac{\theta_2^2}{n_R}}}.$$

Notice that $a, b > 0$ because $\theta_1 < 1 < \theta_2$.

Therefore the power of a BE trial can be also written as

$$P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}},\, a\,T_1^{sup} - a\,T_2^{sup} - b\,T_1^{eq} \leq -t_{\alpha^{eq}}), \quad (3.17)$$

allowing us to evaluate it by simulation by generating values from only three correlated statistics, instead of four.

Going further with the analytic expression, we can formulate the power as a difference of two quadruple integrals as follows:

$$(3.17) = P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}})$$
$$- P(T_1^{sup} \geq t_{\alpha^{sup}},\, T_2^{sup} \geq t_{\alpha^{sup}},\, T_1^{eq} \geq t_{\alpha^{eq}},\, a\,T_1^{sup} - a\,T_2^{sup} - b\,T_1^{eq} \geq -t_{\alpha^{eq}})$$

$$= \frac{2^{1-\nu/2}}{\Gamma(\nu/2)} \int_0^{\infty} \int_{\frac{st_{\alpha^{sup}}}{\sqrt{\nu}} - \Theta_1^{sup}}^{\infty} \int_{\frac{st_{\alpha^{sup}}}{\sqrt{\nu}} - \Theta_2^{sup}}^{\infty} \int_{\frac{st_{\alpha^{eq}}}{\sqrt{\nu}} - \Theta_1^{eq}}^{\infty} \frac{1}{(2\pi)^{3/2} \sqrt{|\det(\rho_{\theta_1})|}}$$

$$\times \exp\left(-\frac{1}{2}(x_1,x_2,x_3)^T\rho_{\theta_1}^{-1}(x_1,x_2,x_3)\right)\, dx_3dx_2dx_1\, s^{\nu-1}e^{-s^2/2}\, ds$$
$$-\frac{2^{1-\nu/2}}{\Gamma(\nu/2)}\int_0^\infty\int_{l_1(s)}^\infty\int_{l_2(s)}^{u_2(s,x_1)}\int_{l_3(s)}^{u_3(s,x_1,x_2)}\frac{1}{(2\pi)^{3/2}\sqrt{|\det(\rho_{\theta_1})|}}$$
$$\times \exp\left(-\frac{1}{2}(x_1,x_2,x_3)^T\rho_{\theta_1}^{-1}(x_1,x_2,x_3)\right)\, dx_3dx_2dx_1\, s^{\nu-1}e^{-s^2/2}\, ds,$$

where ρ_{θ_1} is the 3×3 submatrix of ρ obtained by excluding the second row and column, while the functions at the extremes of the integrals in the second term are defined as

$$l_1(s) = \frac{s}{\sqrt{\nu}}\max\left\{t_{\alpha^{sup}},\, t_{\alpha^{sup}} - \frac{1-b}{a}t_{\alpha^{eq}}\right\} - \Theta_1^{sup},$$
$$l_2(s) = \frac{s}{\sqrt{\nu}}t_{\alpha^{sup}} - \Theta_2^{sup}, \qquad u_2(s,x_1) = x_1 + \frac{s}{\sqrt{\nu}}\frac{1-b}{a}t_{\alpha^{eq}} + \Theta_1^{sup} - \Theta_2^{sup}$$
$$l_3(s) = \frac{s}{\sqrt{\nu}}t_{\alpha^{eq}} - \Theta_1^{eq}, \qquad u_3(s,x_1,x_2) = \frac{a}{b}(x_1-x_2) + \frac{t_{\alpha^{eq}}}{b} + \frac{a}{b}(\Theta_1^{sup} - \Theta_2^{sup}) - \Theta_1^{eq}.$$

This is an alternative exact formula for power that requires one less integration. The downside is that the limits of integration are very complicated.

3.3.4 Sample size reassessment based on the interim estimation of mean and variance

A common adaptation in BE studies is sample size reassessment (or re-estimation), which allows investigators to adjust the planned sample size after the study has started, using updated variance estimates obtained during an interim analysis. In this section, we present a blinded sample size reassessment design for a three-armed trial, inspired by "Approach 1" in Zhu and Sun, 2019. The innovation in our design is that the treatment means are also estimated during the interim, allowing for a more precise re-estimation of the sample size when a certain critical assumption is true.

As in Zhu and Sun, 2019, we assume normally distributed responses with homogeneous variances, that is, $\sigma_T^2 = \sigma_R^2 = \sigma_P^2 = \sigma^2$. By fixing the allocation ratio $n_T = n_R = k\cdot n_P$ for the test, reference and placebo arms, the power becomes a function of σ^2, μ_T, μ_R, μ_P, k, n_P, θ_1, θ_2, α^{sup}, and α^{eq}. The evaluation of the required sample size is obtained by computing the power (3.11) iteratively for increasing values of n_P until the desired power is achieved. The initial sample size is based on an initial estimate of the variance σ_0^2, of the

reference mean $\mu_{R,0}$ and of the placebo mean $\mu_{P,0}$, by assuming a fixed mean ratio μ_T/μ_R contained in the equivalence range (θ_1, θ_2).

After observing outcomes from half of the initial patients, a blinded interim analysis is performed to estimate the variance using the pooled Stage 1 data from all three treatment groups:

$$\hat{\sigma}_1^2 = \frac{1}{n_T + n_R + n_P - 1} \sum_i \left(X_i - \overline{X_{S1}}\right)^2, \tag{3.18}$$

where $\overline{X_{S1}}$ is the overall mean of the Stage 1 data. In Zhu's design, the required sample size to achieve the desired statistical power is then recalculated using the updated estimate of the variance $\hat{\sigma}_1^2$. Zhu and Sun, 2019 shows that $\hat{\sigma}_1^2$ is a conservative estimate, as it tends to be upwardly biased, often leading to larger sample sizes than what is strictly necessary. Although the pooled variance would yield a more accurate estimate, calculating it would require unblinding the data, which is not permitted in our blinded analysis. The trial is then continued until the end using the newly determined sample size.

We propose a similar design with the addition of also estimating the means at the interim analysis, by making use of the following assumption:

$$\mu_T = \mu_R = \mu_P + \delta, \tag{3.19}$$

where $\delta > 0$ is the true treatment difference between reference and placebo. This assumption may be partially justified if a prior pivotal trial showed that the reference treatment was superior to placebo, with a mean difference of δ, and the new treatment is expected to produce very similar effects. This may apply, for example, when the formulation of the test drug is unchanged and only factors such as the manufacturing site, the labeling, or maybe the propellant used for an inhaler device differ. Under this assumption, the interim means approximately satisfy

$$\overline{X_T} \approx \overline{X_R} \approx \overline{X_{S1}} + \frac{n_P}{2n_T + n_P}\delta, \quad \text{and} \quad \overline{X_P} \approx \overline{X_{S1}} - \frac{2n_T}{2n_T + n_P}\delta. \tag{3.20}$$

These values can be computed without breaking the blind, as they depend only on the overall mean, and provide unbiased estimates of μ_T and μ_P if the assumption holds.

Coherently with this assumption, we will also assume $\mu_T = \mu_R$ when computing the initial sample size. In the next section, we use simulations to evaluate the power in scenarios

where the true ratio is below, above, or exactly 1. This allows us to assess the potential gain when the assumption is correct, as well as the magnitude of the error when it is not. Moreover, our design is compared with the approach of Zhu and Sun, 2019 through simulation studies.

To summarise, here is a breakdown of our proposed two-stage sample size reassessment design:

1. **Design stage (before the beginning of the trial):** Determine the initial sample size for the placebo group $n_{P,0}$, based on the preliminary estimates of the mean $\mu_{R,0}$ and variance σ_0, and other nuisance parameters, namely k, δ, θ_1, θ_2, $1-\beta$, α^{sup}, and α^{eq}.

2. **Interim analysis:** Triggered after outcomes from half of the initially planned patients have been observed. Using the Stage 1 data, estimate the reference mean and variance blindly with (3.20) and (3.18), respectively, and update the sample size accordingly. Crucially, the mean estimation requires that the allocation proportions are respected as closely as possible.

3. **Final analysis:** After all patient outcomes are collected, compute the test statistics in (3.7) and (3.8) to perform the superiority and equivalence tests.

3.4 Simulation studies

3.4.1 Study setting

Mimicking some of the quantities raising in BE studies for pulmonary disease drugs, we assume normally distributed responses with homogeneous standard deviation $\sigma = 0.4$, and a sample size ratio $k = 2$ for the active treatments to placebo, i.e., $n_T/n_P = n_R/n_P = 2$. The reference mean is set to $\mu_R = 0.5$ and the superiority effect size is fixed at $\delta = 0.25$. The significance levels for the superiority and equivalence tests are set to $\alpha^{sup} = 0.025$ and $\alpha^{eq} = 0.05$, respectively, and the equivalence bounds are $\theta_1 = 0.8$ and $\theta_2 = 1.25$. The desired power is $1-\beta = 0.8$.

For the mean ratio, alternative scenarios for power evaluation are $\mu_T/\mu_R = 0.9$, 1, and 1.1, while null scenarios for type I error evaluation assume μ_T/μ_R equal to each of the

boundaries of the equivalence range, namely 0.8 and 1.25. To clarify, we are considering a fixed value for $\mu_R = 0.5$ and $\mu_P = \mu_R - \delta = 0.25$, while μ_T varies depending on the true mean ratio.

Different initial estimates of the reference mean $\mu_{R,0}$ and standard deviation σ_0 are considered to assess situations in which the true values are underestimated, accurately estimated, or overestimated: $\mu_{R,0} \in \{\mu_R - 0.1,\ \mu_R,\ \mu_R + 1\}$ and $\sigma_0 \in \{\sigma - 0.1,\ \sigma,\ \sigma + 1\}$. All values are taken from Zhu and Sun, 2019, except for the varying initial estimates of μ_R, which were not considered in their study.

Each simulation study uses $50,000$ replicates in the alternative scenarios and $100,000$ replicates in the null scenarios.

3.4.2 Results of the simulation study

Tables 3.1 to 3.5 compare our approach (Design 1) for sample size reassessment with "Approach 1" from Zhu and Sun, 2019 (Design 2) across three alternative scenarios and two null scenarios.

Each table reports the initial sample size, the average estimated standard deviation, the average estimated mean from Design 1, the average reassessed sample sizes, and the statistical power (or type I error) obtained with both designs. These quantities are calculated under mixed combinations of the initial assumptions for $\mu_{R,0}$ and σ_0, considering accurate, overestimated, or underestimated values. Note that $\mu_{R,1}$ and $\hat{\sigma}_1$ do not depend on their initial estimates and thus remain almost the same across all rows.

Table 3.2 presents simulation results for a scenario that aligns with assumption (3.19). In this case, the mean estimate from Design 1 is unbiased, and the sample size reassessment yields a power level close to the target of 0.80. Power is slightly inflated due to the conservative estimate $\hat{\sigma}_1$ obtained by taking the square root of (3.18). The power inflation is larger in one case (row 7), when the initial standard deviation is overestimated while the initial mean is underestimated. In that case, the initial sample size is much larger than necessary because the treatment variability is assumed to be substantially greater than the treatment effect, and by the interim analysis the trial is already sufficiently powered.

For Design 2, the reassessed sample size is as accurate as Design 1 only when the initial mean estimate is correct. Otherwise, the trial may be noticeably underpowered (not enough patients are enrolled) or overpowered (more patients than necessary are enrolled).

Tables 3.1 and 3.3 show results for scenarios where $\mu_T/\mu_R = 0.9$ and 1.1, respectively. In these cases, assumption (3.19) does not match the actual treatment effects, although the mean ratio still falls within the equivalence range. In these scenarios, both designs consistently exhibit a power that is lower than desired.

Given a fixed mean ratio, Design 1 leads to similar final sample sizes (and thus similar powers) across all initial assumptions. On the other hand, the sample sizes from Design 2 are larger when the initial mean is underestimated and smaller when the initial mean is overestimated. This is a consequence of using the initial assumption $\mu_{R,0}$ also in the interim reassessment. Interestingly, this causes Design 2 to almost meet the power target in Table 3.3 when the mean is underestimated (rows 1, 4 and 7).

It is noteworthy that the power in each row of Table 3.3 is higher than in the corresponding row of Table 3.1, even when the sample size is smaller. This is due to the fact that demonstrating BE is inherently easier when the treatment effect is larger, given the same variance. Indeed, $\mu_T = \mu_R \cdot 1.1 = 0.55$ in Table 3.3, whereas it is $\mu_T = \mu_R \cdot 0.9 = 0.45$ in Table 3.1.

The values in Tables 3.4 and 3.5 were obtained by simulating under the null scenarios $\mu_T/\mu_R = 0.8$ and $\mu_T/\mu_R = 1.25$, respectively. Consistently with the previous findings, Design 1 results, on average, in similar final sample sizes across all initial assumptions, and therefore yields similar type I error rates. The final sample sizes in Design 2, instead, vary depending on $\mu_{R,0}$, leading to more diverse type I error rates. However, in both designs, the type I error rates are usually below the 0.05 level or just slightly inflated. The highest value for Design 1 is 0.0520, in Table 3.5, row 6. The highest value for Design 2 is 0.0526, in Table 3.5, row 7.

3.5 Discussion

In this chapter, we have clarified a source of ambiguity in the statistical literature: the coexistence of two definitions of the non-central t distribution. Although the numerical differences between these definitions are often negligible in practice, using the incorrect

Table 3.1 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 0.9$

			Design 1 (SSR with estimation of μ and σ)				Design 2 (SSR with estimation of σ)		
σ_0	$\mu_{R,0}$	n_0	average $\mu_{R,1}$	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power
$\sigma-0.1$	$\mu_R-0.1$	100	0.48	0.41	130(18.22)	0.47	0.41	186(16.51)	0.59
	μ_R	64	0.48	0.41	131(22.94)	0.47	0.41	119(13.25)	0.44
	$\mu_R+0.1$	45	0.48	0.41	132(27.68)	0.47	0.41	84(11.14)	0.31
σ	$\mu_R-0.1$	176	0.48	0.41	130(13.46)	0.47	0.41	186(12.47)	0.59
	μ_R	113	0.48	0.41	130(17.01)	0.46	0.41	119(9.97)	0.44
	$\mu_R+0.1$	79	0.48	0.41	131(20.54)	0.47	0.41	84(8.36)	0.32
$\sigma+0.1$	$\mu_R-0.1$	275	0.48	0.41	139(4.25)	0.49	0.41	186(9.96)	0.59
	μ_R	176	0.48	0.41	130(13.48)	0.47	0.41	119(7.97)	0.44
	$\mu_R+0.1$	124	0.48	0.41	130(16.20)	0.47	0.41	84(6.64)	0.32

Table 3.2 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.0$

			Design 1 (SSR with estimation of μ and σ)				Design 2 (SSR with estimation of σ)		
σ_0	$\mu_{R,0}$	n_0	average $\mu_{R,1}$	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power
$\sigma-0.1$	$\mu_R-0.1$	100	0.50	0.41	121(16.40)	0.83	0.41	187(16.75)	0.97
	μ_R	64	0.50	0.41	122(20.57)	0.83	0.41	120(13.36)	0.83
	$\mu_R+0.1$	45	0.50	0.41	122(24.94)	0.82	0.41	84(11.20)	0.61
σ	$\mu_R-0.1$	176	0.50	0.41	121(12.32)	0.83	0.41	187(12.56)	0.97
	μ_R	113	0.50	0.41	121(15.41)	0.83	0.41	120(10.11)	0.83
	$\mu_R+0.1$	79	0.50	0.41	121(18.50)	0.83	0.41	84(8.46)	0.62
$\sigma+0.1$	$\mu_R-0.1$	275	0.50	0.41	137(1.48)	0.89	0.41	187(10.05)	0.97
	μ_R	176	0.50	0.41	121(12.26)	0.84	0.41	120(8.07)	0.83
	$\mu_R+0.1$	124	0.50	0.41	121(14.65)	0.83	0.41	84(6.72)	0.62

form can, in general, compromise accuracy. We also reviewed some rudimentary results on the use of the multivariate non-central t distribution for computing power in bioequivalence trials.

To further illustrate its relevance for bioequivalence studies, we presented a simple modification of the blinded sample size reassessment design proposed in Zhu and Sun, 2019. By incorporating the re-estimation of the treatment means at the interim analysis (alongside the usual variance re-estimation) the reassessed sample size can be made more accurate. Our design strongly relies on the assumption that the test and reference treatments have very similar means, and it performs well only when this condition holds. Under this assumption, the proposed reassessment procedure ensures the desired power and type I error rates, even when the initial estimates of the treatment means and variance are inaccurate.

Table 3.3 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.1$

			Design 1 (SSR with estimation of μ and σ)				Design 2 (SSR with estimation of σ)		
σ_0	$\mu_{R,0}$	n_0	average $\mu_{R,1}$	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power	average $\hat{\sigma}_1$	average n_{SSR}(sd)	power
	$\mu_R-0.1$	100	0.52	0.41	114(15.05)	0.55	0.42	190(16.89)	0.74
$\sigma-0.1$	μ_R	64	0.52	0.41	114(18.96)	0.54	0.42	122(13.56)	0.57
	$\mu_R+0.1$	45	0.52	0.41	115(22.75)	0.54	0.41	86(11.43)	0.44
	$\mu_R-0.1$	176	0.52	0.42	113(11.24)	0.55	0.42	190(12.77)	0.74
σ	μ_R	113	0.52	0.42	114(14.12)	0.54	0.42	122(10.17)	0.57
	$\mu_R+0.1$	79	0.52	0.41	114(17.08)	0.55	0.41	86(8.58)	0.44
	$\mu_R-0.1$	275	0.52	0.42	137(0.41)	0.62	0.42	190(10.18)	0.74
$\sigma+0.1$	μ_R	176	0.52	0.42	113(11.21)	0.55	0.42	122(8.18)	0.57
	$\mu_R+0.1$	124	0.52	0.42	114(13.49)	0.55	0.42	86(6.81)	0.44

Table 3.4 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 0.8$

			Design 1 (SSR with estimation of μ and σ)				Design 2 (SSR with estimation of σ)		
σ_0	$\mu_{R,0}$	n_0	average $\mu_{R,1}$	average $\hat{\sigma}_1$	average n_{SSR}(sd)	type I error	average $\hat{\sigma}_1$	average n_{SSR}(sd)	type I error
	$\mu_R-0.1$	100	0.46	0.41	142(20.16)	0.0498	0.41	186(16.48)	0.0503
$\sigma-0.1$	μ_R	64	0.46	0.41	143(25.78)	0.0489	0.41	119(13.27)	0.0504
	$\mu_R+0.1$	45	0.46	0.41	144(30.96)	0.0491	0.41	83(11.11)	0.0452
	$\mu_R-0.1$	176	0.46	0.41	141(15.14)	0.0495	0.41	186(12.49)	0.0510
σ	μ_R	113	0.46	0.41	142(18.99)	0.0499	0.41	119(9.96)	0.0497
	$\mu_R+0.1$	79	0.46	0.41	142(22.97)	0.0501	0.41	83(8.35)	0.0472
	$\mu_R-0.1$	275	0.46	0.41	144(8.74)	0.0503	0.41	186(10.00)	0.0503
$\sigma+0.1$	μ_R	176	0.46	0.41	141(15.07)	0.0502	0.41	119(7.98)	0.0504
	$\mu_R+0.1$	124	0.46	0.41	142(18.22)	0.0497	0.41	83(6.65)	0.0460

Table 3.5 Comparison of two sample size reassessment designs when $\mu_T/\mu_R = 1.25$

			Design 1 (SSR with estimation of μ and σ)				Design 2 (SSR with estimation of σ)		
σ_0	$\mu_{R,0}$	n_0	average $\mu_{R,1}$	average $\hat{\sigma}_1$	average n_{SSR}(sd)	type I error	average $\hat{\sigma}_1$	average n_{SSR}(sd)	type I error
	$\mu_R-0.1$	100	0.55	0.42	105(13.40)	0.0509	0.42	197(17.43)	0.0502
$\sigma-0.1$	μ_R	64	0.55	0.42	106(16.90)	0.0501	0.42	126(14.03)	0.0494
	$\mu_R+0.1$	45	0.55	0.42	106(20.32)	0.0515	0.42	89(11.69)	0.0508
	$\mu_R-0.1$	176	0.55	0.42	105(9.85)	0.0507	0.42	197(13.21)	0.0507
σ	μ_R	113	0.55	0.42	105(12.62)	0.0505	0.42	126(10.55)	0.0505
	$\mu_R+0.1$	79	0.55	0.42	105(15.10)	0.0520	0.42	89(8.84)	0.0497
	$\mu_R-0.1$	275	0.55	0.42	137(0.07)	0.0497	0.42	197(10.52)	0.0526
$\sigma+0.1$	μ_R	176	0.55	0.42	105(9.84)	0.0511	0.42	126(8.46)	0.0502
	$\mu_R+0.1$	124	0.55	0.42	105(11.99)	0.0501	0.42	89(7.04)	0.0505

Conclusions

Across the three chapters of this thesis, we investigated distinct methodological questions in the design and analysis of clinical trials, all motivated by the need to improve decision-making, efficiency, and interpretability in modern study practice. Each chapter addressed a specific challenge — impact of the interim recommendation on the probability of success in group-sequential designs, response-adaptive randomisation in multi-armed trials, and sample size reassessment in bioequivalence studies — but collectively they highlight the importance of sophisticated statistical tools for guiding complex trial decisions.

In the first part of this work, we examined the relationship between the *a priori* probability of study success (PoS) and the conditional probability of success after passing an interim analysis (PoS_{post}). We demonstrated how the futility and efficacy thresholds shape these probabilities and how, in some cases, moderately increasing the futility boundary can meaningfully raise PoS_{post} with only minor loss in unconditional PoS. These findings emphasise the importance of selecting interim rules that not only achieve desired operating characteristics but also support transparent communication with non-statisticians. To our knowledge, this is the first systematic exploration of how interim decisions influence conditional probabilities of success, offering a practical framework for aligning design choices with stakeholder expectations.

The second part of the thesis proposed a new response-adaptive randomisation (RAR) procedure based on weighted information measures for multi-armed trials with continuous outcomes. By integrating Bayesian modelling of both treatment means and variances with a parametric weight functions that emphasises clinically desirable target values, we extended the existing methodology of context-dependent RAR and introduced a novel flexible random allocation rule that can be tuned to balance the exploration-exploitation trade-off, according to the trial objectives. Simulation studies showed that the proposed design can

deliver substantial benefits in both patient allocation and statistical power, particularly when a treatment closely matches prespecified clinical targets. While the method may underperform when the optimal treatment is far from these targets, it offers clear advantages when the objective of the study is to identify the treatment most aligned with specific clinical characteristics rather than simply the best performing one. This highlights the importance of the context of the study during the investigation of experimental treatments.

Finally, in the third part, we clarified a source of ambiguity in the biostatistical literature concerning competing definitions of the non-central t distribution and discussed their implications for power calculations in bioequivalence trials. Building on this theoretical clarification, we proposed a modification of a blinded sample size reassessment design that incorporates (blind) interim re-estimation of treatment means in addition to variance. When the test and reference treatments share the same mean — an assumption typical in bioequivalence studies — this refinement leads to a more accurate reassessment of the sample sizes, achieving the intended power even when the treatment parameters were initially misspecified, while maintaining the type I error rate under the desired level.

Taken together, these contributions provide new insights into the evolving field of adaptive designs in clinical trials. Although the investigated methods address different aspects of trial design, they share a common goal: improving the reliability and interpretability of statistical decision-making in settings where uncertainty is substantial and interim information must be used sensibly.

References

Aziz, M., Kaufmann, E., & Riviere, M.-K. (2021). On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, *22*(14), 1–38.

Azriel, D., Mandel, M., & Rinott, Y. (2011). The treatment versus experimentation dilemma in dose finding studies. *Journal of Statistical Planning and Inference*, *141*(8), 2759–276.

Bacchieri, A., & Cioppa, G. (2007). *Fundamentals of clinical research. bridging medicine, statistics and operations*. Springer.

Bartlett, R., Roloff, D., Cornell, R., Andrews, A., Dillon, P., & Zwischenberger, J. (1985). Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics Journal*, *76*(4), 479–487.

Belis, M., & Guiasu, S. (1968). A quantitative-qualitative measure of information in cybernetic system. *IEEE Transactions on Information Theory*, *14*(4), 593–594.

Biswas, A., Mandal, S., & Bhattacharya, R. (2011). Multi-treatment optimal response-adaptive designs for phase iii clinical trials. *Journal of the Korean Statistical Society*, *40*(1), 33–44.

Blackwelder, W. (1982). Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, *3*(4), 345–353.

Blackwelder, W., & Chang, M. (1984). Sample size graphs for proving the null hypothesis. *Controlled Clinical Trials*, *5*(2), 97–105.

Boulaguiem, Y., Insolia, L., Victoria-Feser, M. P., Couturier, D. L., & Guerrier, S. (2025). Multivariate adjustments for average equivalence testing. *Statistics in Medicine*, *44*(15–17), e10258.

Boulaguiem, Y., Quartier, J., Lapteva, M., Kalia, Y. N., Victoria-Feser, M. P., Guerrier, S., & Couturier, D. L. (2024). Finite sample corrections for average equivalence testing. *Statistics in Medicine*, *43*(5), 833–853.

Carrol, K. (2013). Decision making from phase ii to phase iii and the probability of success: Reassured by "assurance"? *Journal of Biopharmaceutical Statistics*, *23*, 1188–1200.

Caruso, G., & Mozgunov, P. (2024). A response-adaptive multi-arm design for continuous endpoints based on a weighted information measure. *arXiv*, *pre-print*.

Chang, Y., Tsong, Y., Dong, X., & Zhao, Z. (2014). Sample size determination for a three-arm equivalence trial of normally distributed responses. *Journal of biopharmaceutical statistics*, *24*(6), 1190–1202.

Chow, S.-C., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research*. Chapman & Hall/CRC.

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics*, *5*, 305–309.

Chuang-Stein, C., & Kirby, S. (2017). *Quantitative decisions in drug development*. Springer.

Crisp, A., Miller, S., Thompson, D., & Best, N. (2018). Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. *Pharmaceutical Statistics*, *17*, 317–328.

Davit, B., Nwakama, P., Buehler, G., Conner, D., Haidar, S., Patel, D., & Woodcock, J. (2009). Comparing generic and innovator drugs: A review of 12 years of bioequivalence data from the united states food and drug administration. *Annals of Pharmacotherapy*, *43*(10), 1583–1597.

DeMets, D., & Lan, K. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, *13*, 1341–1352.

FDA. (1992). *Guidance on statistical procedures for bioequivalence using a standard two-treatment crossover design*. Food and Drug Administration.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England journal of Medicine*, *317*(3), 141–145.

Gasparini, M., Di scala, L., Bretz, F., & Racine-Poon, A. (2013). Predictive probability of success in clinical drug development. *Epidemiology, Biostatistics and Public Health*, *10*(1), e8760-1–e8760-14.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Springer.

Genz, A., Bretz, F., Miwa, T., & Mi, X. (2025). *R package 'mvtnorm': Multivariate normal and t distributions* [version 1.3-3].

Grieve, A. (2022). *Hybrid frequentist/bayesian power and bayesian power in planning clinical trials*. CRC press.

Grieve, A. (2023). Probability of success and group sequential designs. *Pharmaceutical Statistics*, 1–19.

Hauschke, D., Kieser, M., Diletti, E., & Burke, M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics In Medicine*, *18*, 93–105.

Hauschke, D., Steinijans, V., & Pigeot, I. (2007). *Bioequivalence studies in drug development. methods and applications*. John Wiley & Sons.

Hinds, D., & Sun, W. (2025). An adaptive three-arm comparative clinical endpoint bioequivalence study design with unblinded sample size re-estimation and optimized allocation ratio. *Pharmaceutical Statistics*, *24*(1), 1–23.

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2014). Observer bias in randomized clinical trials with time-to-event outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *International Journal of Epidemiology*, *43*(3), 937–948.

Hu, F., & Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, *34*(28), 3724–3749.

ICH. (2000). *Ich e10 choice of control group in clinical trials - scientific guideline*. ICH Expert Working Group.

Jennison, C., & Turnbull, B. (1999). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC.

Jones, B., Jarvis, P., Lewis, J., & Ebbutt, A. (1996). Trials to assess equivalence: The importance of rigorous methods. *British Medical Journal*, *313*, 36–39.

Kasianova, K., Kelbert, M., & Mozgunov, P. (2021). Response adaptive designs for phase ii trials with binary endpoint based on context-dependent information measures. *Computational Statistics & Data Analysis*, *158*, 107187.

Kasianova, K., Kelbert, M., & Mozgunov, P. (2023). Response-adaptive randomization for multiarm clinical trials using context-dependent information measures. *Biometrical Journal*, *65*(8), 2200301.

Kelbert, M., & Mozgunov, P. (2015). Asymptotic behaviour of the weighted renyi, tsallis and fisher entropies in a bayesian problem. *Eurasian Mathematical Journal*, *6*(2), 6–17.

Kieser, M., & Hauschke, D. (1999). Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics*, *9*(4), 641–650.

Kotz, S., & Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Kunzmann, K., Grayling, M., Lee, K., Robertson, D., Rufibach, K., & Wason, J. (2021). A review of bayesian perspectives on sample size derivation for confirmatory trials. *The American Statistician*, *75*, 424–432.

Lu, J., & Yeh-Fong, C. (2022). Consideration of the adaptive randomization allocation ratio in the presence of treatment group heteroscedasticity in clinical trials. *Journal of biopharmaceutical statistics*, *32*(3), 511–526.

Mozgunov, P., & Jaki, T. (2020a). Improving safety of the continual reassessment method via a modified allocation rule. *Statistics in medicine*, *39*(7), 906–922.

Mozgunov, P., & Jaki, T. (2020b). An information theoretic approach for selecting arms in clinical trials. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(5), 1223–1247.

Mozgunov, P., Jaki, T., & Gasparini, M. (2019). Loss functions in restricted parameter spaces and their bayesian applications. *Journal of Appl Stat.*, *46*(13), 2314–2337.

O'Hagan, A., Stevens, J., & Campbell, M. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*, 187–201.

Phillips, K. (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of pharmacokinetics and biopharmaceutics*, *18*(2), 137–144.

Piantadosi, S. (2017). *Clinical trials: A methodologic perspective*. John Wiley & Sons.

Robertson, D., Lee, K. M., López-Kolkovska, B. C., & S. Villar, S. (2023). Response-adaptive randomisation in clinical trials: From myths to practical considerations. *Statistical Science*, *38*(2), 185–208.

Rosenberger, W. F., & Lachin, J. M. (2016). *Randomization in clinical trials*. John Wiley & Sons.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., & Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, *57*(3), 909–913.

Rothwell, J., Julious, S., & Cooper, C. (2018). A study of target effect sizes in randomised controlled trials published in the health technology assessment journal. *Trials*, *19*.

Rufibach, K., Burger, H., & Abt, M. (2016). Bayesian predictive power: Choice of prior and some recommendations for its use as probability of success in drug development. *Pharmaceutical Statistics*, *15*, 438–446.

Rufibach, K., Jordan, P., & Abt, M. (2016). Sequentially updating the likelihood of success of a phase 3 pivotal time-to-event trial based on interim analyses or external information. *Journal of Biopharmaceutical Statistics*, *26*, 191–201.

Sasabuchi, S. (1988). A multivariate one-sided test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, *42*, 9–19.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*(6), 657–680.

Senn, S. (2007). *Statistical issues in drug development*. John Wiley & Sons.

Shi, H., & Yin, G. (2019). Control of type i error rates in bayesian sequential designs. *Bayesian Anal.*, *14*(2), 399–425.

Smith, A., & Villar, S. (2018). Bayesian adaptive bandit-based designs using the gittins index for multi-armed trials with normally distributed endpoints. *Journal of Appl Stat.*, *45*(6), 1052–1076.

Soch, J. (2024). The book of statistical proofs. Zenodo.

Spiegelhalter, D., Freedman, L., & Blackburn, P. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, *7*, 1–87.

Suhov, Y., Stuhl, I., Sekeh, S. Y., & Kelbert, M. (2016). Basic inequalities for weighted entropies. *Aequationes mathematicae*, *90*, 817–848.

Temple, J., & Robertson, J. (2021). Conditional assurance: The answer to the questions that should be asked within drug development. *Pharmaceutical Statistics*, 1–10.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.

Trippa, L., Lee, E., Wen, P., Batchelor, T., Cloughesy, T., Parmigiani, G., & Alexander, B. (2012). Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of clinical oncology*, *30*(26), 3258–3263.

Tymofyeyev, Y., Rosenberger, W. F., & Hu, F. (2007). Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association*, *102*(477), 224–234.

Villar, S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, *30*(2), 199–215.

Williamson, F., & Villar, S. (2020). A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, *76*(1), 197–209.

Wilson, I., Julious, S., Yap, C., Todd, S., & Dimairo, M. (2025). Response adaptive randomisation in clinical trials: Current practice, gaps and future directions. *Stat Methods Med Res.*, *18*, 9622802251348183.

Woods, S., Sholomskas, M., D.E. an Shear, Gorman, J., Barlow, D., Goddard, A., & Cohen, J. (1998). Efficient allocation of patients to treatment cells in clinical trials with more than two treatment conditions. *American journal of psychiatry*, *155*(10).

Yang, A., & Sun, W. (2019). An exact method for power calculation for a three-arm clinical endpoint bioequivalence study. *International Journal of Statistics and Probability*, *8*(1), 25–39.

Zhang, L., & Rosenberger, W. F. (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, *62*(2), 562–569.

Zhu, L., & Sun, W. (2019). Adaptive clinical endpoint bioequivalence studies with sample size re-estimation based on a nuisance parameter. *Journal of Biopharmaceutical Statistics*, *29*(5), 776–799.