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)

Enhancing decision making in Randomized Controlled Trials via efficient use of internal and external information

By

Marco Ratta

******

Supervisor(s):
Prof. Mauro Gasparini, Supervisor
Doc. Gaëlle Saint-Hilary, Co-Supervisor
Prof. Pavel Mozgunov, Co-Supervisor

Doctoral Examination Committee:
Prof. Annette Kopp-Schneider, Referee, German Cancer Research Center (DKFZ)
Doc. Sebastian Weber, Referee, Novartis

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.

Marco Ratta
2026

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

Acknowledgements

Al termine di questo percorso di dottorato, desidero rivolgere un pensiero di ringraziamento a coloro che mi hanno accompagnato in questi anni. Innanzitutto, un grazie speciale al Prof. Gasparini per la saggezza e i preziosi consigli che hanno orientato il mio cammino. Grazie a Gaëlle, per avermi introdotto al mondo della ricerca e per aver stimolato la mia curiosità fin dalle prime fasi del mio percorso; e a Pavel, per essere stato una guida costante e paziente, per il supporto - accademico e umano - e soprattutto per essere stato capace di trasmettermi la passione per la ricerca. Ringrazio inoltre Donia Skanji, Valentine Barboux e le persone di Servier per avermi accolto nella loro sede a Parigi e per aver contribuito alla mia ricerca con continui ed interessanti scambi di idee durante tutto il percorso di dottorato. Infine, il ringraziamento più importante va alle persone a me più care: ad Alice, mia futura sposa, per essersi presa e per continuare a prendersi cura di me ogni giorno; ai miei genitori, per avermi permesso di coltivare i miei interessi e costruire il mio futuro; ai miei fratelli Valentina e Samuele, per il supporto costante. Un grazie di cuore a Federico, Gianmarco e Luca, per aver condiviso con me le gioie e le fatiche di questo percorso, e alle mie coinquiline Silvia e Matilde, per aver reso Torino un po' più casa.

Abstract

This thesis advances the methodology of Randomized Controlled Trials (RCTs) by developing approaches that enhance efficiency, robustness, and decision-making under uncertainty. Across several studies, we investigate strategies for optimally leveraging both internal and external information—such as early or surrogate outcomes, historical data, and multiple endpoints—to improve trial performance while maintaining strict statistical validity. Key contributions include: (*i*) adaptive group sequential designs integrating surrogate endpoints and predictive metrics; (*ii*) robust dynamic borrowing methods using mixture priors; and (*iii*) frameworks for incorporating benefit–risk assessment into multi-arm, seamless phase II/III trials. The proposed methods are evaluated through theoretical analysis and extensive simulation studies, demonstrating their ability to increase trial efficiency, reduce the risk of premature or erroneous conclusions, and inform regulatory and clinical decision-making. Collectively, these developments provide practical tools to accelerate drug development while upholding scientific rigor and statistical integrity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An overview of clinical drug development

The development of new medicine is a complex, multistage process by which potentially promising compounds discovered in the laboratory are tested and ultimately transformed into safe and efficacious therapies, able to improve patients' life. Clinical drug development represents the core of this process, bridging preclinical discovery and regulatory approval, and provides the empirical basis for claims of safety, dosage, and therapeutic benefit.

The whole drug development process is conventionally structured into two different stages, each characterized by distinct objectives and methodological features: before human testing begins, preclinical studies (*in vitro*, *in vivo* in animal models) are performed in order to establish early evidence of activity of the drug (including pharmac-odynamics and pharmaco-kinetics), and toxicology. Subsequently, only compounds demonstrating acceptable safety profiles and preliminary evidence of efficacy proceed to the so-called *clinical stages*, where they are evaluated in human subjects. Once in the clinical phase, development is typically structured into four stages (Phase I–IV), each with distinct scientific and statistical objectives.

- **Phase I:** The primary goal is to assess safety and tolerability as well as initial pharmaco-kinetic and pharmaco-dynamic profiles. These trials are generally small (20–100 participants), often involving healthy volunteers, and use escalating-dose designs to identify the maximum tolerated dose (MTD). Statistical approaches include model-based or Bayesian dose-escalation methods.

- **Phase II:** At this stage, preliminary evidence of efficacy ("proof of concept") is investigated, safety continues to be assessed, and suitable dose ranges and treatment schedules are identified. Phase II trials typically enroll several hundred patients with the target condition and commonly employ randomization with appropriate control groups. From a statistical perspective, key considerations include the estimation of treatment effect sizes, modeling

of dose-response relationships, and the control of multiplicity when multiple treatment arms (or doses) are evaluated.

- **Phase III:** At this stage, the efficacy of the investigational drug is assessed in so-called *confirmatory trials*, which aim to establish both efficacy and safety in larger and more heterogeneous patient populations, often across multiple centers. These trials are typically randomized, frequently double-blinded, and may enroll thousands of participants. Their results constitute the primary sign of evidence for regulatory approval. From a statistical perspective, key challenges include large-sample inference, multiplicity adjustments for multiple endpoints, non-inferiority or superiority testing, interim analyses, strict control of Type I error rate and maximization of the statistical power.

- **Phase IV:** In this phase, post-marketing studies continue after regulatory approval to monitor long-term safety and effectiveness in real-world populations. These trials may detect rare adverse events or evaluate additional indications. Methodologically, Phase IV often relies on observational data, pragmatic trials, or registry analyses, requiring causal inference methods and advanced approaches to control bias and confounding.

Together, these stages form an iterative framework in which evidence accumulates to support regulatory decision-making. The ultimate responsibility for marketing authorization lies with health authorities, such as the European Medicines Agency (EMA) in Europe, the Food and Drug Administration (FDA) in the U.S. and the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan.

1.2 Randomized Controlled Trials (RCTs)

Randomized Controlled Trials (RCTs) constitute the gold standard for generating rigorous and unbiased evidence regarding the efficacy and safety of novel medical interventions. They play a pivotal role particularly in Phase II and Phase III of drug development, during which high-quality empirical data are required to quantify treatment effects and establish a robust foundation for regulatory approval. Randomization is essential in this context because it balances both observed and unobserved covariates across treatment groups, thereby mitigating confounding and enabling credible causal inference.

In an RCT, participants are randomly assigned to either the treatment group (T) or the control group (C). Suppose that for each patient i belonging to group $k \in \{T, C\}$, a continuous or dichotomous outcome variable Y_k^i is observed and call θ a treatment effect parameter (which may be the difference in mean in case of normally distributed outcomes, the log-hazard ratio in case of survival outcomes or the log-odds ratio in case of binary outcomes), quantifying the relative efficacy of the experimental treatment with respect to the control.

The central research question, whether the treatment demonstrates superiority over control, can

be formulated as a statistical hypothesis test, with the null and alternative hypotheses given by

$$H_0 : \theta \leq 0 \quad \text{and} \quad H_1 : \theta > 0.$$

Statistical inference can be conducted under either the frequentist or the Bayesian paradigm. Since both frameworks are employed in this dissertation, they are outlined below.

1.2.1 Frequentist Inference Approach

Under the frequentist framework, a standard assumption is that the maximum likelihood estimate (MLE) of the treatment effect follows a normal distribution, namely

$$\hat{\theta} \sim N\left(\theta, \mathcal{I}^{-1}(\theta)\right)$$

where $\mathcal{I}^{-1}(\theta)$ is the Fisher information corresponding to the treatment effect parameter, which depend on the sample size of the trial. It follows that a test statistic $Z = \hat{\theta} \cdot \mathcal{I}(\theta)$, following a normal distribution centered at the true value θ and with unit variance can be constructed and the hypothesis test is conducted according to the following decision rule:

- If $Z > \eta$, reject H_0 and conclude that the treatment is superior to control;
- If $Z \leq \eta$, fail to reject H_0 and conclude that superiority cannot be claimed.

The critical threshold η is determined such that

$$P\left(Z > \eta \,\middle|\, \theta = 0\right) = \alpha, \tag{1.1}$$

where α denotes the nominal Type I error rate, i.e. the probability to reject the null hypothesis when the latter holds. Instead the sample size per arm, n_j for $j \in \{C, T\}$, is chosen to ensure that

$$P\left(Z > \eta \,\middle|\, \theta = \Delta\right) = 1 - \beta, \tag{1.2}$$

where β represents the Type II error rate, i.e. the probability to not reject the null hypothesis when the true treatment effect parameter assume a target value $\theta = \Delta$. The quantity $1 - \beta$ corresponds to the statistical power of the trial.

1.2.2 Bayesian Inference Approach

In the Bayesian framework, the parameter of interest θ is treated as random variable characterized by a prior distribution that reflect existing knowledge before data collection. Specifically, prior

distributions $\pi^0(\theta)$ is specified, then upon observing data $\boldsymbol{y_j} = \{y_{j,1}, \ldots, y_{j,n_j}\}$, the posterior distribution is obtained updating the prior distribution with the observed data via Bayes' theorem:

$$f(\theta \mid \boldsymbol{y_C}, \boldsymbol{y_T}) \propto \pi^0(\theta) \times f(\boldsymbol{y_C}, \boldsymbol{y_T} \mid \theta).$$

where $f(\cdot)$ indicates the data likelihood. Once the posterior distributions of θ is obtained, the decision rule for testing superiority is given by:

- If $P(\theta > 0 \mid \boldsymbol{y_C}, \boldsymbol{y_T}) > \eta$, reject H_0 and claim superiority of the treatment;
- If $P(\theta > 0 \mid \boldsymbol{y_C}, \boldsymbol{y_T}) \leq \eta$, fail to reject H_0 and do not claim superiority.

The selection of the probability threshold η plays a critical role in this setting. Specifically, higher values of η correspond to lower rejection rates, and vice versa. To ensure appropriate control over the false positive rejection rate, the threshold η is calibrated—either numerically or, in certain cases, analytically—so that the following condition is satisfied:

$$P(P(\theta > 0 \mid \boldsymbol{y_C}, \boldsymbol{y_T}) > \eta \mid \theta = 0) = \alpha.$$

Similarly, the required sample size per group, n_k for $k \in \{C, T\}$, is determined to control the type II error rates, or equivalently to ensure a minimum power under a target favorable scenario $\theta = \Delta$; namely

$$P(P(\theta > 0 \mid \boldsymbol{y_C}, \boldsymbol{y_T}) > \eta \mid \theta = \Delta) = 1 - \beta.$$

1.3 Adaptive designs

In a traditional framework, the design of a clinical trial is fixed in all its parts prior to the start of the trial, and all statistical analyses (at least for the purposes of primary efficacy analysis) are performed only upon completion of the study.

As opposite to these approaches, the so called *adaptive designs* have been developed, allowing for pre-specified modifications at pre-specified milestones of the trial (referred to *interim analyses*) conducted throughout the course of the study, based on the accumulated information from participants. Such modifications may include, but are not limited to, adjustments to the number of treatment arms, the treatment allocation ratio, or the total sample size [1].

While these adaptations could potentially lead to improved efficiency of the clinical trial, a careful statistical modeling is needed to maintain the integrity of the trial particularly in terms of type I error control.

In this dissertation, we focus on two distinct classes of adaptive designs: the *group sequential designs* (GSD) and the *multi-arm multi-stage designs* (MAMS) incorporating interim treatment

selection. The fundamental concepts underlying these approaches, along with their statistical formalization, are presented in the subsequent subsections.

1.3.1 Group Sequential Designs (GSD)

Group Sequential Designs (GSDs) are a type of adaptive design commonly used in clinical trials and other experimental studies, where data are analyzed at different points in time (referred to as *interim analyses*) before the final analysis. Unlike traditional fixed-sample designs, GSDs allow for the possibility of stopping the trial early based on interim results, either because the treatment is clearly effective (stop for efficacy) or clearly ineffective (stop for futility). A general and broadly used framework for this purpose has been presented by Jennison and Turnbull [2].

Consider a trial in which an experimental treatment T is compared to a control arm C across $k = 1, \ldots, K$ interim analyses. Assume that $\boldsymbol{y}_{C,k} = \{y_{C,k,1} \ldots, y_{C,k,n_{Ck}}\}$ are the accumulated data for the $n_{C,k}$ patients enrolled in the control arm up to stage k, while $\boldsymbol{y}_{T,k} = \{y_{T,k,1} \ldots, y_{T,k,n_{Tk}}\}$ are the accumulated data for the $n_{T,k}$ patients enrolled in the control arm up to stage k. Group Sequential Designs are constructed under the assumption that at each interim look k, the estimate of the treatment effect parameter θ, namely $\hat{\theta}^k = g(\boldsymbol{y}_{C,k}, \boldsymbol{y}_{T,k})$ follows a normal distribution with mean the true treatment effect θ and variance which depends on the amount of information available at the k-th interim analysis; namely

$$\hat{\theta}_k \sim \mathcal{N}\left(\theta, \mathcal{I}_k^{-1}(\theta)\right) \qquad k = 1, \ldots, K. \tag{1.3}$$

Moreover, the vector of treatment effect estimates across the interim analyses, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)^\top$, follows a multivariate normal distribution with covariances determined by the accumulated Fisher information:

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_K \end{pmatrix} \sim \mathcal{N}_K \left(\begin{pmatrix} \theta \\ \vdots \\ \theta \end{pmatrix}, \begin{pmatrix} \mathcal{I}_1^{-1}(\theta) & \mathcal{I}_1^{-1}(\theta) & \cdots & \mathcal{I}_1^{-1}(\theta) \\ \mathcal{I}_1^{-1}(\theta) & \mathcal{I}_2^{-1}(\theta) & \cdots & \mathcal{I}_2^{-1}(\theta) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{I}_1^{-1}(\theta) & \mathcal{I}_2^{-1}(\theta) & \cdots & \mathcal{I}_K^{-1}(\theta) \end{pmatrix} \right). \tag{1.4}$$

Under the latter assumption, different strategies have been proposed to early terminate the trial.

Sequential hypothesis testing
In this setting, the null hypothesis $H^0 : \theta \leq 0$ can be tested versus the alternative hypothesis $H' : \theta > 0$ at each interim look by comparing the Z-statistic at the k-th interim analysis $Z_k = \hat{\theta}_k \mathcal{I}_k$ with some critical value u_k and l_k, so that

- if $Z_k > u_k$, then H^0 is rejected and the trial is stopped for efficacy
- if $Z_k < l_k$, then H^0 is accepted and the trial is stopped for futility

– if $l_k \leq Z_k \leq u_k$, then a decision is not made and the trial is continued to stage $k+1$

In this context the overall probability to reject the null hypothesis can be written as the probability to reject H^0 at any interim analysis or at the final analysis, and accordingly type I error and type II errora can be expressed as

$$\begin{aligned} \alpha &= \sum_{k=1}^{K} P\big(\text{reject } H^0 \text{ at stage } k \mid \theta = 0\big) \\ \beta &= \sum_{k=1}^{K} P\big(\text{not reject } H^0 \text{ at stage } k \mid \theta = \Delta\big) \end{aligned} \tag{1.5}$$

Based on Equation 1.5, the critical values u_k and l_k are determined in order to ensure that the type I error is controlled respectively at the nominal level α under the null hypothesis, and the target statistical power $1-\beta$ is achieved under the target alternative hypothesis.

Among the different strategies proposed in literature to set the critical values some common choices are the Pocock approach [3], where the critical values are set to be equal at each interim stage, and the O'Brien and Fleming approach [4], where the critical values are set to be more stringent at earlier stages and progressively less stringent towards later stages of the trial. A more general approach comprising the latter two is the *α (respectively β spending function approach* [5], which consists in splitting the nominal α (respectively β) level into k levels α_k, so that $\sum_{k=1}^{K} \alpha_k = \alpha$, and setting the critical bounds u_k and l_k so that

$$\begin{aligned} P\left(\text{reject } H^0 \text{ at stage } k \mid H^0\right) &= \alpha_k \qquad \forall k = 1 \ldots, K \\ P\left(\text{accept } H^0 \text{ at stage } k \mid H^{'}\right) &= \beta_k \qquad \forall k = 1 \ldots, K \end{aligned}$$

Another approach proposed in the literature for controlling the family-wise error rate (FWER) is the so-called *combination test*. This method involves computing stage-wise p-values using only the data collected at each stage, and then, at each stage, calculating a global p-value by combining the stage-wise p-values.

Alternative approaches

Early termination of the trial either for efficacy or for futility may be planned under alternative decision rules. Some common alternatives are:

– *Conditional Power (CP):* Conditional Power [6] is the probability of ultimately rejecting the null hypothesis, given the interim data and assuming a specific value for the treatment effect in the remainder of the study. It can be expressed as

$$\text{CP} = P\big(\text{Reject } H^0 \mid \{Y_k\}\ ;\ \theta = \theta^*\big)$$

where θ^* is a specific assumed value for the treatment effect parameter. Two common choices are setting it equal to the interim maximum likelihood estimate of the treatment effect or to the pre-specified target treatment effect used for sample size calculation; the first case is referred to as *observed Conditional Power*, while the second is referred to as *assumed Conditional Power*.

– *Predictive Probability of Success (PPoS):* Predictive Probability of Success [7, 8](also known as *assurance* [9]) is a Bayesian analogue of conditional power and represents the posterior probability of trial success, given the interim data and integrating over the uncertainty in the treatment effect. Rather than assuming a fixed effect size, PPoS incorporates the posterior distribution of the treatment effect to account for current evidence and prior information. It is formally defined as

$$\text{PPoS} = \int P\big(\text{Reject } H^0 \mid \theta\big) \cdot p\left(\theta \mid \boldsymbol{y}_{C,k}, \boldsymbol{y}_{T,k}\right) d\theta,$$

where $p\left(\theta \mid \boldsymbol{y}_{C,k}, \boldsymbol{y}_{T,k}\right)$ is the posterior distribution of the treatment effect θ, and $P(\text{Reject } H^0 \mid \theta)$ is the conditional probability of rejecting the null hypothesis at the final analysis given the true treatment effect θ.

1.3.2 Multi-arm multi-stage designs (MAMS)

Group Sequential Designs (GSD), as previously introduced, provide a methodological framework for early stopping in two-arm clinical trials. However, in many applied settings, particularly in clinical research, it is common to evaluate multiple experimental treatments concurrently against a shared control. The Multi-Arm Multi-Stage (MAMS) design generalizes the principles of GSD to such settings, accommodating comparisons between several experimental arms and a common control across multiple interim analyses.

Consider a trial with J experimental treatments, denoted by $T_1, \ldots, T_J$, evaluated against a common control treatment T_0 over K stages. At each interim stage $k = 1, \ldots, K$, data are collected from patients allocated to each treatment arm. Let $\boldsymbol{y}_{j,k} = \{y_{j,k,1} \ldots, y_{j,k,n_j^k}\}$ represent the data observed in treatment arm $j \in \{0, \ldots, J\}$ up to stage k, where n_j^k denotes the cumulative number of participants in arm j at the k-th analysis (and the index $j = 0$ is usded for the control arm).

Let θ_j represent the treatment effect parameter quantifying the difference between experimental arm $j \geq 1$ and the control. At each interim analysis k, an estimator $\hat{\theta}_j^k = g(\boldsymbol{y}_{j,k}, \boldsymbol{y}_{0,k})$ is computed for each experimental arm, where $g(\cdot)$ is a suitable function of the observed outcomes in arm j and the control arm. Under the assumptions of normality and sufficiently large sample sizes, the asymptotic distribution of the estimators is given by

$$\hat{\theta}_j^k \sim \mathcal{N}\left(\theta_j, \left[\mathcal{I}_j^k(\theta_j)\right]^{-1}\right), \quad \text{for } j = 1, \ldots, J,\ k = 1, \ldots, K, \tag{1.6}$$

where $\mathcal{I}_j^k(\theta_j)$ denotes the Fisher information accumulated for treatment arm j at stage k. The trial involves the simultaneous testing of J individual null hypotheses of the form:

$$H_{0j} : \theta_j \leq 0 \quad \text{versus} \quad H_{1j} : \theta_j > 0, \quad \text{for } j = 1, \ldots, J.$$

A central inferential challenge in this context is the control of the *familywise error rate* (FWER), defined as the probability of falsely rejecting at least one true null hypothesis, at a prespecified significance level α. There are two types of control of the FWER in this setting: *weak control* and *strong control*. *Weak control* refers to controlling the family-wise error rate under the *global null hypothesis*—that is, when all null hypotheses ($\bigcap_j H_{0j}$) are true. In contrast, *strong control* ensures error rate control under any configuration of true and false null hypotheses—i.e., for any case where at least one individual null hypothesis is true ($\bigcup_j H_{0j}$).

Within the MAMS framework, we will focus on the design proposed by Stallard and Todd [10], which incorporates interim treatment selection withing a GSD. Particularly, at the first stage, a vector of test statistics $\mathbf{Z}_1 = (Z_{1k}, \ldots, Z_{Jk})$ is computed, where each component is defined as $Z_{jk} = \hat{\theta}_j^k \cdot \mathcal{I}_j^k$. Based on these statistics, a pre-specified selection rule is applied to identify the most promising treatment arm, while the remaining arms are dropped from further consideration. The selected treatment is then subjected to continued evaluation in subsequent stages. If, for any stage $k > 1$, the test statistic Z_{Sk} for the selected treatment exceeds a predefined efficacy boundary η_k, the corresponding null hypothesis is rejected.

This approach has been further extended by Magirr et al. [11], who generalized the selection mechanism to permit the retention of multiple treatment arms at interim stages, rather than a single best arm.

In both methodologies, the treatment selection decision is assumed to rely solely on the interim test statistics derived from the primary endpoint. In Chapter 5, we propose an extension to this framework by incorporating an alternative selection rule that accommodate multiple endpoints, thereby enhancing the flexibility and applicability of MAMS designs in multi-dimensional outcome settings.

1.4 Historical Borrowing

In all the methods discussed so far, inference relies solely on data collected within the trial itself—meaning that hypothesis testing is based exclusively on information from patients enrolled in the current study. However, in certain settings, such as trials for rare diseases or pediatric populations, recruiting a sufficient number of participants can be extremely difficult or even

ethically problematic. In these cases, incorporating information from *external* sources may offer a valuable alternative.

Historical borrowing refers to the incorporation of external or historical control data into the analysis of a current clinical trial, with the aim of improving efficiency or reducing the required sample size. Within the Bayesian framework, borrowing is achieved by specifying informative prior distributions for model parameters based on historical evidence. A central challenge in this context is the potential for systematic differences between the historical and current data (referred to as *drift* or *prior-data conflict*) which can lead to biased or misleading inference if not properly addressed.

Several Bayesian approaches have been developed to mitigate this issue. A review of the most widespread methodologies can be found in [12] and [13]. Among them, some alternatives are *Power priors* [14, 15] and *commensurate priors* [3]. *Power priors* incorporate historical data through a likelihood raised to a power parameter, allowing the researcher to control the degree of borrowing based on prior knowledge or empirical evidence. *Commensurate priors*, on the other hand, introduce a hierarchical structure in which the similarity between historical and current data is modeled explicitly via a variance or precision parameter, enabling dynamic and data-driven borrowing.

A more recent and principled class of methods is based on *Robust Mixture Priors (RMPs)* [16, 17], which address prior-data conflict by combining informative and non-informative components within a mixture prior formulation. Let θ denote the parameter of interest (e.g., a treatment effect or control parameter), and let $\pi_H(\theta)$ represent a prior distribution for such parameter derived from historical data (for example via a meta-analytic predictive approach [16]). The robust mixture prior is defined as

$$\pi_{\text{RMP}}(\theta) = w \cdot \pi_H(\theta) + (1 - w) \cdot \pi_0(\theta), \tag{1.7}$$

where $\pi_0(\theta)$ denotes a vague or weakly informative prior reflecting skepticism about the relevance of the historical information, and $w \in [0, 1]$ is a fixed weight that typically encodes the prior belief in the exchangeability between historical and current data sources. This formulation ensures that, in the presence of prior-data conflict, posterior inference is primarily driven by the non-informative component, thereby protecting against bias. Conversely, when the historical and current data are in agreement, the informative component contributes substantially to the posterior, allowing efficient borrowing. This adaptive behaviour is commonly referred to as *dynamic borrowing* and it is reflected from a mathematical standpoint in the posterior distribution for θ which is again a mixture of normal distributions with *updated* weights, which depend on the degree of similarity the informative component of the RMP and and the concurrent data.

This approach will be applied in Chapters 3 and 4 of the present dissertation, while some novel theoretical insights on this methodology will be explored in Chapter 5.

1.5 Dissertation outline

In Chapter 2, we investigate a two-arm group sequential design in which a primary efficacy endpoint is jointly monitored with a surrogate endpoint. An early interim analysis for futility is incorporated, guided by the trial's predictive probability of success (PPoS). The objective of this work is to extend the *surrogate prior* historical borrowing approach introduced by Saint-Hilary *et al.* within the context of group-sequential designs, thereby enabling its use in the computation of PPoS and improving interim decision-making.

In Chapter 3, we consider a Bayesian group-sequential design in which a primary efficacy endpoint is jointly monitored alongside a surrogate endpoint. An interim analysis is incorporated with the objective of supporting an application for *accelerated approval*. We propose a dual-criterion approach for early accelerated approval, integrating evidence of efficacy on the surrogate endpoint with a robust predictive probability of success (PPoS) for the trial. Furthermore, we investigate how the integration of diverse types of external data can enhance the operating characteristics of the trial.

In Chapter 4, we consider a two-stage multi-arm multi-stage (MAMS) design in which an interim analysis is planned with the objective of selecting a single treatment to advance in the trial. The aim of this work is to develop a statistical framework integrating a multi-criteria decision analysis (MCDA) approach at the treatment selection stage, thereby incorporating safety considerations and reducing the risk of advancing toxic treatment arms.

In Chapter 5, we investigate the interplay between the hyper-parameters within the Robust Mixture Prior (nRMP) Bayesian dynamic borrowing framework. We theoretically prove and analytically demonstrate that the operating characteristics of a trial employing nRMPs are determined by the joint selection of the prior distribution's hyper-parameters, and particularly we find that mamy different combination of the latter may produce very similar posterior inference. Building on this result, we propose a novel, more appropriate interpretation of these hyper-parameters and introduce a practical elicitation routine to guide their specification.

In Chapter 6, we conclude the dissertation with a summary discussion of the findings and outline potential directions for future research.

Chapter 2

Futility interim analysis based on probability of success using surrogate endpoint

2.1 Introduction

Drug development is a long, costly and high-risk process that often requires innovative strategies to optimize the design of clinical trials. The urgency to deliver promptly new medicines to patients pushes sponsors and regulators to constantly rethink drug development, especially in areas of serious and life-threatening diseases with unmet medical needs. Hence, there is a high need for statistical methods making use of all available information in order to improve and accelerate the decision-making process. Over the past decades, interim analyses have become a classical way to speed-up drug development and a large literature is now accessible. In parallel, leveraging prior knowledge from former studies and literature is increasingly considered, especially through Bayesian statistics.

In 1986, Spiegelhalter et al. [7] introduced the concept of predictive power, also known as predictive probability of success (PPoS), in the context of an interim analysis. This was further elaborated by O'Hagan et al. [9], who refers to it with the term "assurance" and by Gasparini et al.[8] and Rufibach et al. [18] described a general framework to update sequentially the PPoS for time-to-event trials taking advantage of both external information coming from parallel trials and internal information arising from interim analysis.

However, when dealing with group sequential trials with long-term clinical endpoint as primary endpoint, reaching the information fraction for the desired interim analysis can take a long time. Therefore, using the long-term clinical endpoint of interest sometimes leads to one of the two following situations: either the interim analysis has poor operating characteristics and is not informative enough because it occurs too early in the study course to make an informative

decision, or the interim analysis has limited value because it occurs too late, when for instance all patients have already been enrolled. One way to address this issue is to use a surrogate endpoint Surrogate endpoint can be "a laboratory measurement or a physical sign used as a substitute for a clinically meaningful outcome that measures directly how a patient feels, functions or survives" [19]. In oncology, Progression-Free Survival (PFS) has been demonstrated to be a valid surrogate for Overall Survival (OS) in many cancer settings and, in particular, in metastatic colorectal cancer (mCRC), surrogacy between PFS and OS has been proven using different methodologies, for example, using Bayesian meta-analytic regression [20] or via esti- mation of the correlation parameters [21].

In 2019, Saint-Hilary et al. [22] proposed an approach to predict the probability of success of a future clinical trial based on the observed results of past trials and using a surrogate endpoint to predict the success based on the long-term clinical endpoint of interest. Quan et al. [23] extended the methodology of Saint-Hilary et al. by using a bivariate normal prior accounting for the within trial correlation between surrogate and long-term endpoints, and Zhang et al. [24] further extended the reasoning to trials with multiple co-primary endpoints. However, while all the above-mentioned approaches focus on the prediction of the PPoS for a future trial using information coming from past trials on the same treatment, none of them extended the methods to the case of group sequential designs.

In this publication, extending the method from Saint-Hilary et al. [22], we propose to use the PPoS of a trial as a criterion for a futility interim analysis, based on available evidence from both the surrogate endpoint and the long-term clinical endpoint of interest at the time of the interim analysis.

Let us assume that the primary endpoint (also called clinical endpoint) is a long-term endpoint and suppose that a relationship between a surrogate endpoint and the final endpoint has been established from previous clinical studies. As only little information on the final endpoint will be observed at the time of an early interim analysis, the surrogate endpoint and its documented relationship with the final endpoint can be used in order to enrich the information about the final endpoint. In a Bayesian framework [25], it is possible to derive an informative prior for the final endpoint from historical data on the final and surrogate endpoints and interim results on the surrogate endpoint. This informative prior is called the *surrogate prior* [22]. The available evidence on the final endpoint at the interim analysis is finally integrated to calculate the PPoS of the trial, which is used as a stop criterion for futility. This approach is summarized in Figure 2.1.

In some cases, though, leveraging historical knowledge may introduce a risk of bias due to the discordance between past and current information. Indeed, if the relationship between the surrogate and final endpoints established from historical data does not hold in the trial of interest (i.e., data on the surrogate and primary endpoint are conflicting), using a surrogate informative prior may be misleading, impacting the PPoS accuracy and accordingly the operating characteristics. Several methods intended to address prior data conflict exist in the literature [17, 16, 15]. For this publication we use a robust mixture prior approach [16] which combines

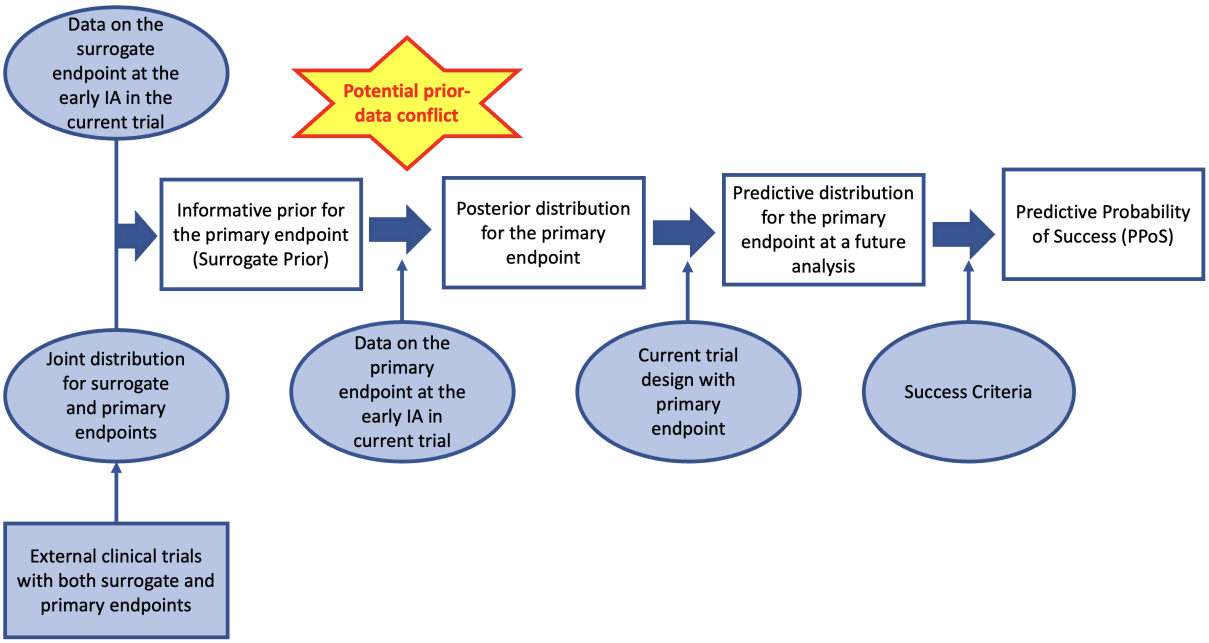


Fig. 2.1 Proposed approach to compute the predictive probability of success (PPoS) of the trial based on the primary endpoint, using data on the surrogate endpoint at an interim analysis.

the surrogate prior with a vague component. The intent is to borrow the most information from historical data when the observed data on the final endpoint at the interim analysis is consistent with the 'surrogate prior', and to borrow less information otherwise.

The main objectives of this paper are to assess the benefit of using the surrogate prior to inform decision making at an early interim futility analysis and to investigate how a mixture prior approach can be used to deal with different levels of prior-data conflict.

After presenting the methodology in Section 2.2, reminding the work from Saint-Hilary et al and specifying how it is extended to the interim futility analysis context, we present in Section 2.3 a real case study for a phase III trial in metastatic colorectal cancer, which design was approved by regulatory authorities. Finally, we discuss the interest and potential further development of this approach in Section 2.4.

2.2 Methodology

The methodology described in this Section focuses on group sequential designs with one futility interim analysis, for the sake of simplicity. The case-study, on the other hand, corresponds to a group sequential design with three interim analyses, which introduces additional complexities beyond the scope of this proposed methodology. Therefore, the detailed information on the case-study methodology can be found in the supplementary material.

2.2.1 PPoS based on the primary endpoint only (reminders and notations)

Consider a clinical trial designed to compare two treatments with the conventional frequentist approach. Let θ denote the corresponding true treatment difference on the primary endpoint of interest. The primary objective is then defined as the rejection of the null hypothesis $H^0 : \theta \leq 0$ against the alternative, $H^1 : \theta > 0$

In most cases, the estimate (indexed by F for final) of the primary endpoint at the final analysis can be asymptotically approximated by a normal distribution with:

$$\hat{\theta}_F \sim \mathcal{N}\left(\theta, \sigma_F^2\right) \tag{2.1}$$

where its sampling variance is, supposed to be known. For example, this could be the difference between means, the log odds ratio, the log hazard ratio, or the log rate ratio if the final endpoint is a continuous variable, a binary or ordinal variable, a time-to-event variable, or a count variable, respectively.

Suppose now that the trial design includes an interim analysis for futility based on the primary endpoint, and that the treatment effect at interim analysis is summarized with

$$\hat{\theta}_{IA} \sim \mathcal{N}\left(\theta, \sigma_{IA}^2\right) \tag{2.2}$$

where the subscript IA stands for interim analysis, $\hat{\theta}_{IA}$ is the point estimate of the treatment effect on the primary endpoint and σ_{IA}^2 its variance, supposed to be known. In a Bayesian framework, using a normal conjugate prior approach with a vague prior $\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$, the posterior distribution of θ after the interim analysis is:

$$\theta \sim \mathcal{N}(\theta_p, \sigma_p^2) \tag{2.3}$$

where $\theta_p = \frac{\hat{\theta}_{IA}\sigma_0^2 + \theta_0\sigma_{IA}^2}{\sigma_0^2 + \sigma_{IA}^2}$, and $\sigma_p^2 = \frac{\sigma_0^2\sigma_{IA}^2}{\sigma_0^2 + \sigma_{IA}^2}$(indexed by p for posterior). Notice that θ_p can equivalently be seen as the weighted average of $\xi\hat{\theta}_{IA} + (1-\xi)\theta_0$, with weight $\xi = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_{IA}^2}$.

In Saint-Hilary et al. [22], the objective was to predict the results of a future trial from the results of past trials. The context of an interim analysis is however different, since it has to account for the within trial correlation that exists between interim analysis and final analysis.

Under the assumption of asymptotically independent increment in the test statistics of the group sequential design [2, 26], the estimate of the treatment effect at the end of the trial can be written as follows:

$$\hat{\theta}_F = t_{IA}\hat{\theta}_{IA} + t_{F-IA}\hat{\theta}_{F-IA} \tag{2.4}$$

where $\hat{\theta}_{F-IA}|\theta \sim \mathcal{N}(\theta, \sigma^2_{F-IA})$, is the incremental estimate of θ using the information observed between interim and final analyses, t_{IA} and t_{F-IA} are the information fractions at the interim analysis and after the interim respectively, and $\hat{\theta}_{F-IA}$ and $\hat{\theta}_{IA}$ are supposed independent. From the posterior distribution of θ, we obtain the predictive distribution of $\hat{\theta}_{F-IA}$:

$$\hat{\theta}_{F-IA} \sim \mathcal{N}\left(\theta_p, \sigma^2_p + \sigma^2_{F-IA}\right) \tag{2.5}$$

After the interim analysis, the predictive probability of rejecting H^0 at the final analysis is:

$$\begin{aligned} \text{PPoS} &= P\left(\hat{\theta}_F > s \middle| \hat{\theta}_{IA}\right) = P\left(t_{IA}\hat{\theta}_{IA} + t_{F-IA}\hat{\theta}_{F-IA} > s \middle| \hat{\theta}_{IA}\right) = \\ &= P\left(\hat{\theta}_{F-IA} > \frac{s - t_{IA}\hat{\theta}_{IA}}{t_{F-IA}} \middle| \hat{\theta}_{IA}\right) = 1 - \Phi\left(\frac{\frac{s-t_{IA}\hat{\theta}_{IA}}{t_{F-IA}} - \theta_p}{\sqrt{\sigma^2_p + \sigma^2_{F-IA}}}\right) \end{aligned} \tag{2.6}$$

where s is the critical value to reach statistical significance at the end of the trial.

2.2.2 PPoS based on the surrogate endpoint only

Within the same clinical trial, let γ denote the true treatment difference on the surrogate endpoint of interest (short term observed endpoint). After observing the interim estimate $\hat{\gamma}_{IA}$ with its sampling variance δ^2_{IA}, using a Bayesian approach with a (vague) normal prior $\gamma \sim \mathcal{N}(\gamma_0, \delta^2_0)$, and the sampling distribution $\hat{\gamma}_{IA} \sim \mathcal{N}(\gamma, \delta^2_{IA})$, we obtain the posterior distribution:

$$\gamma \sim \mathcal{N}(\gamma_p, \delta^2_p) \tag{2.7}$$

where $\gamma_p = \frac{\hat{\gamma}_{IA}\sigma^2_0 + \gamma_0\delta^2_{IA}}{\delta^2_0 + \delta^2_{IA}}$, and $\delta^2_p = \frac{\delta^2_0\delta^2_{IA}}{\delta^2_0 + \delta^2_{IA}}$(indexed by p for posterior).
We wish to use this information to calculate the PPoS of the final analysis based on the primary endpoint. Suppose now that historical data can be leveraged to use information collected in the past on both surrogate and final endpoints. As in Saint-Hilary et al. [22], a Bayesian meta-analytic approach for the evaluation of surrogate endpoints developed by Daniels and Hughes [27], which uses summary data from multiple clinical trials, allows to establish a linear relationship between the final and the surrogate endpoints.

Consider N external randomized clinical trials (indexed by subscript k) where the surrogate and the final endpoints were evaluated. For each trial, θ_k and γ_k denote the true treatment effect on the final endpoint and the surrogate endpoint respectively, with $\hat{\theta}_k$ and $\hat{\gamma}_k$ their estimates; δ^2_k and σ^2_k indicate their sampling variances (on the surrogate and primary endpoint respectively) and ρ the between endpoints correlation (assumed to be known, see discussion in Saint-Hilary et al. [22]).

Assuming the following linear relationship between θ_k and γ_k across all N trials:

$$\theta_k \big| \gamma_k, a, b, \tau \sim \mathcal{N}\left(a + b\gamma_k, \tau^2\right) \tag{2.8}$$

where a, b and τ^2 are the intercept, the slope and the variance estimated by the meta-analytic regression, the model can be generalized as follows:

$$\begin{bmatrix} \hat{\theta}_k \\ \hat{\gamma}_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} a + b\gamma_k \\ \gamma_k \end{bmatrix}, \begin{bmatrix} \sigma_k^2 + \tau^2 & \rho\sigma_k\delta_k \\ \rho\sigma_k\delta_k & \delta_k^2 \end{bmatrix} \right) \tag{2.9}$$

Considering Formula 2.8 for $k = 1, \ldots, N$, and $\gamma = \mathcal{N}(\gamma_p, \delta_p^2)$, the distribution of the treatment effect on the final endpoint conditional on the regression parameters is obtained as:

$$\theta \big| a, b, \tau \sim \mathcal{N}\left(a + b\gamma_p, \tau^2 + b^2\delta_p^2\right) \tag{2.10}$$

The hierarchical model employed assumes a common slope parameter (b) across trials, while for the intercept holds $a_k \sim N(a, \sigma^2)$, i.e. it is treated as a random effect to account for between-trial heterogeneity. This pooling of the slope is a modeling choice aimed at maximizing the information borrowed from historical data, which substantially contributes to the observed gains in precision for the relationship between surrogate and primary endpoint. By allowing the intercept to be effectively down-weighted through the hierarchical variance component (τ^2), the model maintains flexibility in capturing baseline differences in risk across different trial populations.

Then, integrating over the joint distribution of the regression parameters $f_{a,b,\tau}(x, y, z)$ (with x, y, z representing the integration variables corresponding to a, b and τ respectively) with domain $\Omega = [-\infty, +\infty] \times [-\infty, +\infty] \times [0, +\infty]$, we obtain an informative prior for θ, called the surrogate prior

$$\pi_\theta^S(\cdot) = \int_\Omega f_{\theta|a,b,\tau}(\cdot) f_{a,b,\tau}(x, y, z) d(x, y, z) \tag{2.11}$$

which, since we assume no data regarding the final endpoint are available yet, coincides with the posterior distribution. The predictive distribution of the parameter for the remainder of the trial, conditional on interim results, is therefore

$$h_{\hat{\theta}_{F-IA}}^S(\cdot) = \int_\Omega f_{\hat{\theta}_{F-IA}|a,b,\tau}(\cdot) f_{a,b,\tau}(x, y, z) d(x, y, z) \tag{2.12}$$

where $f_{\hat{\theta}_{F-IA}|a,b,\tau}$ is the density of $\hat{\theta}_{F-IA}$ given the regression parameters and is normally distributed with mean $a + b\gamma_p$ and variance $\tau^2 + b^2\delta^2 + \sigma_{F-IA}^2$.

The Predictive Probability of Success (PPoS) at the interim analysis can then be written as follows:

$$\text{PPoS}^{IA} = P\left(\hat{\theta}_{F-IA} > \eta \middle| \hat{\theta}_{IA}\right) = \int h^S_{\hat{\theta}_{F-IA}}(t)dt \tag{2.13}$$

where $\eta = \frac{s - t_{IA}\hat{\theta}_{IA}}{t_{F-IA}}$.

2.2.3 PPoS based on the surrogate and final endpoints

As explained in Saint-Hilary et al. [22], the information coming from the surrogate endpoint can be combined with the data collected on the final endpoint, to obtain the following posterior distribution noted

$$g^S_\theta(\cdot) = \frac{f_{\hat{\theta}|\theta}(d)\pi^S_\theta(\cdot)}{\int f_{\hat{\theta}|\theta=t}(d)\pi^S_\theta(t)dt} \tag{2.14}$$

where d is the observed outcome of $\hat{\theta}$. From this posterior distribution and the distribution of $\hat{\theta}_{F-IA}|\theta$, the predictive distribution of $\hat{\theta}_{F-IA}$ is obtained below:

$$h^S_{\hat{\theta}_{F-IA}}(\cdot) = \int f_{\hat{\theta}_{F-IA}|\theta=t}(\cdot)g^S_\theta(t)dt \tag{2.15}$$

and the predictive probability of success of the trial after the interim analyses is obtained as in the previous subsection.

2.2.4 Prior-data conflict

As suggested in Saint-Hilary et al. [22], we use a mixture prior [16] to down weight the information coming from the surrogate endpoint in case of prior-data conflict. We consider a weighted sum of a vague normal prior π^V_θ and the surrogate prior π^S_θ. The mixture prior distribution π^M_θ (indexed by M for mixture) is defined as:

$$\pi^M_\theta(\cdot) = (1-w)\pi^V_\theta(\cdot) + w\pi^S_\theta(\cdot) \tag{2.16}$$

where w is the prior probability that the assumption of surrogacy holds in our clinical trial. The predictive probability of success is then calculated as:

$$\text{PPoS}^M = (1-\tilde{w})\text{PPoS}^V + \tilde{w}\text{PPoS}^S \tag{2.17}$$

where $\tilde{w}$ is the posterior probability calculated as:

$$\tilde{w} = \frac{w h^S_{\hat{\theta}}(d)}{(1-w) h^V_{\hat{\theta}}(d) + w h^S_{\hat{\theta}}(d)} \tag{2.18}$$

with $h^V_{\hat{\theta}}(\cdot)$ being the density of the prior-predictive distribution of $\hat{\theta}$ based on the vague prior, $\mathcal{N}(\theta_0, \sigma_0^2 + \sigma^2)$, and $h^S_{\hat{\theta}}(\cdot)$ the density of the prior-predictive distribution based on the surrogate prior.

2.2.5 Scenario Plausibility Metric

The performances of the proposed methodology will be assessed in a simulation study, using different scenarios for the treatment effect on the surrogate endpoint and the final endpoint (one scenario corresponds to a pair of assumed true values denoted by $\gamma^\#$ and $\theta^\#$ for the surrogate and the final endpoint respectively). To assess the impact of a prior-data conflict, these scenarios will be chosen so that they are not equally plausible given the relationship between endpoints established from historical data.

For a better interpretation of the simulation results, we propose a "scenario plausibility metric" (SPM) to quantify the extent of the prior-data conflict. Given the conditional distribution in Equation 2.10 and assuming the treatment effect on the surrogate endpoint is equal to $\gamma^\#$, the SPM is defined as the probability of obtaining a value at least as extreme as $\theta^\#$ for the scenario to be evaluated. This metric can be expressed as follows:

$$\begin{aligned} \mathrm{SPM}(\gamma^\#, \theta^\#) = 2 \min \Bigg[& \int_{-\infty}^{\theta^\#} \int_\Omega f_{\theta|\gamma^\#,a,b,\tau}(t) f_{a,b,\tau}(x,y,z) d(x,y,z) dt \,, \\ & \int_{\theta^\#}^{+\infty} \int_\Omega f_{\theta|\gamma^\#,a,b,\tau}(t) f_{a,b,\tau}(x,y,z) d(x,y,z) dt \Bigg] \end{aligned} \tag{2.19}$$

If the considered scenario is plausible according to the linear relationship obtained from historical data, then $\theta^\#$ will be close to the expected value of $\theta|\gamma^\#, a, b, \tau$ and $\mathrm{SPM}(\gamma^\#, \theta^\#)$ will be close to 1. If the considered scenario is not plausible, i.e. the assumed treatment effect on the final endpoint is far from what would be expected given the assumed treatment effect on the surrogate endpoint, then the plausibility metric will be low.

In the following sections, we will use the terminology "minor", "moderate" and "large" prior-data conflict referring to scenarios which SPM is respectively greater than 0.8, between 0.2 and 0.8 and lower than 0.2

It is worth highlighting at this stage that the linearity assumption between endpoints is not strictly necessary to define SPM. The latter, instead, is well defined as long as the distribution of the

primary endpoint conditional on the surrogate endpoint can be found, thus not depending on which specific methodology is used to derive it.

2.3 Case Study In Oncology

The approach described in Section 2.2 was recently implemented for a phase III trial of Futuximab/Modotuximab in combination with Trifluridine/Tipiracil versus Trifluridine/Tipiracil single agent in participants with previously treated metastatic colorectal cancer (COLSTAR [28]). The primary endpoint of the study is Overall Survival (OS), and Progression-Free Survival (PFS) is a secondary endpoint. OS is defined as the time elapsed from the first experimental drug intake to death and PFS is defined as the time elapsed from the date of randomization to disease progression or death, whichever occurs first.

PFS has been demonstrated to be a good candidate for a surrogate endpoint in advanced colorectal cancer [21, 22]. For this case-study, OS will be called the final endpoint and PFS the surrogate endpoint.

An initial study design was based on a classical group sequential design with two interim analyses (IA1, with a futility analysis, and IA2, with both futility and efficacy analyses) using O'Brien Fleming alpha-spending function (the Lan-DeMets method) for efficacy boundaries determination and Hwang-Shih-Decani (non-binding) beta-spending function with γ parameter equal to -1.5 for futility boundary determination. A total of 500 patients needs to be randomized in a 1:1 ratio and followed-up until at least 383 OS events are observed to ensure 90% power to detect a true hazard ratio (HR) ≤ 0.71 (median OS: 8.5 vs 12 months) using a log-rank test with an overall 1-sided significance level of 0.025. Assuming enrollment over 21 months with a ramp-up to a maximum of 30 patients per month and a lost-to-follow-up rate of 7% across both treatment arms, the final analysis expected to occur 37.5 months after first patient randomization.

As described in Table 2.1, the IA1 occurs 17.5 months after first patient in, and by that time 382 patients should already be included. It was then decided to modify the design (still before the start of the study) and to use the methodology presented in this paper in order to add an early futility interim analysis (IA0) when 83 PFS events are observed. This will occur approximately 10 months after the first patient in, and at that time, approximately 34 OS events should be observed, and 168 patients should be included. IA0, based on both PFS and OS, allows to stop the trial for futility approximately 7 months earlier than IA1 under the null hypothesis and avoids the inclusion of more than two hundred patients if the drug is ineffective.

The first step was to leverage historical data in order to establish a linear relationship between PFS and OS, with a Bayesian meta-analytic approach. For this purpose, we used the systematic literature review (SLR) published in 2018 by Arnold et al. [29]. This SLR focuses on clinical trials beyond the second line in patients with mCRC. In addition to the relevant trials reported in the SLR (randomized phase II or phase III studies, [30–39]), 3 additional randomized trials in

Table 2.1 Group sequential design of the case study.

Analysis	IF	N° events (OS/PFS)	Time since FPI (months)	Boundaries		N° patients enrolled
				Efficacy boundary	Futility boundary	
Interim 0 (IA0)	0.089	34/83	10	NA	PPoS < X%	168
Interim 0 (IA1)	0.334	128/-	17.5	NA	$HR(OS) = 1.024$	382
Interim 0 (IA2)	0.666	255/-	24	$HR(OS) = 0.73$	$HR(OS) = 0.883$	500
Final (FA)	1	383/-	37.5	$HR(OS) = 0.816$	$HR(OS) = 0.816$	500

advanced colorectal cancer were included (PFEIFFER [40], FRESCO [41] and IMBLAZE370 [42]). In total, fifteen randomized clinical trials were selected as they evaluate both PFS and OS in advanced metastatic colorectal cancer. The main results and characteristics of these trials are summarized in Table A1.1 of the Appendix.

Under the proportional hazard assumptions, we assume that the log Hazard Ratio (log(HR)) asymptotically follows a normal distribution; moreover we assume that there exists a linear relationship between the log(HR) for OS and for PFS as described in Equation 2.10, with $\theta = \log(\text{HR(OS)})$ and $\gamma = \log(\text{HR(PFS)})$.

For the proposed Bayesian meta-analytic regression, improper uniform priors have been used for the regression parameters a, b, and an approximation of the parameters' posterior distribu- tion has been made simulating 150,000 Markov Chain Monte Carlo (MCMC) iterations including a 50,000 iteration burn-in. The analyses were carried out using R version 3.6.1 [43] and Rstan package version 2.18.1 [44].

Figure 2.2 (A) presents the regression line with its 95% credibility interval – obtained from fitting the meta-analytic model and assuming for the correlation between surrogate and long-term endpoint – and the 15 historical studies, sized proportionally to their precision. Summary statistics on the regression parameters are provided in the inset table. Notice that according to the methodology in the study by Daniels and Hughes [17], a perfect surrogacy should imply $a = 0, b \neq 0, \tau = 0$. Although this is hardly achievable in practice, in our context, it can be seen that PFS is estimated to be a good surrogate of OS, since the credibility interval for the slope parameter b does not include the value 0 and the credibility interval for the intercept parameter a includes the value 0.

2.3.1 Investigated designs

Once the linear relationship between log(HR) on OS and PFS is established, the surrogate prior and the observed values of log(HR) on OS and PFS are used to calculate the PPoS of the trial, as described in Section 2.2 and in supplementary material.

The futility criterion at IA0 is based on the PPoS, i.e. the predictive probability for the primary analysis to be statistically significant at IA2 or at the final analysis, conditional on not stopping at IA1 for futility.

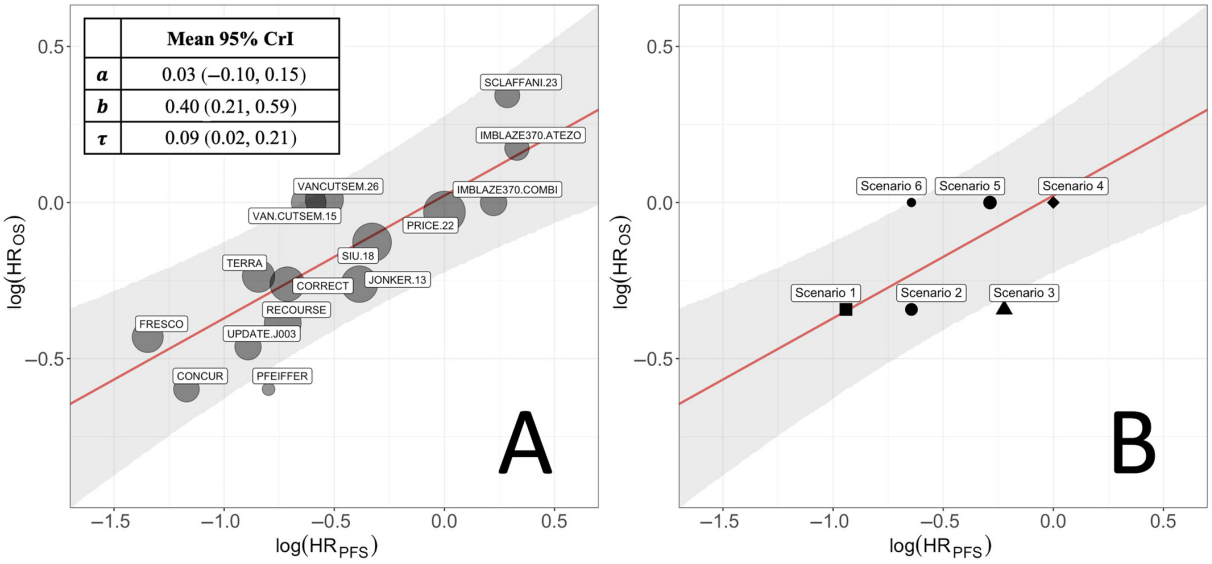


Fig. 2.2 (A) Meta-analytic model fitted to the 15 historical studies, on PFS and OS endpoints. (B) Scenarios for the simulation study.

Several (non-binding) futility boundaries ranging from $X = 0\%$ to $X = 50\%$ are explored for the two designs as follows:

- if PPoS at IA0 <= X% it is recommended to stop the trial for futility.
- if PPoS at IA0 > X% it is recommended that the trial continues

The heatmap displayed in Figure 2.3 presents the PPoS for many possible observed values of HRs on OS and PFS at IA0, and illustrates the decision rule and the futility zone at IA0 for a futility boundary based on a PPoS<10%. It allows to summarize the combinations of HR values on OS and PFS that will lead to an early termination at IA0, which is convenient for the protocol preparation as well as for the communication with the clinical team.

In order to evaluate the added-value of the early futility analysis using PFS as a surrogate endpoint, we compare the operating characteristics of a "vague prior design" (VAGUE), where a vague prior distribution is used for the treatment effect parameter, to a "surrogate prior design" (SURR), where the surrogate prior is used for the treatment effect parameter.

As a second step, in order to address a potential prior-data conflict, we also explore the operating characteristics of a third design, named "robust surrogate prior design" (R-SURR), where a mixture of a vague prior and the surrogate prior (with specific weight w) is used, as detailed in Section 2.2.4.

We remark at this stage that OS data only are used in the "vague prior design", whereas both PFS and OS data are used in the "surrogate prior design" as well as in the "robust surrogate prior design."

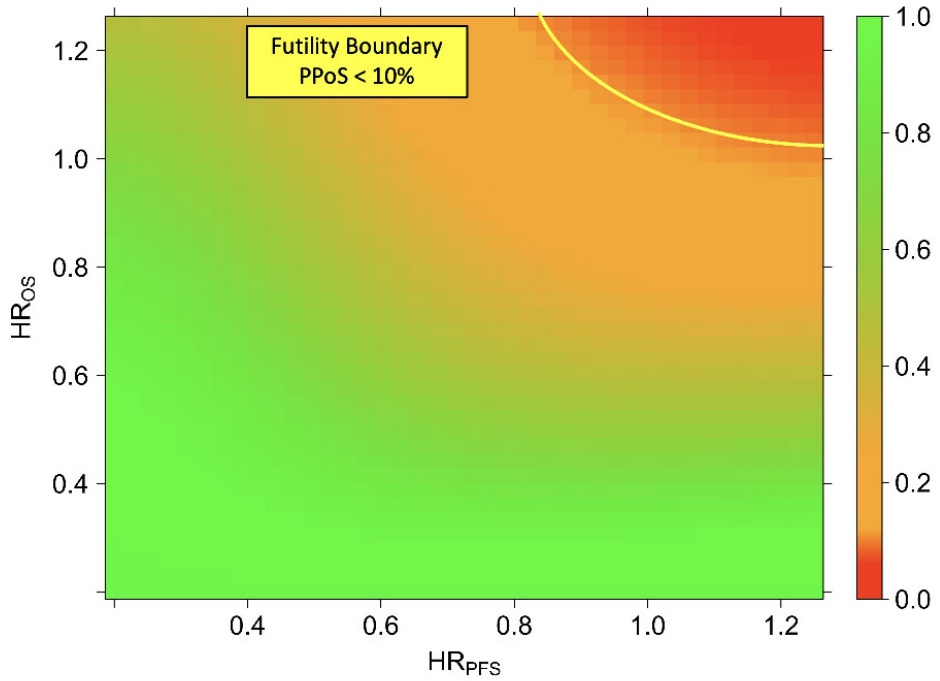


Fig. 2.3 PPoS and decision rule at interim IA0: if at the time of the interim analysis PPoS < X%, then it is recommended that the trial stops for futility. The yellow line indicates a possible choice for the decision boundary with X=10%.

The impact of the timing of the early interim analyses IA0 on the 3 approaches is discussed in the supplementary material, where a further design is investigated with the information fraction of IA0 equal to 0.2.

2.3.2 Simulation plan

Data were simulated in 6 scenarios presented in Table 2.2 and Figure 2.2 (B).

Table 2.2 Simulation scenarios, with scenario plausibility metrics (SPM, defined in Section 2.2.5) assessing the extent of prior-data conflict.

True HR(OS) = 0.71						**True HR(OS) = 1**					
Sc. 1 Minor conflict (SPM=0.97)		**Sc. 2 Moderate conflict (SPM=0.25)**		**Sc. 3 Large conflict (SPM=0.03)**		**Sc. 4 Minor conflict (SPM=0.84)**		**Sc. 5 Moderate conflict (SPM=0.33)**		**Sc. 6 Large conflict (SPM=0.05)**	
θ	γ	θ	γ	θ	γ	θ	γ	θ	γ	θ	γ
0.71	0.39	0.71	0.525	0.71	0.80	1	1	1	0.75	1	0.525

In the first set of scenarios (scenarios 1, 2 and 3), the correct decision is to continue the trial at IA0, since the simulated true hazard ratio for OS is assumed to be 0.71 (alternative hypothesis H^1). In the second set of scenarios (sc.4, sc.5 and sc.6), the correct decision is to stop the trial at IA0, since the simulated true hazard ratio for OS is equal to 1 (null hypothesis H^0). Each set includes a scenario with minor prior-data conflict, a scenario with a moderate prior-data conflict

and a scenario with a large prior-data conflict, and the conflict extent is assessed in terms of SPM as described in Section 2.2.5.

For each scenario, 100,000 clinical trials are simulated by generating observed values of HRs for OS and PFS $(\hat{\gamma}, \hat{\theta})$ at IA0, IA1, IA2 and FA. The sampling variances of the log(HRs) were approximated by the delta-method formula $\sigma^2 = 4/n_{PFS}$ and $\delta^2 = 4/n_{OS}$, with n_{OS} and n_{PFS} being the number of events for OS and PFS, respectively. Notice that no individual patient data are generated in this simulation scheme; instead the treatment effect estimates for the surrogate and primary endpoints are directly generated from their asymptotic normal distribution.

The operating characteristics of a design are assessed using the following metrics:

- the probability to continue the trial after the futility analysis IA0 (referred to as *GO Probability*)
- the probability to reach statistical significance under H^1 (referred to as *power*)
- the probability to reach statistical significance under H^0 (referred to as *type I error*)

2.3.3 Results

Figures 2.4 and 2.5 show the probability to continue the study after IA0, respectively, for scenarios 1, 2, 3 (effective treatment) and 4, 5, 6 (ineffective treatment). Figure 2.6 shows the study power for scenarios 1, 2, and 3 (effective treatment), and Figure 2.7 shows the type I error for scenarios 4, 5, and 6 (ineffective treatment). For all the plots, the futility boundaries range from 0%to 50% and the performances of the three designs described in Section 2.3.3 are assessed, assuming values of $w = 0.3, 0.5, 0.7$ for R-SURR design.

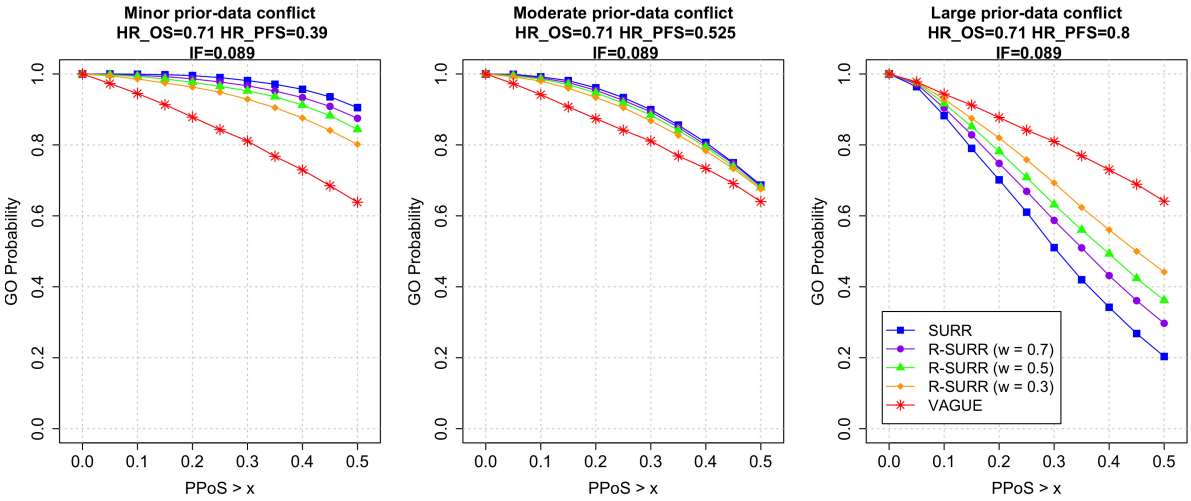


Fig. 2.4 Probability to continue at IA0 depending on futility boundaries for effective treatments (scenarios 1, 2, and 3).

We first compare the "vague prior design" (VAGUE) and "surrogate prior design" (SURR).

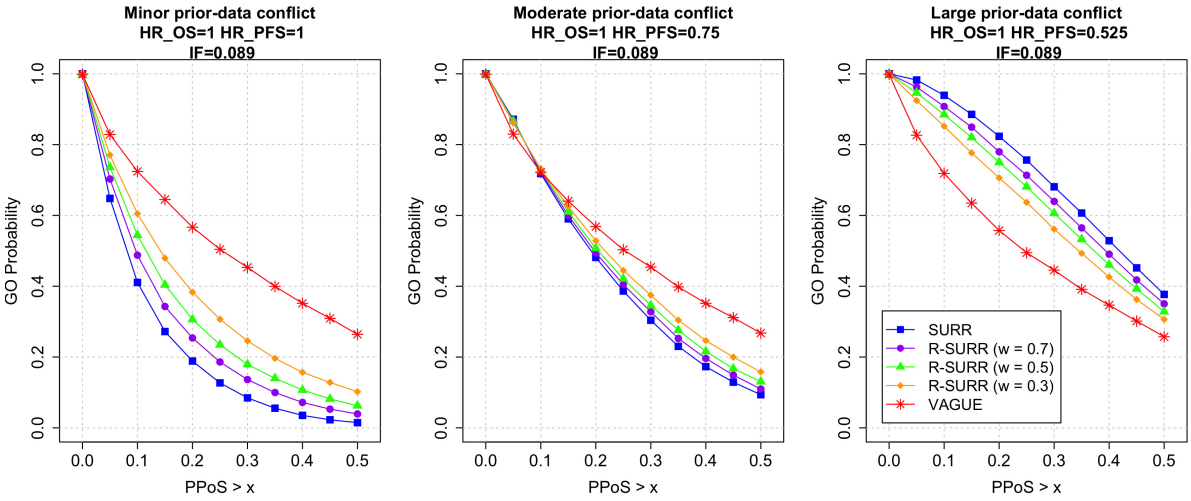


Fig. 2.5 Probability to continue at IA0 depending on futility boundaries for ineffective treatments (scenarios 4, 5, and 6).

Under scenarios 1, 2, 4, and 5 (i.e., in case of minor or moderate prior-data conflict), the probability to make the appropriate decision at IA0 is maximized when using the SURR design compared with the VAGUE design.

Indeed, when HR(OS)=0.71, compared with VAGUE design, the probability to continue the study with SURR design is increased by up to 27% in case of no prior data conflict (scenario 1), and by up to 10% in case of moderate conflict (scenario 2).

When HR(OS)=1, the risk to continue the study with SURR design is decreased by up to 37% in case of no prior data conflict (scenario 4) and up to 18% in case of moderate conflict (scenario 5), depending on the futility boundaries. Also, SURR design offers better performances in terms of type I error with respect to VAGUE design, decreasing the error rates up to 1.4% in case of minor prior-data conflict and up to 0.8% in case of moderate prior-data conflict.

On the other hand, in case of large prior data conflict (scenarios 3 and 6), since the log-linear relationship does not hold for the current trial, using the surrogate prior (SURR design) to predict data on the clinical endpoint may be misleading, increasing the risk of making an incorrect decision at IA0 by up to 40% in scenario 3, and by up to 10% in scenario 6.

Table A.2 in the Supplementary Material illustrates the gains in posterior effective sample size (ESS) at early interim analyses offered by the SURR and R-SURR designs. Across all evaluated scenarios, the expected posterior ESS at the interim stage increases with the weight assigned to the informative component of the mixture prior. However, some variations across scenarios persist, driven by two primary mechanisms: first, the surrogate prior exhibits sensitivity to the number of progression-free survival (PFS) events observed at the interim, which fluctuates according to the true surrogate treatment effect; and second, the posterior distribution under a robust mixture prior is influenced by the degree of observed prior-data conflict between concurrent and historical data. Notably, these factors may influence the posterior ESS in opposite directions, often resulting in a compensatory effect. In effective scenarios (Sc. 1–3), for instance,

increased drift heightens prior-data conflict—which tends to diminish the ESS—while the larger volume of PFS data resulting from a higher HR(PFS) acts to increase it. Conversely, in null scenarios (Sc. 4–6), the increase in drift is accompanied by lower surrogate treatment effects, which simultaneously reduces ESS through conflict and increases it through the reduction in HR(PFS).

In summary, the surrogate prior outperforms the noninformative prior except in case of major data conflict. Hence, if at the design stage, it is decided to use the surrogate prior (SURR design) but the risk of prior-data conflict in the trial cannot be reasonably excluded, it may be desirable to downweight the impact of the informative prior on the study results, and using a robustified surrogate prior (R-SURR design) would be preferred.

As expected, the probability to continue after IA0 with R-SURR design lies between those of VAGUE design and SURR design. In particular, when w is large, the probabilities to continue after IA0, the power and the type I error for R-SURR design get closer to those of SURR design. On the other hand, when w tends toward 0, the results for R-SURR design get closer to those for VAGUE design.

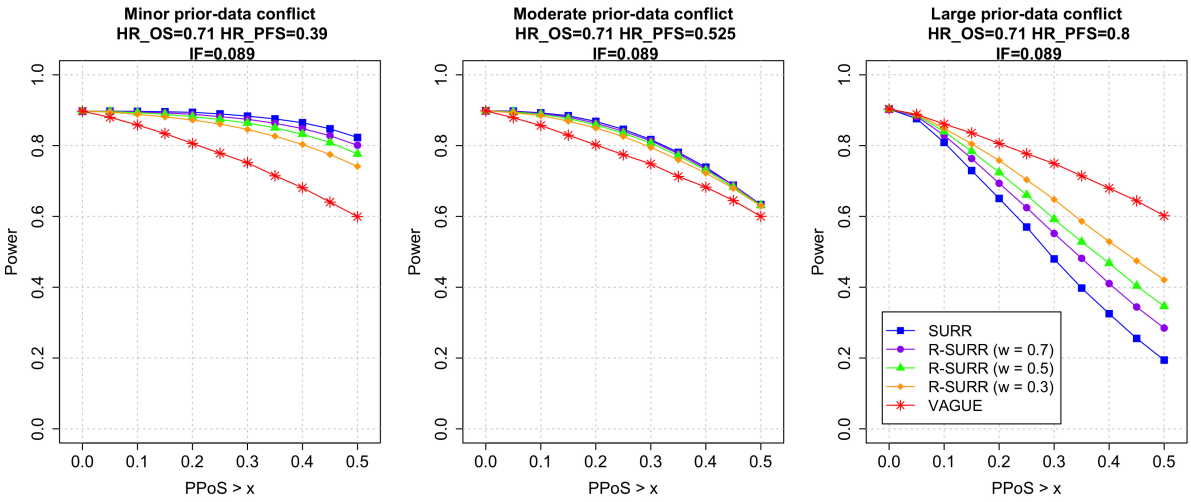


Fig. 2.6 Power depending on futility boundaries (scenarios 1, 2, and 3).

When HR(OS)=0.71 and minor or moderate prior-data conflict is simulated (scenarios 1 and 2), the benefits to include prior information are fairly maintained when using the mixture prior. On the contrary, with large prior data conflict (scenario 3), the probability to continue at IA0 is substantially increased, resulting in a higher power.

Accordingly, when HR(OS)=1 and minor or moderate prior-data conflict is simulated (scenarios 4 and 5), the risk to continue the trial after IA0 with R-SURR design (as compared with SURR design) is only slightly increased, leading to a minor impact on the type 1 error, while it is maintained in case of large prior-data conflict (scenario 6).

In summary, we can conclude that SURR design improves the operating characteristics as compared with VAGUE design in case of minor and moderate prior data conflict, but performs

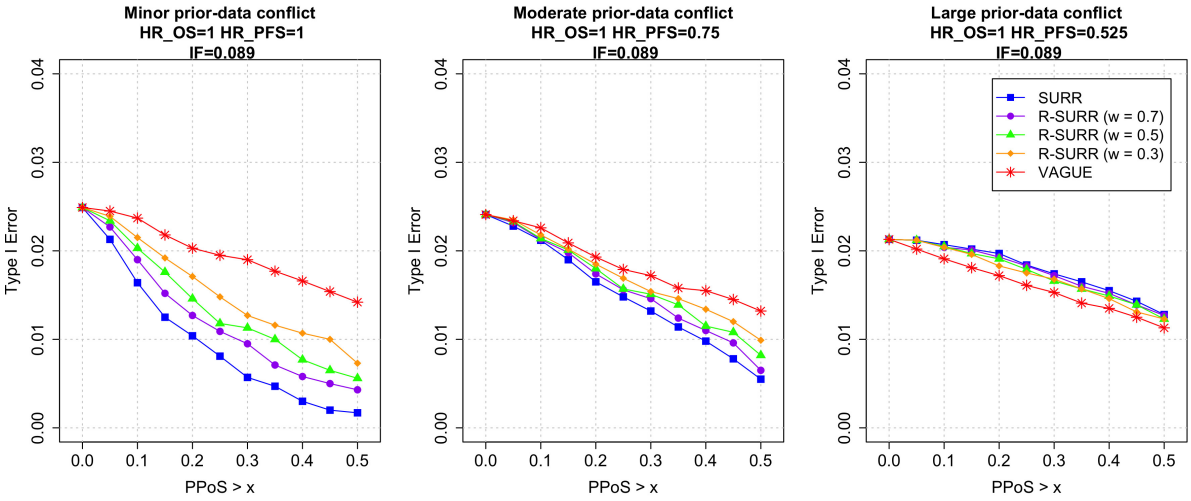


Fig. 2.7 Type I error depending on futility boundaries (scenarios 4, 5, and 6).

poorly in case of large inconsistency between prior and current data. R-SURR design, based on a mixture prior of vague and surrogate prior, fairly combines the improvement in the operating characteristics in case of minor and moderate prior-data conflict with a limited loss of performance in case of large prior-data conflict.

It should be noted that the scenarios are not equally likely, given the available evidence. Indeed, while large values of SPM in scenarios 1 and 4 (0.97 and 0.84, respectively) indicate plausible situations, low SPM as in scenarios 3 and 6 (0.03 and 0.05, respectively) correspond to very unlikely situations. It can then be concluded that the SURR design and R-SURR design have good performances in the most plausible scenarios.

2.4 Discussion

In the context of group sequential trials with long-term primary endpoints, the utilization of short-term endpoints as surrogate endpoints is increasingly common. These surrogate endpoints are employed to predict the clinical efficacy of the primary endpoint of interest, providing valuable information for decision-making throughout all phases of a clinical trial.

In this article, extending the methodology presented in Saint-Hilary et al. [22], we proposed a comprehensive approach to take advantage of historical data on the relationship between the surrogate and the primary endpoint of interest to build a futility rule at an early interim analysis of a Phase III randomized clinical trial. Once the relationship between endpoints is established, using a meta-analytic approach, an informative prior distribution for the primary endpoint (called surrogate prior) is derived at the early interim analysis using data on the surrogate endpoint. It is updated with current data on the clinical endpoint, if available. The PPoS of the trial is then calculated and used to determine futility. We showed that using the surrogate prior, eventually combined with a robust component, leads to a substantial improvement both in terms of power

and type I error as compared with using a vague prior (or no prior information) in most of the scenarios.

The methodology outlined in the manuscript integrates data from both a surrogate endpoint and the primary endpoint in the early futility stopping rule. However, since only limited data may be accessible on the primary endpoint, it prompts discussion regarding the extent to which utilizing these data are beneficial and whether surrogate data alone suffice for decision-making. The rationale for combining these two sources of information relies on two primary justifications: first, incorporating data on the primary endpoint contributes to reducing the variance of the predictive distribution utilized for the PPoS computation; second, even a minimal amount of primary endpoint data is useful for evaluating the consistency between surrogate and primary endpoints. Although possible [8], sole reliance on the available surrogate endpoint data could result in important biases in the treatment effect estimation if the surrogacy assumption is invalid for the current trial.

Notice that the patient-level correlation is not explicitly taken into account in our model. While in Saint-Hilary et al. [22], this choice was justified by the fact that surrogate prior is computed in a Phase II trial to predict the outcome of a future Phase III with different patients, in our setting a different choice may be considered. In fact, since interim data are available for both the surrogate and the primary endpoints, the patient-level cor- relation between endpoints could be estimated to inform later stages of the trial [45]. We propose to explore this option, that is, by designing a bivariate model to jointly model surrogate and primary endpoints at the individual-patient level, as a further development of this work.

Since the linear relationship between the surrogate and primary endpoints, on which the surrogate prior relies, is derived from a meta-analytic approach, it typically necessitates a large amount of data to ensure sufficient precision, and careful attention must be paid to the selection of studies in the meta-analytic model to avoid bias. To this purpose, when available, the use of SLRs collecting the results of all historical trials in a certain indication is the optimal approach, as presented in the case study. Sensitivity analyses should also be conducted using different selection criteria to ensure the robustness of the surrogacy effect to study variations. From a practical point of view, we acknowledge that performing a meta-regression is not always feasible, due to the lack of reported trials or to the lack of comparability between them. Other options may be considered, such as estimating the relationship between endpoints through an elicitation process if the amount (or the quality) of historical information is not sufficient [46–49]. It should be noted that even though a linear model fits well the data in our case, the methodology described in Section 2.2.2 can be easily adapted to a nonlinear dependency between endpoints.

In addition, we acknowledge that the credibility of the relationship between surrogate and primary endpoint would benefit from a rigorous model checking. Specifically, we recommend the use of historical datasets used for the meta-analysis to perform posterior checks. For instance, this can be done by comparing the observed point estimates of the treatment effects of the primary endpoint in historical trials with the predictive distribution generated by the model; verifying

whether they lay in the credibility range identified by the meta-analytic model (gray band in Figure 2.2. As a second good practice, we advise to perform a leave-one-out analysis, fitting the model multiple times leaving each time one of the N trials out of the analysis: if the estimation of the regression coefficients remains similar across the different analyses, this would strengthen the belief in the surrogacy relationship. For the presented meta-analytic regression, this has been shown in Figure A.5 of the Supplementary Material.

In our work, among various options, the mixture prior approach proposed by Schmidli et al. [16] is used to address a potential discrepancy that may arise between historical and current data. This methodology requires the prior weight of the mixture, as well as the distribution of the vague component, to be chosen at the design stage. While giving a large prior weight to the vague component may be beneficial in mitigating the bias in case of prior-data conflict, it may be suboptimal in case of minor or moderate prior-data conflict. The choice of the mixture prior weight in our study would be made based on the operating characteristics of the design. However, it could also include clinical considerations, for example, via an experts' formal elicitation process [50] or using empirical Bayes-based methods [51, 52]. Also, choosing a too informative robust component in the mixture prior may lead to a limited borrowing of historical information even in the best-case scenario of no prior-data conflict, while selecting a too vague one could lead to an excessive borrowing in case of large prior-data conflict. A unit-information robust mixture component has been employed in our analysis so that it is approximately equivalent to one patient's worth of information.

Alternative methods for leveraging historical data, such as the power prior [15], may also be used to build an informative prior from historical data on a surrogate endpoint, and they could be applicable in our context.

For simplicity, the actual study COLSTAR [28] that motivated this work was designed using SURR design, which incorporates the non-robustified surrogate prior. This decision was motivated by the favorable operating characteristics observed and the perceived minimal likelihood of prior-data conflict. Following the submission and approval of the protocol by regulatory authorities, the study started. It was subsequently halted due to strategic reasons unrelated to the study design.

In the proposed design, an early interim analysis for futility was planned at IF=0.089. This choice has been made so that 83 PFS events are observed at the time of the early interim analysis, which provides a good precision in the estimation of the surrogate prior while conducting the interim analysis early enough to substantially save patients if futility is declared. However, different choices for the timing of the interim analysis are possible, as described in the literature, for example, based on the minimization of the trial's expected (or maximum) sample size under the null hypothesis [53–56] or on the joint optimization of the power loss and the probability of correct futility stop [57, 58]. Minimization of the expected study length and the expected number of observed events has also been explored under the specific assumption of non-proportional hazard [59].

As a general principle, we recommend performing sensitivity analyses, assessing the operating characteristics in various settings, varying the prior weights, the distribution of the robust component, the information fraction of the interim analysis, and any parameter that is expected to have an impact on the design behavior.

The methodology described in the article is applied to a trial in mCRC, where the surrogacy of PFS for OS has been demonstrated in the literature [20, 21]. It may generally be employed in any setting where a surrogate endpoint is acknowledged for the primary endpoint of interest, for example, in lung cancer [60], where both PFS and disease-free survival (DFS) have been demonstrated to be good surrogates for OS, or in prostatic cancer where metastasis-free survival (MFS) has been established to be a good surrogate for OS [61]. Note that the validation of the surrogacy assumption is still a topic of ongoing discussions, since no unique definition of surrogate endpoint exists. A review of the main available methodologies to assess surrogacy in cancer settings is proposed by Buyse et al. [62].

A natural continuation of the proposed work is the extension of the methodology to early efficacy interim analysis. However, although from a technical perspective the methodology would have similarities, the use of surrogate endpoints for early treatment approval poses significant hurdles and is likely to be challenged by regulatory authorities. The main challenge is related to the uncertainty in the predictive value of surrogates, which may potentially lead to an overestimation of the treatment benefit (and consequently the potential commercialization of non-effective treatments). The need for expedited treatment access should be balanced with the requirement of appropriate proof of treatment effectiveness, based on substantial and convincing evidence primarily grounded on primary endpoint data.

Other future extensions of the current work may consist of employing the surrogate prior in other adaptive designs, for example, including sample size reassessment. More than one surrogate endpoint, as in Sant-Hilary et al. [22] and Zhang, Lin, and Liu [24], could also be considered to build the surrogate prior.

In conclusion, this manuscript highlights the advantage of incorporating prior information on short-term endpoints in group sequential trials with long- term clinical endpoints to optimize decision-making and enhance the efficiency of clinical trials.

Chapter 3

Dual-criterion approach incorporating historical information to seek accelerated approval with application in time-to-event group sequential trials

3.1 Introduction

Developing a new drug - from early phases to commercialization - is an extensive journey that requires substantial economical resources and time. This lengthy process is essential for rigorously evaluating many clinical aspects to guarantee the safety and efficacy of experimental treatments. However, accelerating drug developments becomes imperative for drugs filling unmet medical needs, where saving time may seriously impact the population survival.

To address this issue, Group Sequential Designs [63, 2, 64] have been widely used in the last decades, allowing for decision making at pre-specified milestones during the study. There are many benefits of including interim analyses in a clinical trial: they permit the early termination of a trial, potentially reducing the number of patients exposed to an ineffective drug, or shortening the study duration in case of overwhelming efficacy. This provides the opportunity for patients with medical needs to receive an effective treatment earlier.

Interim analyses may have a sensible impact in reducing the total number of randomized patients and the study duration. However, due to the long time duration from randomization to endpoint observation in specific context such as in oncology, the number of events collected on the endpoint of interest at such an early time point is often not sufficiently large to make an informed decision on stopping or continuing the study. As a consequence, there is a growing interest for *surrogate endpoints*, defined as *biomarkers, laboratory measurements, radiographic images, physical signs or other measure allowing to predict clinical benefit* [65, 19]. These surrogate

endpoints are linked to the primary endpoint of interest but can be observed in a shorter time frame.

The latter consideration, along with the urgency of delivering a prompt solution to life threatening diseases, let the US Food and Drug Administration (FDA) to institute the *Accelerated Approval* regulations in 1992 [65], a special program to give an early approval based on a *surrogate endpoint*. Despite the use of surrogate endpoints has many practical advantages, their consistency with the primary endpoint of interest still needs to be supported by actual data; as a result, once the Accelerated Approval is obtained - based on scientific relevance supporting the treatment efficacy on the surrogate endpoint - the company is required to provide valuable evidence of clinical benefit on the primary endpoint (under penalty of withdrawal of the product).

In a recent draft guidance by the FDA titled *Clinical Trial Considerations to Support Accelerated Approval of Oncology Therapeutics - Guidance for Industry* [66] (2023), two ways are detailed to conduct a clinical trial supporting an application for Accelerated Approval: (i) a *two-trial approach* where one trial is conducted using a surrogate endpoint to support Accelerated Approval and a second confirmatory trial is conducted to verify clinical benefit on the long term primary endpoint, and (ii) a *one-trial approach* where a single randomized controlled trial is conducted both to support Accelerated Approval and confirm clinical benefit. In particular, referring to the latter approach, two important points are that " *the protocol should specify a plan to strongly control the overall false positive rate (type-I error) for the endpoint supporting Accelerated Approval and the endpoint supporting verification of clinical benefit"* and *"the trial sample size should be chosen so that it has adequate power to detect a clinically meaningful and statistically significant improvement in both the endpoints for Accelerated Approval and verification of clinical benefit"*.

Using a surrogate endpoint as a key endpoint for Accelerated Approval may seem a natural choice due to its wider and ready availability, however the validity of the surrogacy assumption and its quantification might not be easy to assess, potentially leading to incorrect decision making when surrogacy is incorrectly assumed. A review of the main methods for testing surrogacy is presented in [62]. In oncology, Progression-Free Survival (PFS) has been demonstrated to be a valid surrogate for Overall Survival (OS) in many different cancer settings [67]. In particular, the surrogacy of PFS for OS in Metastatic Colorectal cancer (mCRC) has been demonstrated using different methodologies, e.g. using Bayesian meta-analytic regression [20] or via estimation of the correlation parameters [21].

Even if Accelerated Approval regulation has been vastly used since the program initiation, resulting in the Accelerated Approval of 192 drugs [68] (66 of which ongoing) as of October 2023, still 26 out of the 122 treatments that underwent confirmatory analysis failed to meet post-marketing requirements, leading to a withdrawal of their approval by the FDA [69, 70]. Aside from safety reasons, this is mainly due to the lack of correlation between surrogate and primary endpoints. This shows that improving the current practice for assessing the Accelerated Approval criterion - based solely on testing a surrogate endpoint - is desirable. It may be beneficial not

only from a regulatory and patients' perspective, by preventing non effective treatments from entering the market erroneously, but also from a sponsor perspective, by increasing the chance that an Accelerated Approval is confirmed into a full approval.
In this work, a dual-criterion approach for Accelerated Approval in the one-trial approach is proposed, reinforcing the standard testing on the surrogate endpoint (first criterion) with a predictive criterion on the primary endpoint based on the *Predictive Probability of Success* (PPoS), a.k.a. assurance (second criterion) [71, 8, 9]. In our context, we are interested in predicting the probability of study success given partial data available at the time of the interim analysis. Even though in principle PPoS can be indiscriminately employed both in frequentist and Bayesian designs, the latter is used in the current work, with the main advantage of allowing naturally for the incorporation of historical information through a proper definition of the prior distributions for the parameters of the model.

We will explore how PPoS at the time of the interim analysis can be informed by historical information, namely (i) via historical control borrowing and (ii) making use of a documented relationship between treatment effects on the surrogate and the primary endpoints.

The remainder of this article is organized as follows. Section 3.2 presents a detailed description of the methods. In Section 3.3, a motivating case study is introduced. Section 3.4 reports a simulation study comparing the single-criterion and dual-criterion approaches. In Section 3.5, the incorporation of historical information is discussed, and its added value within the dual-criterion framework is investigated through a simulation study. Section 3.6 provides a more comprehensive evaluation of the Bayesian design using Bayesian metrics. Finally, Section 3.7 offers a discussion and outlines potential extensions of this work.

3.2 Methodology

3.2.1 Single-criterion one-trial approach for Accelerated Approval (SCA)

Consider a randomized clinical trial, in which a new promising treatment is compared with a placebo or a standard of care using time-to-event endpoints. Suppose that a short(er) time endpoint - here, Progression Free Survival (PFS) - and a long(er) time endpoint - here, Overall Survival (OS) - are monitored along the trial, and consider PFS as a surrogate endpoint for OS, which is the primary endpoint of interest. A Group Sequential Design (GSD) [63, 2, 64] is employed, where a certain number of analyses in the set $\mathcal{I} = \{1, \ldots, I+1\}$ (including I interim ones, and final one $I+1$) - are planned at pre-specified information fractions (IF) on the primary endpoint. Moreover, let us suppose that, among the I interim analyses, some of them in the set $\mathcal{I}_{\text{AA}} \subset \mathcal{I}$ can lead to an Accelerated Approval request, and some of them in the set $\mathcal{I}_{\text{FA}} \subseteq \mathcal{I}$ can lead to Full Approval request.
Let $r^k_{i,j}$ and $E^k_{i,j}$ be number of events occurred and the total exposure times at the i-th stage of the trial (i.e. $i = 1$ for the first interim look), for the j-th endpoint ($j \in \{\text{PFS, OS}\}$) in the k-th

arm ($k \in \{\text{C, T}\}$, where C and T stands for *control* and *treatment* respectively); and let us define $\Delta_{i,j} = \left\{ r^k_{i,j}, E^k_{i,j},\ k = C, T \right\}$ as the generic set of data available on the j-th endpoint at the i-th interim analysis.

Note that the total exposure time is defined as the sum of the individual exposure times across all patients, and each individual exposure time corresponds to the duration from randomization to either the occurrence of the endpoint or the end of the study, whichever occurs first.
Suppose the two time-to-event endpoints are exponentially distributed for both the control and treatment arms, and assume that the proportional hazards assumption holds. Let γ and θ be the hazard ratios respectively on the surrogate (PFS) and on the primary (OS) endpoints, and let λ^C_{PFS} and λ^C_{OS} be the control hazards on the surrogate and on the primary endpoints respectively. The number of events on OS conditional on the model parameters has Poisson distribution

$$r^C_{i,\text{OS}} | \lambda^C_{\text{OS}} \sim \text{Poisson}\left(\lambda^C_{\text{OS}} E^C_{i,\text{OS}}\right) \qquad r^T_{i,\text{OS}} | \theta, \lambda^C_{\text{OS}} \sim \text{Poisson}\left(\theta \lambda^C_{\text{OS}} E^T_{i,\text{OS}}\right) \tag{3.1}$$

and similarly the number of PFS events at interim analyses

$$r^C_{i,\text{PFS}} | \lambda^C_{\text{PFS}} \sim \text{Poisson}\left(\lambda^C_{\text{PFS}} E^C_{i,\text{PFS}}\right) \qquad r^T_{i,\text{PFS}} | \gamma, \lambda^C_{\text{PFS}} \sim \text{Poisson}\left(\gamma \lambda^C_{\text{PFS}} E^T_{i,\text{PFS}}\right). \tag{3.2}$$

In a standard one-trial approach, the study could be designed so that Full Approval is requested if efficacy is achieved on the primary endpoint either at any of the interim analysis in the set $\mathcal{I}_{\text{FA}}$ (including the final analysis $I+1$), while Accelerated Approval (AA) is requested if clinical benefit is achieved on the surrogate endpoint at any interim analysis in the set $\mathcal{I}_{\text{AA}}$.
Let us define the "double null hypothesis" as the configuration where there is no treatment effect either on the surrogate nor on the primary endpoint (e.g. $\theta = \gamma = 1$), while we define an "alternative hypothesis" as the configuration where $\theta = \theta^{\#} < 1$ and $\gamma = \gamma^{\#} < 1$ (where $\theta^{\#}$ and $\gamma^{\#}$ represent the target hazard ratios on the primary and surrogate endpoints, respectively). Let us define moreover "partial null scenarios" as the configurations where $\theta = 1$ and $\gamma < 1$, meaning that the treatment is not effective on the primary endpoint but has some effect on the surrogate.
Let $\pi^0_\theta(\cdot)$, $\pi^0_\gamma(\cdot)$, $\pi^0_{\lambda^C_{\text{OS}}}(\cdot)$ and $\pi^0_{\lambda^C_{\text{PFS}}}(\cdot)$ be the prior densities for the model parameters, properly chosen in order to reflect prior available information (including the use of vague priors when no prior information is available).
In a Bayesian framework, the success criterion for requesting a Full Approval at the i-th interim analysis is defined as

$$\mathbb{P}\left(\theta < 1 \,\middle|\, \Delta_{i,\text{OS}} \,;\, \pi^0_{\lambda^C_{\text{OS}}}, \pi^0_\theta\right) > \eta^{\text{OS}}_{i,\text{ eff}} \tag{3.3}$$

and the similarly success criterion for requesting an Accelerated Approval at the i-th analysis is defined as

$$\mathbb{P}\left(\gamma < 1 \,\middle|\, \Delta_{i,\text{PFS}} \,;\, \pi^0_{\lambda^C_{\text{PFS}}}, \pi^0_\gamma\right) > \eta^{\text{PFS}}_{i,\text{ eff}} \tag{3.4}$$

where $\eta^{\text{OS}}_{i,\text{ eff}}$ and $\eta^{\text{PFS}}_{\text{i, eff}}$ are the probability thresholds to claim efficacy respectively on the primary endpoint and on the surrogate endpoint at the i-th stage of the trial. Early stops for futility

at the i-th stage of the trial may be also possible if $\mathbb{P}\left(\theta < 1 \;\middle|\; \Delta_{i,\mathrm{OS}} \,;\, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta\right) < \eta^{\mathrm{OS}}_{i,\,\mathrm{fut}}$. All the probabilities are computed with respect to the posterior distributions for θ and γ, which corresponding posterior densities are denoted by $g_\theta(\cdot|\Delta_{i,\mathrm{OS}}, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta)$, and $g_\gamma(\cdot|\Delta_{i,\mathrm{PFS}}, \pi^0_{\lambda^C_{\mathrm{PFS}}}, \pi^0_\gamma)$. According to the recommendations in [66], the above mentioned probability thresholds should be calibrated in order to control the overall type I error under the double null scenario.

3.2.2 Dual-criterion one-trial approach for Accelerated Approval (DCA)

In the previous section, a single-criterion one-trial approach for AA was based on PFS data collected at the time of the interim analysis. However, a number of events on OS is likely to be available at these times and may be employed in order to generate more convincing evidence that the experimental treatment has a positive benefit-risk (which is not only based on statistical but also clinical aspects).

Let us assume that some evidence on OS is available at the time of the interim analysis i (among the ones targeted for Accelerated Approval request), with its posterior density function $g_\theta(\cdot|\Delta_{i,\mathrm{OS}}, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta)$, but that not enough evidence is available for an early decision regarding a Full Approval request.

The Predictive Probability of Success (PPoS) of the current trial at the i-th interim analysis is defined as the probability that the study demonstrates efficacy on OS at any future analysis (among the ones targeted for full approval), conditional on the partial information collected at the i-th interim analysis, which is

$$\mathrm{PPoS}_i = \sum_{\substack{h \in \mathcal{I}_{\mathrm{FA}} \\ h \geq i+1}} \mathrm{PPoS}_{i,h} \tag{3.5}$$

where $\mathrm{PPoS}_{i,h}$ denotes the predictive probability computed at the i-th interim analysis that the trial is successful at the h-th interim look and not before, which is

$$\begin{aligned} \mathrm{PPoS}_{i,h} = \int_0^{+\infty} \int_{\Omega^m} 1\Bigg\{ & \mathbb{P}\left(\theta < 1 \;\middle|\; \tilde{\Delta}_{h,\mathrm{OS}} \,;\, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta\right) > \eta^{\mathrm{OS}}_{h,\,\mathrm{eff}} \quad \text{AND} \\ & \bigcap_{k \in \mathcal{I}_{\mathrm{FA}}\, i+1 \leq k \leq h-1} \mathbb{P}\left(\theta < 1 \;\middle|\; \tilde{\Delta}_{k,\mathrm{OS}} \,;\, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta\right) < \eta^{\mathrm{OS}}_{k,\,\mathrm{eff}} \Bigg\} \times \\ & f_{\Delta_{ih}}\left(\tilde{\Delta}_{i+1,\mathrm{OS}}\,,\,\ldots\,,\,\tilde{\Delta}_{h,\mathrm{OS}} \;\middle|\; \Delta_{i,\mathrm{OS}}, \theta = t\right) g_\theta\left(t \;\middle|\; \Delta_{i,\mathrm{OS}}, \pi^0_{\lambda^C_{\mathrm{OS}}}, \pi^0_\theta\right) \mathrm{d}\Delta_{ih}\mathrm{d}t \end{aligned} \tag{3.6}$$

The notation $\tilde{\Delta}_{*,\mathrm{OS}} = \left\{\tilde{r}^{\mathrm{C}}_{*,\mathrm{OS}}, \tilde{r}^{\mathrm{T}}_{*,\mathrm{OS}}, \tilde{E}^{\mathrm{C}}_{*,\mathrm{OS}}, \tilde{E}^{\mathrm{T}}_{*,\mathrm{OS}}\right\}$ refers to the predictive data at *-th stage in the domain $\Omega = \mathbb{N}^2 \times \mathbb{R}^2$, and $f_{\Delta_{ih}}$ represents the multivariate predictive data distribution at all the future stages of the study between the i-th (exluded) and the h-th (included) in the domain $\Omega^m = \Omega \times \Omega \times \cdots$ (m times, where m is the number of analyses targeted for Full Approval request between the i-th and the h-th analysis). The probability is computed with respect to the posterior

distribution for the treatment effect parameter on the primary endpoint $g_\theta(\cdot|\Delta_{i,\text{OS}}, \pi^0_{\lambda^C_{\text{OS}}}, \pi^0_\theta)$ at the i-th look.

A modification of the single criterion for Accelerated Approval at the i-th interim look is proposed here by supplementing the condition on the surrogate endpoint in Equation (3.4) with a predictive criterion on the primary endpoint

$$\text{PPoS}_i > \eta_i^{\text{PPoS}} \tag{3.7}$$

where η_i^{PPoS} may be chosen depending on the desired degree of confidence needed in the prediction. In the following sections, the two conditions in Equation (3.4) and Equation (3.7) will be referred as *PFS criterion* and *PPoS criterion* respectively. Note that, unlike the single-criterion approach, where the analysis for Accelerated Approval request was based solely on meeting the PFS criterion specified in Equation 3.4, the proposed dual-criterion approach requires that both criteria in Equation 3.4 and Equation 3.7 are simultaneously satisfied in order for the sponsor to request Accelerated Approval. This requirement is more stringent, but it also provides greater assurance on the efficacy of the experimental drug, hence greater confidence in a full approval.

The schematic of the example design with a single interim analysis for either a Full Approval or a Accelerated Approval is given in Figure 3.1. At the time of the interim analysis, after a pre-specified number of events have occurred on the primary endpoint, data on the primary endpoint are analyzed and a first decision regarding early stop is made based on the posterior distribution for θ at this stage. If no final decision can be made - e.g. insufficient evidence is provided by interim data to either support or reject the null hypothesis - then an Accelerated Approval analysis is performed based on the combined evidence from interim data on both the surrogate and primary endpoints. This analysis utilizes the dual-criterion expressed in Equation (3.4) and (3.7), and only if the two criteria (namely the statistical significance in the surrogate treatment effect and a PPoS above the pre-specified threshold) are jointly satisfied then Accelerated Approval is requested. The study could continue until the final analysis, where a final decision regarding requesting a Full Approval is made based only on the posterior distribution for θ, obtained using updated information on the primary endpoint. At this stage, we note that the use of primary endpoint data during the interim analysis necessarily results in unblinding of the trial. Consequently, an independent third-party statistician is required to preserve the integrity of the study.

3.2.3 Control of error rates

In the design proposed in Section 3.2.2, an approval may be requested for the treatment either at one of the interim analyses times (Full Approval or Accelerated Approval) or at the final analysis time (Full Approval only). Multiple hypothesis testing due to the interim analyses and the two criteria for Accelerated Approval may lead to type I error inflation, hence multiplicity adjustments are needed in order to control the family-wise error rate (FWER).

In this context, it is important to distinguish between two types of error: *(i)* the risk of incorrectly

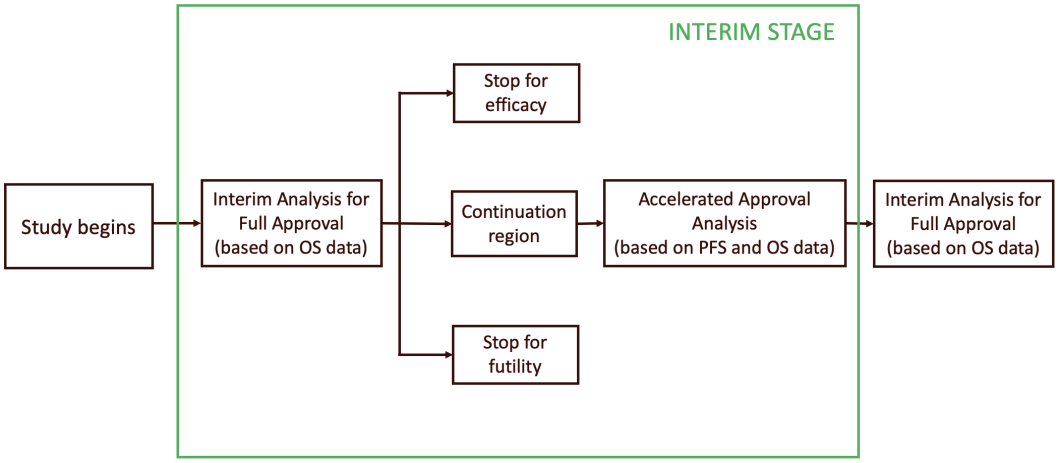


Fig. 3.1 Illustration of the proposed trial pathway. One interim analysis - assessing both efficacy and futility - is presented here for Full Approval based only on evidence on OS; if no decision to stop is made, an Accelerated Approval analysis is performed based on the dual-criterion approach using available evidence on both surrogate and primary endpoints. Only if the dual-criterion is satisfied, the Accelerated Approval is requested. Regardless of the outcome of the Accelerated Approval analysis, if the study is not stopped for efficacy or futility, it continues until the Final Analysis, where a decision is made to request Full Approval based on OS data only.

requesting Accelerated Approval, and *(ii)* the risk of incorrectly requesting Full Approval. Recall that the latter corresponds to a conventional hypothesis test on the primary endpoint. In contrast, the decision to request Accelerated Approval is based on the joint fulfillment of two criteria, one of which involves hypothesis testing on the surrogate endpoint. Therefore, rejecting the null hypothesis on the surrogate endpoint ($\gamma = 1$) when it is in fact true is not, in isolation, to be considered an error, since it does not automatically lead to an Accelerated Approval request unless the second criterion is also satisfied.

We define FWER (also referred as *"Global Type I error"* from now on) in our setting as the probability to be positive in at least one between Full Approval and Accelerated Approval analysis at any stage of the trial (interim or final) when there is no treatment effect on the primary endpoint ($\theta = 1$). This quantity should be interpreted as the overall risk of incorrect decision-making, either through an incorrect application for Accelerated Approval or an incorrect request for Full Approval. However, since a global type I error can arise through multiple pathways, it is useful to distinguish among the specific sources of error that contribute to this overall risk.
Let us define the following quantities:

- **Full Approval rate (denoted as $\alpha_\theta^{\mathbf{FA}}$):** the probability to reject the null hypothesis on the primary endpoint (either at any of the interim or at the final analysis), regardless of whether Accelerated Approval is requested or not at any previous stage. Note that this quantity does not depend on the treatment effect on the surrogate endpoint (see Equation (3.3)). Notice that when there is no effect on the primary endpoint ($\theta = 1$), this probability represents the risk of incorrectly request Full Approval.

- **Accelerated Approval rate (denoted as $\alpha^{\mathbf{AA}}_{\gamma,\theta}$):** the probability to fulfill the two criteria for Full Approval analysis based on the dual-criterion approach (Equations (3.4) and (3.7)); it depends on both the true treatment effect on the surrogate endpoint γ and the true treatment effect on the primary endpoint θ. Notice that when there is no effect on the primary endpoint ($\theta = 1$), this probability represents the risk of incorrectly request an Accelerated Approval at any interim analysis.

- **Confirmed Accelerated Approval rate (denoted as $\alpha^{\mathbf{CAA}}_{\gamma,\theta}$):** the probability that the dual-criterion for Accelerated Approval has been fulfilled (at any interim analysis), and that the criterion for Full Approval (at any subsequent analysis) is also fulfilled; it depends on both the true treatment effect on the surrogate endpoint γ and the true treatment effect on the primary endpoint θ. Notice that when there is no effect on the primary endpoint ($\theta = 1$), this probability may be intended as the risk of incorrectly requesting an Accelerated Approval at interim, and Full Approval at any subsequent analysis.

- **Global Approval rate (denoted as $\alpha^{\mathbf{G}}_{\gamma,\theta}$):** it is the probability to be positive in at least one among Accelerated Approval analysis and Full Approval analysis at any of the interim analysis or at the final analysis; it depends on both the true treatment effect on the surrogate endpoint γ and the true treatment effect on the primary endpoint θ. Notice that when there is no effect on the primary endpoint ($\theta = 1$), this probability may be intended as the risk of incorrectly requesting at least one among Accelerated Approval and Full Approval, which represents the global type I error rate.

Note that both Accelerated Approval and Full Approval analysis contribute to the global Approval rate, then the following decomposition holds:

$$\alpha^{\mathrm{G}}_{\gamma,\theta} = \alpha^{\mathrm{AA}}_{\gamma,\theta} + \alpha^{\mathrm{FA}}_{\theta} - \alpha^{\mathrm{CAA}}_{\gamma,\theta} \tag{3.8}$$

Notice that the minus sign arises from the inclusion-exclusion principle, which states that for any two events A and B, the probability that at least one occurs is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In our context, the *global type I error* is defined as the probability that a false positive conclusion is reached in at least one of the two analyses, Accelerated Approval or Full Approval. As such, the probability of a type I error in either analysis is expressed as the sum of the type I error probabilities for the individual analyses (nemely $\alpha^{\mathrm{AA}}_{\gamma,\theta}$ and $\alpha^{\mathrm{FA}}_{\theta}$), minus the probability that both yield false positive results (namely $\alpha^{\mathrm{CAA}}_{\gamma,\theta}$). The subtraction term ensures that this last term $\alpha^{\mathrm{CAA}}_{\gamma,\theta}$ is not counted twice, as it is already included in both $\alpha^{\mathrm{AA}}_{\gamma,\theta}$ and $\alpha^{\mathrm{FA}}_{\theta}$.
Since $\alpha^{\mathrm{AA}}_{\gamma,\theta}$ depends on the two criteria in Equations (3.4) and (3.7), then we also define:

- **PFS Accelerated Approval rate (denoted as $\alpha^{\mathbf{AA\text{-}PFS}}_{\gamma}$):** the probability to meet PFS criterion for Accelerated Approval, i.e. the probability to claim statistical significance in treatment effect on the surrogate endpoint, when the latter is equal to γ.

- **OS Accelerated Approval rate (denoted as $\alpha_{\gamma,\theta}^{\text{AA-PPoS}}$):** the probability to meet the PPoS criterion for Accelerated Approval, which is the probability to have a high PPoS on the primary endpoint at the interim analysis.

Note that the relationship between $\alpha_{\gamma,\theta}^{\text{AA}}$, $\alpha_{\theta,\gamma}^{\text{AA-PPoS}}$ and $\alpha_{\gamma}^{\text{AA-PFS}}$ depends on the patient level correlation between the surrogate and primary endpoints, in particular the following holds

$$\alpha_{\gamma}^{\text{AA-PFS}}\alpha_{\gamma,\theta}^{\text{AA-PPoS}} \leq \alpha_{\gamma,\theta}^{\text{AA}} \leq \min\left(\alpha_{\gamma}^{\text{AA-PFS}},\, \alpha_{\gamma,\theta}^{\text{AA-PPoS}}\right). \tag{3.9}$$

From Equation (3.8) and (3.9), it follows that

$$\alpha_{\gamma,\theta}^{\text{G}} < \alpha_{\gamma}^{\text{AA-PFS}} + \alpha_{\theta}^{\text{FA}} \tag{3.10}$$

For requesting an Accelerated Approval, a standard requirement imposed by health authorities [66] is that Global type I error is maintained under the *double null scenario* $\theta = \gamma = 1$ [66] under a pre-specified level ω^{G}. This control depends directly on the choice of the probability thresholds $\eta_{i,\,\text{eff}}^{\text{PFS}}$ and $\eta_{i,\,\text{fut}}^{\text{PFS}}$ (which contribute to $\alpha_{\gamma=1}^{\text{AA-PFS}}$), $\eta_{i,\text{eff}}^{\text{OS}}$ and $\eta_{i,\text{fut}}^{\text{OS}}$ (which contribute to $\alpha_{\theta=1}^{\text{FA}}$) and η^{PPoS} (which contributes to $\alpha_{\gamma=1,\theta=1}^{\text{AA-PPoS}}$). Many distinct combinations of the latter may be employed so that $\alpha_{\gamma=1,\theta=1}^{\text{G}} < \omega^{\text{G}}$.
In our context, exploiting the inequality in Equation (3.10) in order to control the global type I error rate under the double null scenario, we propose to split ω^{G} between $\alpha_{\gamma=1}^{\text{AA-PFS}}$ and $\alpha_{\theta=1}^{\text{FA}}$ (so that $\alpha_{\gamma=1}^{\text{AA-PFS}} + \alpha_{\gamma=1,\theta=1}^{\text{FA}} = \omega^{\text{G}}$), choosing accordingly the probability thresholds for PFS testing ($\eta_{i,\,\text{eff}}^{\text{PFS}}$, $\eta_{i,\,\text{fut}}^{\text{PFS}}$) and for OS testing ($\eta_{i,\text{eff}}^{\text{OS}}$ and $\eta_{i,\text{fut}}^{\text{OS}}$) based on any standard GSD rule e.g. alpha-spending functions [5]. Under this splitting strategy, no portion of the nominal level ω^{G} is allocated to the PPoS criterion; therefore, the same allocation can also be applied to the single-criterion approach. It is important to note that the fact that no portion of ω^{G} is allocated to the PPoS criterion does not imply that the criterion defined in Equation 3.7 is without impact. Rather, the PPoS criterion is applied in addition to the hypothesis test on the surrogate endpoint, with the goal of *reinforcing* the evidence in support of an Accelerated Approval request. This strategy also ensures control of the global type I error rate, even in the presence of potential misspecification of the PPoS model.

In principle, a control of the global type I error could be achieved also by splitting ω^{G} exploiting Equation (3.8). However, not including $\alpha_{\gamma=1,\theta=1}^{\text{AA-PPoS}}$ in the split driven by Equation (3.10) has two main advantages: first it leads to type I error rates strictly below the nominal level (since it relies on a strict inequality), second it assures that the Global type I error is maintained not depending on predictions, but rather based solely on concurrent data. We acknowledge that this approach is conservative and does not fully utilize the nominal level ω^{G}. However, the rationale behind this choice is that allocating a portion of the nominal level explicitly to the PPoS criterion may be difficult to justify to regulatory authorities, given the potential bias arising from the limited amount of primary endpoint data typically available at the time of the interim analysis.
As a consequence of this choice the control of the FWER under the double null scenario is

guaranteed for any value of η^{PPoS}, which then remains to be set.
Alternatively, we propose calibrating the η^{PPoS} threshold based on the preservation of the FWER under partial null configurations. We formally define a "safeguard scenario" as a specific partial null setting $[\theta = 1, \gamma = \gamma^*]$ where stringent control of the false positive rate is desirable (to be discussed internally and with health authorities). Within this framework, the PPoS threshold is determined as follows:

$$\eta_*^{\text{PPoS}} = \arg \min_{\eta^{\text{PPoS}} \in [0,1]} \alpha^G_{\gamma=\gamma^*,\theta=1} \quad \text{s.t.} \quad \alpha^G_{\gamma=\gamma^*,\theta=1} \leq \omega^{\text{SG}} \tag{3.11}$$

where ω^{SG} represents a pre-specified target level for the false positive rate under the defined safeguard scenario. Note that the further "*safeguard scenario*" is from the double null scenario (e.g. a safeguard scenario with $\gamma^* << 1$), the higher η^{PPoS} and vice-versa. Additionally, for a given treatment effect on the surrogate endpoint, a lower value of η^{PPoS} leads to a higher rate of Accelerated Approval requests, which is favorable if the treatment is actually effective, but implies an increase in incorrect Accelerated Approval rates if the treatment has no effect on the primary endpoint. On the contrary, high values of η^{PPoS} would decrease the number of incorrect Accelerated Approval requests in case of non effective treatments, but may limit the number of Accelerated Approvals in case of effective ones.

We note that, within the framework considered in this work, an Accelerated Approval (AA) request, even if granted by the regulatory authority, does not guarantee eventual Full Approval (FA). As a result, the parameter η^{PPoS} does not influence the type I error rate associated with Full Approval. Nevertheless, high values of η^{PPoS} are recommended in order to better align AA requests with FA request, thereby increasing the probability that an Accelerated Approval ultimately leads to Full Approval.

3.2.4 Specification of prior distributions

In the single criterion one-trial approach (SCA approach) proposed in Section 3.2.1, we first propose to use weak priors in the Full Approval analysis for the hazard ratio θ and the control hazard $\lambda^{\text{C}}_{\text{OS}}$, as well as in the Accelerated Approval analysis for γ and $\lambda^{\text{C}}_{\text{PFS}}$. This ensures that decision-making is almost entirely driven by concurrent data, while achieving almost equivalence between Bayesian and frequentist analyses (e.g. based on the log-rank test). Consistently with the specification mentioned in [72], the following prior distributions are used:

$$\begin{aligned} \lambda^{\text{C}}_{\text{OS}} &\sim \text{Lognormal}\left(0, 10^2\right) \\ \lambda^{\text{C}}_{\text{PFS}} &\sim \text{Lognormal}\left(0, 10^2\right) \\ \gamma &\sim \text{Lognormal}\left(0, 2^2\right) \\ \theta &\sim \text{Lognormal}\left(0, 2^2\right) \end{aligned} \tag{3.12}$$

The standard deviations in the prior distributions for γ and θ are set such that their *effective sample sizes* are equal to one. Conceptually, this implies that the Fisher information conveyed by each prior distribution is equivalent to that obtained from observing a single patient per group.

In the novel dual-criterion approach (DCA) detailed in Section 3.2.2, the PPoS criterion is introduced to strengthen the Accelerated Approval analysis by incorporating data on the primary endpoint. In this approach, for sake of first evaluation, the same prior distributions employed for the SCA approach in Formulas (3.12) are used for θ, γ, $\lambda_{\text{OS}}^{\text{C}}$ and $\lambda_{\text{PFS}}^{\text{C}}$. This choice may be sensible for example when no historical information is available for any of the model parameters, or if it is believed that the available prior information is significantly different from what is expected to be observed in the current trial. We acknowledge that this assumption is often not reflective of typical confirmatory settings, which are usually conducted when some prior information on the model parameters is already available. Nonetheless, the proposed approach remains a valid option in cases where the sponsor opts (or the regulator asks) not to incorporate such information at the *analysis* stage of the trial. This does not imply that existing knowledge should be disregarded entirely; rather, external data may still play an important role at the *design stage*, for example, as inputs for sample size calculation or other design-related decisions. The use of historical information to inform the prior distribution employed in the computation of the PPoS is examined in Section 3.5.

3.3 Case study

3.3.1 Motivating example

In this section the proposed methodology will be applied in the context of a phase III trial in metastatic colorectal cancer (mCRC). Although all the data used for this example are fictive, the design assumptions made for this case study are inspired by a real study.

The primary endpoint of our case study is Overall Survival (OS), defined as the time from randomization to death, and the secondary (surrogate) endpoint is Progression Free Survival (PFS), defined as the time from randomization to disease progression or death (whichever happens first). The hazard ratio (HR) is used as a measure of the treatment effect for both endpoints.

The trial compares the experimental treatment to a control using a 1:1 randomization. The global type I error $\alpha_{\gamma=1,\theta=1}^{\text{G}}$, i.e. the overall probability to requesting a marketing approval for a non-effective treatment (either via Accelerated or Full Approval) must be controlled at a level $\omega^{\text{G}} = 2.5\%$ one-sided, and an equally weighted Bonferroni split between $\alpha_{\gamma=1}^{\text{AA-PFS}}$ and $\alpha_{\theta=1}^{\text{FA}}$ is chosen according to Equation (3.10). This implies that half the nominal level $\omega^{\text{G}} = 2.5\%$ is assigned to the probability to apply for Accelerated Approval and Full Approval.

Assuming a maximum of 500 patients can be enrolled in the study and an accrual rate of 30

patients per month, supposing a median OS of 8.5 months for the control arm and targeting a 29% reduction in OS on the treatment arm ($\theta = 0.71$, corresponding to 3.5 months increase in median OS from baseline), and a 1.25% Full Approval type I error one-sided, a total of 424 events is required to achieve 90% power. Computation of the sample size has been performed using the R [73] package 'rpact' [74].

One single interim analysis is planned - both to test treatment efficacy on the primary endpoint and to assess the Accelerated Approval criteria - after 84 (20%) OS events are observed. Assuming a median PFS of 2.1 months for the control arm and a 47.5% reduction in PFS ($\gamma = 0.525$, corresponding to 1.9 months increase in median PFS from baseline) , 170 PFS events are expected at the time of the interim analyses with a marginal power (probability to get a statistical significant surrogate treatment effect) of 97.5% .

An O'Brien-Fleming spending function is chosen to set the probability thresholds for efficacy on the primary endpoint at the time of the interim analysis and at the time of the final analysis, which are respectively $\eta^{\text{OS}}_{\text{I, eff}} = 0.9999$ and $\eta^{\text{OS}}_{\text{I, eff}} = 0.9875$. No futility interim analyses are set for the sake of simplicity ($\eta^{\text{OS}}_{\text{I, fut}} = 0$). According to the Bonferroni split of $\alpha^{\text{G}}_{\gamma=1,\theta=1}$ between the two endpoints PFS and OS, a threshold $\eta^{\text{PFS}}_{\text{I, eff}} = 0.9875$ is set to keep $\alpha^{\text{AA-PFS}}_{\gamma=1}$ below its nominal level 1.25% under $\gamma = 1$ (and accordingly the FWER $\alpha^{\text{G}}_{\gamma=1,\theta=1}$ below $\omega^{\text{G}} = 2.5\%$ under the double null scenario of γ=1 and θ=1). A threshold for the PPoS criterion $\eta^{\text{PPoS}}_{*} = 0.91$ is moreover obtained from numerical simulations in order to control the global type I error rate $\alpha^{\text{G}}_{\gamma=\gamma^*,\theta=1}$ at $\omega^{\text{SG}} = 2.5\%$ level under the safeguard scenario of γ^*=0.525 and θ=1 (corresponding to a situation where the treatment is not effective on the primary endpoint but it exhibits the target treatment effect on the surrogate endpoint)). The threshold is computed assuming that the study data for the control arm follow the design assumptions (i.e. median OS of 8.5 months).

3.3.2 Analysis

To illustrate the practical implementation of the proposed approach, we present an analysis based on a fictitious trial, for which the data have been generated numerically. Specifically, in this example, progression-free survival (PFS) and overall survival (OS) data are simulated under the assumption of no treatment effect on the primary endpoint ($\theta = 1$) and a moderate treatment effect on the surrogate endpoint ($\gamma = 0.6$).

Assume that, for an ongoing phase III trial, the following data are available at the time of the interim analysis:

- For the surrogate endpoint: $r^{C}_{1,\text{PFS}} = 104$, $E^{C}_{1,\text{PFS}} = 283$, $r^{T}_{1,\text{PFS}} = 88$, $E^{T}_{1,\text{PFS}} = 356$;
- For the primary endpoint: $r^{C}_{1,\text{OS}} = 48$, $E^{C}_{1,\text{OS}} = 495$, $r^{T}_{1,\text{OS}} = 36$, $E^{T}_{1,\text{OS}} = 560$.

A summary of the results of the analysis is in Table 3.1.

Testing the treatment efficacy on the primary endpoint at the interim analysis, we get $\mathbb{P}\Big(\theta <$

$1 \Big| \Delta_{1,\text{OS}} ; \pi^0_{\lambda^C_{\text{OS}}}, \pi^0_\theta \Big) = 0.967$, which is lower than the pre-specified threshold $\eta^{\text{OS}}_{1,\text{eff}} = 0.9999$. Since not enough evidence is provided to stop early for efficacy, we proceed with the Accelerated Approval analysis. Testing the treatment efficacy on the surrogate endpoint at the interim analysis, we get $\mathbb{P}\Big(\gamma < 1 \Big| \Delta_{1,\text{PFS}} ; \pi^0_{\lambda^C_{\text{PFS}}}, \pi^0_\gamma \Big) = 0.998$, which is greater than the pre-specified threshold $\eta^{\text{PFS}}_{1,\text{eff}} = 0.9875$. As a consequence, the PFS criterion is satisfied (the treatment seems effective in reducing the risk on PFS), and an Accelerated Approval request is recommended using the single-criterion approach (SCA).

Testing the PPoS criterion on the primary endpoint, we obtain a PPoS of 0.793. Although the PFS criterion is satisfied, the PPoS is not greater than the pre-specified threshold of $\eta^{\text{OS}}_{\text{PPoS}} = 0.91$, hence, there is not enough evidence to recommend an Accelerated Approval according to the dual-criterion approaches, and further data are needed to make a decision.

At the end of the trial, when 424 planned events on OS have been observed, let's assume that we observe on the primary endpoint $r^C_{2,\text{OS}} = 215$, $E^C_{2,\text{OS}} = 2600$, $r^T_{2,\text{OS}} = 209$, $r^T_{2,\text{OS}} = 2836$. Testing the treatment efficacy on the primary endpoint at the final analysis, we get $\mathbb{P}\Big(\theta < 1 \Big| \Delta_{2,\text{OS}} ; \pi^0_{\lambda^C_{\text{OS}}}, \pi^0_\theta \Big) = 0.88$, which is lower than the pre-specified threshold $\eta^{\text{OS}}_{2,\text{eff}} = 0.9875$, hence not enough evidence against the null hypothesis is provided and a Full Approval cannot be requested.

This example shows the added value of our methodology: while relying on data on the surrogate endpoint only would have been misleading (bringing us to an incorrect Accelerated Approval request), reinforcing the Accelerated Approval request criteria with the PPoS criterion helped us in avoiding the wrong Accelerated Approval request for an ineffective treatment.

Interim Analysis (IF=0.2)						**Final Analysis**	
$\mathbb{P}(\gamma < 1)$	$\eta^{\text{PFS}}_{\text{I,eff}}$	$\mathbb{P}(\theta < 1)$	$\eta^{\text{OS}}_{\text{I,eff}}$	**PPoS**	η^{PPoS}	$\mathbb{P}(\theta < 1)$	$\eta^{\text{OS}}_{\text{II,eff}}$
0.998	0.9875	0.967	0.9999	0.793	0.9	0.88	0.9875

Table 3.1 Summary of the case study analysis.

3.4 Numerical evaluation

This section presents a simulation study designed to evaluate the approach introduced in Section 3.2 across a range of scenarios. The primary objective is to evaluate and compare the performance of the Dual-Criterion Approach (DCA) relative to the Single-Criterion Approach (SCA) across various parameter settings, with particular emphasis on potential deviations from the design assumptions regarding the control parameter and the surrogate treatment effect parameter.

3.4.1 Setting

The design assumptions, as well as the probability thresholds used for decision making and the available historical information, are the same as in Section 3.3.1.

The performance of the single-criterion approach (SCA) and the dual-criterion approach (DCA), introduced in Section 3.2, is evaluated across 12 scenarios (Table 3.2). The model parameters λ_{OS}^{C}, γ, and θ are systematically varied to represent possible deviations from the design assumptions.

Effective treatments ($\theta = 0.71$) are denoted by the letter "A" (standing for *alternative*), whereas non-effective treatments ($\theta = 1$) are denoted by the letter "N" (standing for *null*). Scenarios labeled with the number "0" represent situations of agreement between the design assumptions and the concurrent data in terms of treatment effects γ and θ, while scenarios labeled with other indices correspond to deviations of the concurrent data from the design assumptions.

For each main scenario, three sub-scenarios are further analyzed, varying the control parameter. In particular, scenarios labeled "LOW" and "HIGH" correspond respectively to inferior (median(OS) = 7 months) and superior (median(OS) = 10 months) concurrent controls with respect to the design assumptions, whereas unlabeled scenarios indicate perfect agreement between the control parameter and the design assumptions. It is worth noting that scenario N1 represents the previously defined *safeguard scenario* used to calibrate the threshold for the PPoS criterion, η^{PPoS}.

For this analysis, data were generated assuming no patient-level correlation between surrogate and primary endpoints; the same analysis made assuming 0.45 is presented in the Supplementary Material.

For each of the 12 scenarios, 1000 trials are simulated and results are obtained making use of an approximation of the posterior distributions for the model parameters obtained via Markov Chain Monte Carlo (MCMC) obtain using the R [73] package RJags [75].

	Scenarios											
	A0 LOW	**A1 LOW**	**N0 LOW**	**N1 LOW**	**A0**	**A1**	**N0**	**N1**	**A0 HIGH**	**A1 HIGH**	**N0 HIGH**	**N1 HIGH**
γ	0.39	0.75	1	0.525	0.39	0.75	1	0.525	0.39	0.75	1	0.525
θ	0.71	0.71	1	1	0.71	0.71	1	1	0.71	0.71	1	1
median(OS)	7	7	7	7	8.5	8.5	8.5	8.5	10	10	10	10

Table 3.2 Considered scenarios: for three different median OS on current control, 5 scenarios - 2 for effective treatment (listed with the letter A) and 3 for non effective treatments (listed with the letter N) are simulated varying γ, θ and λ_{C}^{OS}. Median OS for the control is retrieved by the formula median(OS) = $\log(2)/\lambda_{C}^{OS}$, which is valid for exponential OS.

3.4.2 Evaluation metrics

The two approaches under comparison, namely the single-criterion approach (SCA) and the dual-criterion approach (DCA), are evaluated according to the following performance metrics:

- Accelerated Approval Rate (AAR) which is approximated as the fraction of the total simulated trials which is positive in the Accelerated Approval analysis.

$$\text{AAR} = \frac{\#\text{ Accelerated Approvals}}{\#\text{ Trials Simulated}}$$

- Confirmation Rate (CR) which is approximated as the fraction of the simulated trials passing the Full Approval Analysis at the final analysis in Equation (3.3) among the ones which pass the Accelerated Approval Analysis.

$$\text{CR} = \frac{\#\,(\text{Accelerated Approval} \cap \text{Full Approval})}{\#\text{ Accelerated Approval}}$$

- Full Approval Rate (FAR) which is defined as the fraction of the total simulated trial which is positive in the Full Approval Analysis (either at the interim or at the final stage) in Equation (3.3).

$$\text{FAR} = \frac{\#\text{ Full Approval}}{\#\text{ Trials Simulated}}$$

- Global type I error rate (G-t1E) which is approximated - only for ineffective treatments ($\theta = 1$) - as the fraction of the total simulated trial which passes at least one among Full Approval Analysis in Equation (3.3) or Accelerated Approval Analysis.

$$\text{G-t1E} = \frac{\#\,(\text{Full Approval} \cup \text{Accelerated Approval})}{\#\text{ Trials Simulated}}$$

3.4.3 Results

Table 3.3 provides a comprehensive summary of the results. Each row corresponds to a specific scenario defined in Table 3.2. The first column, labeled *Scenario*, identifies the respective scenario, while the subsequent columns report the four performance metrics considered, namely, the *Accelerated Approval Rate*, *Confirmation Rate*, *Full Approval Rate*, and *Global Type I Error Rate*. For each metric, the results are presented under two methodological frameworks: the *Single-Criterion Approach* (SCA) and the *Dual-Criterion Approach* (DCA).

In terms of AAR, results show that under the SCA approach, the probability of passing the analysis for an Accelerated Approval request is consistently high whenever the surrogate endpoint shows meaningful treatment effects, reaching almost 100% in settings where a strong surrogate treatment effect is shown (A0 and N1 scenarios) and remaining between 39% and 47% under moderate effects (A1 scenarios). In null scenarios (N0), it closely aligns with the nominal level of

Table 3.3 Comparison between single-criterion approach (SCA) and dual-criterion approach (DCA)

Scenario	Accelerated Approval Rate		Confirmation Rate		Full Approval Rate		Global type I Error Rate	
	SCA	DCA (no borrow)	SCA	DCA (no borrow)	SCA	DCA (no borrow)	SCA	DCA (no borrow)
A0 LOW	99.8	40.9	91.3	98.8	91.3	91.3	–	–
A1 LOW	39.5	15.8	91.6	98.1	91.3	91.3	–	–
N0 LOW	1.1	0.0	–	–	1.2	1.2	2.3	1.2
N1 LOW	96.8	1.6	–	–	1.2	1.2	96.9	2.6
A0	100	44.1	91.1	98.2	91.1	91.1	–	–
A1	42.8	19.4	91.8	97.9	91.1	91.1	–	–
N0	1.4	0.0	–	–	1.2	1.2	2.6	1.2
N1	97.6	1.5	–	–	1.1	1.1	97.7	2.3
A0 HIGH	100	45.6	90.8	98.2	90.8	90.8	–	–
A1 HIGH	47.3	21.8	91.1	97.7	90.8	90.8	–	–
N0 HIGH	1.0	0.0	–	–	1.2	1.2	2.2	1.2
N1 HIGH	99.1	1.8	–	–	1.2	1.2	99.1	2.5

$\omega^{G}/2 = 1.25\%$. When moving to the DCA, however, the inclusion of the Probability of Success (PPoS) criterion, based on partially observed primary endpoint data, substantially lowers the probability of meeting the conditions for an Accelerated Approval request in all configurations, often reducing it by half or more compared with the SCA (e.g., 100% vs 45.6% in A0-HIGH, 42.8% vs 19.4% in A1). This reduction is a direct consequence of the additional evidentiary requirement introduced by the PPoS, which demands a consistent signal of efficacy on both the surrogate and the primary endpoint, even when the latter is only partially observed. As a result, the DCA acts as a more conservative filter at the interim stage, considerably limiting the number of trials that would proceed to an Accelerated Approval request.

This more stringent decision rule also affects the consistency between the interim and final analyses, as reflected by the Confirmation Rate (CR). While the SCA yields CR values around 91%, indicating that roughly one in ten trials that requested Accelerated Approval would eventually fail to meet the requirements for Full Approval request, the DCA increases the CR to approximately 98% across all scenarios. This improvement means that when a trial passes the interim analysis under the DCA, it is much more likely to ultimately meet the conditions for Full Approval. In other words, the DCA substantially enhances the reliability of the process, ensuring that interim decisions based on the surrogate and predictive information are more coherent with the final primary endpoint results. As expected, the Full Approval Rate (FAR) itself remains virtually unchanged between the two methods, around the nominal level of 90% in alternative scenarios and around 1.25% under the null, since both rely on the same confirmatory analysis of the primary endpoint. What differentiates the two approaches, therefore, is not the ultimate probability of final approval, but the coherence between the intermediate and final phases: the SCA produces more early requests, with a higher chance of later rejection, whereas the DCA results in fewer, but more reliable, requests.

In terms of G-t1E, under the SCA, the global type I error is around the nominal level of 2.5% under the the double null scenarios (with minor deviations due to simulation error) but becomes severely inflated under the partial null configurations (N1), reaching values above 95%. This inflation arises from the inconsistency between the large treatment effect observed on the

surrogate endpoint and the absence of a corresponding effect on the primary endpoint, which ultimately results in a high probability of erroneously proceeding with an Accelerated Approval request when the treatment provides no true benefit on the primary endpoint. In contrast, the DCA maintains tight control of the FWER across all scenarios, consistently keeping it close to the nominal level of 2.5% under the *safeguard scenarios* (N1-LOW, N1, N1-HIGH) and below the nominal level under the double null scenarios (N0-LOW, N0, N0-HIGH). This demonstrates the robustness of the dual-criterion rule in preventing false requests for early approval and maintaining overall statistical integrity.

Notably, in both approaches, and in the presence of a surrogate treatment effect, an increase in the AAR is observed for larger values of the median control OS (e.g., 40.9% in scenario A0-LOW, 44.1% in scenario A0, and 45.6% in scenario A0-HIGH). This can be attributed to the fact that higher survival in the control arm delays the timing of the interim analysis, which is consequently performed after a greater number of PFS events have accrued, ultimately yielding a more precise estimation of the surrogate treatment effect.

Overall, these results indicate that the DCA framework offers clear advantages in aligning the outcomes of the Accelerated Approval analysis with those of the Full Approval analysis, thereby improving the coherence between Accelerated Approval requests and subsequent Full Approval requests. By integrating partial information from the primary endpoint through the PPoS, the DCA limits the number of incorrect or premature requests while ensuring that those that do proceed are more likely to be confirmed at the end of the study. However, this greater reliability comes at the cost of a marked reduction in the probability of passing the Accelerated Approval analysis. The high uncertainty associated with the PPoS, due to the limited amount of primary endpoint data available at the interim stage, makes the DCA highly conservative in identifying trials suitable for early submission.

In the next section, we explore how incorporating historical data can mitigate this limitation by improving the precision of PPoS estimation, thereby enhancing the efficiency of the DCA while preserving its robustness.

3.5 Augmenting DCA via historical information borrowing

In the new framework detailed in Section 3.2.2, the PPoS criterion is introduced to strengthen the Accelerated Approval analysis by incorporating data on the primary endpoint. However, the number of events on the primary endpoint at the time of the interim analysis is likely to be small, and this may lead to a poor estimation of the treatment effect (due to the high sampling variance), thus limiting the benefit of the PPoS criterion itself. To avoid this risk, different types of informative priors can be considered, aiming to improve the parameter estimation and to enhance the Accelerated Approval request.

In the following paragraphs, building upon existing methodologies described in the literature,

we propose two distinct approaches for incorporating historical information, namely to borrow information on the control parameter λ^{C}_{OS} and on the treatment effect parameter θ.

3.5.1 Borrowing historical control information

Let us assume that data for the control arm are available, e.g. from a literature review, on the primary endpoint in H historical trials. Suppose that we want to leverage information from them to inform the control parameter for the current study in the PPoS estimation. Adopting the notations presented in Section 3.2.1 Equation (3.2), and using the left superscript referring to the h-th historical trial, we have:

$$ {}^{h}r^{C}_{OS} \mid {}^{h}\lambda^{C}_{OS} \sim \text{Poisson}\left({}^{h}\lambda^{C}_{OS}\, {}^{h}E^{C}_{OS}\right) \qquad h = 1, ..., H \tag{3.13} $$

where in this context ${}^{h}r^{C}_{OS}$ and ${}^{h}E^{C}_{OS}$ represent respectively the number of events and the total exposure time at the *final analysis* of the h-th historical trial, not depending on the number of interim looks.
Following the meta-analytic predictive approach (MAP) detailed by Roychoudhury and Neuenschwander in [72], we assume that the historical control hazard and the current control hazard are drawn from the same lognormal distribution:

$$ \lambda^{C}_{OS}, {}^{1}\lambda^{C}_{OS}, \dots, {}^{H}\lambda^{C}_{OS} \mid \mu, s \sim \text{Lognormal}\left(\mu, s^2\right) \tag{3.14} $$

where μ represents the across trial mean parameter and s represents the between trial variability. Assuming no prior information is available for the latter parameters, the following weakly informative priors are used:

$$ \mu \sim \text{Normal}\,(0, 1) \qquad s \sim \text{Half-normal}\,(0, 0.5) \tag{3.15} $$

where the prior distribution for s is chosen accordingly to [72] in order to allow a wide range of heterogeneity scenarios a priori. The choice of the normal prior distribution for μ is again based on [72], and the variance parameter is arbitrarily set to reflect lack of prior information regarding the model parameter.
From the above hierarchical model, a distribution $\pi^{\text{MAP}}_{\lambda^{C}_{OS}}$ for the current control hazard can be obtained and used to inform PPoS in Equation (3.5). Note that only historical data are used for the construction of the MAP distribution.
Although borrowing historical information may be useful for improving the posterior estimation of the model parameters, it may still happen that concurrent data are inconsistent with historical ones. In this situation - known either as *prior-data conflict* or *drift* - historical information should ideally be discounted, and estimation should be only driven by concurrent data. To this purpose, we use a mixture prior approach [16] which consists in combining the informative MAP prior with a vague prior in a mixture distribution. The resulting robust meta analytic prior (rMAP)

takes the following form:

$$\pi^{\text{rMAP}}_{\lambda^C_{\text{OS}}} = w_h \pi^{\text{MAP}}_{\lambda^C_{\text{OS}}} + (1 - w_h)\pi^{v}_{\lambda^C_{\text{OS}}} \tag{3.16}$$

where $\pi^{v}_{\lambda^C_{\text{OS}}}$ is the vague component of the mixture prior and w_h is the prior weight on the informative component, reflecting the prior belief about the exchangeability between the control effect estimated from historical data and the control effect estimated from current data. Note that the term *robust* in the context of bayesian dynamic borrowing context refers to reduced sensitivity to potential inconsistencies between the prior and the observed data, thereby yielding more reliable posterior inference.
For this approach, the weak priors from the SCA approach in Equation (3.12) are used for θ, γ and λ^C_{PFS}.

For data borrowing on the control arm, OS data from 3 historical trials are supposed to be available with a sample size of 270 patients each, in particular the observed number of OS events ${}^1r^C_{\text{F,OS}} = 87$, ${}^2r^C_{\text{F,OS}} = 80$, ${}^3r^C_{\text{F,OS}} = 76$, and the related total exposure times ${}^1E^C_{\text{F,OS}} = 950$, ${}^2E^C_{\text{F,OS}} = 983$, ${}^3E^C_{\text{F,OS}} = 1050$ (defined as the sum of all patients' exposure times), expressed in months. Note that the data associated with the three historical trials used for control borrowing are fictitious and were generated so that the prior distribution for the control parameter $\lambda^{\text{C}}_{\text{OS}}$ match with a median OS of 8.5 months. A prior weight $w_h = 0.9$ is used in this analysis, reflecting high confidence in the relevance of historical control data for our trial and a unit-information prior (UIP) is used as robustification component.

3.5.2 Borrowing historical information from HR(PFS)-HR(OS) relationship

Consider H' historical randomized clinical trials, where $\hat{\gamma}_{h'}$ and $\hat{\theta}_{h'}$ are the estimates of the true parameters $\gamma_{h'}$ and $\theta_{h'}$ for the treatment effects on PFS and OS respectively in each historical trial h' and are available together with their sampling variances $\delta_{h'}$ and correlations $\rho_{h'}$.
Referring to the methodology detailed in [22] - relying on the meta analytic approach proposed in [27] - we assume that a linear relationship holds between $\log(\gamma)$ and $\log(\theta)$ (although the methodology may be adapted for other transformations if needed). We consider the following bi-variate normal model

$$\begin{pmatrix} \log(\hat{\theta}_{h'}) \\ \log(\hat{\gamma}_{h'}) \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} a + b \cdot \log(\gamma_{h'}) \\ \log(\gamma_{h'}) \end{pmatrix}, \begin{pmatrix} \sigma^2_{h'} + \tau^2 & \rho_{h'}\sigma_{h'}\delta_{h'} \\ \rho_{h'}\sigma_{h'}\delta_{h'} & \delta^2_{h'} \end{pmatrix} \right] \tag{3.17}$$

The joint posterior distribution $f_{a,b,\tau}(\cdot)$ of the parameters a, b and τ - representing respectively the intercept, the slope and the between trial variability - can be estimated via meta-analytic regression (see [22] for details). Conditional on the regression parameters, the distribution of the treatment effect on the primary endpoint can then be obtained from the treatment effect on the

surrogate endpoint as

$$\log(\theta) \mid a, b, \tau \sim \text{Normal}\left(a + b \cdot \log(\gamma), \tau^2\right) \tag{3.18}$$

We note $f_{\log(\theta) \mid a,b,\tau}(\cdot)$ its density. At the interim analysis, the posterior distribution of γ is estimated from the data available on the surrogate endpoint at this stage.

An informative prior distribution for the primary endpoint θ, called *surrogate prior*, is obtained by integrating Equation (3.18) over the joint distribution of the regression parameters as follows

$$\pi^{S}_{\log(\theta)}(\cdot) = \int f_{\log(\theta) \mid a,b,\tau}(\cdot)\, f_{a,b,\tau}(x, y, z)\, dxdydz \tag{3.19}$$

A *robustification* of this surrogate prior is used to handle prior-data conflicts by combining the distribution in (3.19) with a vague component in a mixture distribution, and the resulting robust surrogate prior can be written as

$$\pi^{\text{rSURR}}_{\log(\theta)} = w_s \pi^{S}_{\log(\theta)} + (1 - w_s)\pi^{v}_{\log(\theta)} \tag{3.20}$$

where $\pi^{v}_{\log(\theta)}$ is the vague component of the mixture prior and w_s is the informative prior weight, e.g. a probability measure of the prior confidence in the estimated relationship between the surrogate and the primary endpoints.
For this approach, the weak priors from the SCA approach in Equation (3.12) are used for γ, λ^{C}_{OS} and λ^{C}_{PFS}.

Note that although a linear relationship between a transformation of the treatment effects on the surrogate and primary endpoints is assumed in the current formulation, the methodology is flexible and can accommodate deviations from this assumption. Specifically, the approach can be extended to incorporate any functional relationship between the parameters by modifying the mean of the marginal distribution for the treatment effect on the primary endpoint accordingly.

3.5.3 Historical Data

For data borrowing on the control arm, OS data from 3 historical trials are supposed to be available with a sample size of 270 patients each, in particular the observed number of OS events ${}^{1}r^{C}_{\text{F,OS}} = 87$, ${}^{2}r^{C}_{\text{F,OS}} = 80$, ${}^{3}r^{C}_{\text{F,OS}} = 76$, and the related total exposure times ${}^{1}E^{C}_{\text{F,OS}} = 950$, ${}^{2}E^{C}_{\text{F,OS}} = 983$, ${}^{3}E^{C}_{\text{F,OS}} = 1050$ (defined as the sum of all patients' exposure times), expressed in months. Note that the data associated with the three historical trials used for control borrowing are fictitious and were generated so that the prior distribution for the control parameter $\lambda^{\text{C}}_{\text{OS}}$ match with a median OS of 8.5 months. A prior weight $w_h = 0.9$ is used in this analysis, reflecting high confidence in the relevance of historical control data for our trial and a unit-information prior (UIP) is used as robustification component.

In order to borrow information on the relationship between PFS and OS, the same Bayesian meta

analytic approach used in [76] is used, fitting the model in Equation (3.17) with historical data on the log-transformation of the treatment effect parameters on the two endpoints. To this end, a total of 15 randomized trials beyond the second line in mCRC were employed (12 of which taken from a systematic literature review of 2018 by Arnold [29] and 3 additional relevant trials [40–42]), evaluating both PFS and OS in similar populations (even though not testing the same drug).

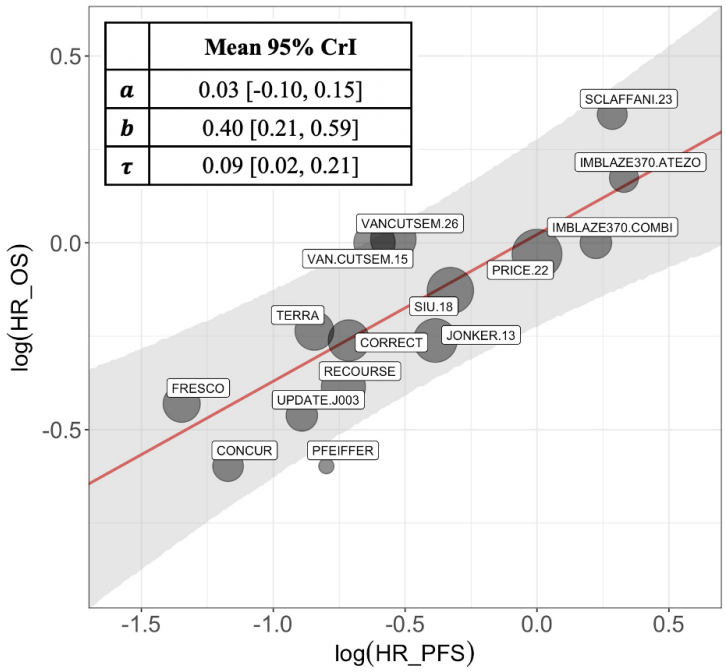


Fig. 3.2 Meta-regression to establish a log-linear relationship between γ and θ in mCRC. In red: the regression line (with its credibility bounds in grey). The sizes of the bubbles are proportional to the inverse of the standard errors of the estimated log hazard ratio on OS.

The estimates of the hazard ratios and their variability on both PFS and OS were used to build the model in Equation (3.17) (Details are provided in Supplementary Material). A prior weight $w_s = 0.9$ is used in this analysis, reflecting a high confidence in the relevance of the estimated relationship estimated from historical information for our trial and a unit-information prior (UIP) is used as robustification component. The Bayesian model was fitted using 5 chains of 100 000 Markov Chain Monte Carlo sampling iterations (preceded by 50.000 warm up iterations) with the R [73] package RStan [44]. We assumed a correlation coefficient $\rho_{h'} = 0.05$ between the treatment effects on the two endpoints for all the studies (see discussion in [77, 22] for more details); and improper vague prior distribution are used for the regression coefficients. The posterior medians of the regression coefficients a, b and τ with their credibility intervals are provided in Figure 3.2, together with a representation of the fitted regression line.

3.5.4 Revised simulation results

In this simulation study, we investigate the effect of incorporating historical information in the computation of the Predictive Probability of Success (PPoS) criterion within the dual-criterion approach (DCA). Two variants of the DCA are considered: the first employs a non-informative prior distribution for the PPoS calculation (hereafter referred to as *DCA (no borrow)*), while the second utilizes an informative prior distribution (hereafter referred to as *DCA (borrow)*). In the latter, historical borrowing is applied both to the control parameters, as described in Section 3.5.1, and to the treatment effect parameters, as described in Section 3.5.2.

The evaluation is conducted under the same scenarios and operating characteristics (OCs) considered in the simulation study presented in Section 3.4. It is important to note that, in the present context, the selected scenarios acquire an interpretation in terms of alignment with *historical information*. Specifically, scenarios labeled with the identifier "0" represent alignment between the concurrent data and the meta-analytical relationship between HR(PFS) and HR(OS), whereas scenarios labeled with the identifier "1" correspond to a situation of *prior-data conflict* (or *drift*) between the concurrent data and the meta-analytical relationship. Similarly, scenarios denoted by the labels "LOW" and "HIGH" indicate a *drift* between concurrent and historical control data, while a basic agreement between concurrent and historical controls is considered in scenarios A0,A1,N0 and N1. A graphical representation of the scenarios considered, with respect to their alignment with historical information, is provided in the Supplementary Material.

Results

Table 3.4 shows the results corresponding to the comparison between the dual-criterion approach without using historical information (*DCA (no borrow)*) and using historical information (*DCA (borrow)*).

Table 3.4 Comparison between the Dual-Criterion Approach without historical borrowing (*no borrow*) and with historical borrowing (*borrow*).

Scenario	Accelerated Approval Rate		Confirmation Rate		Full Approval Rate		Global type I Error Rate	
	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)
A0 LOW	40.9	64.1	98.8	97.7	91.3	91.3	–	–
A1 LOW	15.8	13.2	98.1	97.7	91.3	91.3	–	–
N0 LOW	0.0	0.0	–	–	1.2	1.2	1.2	1.2
N1 LOW	1.6	1.2	–	–	1.2	1.2	2.6	2.2
A0	44.1	73.0	98.2	97.5	91.1	91.1	–	–
A1	19.4	18.8	97.9	98.9	91.1	91.1	–	–
N0	0.0	0.0	–	–	1.2	1.2	1.2	1.2
N1	1.5	1.5	–	–	1.1	1.1	2.3	2.2
A0 HIGH	45.6	79.2	98.2	96.5	90.8	90.8	–	–
A1 HIGH	21.8	22.1	97.7	98.6	90.8	90.8	–	–
N0 HIGH	0.0	0.0	–	–	1.2	1.2	1.2	1.2
N1 HIGH	1.8	2.0	–	–	1.2	1.2	2.5	2.8

In the alternative scenarios, where no drift exists between the historical relationship linking HR(PFS) and HR(OS) and the concurrent data (scenarios A0-LOW, A0, A0-HIGH), the inclusion of historical information in the *borrow* approach proves beneficial. In these situations, historical borrowing leads to a noticeable increase in the Accelerated Approval rate compared with the *no borrow* approach (e.g. 64.1% vs 40.9% in scenario A0-LOW, 73.0% vs 44.1% in scenario A0-HIGH and 79.2% vs 45.6% in scenario A0-LOW).

This improvement arises because the prior informed by the historical relationship reduces uncertainty in the posterior probability of success (PPoS). Consequently, the predictive distribution of the treatment effect on the primary endpoint becomes narrower and more precise, thereby increasing the probability of meeting the PPoS criterion required for Accelerated Approval. However, the magnitude of the increase in AAR depends on the degree of prior-data conflict between concurrent and historical control data. For instance, an approximately 30% increase in AAR is observed when no drift is present, a 35% increase when the current control outperforms the historical control, and a 25% increase when the current control underperforms relative to the historical control. The reason of this is that when the concurrent control data are superior to the historical data (scenario A0-HIGH), the borrowing process tends to *underestimate* the control parameter, leading to an *overestimation* of the treatment effect. Conversely, when the concurrent control data are inferior to the historical data, the prior distribution on the control parameter tends to *overestimate* the control parameter, resulting in an *underestimation* of the treatment effect.

Conversely, in scenarios A1-LOW, A1, and A1-HIGH, where a *negative* drift is observed between the concurrent and historical data (indicating that the current treatment effect on the primary endpoint is *smaller* than that predicted by the meta-analytic relationship for the corresponding surrogate treatment effect) the prior derived from the meta-analytic association between HR(PFS) and HR(OS) becomes biased toward *lower* treatment effects. Consequently, under these scenarios, the advantages of incorporating the *surrogate prior* are lost, and accordingly the AAR obtained with the *borrow* approach is closely aligned with that of the *no-borrow* approach, with only minor differences attributable to the bias introduced through historical control borrowing.

Under the partial null scenarios, and in the absence of drift between the concurrent and historical control data (scenario N1), the *borrow* approach yields the same AAR as the *no-borrow* approach. In this case, although the surrogate prior is biased toward higher treatment effects due to discrepancies between the meta-analytic relationship and the current data, the reduction in predictive variance achieved through borrowing is offset by the influence of the prior bias, resulting in an equivalent AAR. Conversely, when the concurrent controls are inferior to the historical controls (scenario N1-LOW), the bias introduced by the prior on the control parameter leads to a decrease in AAR (1.2% vs 1.6%). In contrast, when the concurrent controls are superior to the historical controls (scenario N1-HIGH), the same mechanism results in an inflation of AAR (2.0% vs 1.8%). Under the double null scenarios (N0-LOW, N0, N0-HIGH), both approaches produce identical results, with an AAR equal to zero.

As expected, both approaches yield identical results in terms of FAR, since this metric depends solely on the analysis of the primary endpoint within the concurrent data, which is identical across methods. Consequently, any differences in G-t1E between the *borrow* and *no-borrow* approaches arise exclusively from differences in their respective Accelerated Approval rates (AAR). Accordingly, in scenarios where a decrease in AAR is observed, a corresponding reduction in G-t1E is also noted (e.g., from 2.6% to 2.2% in scenario N1-LOW). Conversely, in scenarios where an increase in AAR occurs, a corresponding inflation in G-t1E is observed (e.g., from 2.5% to 2.8% in scenario N1-HIGH).

In terms of CR, *borrow* and *no borrow* display comparable performance, with only minor variations observed across scenarios. Both approaches demonstrate a high likelihood that treatments granted Accelerated Approval would ultimately confirm efficacy at the Full Approval stage, indicating overall robustness of the dual-criterion design.

Overall, the comparison between *borrow* and *no borrow* highlights the trade-off between stability and efficiency. The *no borrow* approach, relying exclusively on concurrent data, provides consistent but conservative estimates, while the *borrow* approach, through the use of historical and surrogate information, enhances decision-making power and improves Accelerated Approval rates under concordant conditions. However, the latter's performance depends critically on the compatibility between historical and current evidence, as greater inconsistency can attenuate its benefits or even reduce accuracy in decision-making.

3.6 Sensitivity Analysis

3.6.1 Motivation

When evaluating a given trial design, health authorities typically request an assessment of the *frequentist* operating characteristics (OCs) under pre-specified scenarios, such as the null scenario for type I error control and the alternative scenario for power evaluation. However, particularly in settings where historical borrowing is incorporated into the design, it is a standard regulatory requirement [78, 79] to report frequentist OCs under additional scenarios that may deviate from both the design assumptions and the historical data sources. These evaluations are used by regulators to assess the robustness of the design with respect to potential violations of the assumptions underlying the use of external information.

It is acknowledged, however, that not all scenarios are equally plausible [80]. For instance, a scenario assuming no treatment effect on the primary endpoint but a very large effect on the surrogate endpoint (e.g., $\gamma = 0.2$, $\theta = 1$) is less plausible than the double null scenario ($\gamma = 1$, $\theta = 1$), particularly if the surrogacy assumption between PFS and OS holds. Similarly, observing a median OS of 2 months for the current control arm is less credible than observing a median OS of 8 months, given that a median OS of 8.5 months is assumed by design.

For this reason, Best et al. [80] advocate for the use of appropriate Bayesian metrics when evaluating Bayesian designs. These metrics involve averaging the frequentist OCs, each computed under a specific configuration of true parameters, over a so-called *design prior*, which represents a prior distribution reflecting the relative plausibility of different parameter values. In this framework, each scenario contributes to the overall metric proportionally to its plausibility, as defined by the design prior. As a result, a high type I error rate under an implausible scenario has a limited impact on the overall evaluation metric, whereas the same error rate under the design assumption would contribute more significantly.

Note that, differently from the *analysis prior* - which synthesizes all available information regarding the model parameter and is employed in the actual analysis of the trial - the *design prior* represents an assumption regarding the distribution of the true parameters and is uniquely used for design evaluation (hence may or may not be consistent with the prior knowledge regarding the true parameter).

In this section, an extensive simulation study is conducted to *i)* assess the frequentist operating characteristics (OCs) of the proposed approaches across a *continuous* range of possible scenarios, thereby enabling a *quantitative* evaluation of the impact of prior-data conflict on these OCs, and *ii)* examine several *global* properties of the Bayesian methods, including the Bayesian OCs introduced by Best et al. [80], to gain a deeper understanding of the trade-off between the overall benefits and risks associated with the Bayesian designs under consideration.

3.6.2 Setting

Two analysis are performed: in the first, effective treatments are tested under the alternative hypothesis θ=0.71; in the second, non effective treatments are tested under the null hypothesis θ=1. For both analyses, the operating characteristics are simulated in a 20 by 20 grid of scenarios, which are set by varying $\log(\gamma)$ and median OS for concurrent controls in the set of equispaced values in $\Gamma \times \Lambda$, where $\Gamma = [-2, 0.5]$ and $\Lambda = [3, 16]$ (months). The transformation of Λ on the hazard scale $\Lambda^* = \frac{\log(2)}{\Lambda}$ (following from the assumption of exponentially distributed λ_{OS}^{C}) will be used to derive the prior distribution for the control parameter. The extreme values of the simulation grid are chosen considering that:

- All scenarios where the median OS for concurrent controls is lower then 2.1 months correspond to situations in which average PFS is longer than OS (which is impossible since all OS events (deaths) are also PFS events);
- Values greater than 16 months represents very unlikely scenarios in that disease (we remind that a median OS for concurrent control of 8.5 months was assumed by design);
- For γ, the extreme values correspond to hazard ratios of 0.135 and 1.65, which are considered the thresholds below (respectively, above) which it is implausible to observe

values (we remind that a treatment effect for the surrogate endpoint $\gamma = 0.525$ was assumed by design).

In order to understand the impact of the prior weights in the proposed approaches, the above analyses are performed under different choices of w_h and w_s in the set $\mathcal{W} = (0.1, 0.3, 0.5, 0.7, 0.9)$.

3.6.3 Choice of design priors

In the context of this work, two distinct types of historical borrowing are considered: one on the hazard parameter for the control arm, λ_{OS}^{C}, and the other on the treatment effect on the primary endpoint, θ. Consequently, *design priors* must be specified for both parameters.

For the historical control parameter λ_{OS}^{C}, we use the MAP prior $\pi_{\lambda_{\text{OS}}^{C}}^{\text{MAP}}$ described in Section 3.5.1 as design prior, meaning that the distribution of assumed values for the control hazard employed for design evaluation is consistent to the prior assumption about the control parameter used at the analysis stage. This choice is motivated by the fact that $\pi_{\lambda_{\text{OS}}^{C}}^{\text{MAP}}$ represents the most plausible assumption on the control parameter so far.

On the other hand, historical data introduced in Section 3.5.2 - even though representing treatment effects - only inform the *relationship* between γ and θ, but they provide no direct information on the treatment effect of the concurrent treatment. A modification of the procedure described in Section 3.5 can be used in order to determine a *design prior* for the parameter γ.
Let suppose that a dual representation of the bi-variate model in Equation (3.17) holds:

$$\begin{pmatrix} \log(\hat{\gamma}_{h'}) \\ \log(\hat{\theta}_{h'}) \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \tilde{a} + \tilde{b} \cdot \log(\theta_{h'}) \\ \log(\theta_{h'}) \end{pmatrix}, \begin{pmatrix} \sigma_{h'}^2 + \tilde{\tau}^2 & \rho_{h'} \sigma_{h'} \delta_{h'} \\ \rho_{h'} \sigma_{h'} \delta_{h'} & \delta_{h'}^2 \end{pmatrix} \right] \tag{3.21}$$

where in this case the marginal relative to the treatment effect parameter on the surrogate endpoint is expressed in terms of the treatment effect on the primary endpoint. Once posterior distributions for the regression parameters $\tilde{a}$, $\tilde{b}$ and $\tilde{\tau}$ are obtained, a distribution for $\log(\gamma)$ conditional on the regression coefficients takes the following form

$$\log(\gamma) \mid \tilde{a}, \tilde{b}, \tilde{\tau} \sim \text{Normal}\left(\tilde{a} + \tilde{b} \cdot \log\left(\theta^{\#}\right), \tilde{\tau}^2\right) \tag{3.22}$$

where $\theta^{\#}$ may represent the most likely treatment effect on the primary endpoint (e.g. the alternative hypothesis). The *design prior* for $\log(\gamma)$ is finally obtained by marginalising the conditional distribution in Equation (3.22) over the joint distribution of the regression parameters:

$$p_{\log(\gamma)}(\cdot) = \int f_{\log(\gamma) \mid \tilde{a}, \tilde{b}, \tilde{\tau}}(\cdot)\, f_{\tilde{a}, \tilde{b}, \tilde{\tau}}(x, y, z)\, dxdydz. \tag{3.23}$$

Note that $\theta^{\#}$ is a fixed value in this context, however a *design prior* distribution for $\theta^{\#}$ can be used in principle, even though it would make the derivation of the design prior for $\log(\gamma)$ more complex.

3.6.4 Evaluation metrics

For the current analysis first the G-t1E will be evaluated for each point of the grid in order to analyze the impact of median(OS) for concurrent controls and the treatment effect on the surrogate endpoint $\log(\gamma)$ on the false positive rate. Moreover, three different global metrics are proposed:

- Maximum Global type I Error Rate, defined as the maximum probability to pass either Accelerated Approval Analysis or Full Approval analysis under the null hypothesis of no treatment effect on the primary endpoint, among the scenarios simulated in the grid.

$$\max(\text{G-t1E}) = \max_{\substack{x\in\Gamma \\ y\in\Lambda^*}} \Big[\text{G-t1E}\,(x,y)\Big] \tag{3.24}$$

- Average Global type I error rate, defined as the probability to pass either Accelerated Approval Analysis or Full Approval analysis under the null hypothesis of no treatment effect on the primary endpoint, averaged over the *design priors* defined in Section 3.6.3:

$$\text{avg(G-t1E)} = \int_{\Lambda^*}\int_{\Gamma} \text{G-t1E}\,(x,y)\cdot \pi^{\text{MAP}}_{\lambda^C_{\text{OS}}}(y)\, p_{\log(\gamma)}(x)\,dx\,dy \tag{3.25}$$

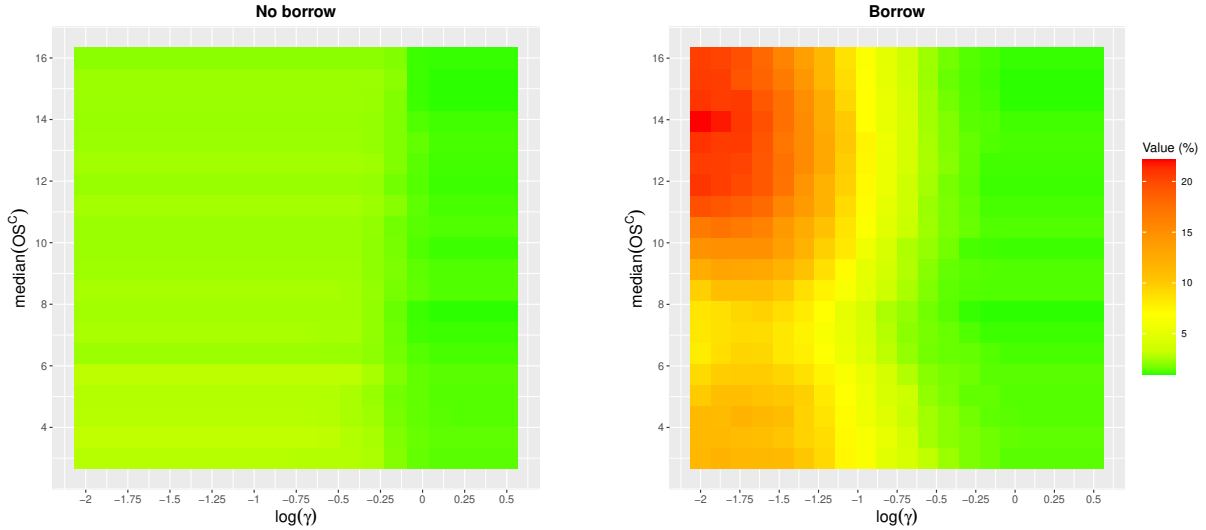
- Average Accelerated Approval Power, defined as the probability to pass Accelerated Approval Analysis under the alternative hypothesis, averaged over the *design priors* defined in Section 3.6.3:

$$\text{avg(AA-Pow)} = \int_{\Lambda^*}\int_{\Gamma} \text{AA-Pow}\,(x,y)\cdot \pi^{\text{MAP}}_{\lambda^C_{\text{OS}}}(y)\, p_{\log(\gamma)}(x)\,dx\,dy \tag{3.26}$$

where $\text{G-t1E}\,(x,y) = \mathbb{P}\left[\text{AA}\cup\text{FA}\;\middle|\;\gamma = x,\ \lambda^C_{\text{OS}} = y,\ \theta = 1\right]$ and
$\text{AA-Pow}\,(x,y) = \mathbb{P}\left[\text{AA}\;\middle|\;\gamma = x,\ \lambda^C_{\text{OS}} = y,\ \theta = 0.71\right]$.
For the computation of the bi-variate integrals in Equations (3.25) and (3.26): a kernel estimation of the density functions corresponding to the design priors is obtained (using the function *density* of R [73]); $\text{G-t1E}\,(x,y)$ and $\text{AA-Pow}\,(x,y)$ are estimated by the proportion of trials meeting the criteria out of 3000 simulated trials in the grid of scenarios; and the trapezoid rule is employed in order approximate the integrals over the grid.

3.6.5 Results

Figure 3.3 illustrates how the global type I error rate varies as a bivariate function of the logarithm of the surrogate treatment effect (x-axis) and the median overall survival (OS) in the control arm (y-axis). Under the assumption of an exponential distribution for OS, the median OS is directly related to the control hazard rate λ_{OS}^{C} through the relationship median(OS) = $\log(2)/\lambda_{\text{OS}}^{C}$. The choice of representing the y-axis in terms of the median OS, rather than the hazard rate, is intended to enhance interpretability for applied readers. The heatmap uses color shading to represent the same quantity, i.e. the simulated global type I error rate at each combination of surrogate effect and median OS. Namely, green regions of the space represent values of G-t1E equal or below the nominal level, while from yellow to red regions represent increasing level of the latter metric.



(a) (b)

Fig. 3.3 Global type I error rate under different pairs [$\log(\gamma)$, median(OS^{C})] in the simulation grid. Prior weights for historical borrowing on the concurrent control parameter λ_{OS}^{C} and the surrogate treatment effect γ are set to $w_h = 0.9$ and $w_s = 0.9$.

Concerning the *no borrow* approach, variations in the median overall survival (OS) of the current control arm exert only a minor influence on the G-t1E, which remains approximately 2.5% across all scenarios considered. A slight decrease in G-t1E is observed when $\gamma \approx 0$ as the median control OS increases. In this setting, higher control-arm survival results in a greater number of progression-free survival (PFS) events being available at the interim analysis, thereby enhancing the estimation of the surrogate treatment effect, γ, and consequently reducing the risk of an incorrect Accelerated Approval. It should be noted that this pattern manifests only when the surrogate treatment effect is small, as a larger number of observed PFS events can meaningfully improve the estimation. For higher surrogate treatment effects, the PFS criterion is consistently satisfied, rendering this effect negligible. Since no historical data are utilized to estimate the relationship between surrogate and primary endpoints, the PPoS criterion remains unaffected by

$\log(\gamma)$. Consequently, variations of G-t1E along the x-axis are confined to values near $\log(\gamma) \approx 0$, with G-t1E approaching zero for $\gamma > 0$.

Regarding the *borrow* approach, distinct patterns emerge in G-t1E, as summarized below:

- Similar to the *no borrow* approach, G-t1E approaches zero when $\gamma > 0$. This behavior arises from the PFS criterion, which is consistently unmet in the absence of a treatment effect on the surrogate endpoint.
- The influence of borrowing information on the control parameter is observed along the Y-axis: for a given value of γ, when the current control is superior (median OS > 8.5 months) relative to historical controls, G-t1E increases, reaching a maximum when the median OS of the concurrent control is approximately 13 months. Conversely, for inferior current controls, G-t1E decreases, attaining a minimum around a median OS of 8 months. Beyond these thresholds, the impact of prior-data conflict is mitigated by the robust component of the mixture prior, resulting in a decrease in G-t1E for median OS exceeding 13 months and an increase in G-t1E for median OS below 8 months.
- The influence of borrowing information on the control parameter is also observed along the X-axis: for a fixed value of median(OS^C), when the drift between concurrent data and the meta-analytic relationship between HR(PFS) and HR(OS) is small ($-0.5 < \log(\gamma) < 0$), G-t1E is reduced due to the increased precision afforded by the surrogate prior. In contrast, as prior-data conflict increases ($\gamma < -0.75$), G-t1E inflates, reaching a maximum around $\log(\gamma) = -1.75$. A subsequent decrease in G-t1E for more extreme values of γ is attributable to the robust component of the mixture prior, which downweights prior information in the presence of substantial drift.

Since the OCs within the *borrow* approach exhibits substantial variations across the parameter space, the assessment of *global* metrics, as proposed in section 3.6.4, is convenient in the assessment of benefit and risks of Bayesian designs, particularly with respect to the choice of the borrowing parameters.

In the Supplementary Material the maximum G-t1E obtained under different choices of the mixture weights w_h and w_s is presented. The results indicate that higher mixture weights are associated with an increase in the maximum G-t1E. Specifically, when both weights are small (e.g., $w_h = 1$, $w_s = 1$), the maximum G-t1E is approximately 4%, whereas for larger weights (e.g., $w_h = 9$, $w_s = 9$), the maximum G-t1E exceeds 20%. Notably, the mixture weight corresponding to the surrogate prior (w_s) exerts a stronger influence on the maximum G-t1E.

Figure 3.4 (panels (a) and (b)) displays the average G-t1E, denoted as avg(G-t1E), for both the *no borrow* and *borrow* approaches. For the former, this metric remains at the nominal level of 2.5% (slight below, due to the residual impact of the unit-information prior centered at the null hypothesis), with no variation across different pairs of (w_h, w_s), as no information borrowing is implemented. For the latter, an increase in avg(G-t1E) is observed with increasing values of

w_s (ranging from 2.5% to approximately 4%), while a slight decrease is noted as w_h increases. Notably, although a mild inflation relative to the nominal level is detected, the avg(G-t1E) remains relatively low. This outcome reflects the fact that although large values of G-t1E are possible, the highest increases in G-t1E occur in regions of the parameter space that have low probability under the *design prior* employed, thus only slightly impact in the averaging process.

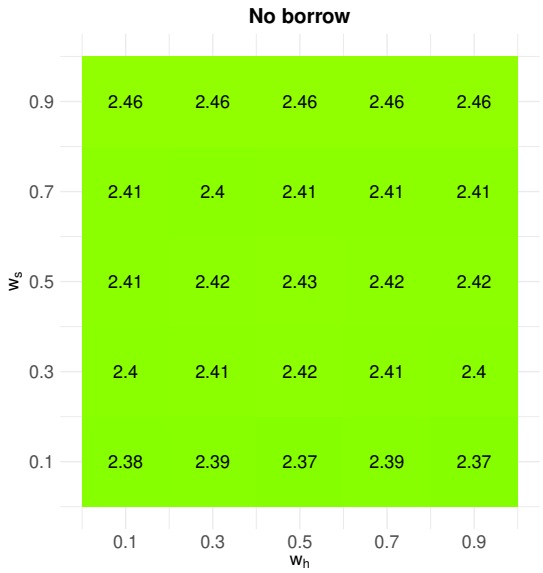
Figure 3.4 (panels (c) and (d)) presents the average Accelerated Approval power, denoted as avg(AA-pow), for both the *no borrow* and *borrow* approaches. In the *no borrow* case, avg(AA-pow) is approximately 43%, with no variation across different combinations of (w_h, w_s), consistent with the absence of information borrowing. Conversely, under the *borrow* approach, avg(AA-pow) increases with larger values of w_h and w_s, reaching up to 65% when strong borrowing is applied ($w_h = w_s = 0.9$), while lower values (around 47%) are observed under minimal borrowing ($w_h = w_s = 0.1$). These results emphasize the added value of incorporating historical borrowing in enhancing the evidence supporting an Accelerated Approval request.

In summary, borrowing information within the dual-criterion framework entails both benefits and risks. Strong borrowing is associated with higher average Accelerated Approval rates under the alternative scenario compared with the *no borrow* approach; however, it also carries certain risks, particularly in terms of increases in both the maximum global type I error and the inflation of the average global type I error rate. While we acknowledge the importance of evaluating standard operating characteristics under fixed scenarios, we advocate that the assessment of Bayesian metrics represents a valuable tool for quantifying the benefits and risks of Bayesian designs, thus ultimately facilitating informed discussions between sponsors and regulators.

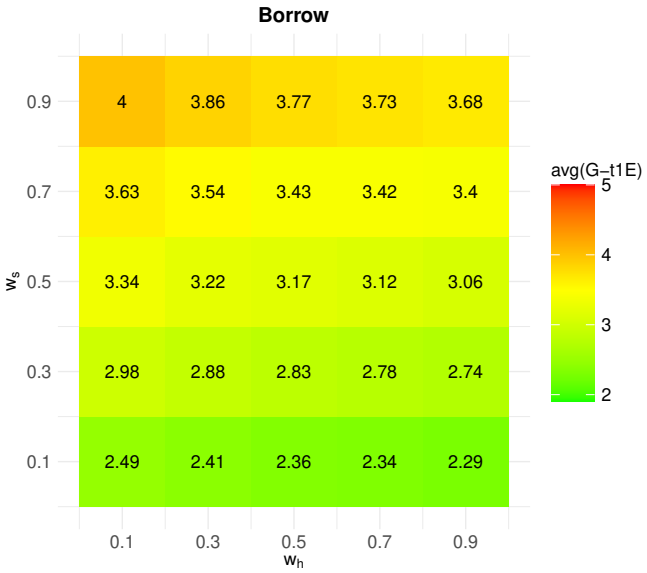
3.7 Discussion

In the last decades, the increased need to deliver effective treatments against life threatening diseases to the market has led various health authorities to allow for Accelerated Approvals, giving the possibility to promising treatments to enter the market earlier if enough evidence supports their efficacy.

In this context, we proposed a novel approach for Accelerated Approval (AA) interim analyses within phase III group sequential designs. With this approach, treatment efficacy on a short-term surrogate endpoint is tested along with the predictive probability of study success (PPoS) on the long-term primary endpoint. Different approaches are proposed to inform PPoS by (i) leveraging historical data on the control arm and (ii) borrowing from a documented relationship between surrogate and primary endpoints, derived from a meta-regression on historical trials. For historical controls borrowing, we relied on the methodology described in [72]. For incorporating historical information on the endpoints' relationship, we followed the methodology in [22, 76], where an informative prior (surrogate prior) for the primary endpoint, derived by combining partial data on the surrogate endpoint and the regression parameters, is updated by data on the

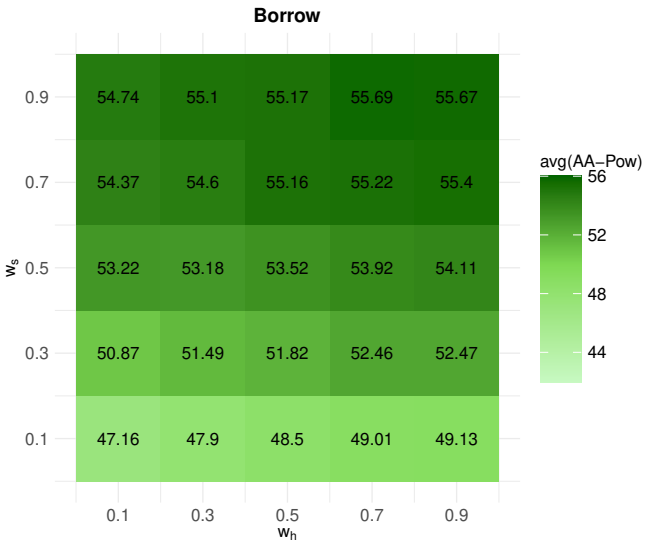


(a) Average Global type I error without borrowing.



(b) Average Global type I error with borrowing.

(c) Average Accelerated Approval Power without borrowing.



(d) Average Accelerated Approval Power with borrowing.

Fig. 3.4 Comparison of (a–b) the Average Global type I error avg(G-t1E) and (c–d) the Average Accelerated Approval Power avg(AA-Pow), computed for different pairs of the prior mixture weights (w_h, w_s) in the set $\mathcal{W} = (0.1, 0.3, 0.5, 0.7, 0.9)$, with and without borrowing.

primary endpoint available at the interim analysis.

Numerical results indicate that reinforcing efficacy testing on the surrogate endpoint through a predictive criterion based on the PPoS within the dual-criterion approach effectively reduces the probability of requesting Accelerated Approval for non-effective treatments. However, this improvement comes at the cost of a substantial reduction in the probability of meeting the criteria for an Accelerated Approval request, which is suboptimal from a sponsor's perspective.

In this context, a significant improvement within the dual-criterion framework is attained by

incorporating historical information from multiple sources, namely borrowing control information from past trials and borrowing on the treatment effect scale using a meta-analytic historical relationship between surrogate and primary endpoint. Notably, borrowing on the treatment effect scale has a greater impact than borrowing on the control arm, suggesting that this second approach alone may achieve comparable benefit.

This approach has been shown to preserve a high probability of satisfying the Accelerated Approval criteria in scenarios where concurrent and historical data are consistent, while simultaneously maintaining a low probability of requesting Accelerated Approval for ineffective treatments. These advantages have been proven also through the assessment of Bayesian operating characteristics proposed by Best et al. [80].

To address potential issues due to prior-data conflict (in both historical sources considered), we chose a robust mixture approach as presented in [16]. With this approach, the prior weight (reflecting the prior belief in the relevance of historical data for our current trial) needs to be determined. In this work, we provide a comparative analysis of the operating characteristics under different choices of weights, which could be useful for the final decision. Other approaches may be considered to find an appropriate prior weights, such as using empirical Bayes-based methods as proposed in [51] and [52]. Also, other methodologies for incorporating historical data have been presented in literature and could be considered for the approach proposed herein, such as power priors [15], commensurate priors [81] or more recently elastic priors [82].

In our work, a conservative choice for the decision thresholds is proposed so that the global type I error under the double null scenario is protected non-depending on the threshold for the predictive criterion, which is used to achieve further protection under a safeguard partial null scenario. Because of that, the global type I error is strictly below the nominal level, however less conservative options may be explored e.g. linking the selection of η^{PPoS} directly on the protection of the global type I error under the double null scenario. Moreover in our proposal an equal allocation of the nominal level for global type I error rate is proposed for the Full Approval analyses and the PFS criterion in the Accelerated Approval analyses; other types of allocation may be considered, e.g. allocating a larger portion to the early Accelerated Approval analysis if there is a high confidence in the relevance of the surrogate endpoint, or, conversely, placing more weight on the Full Approval analysis to avoid incorrect Accelerated Approvals.

A potential element of concern within our setting is trial integrity and the potential disclosure of primary endpoint information through PPoS criterion. It is important to emphasize that the interim analyses are intended to be conducted by an Independent Data Monitoring Committee (IDMC). As a consequence, the specific PPoS values would remain internal to the decision-making body, and only the final recommendation to seek Accelerated Approval would be shared, consistently with well established GSD practices. In cases where these procedural safeguards are not feasible, or if a more conservative approach to data disclosure is preferred, our Single-Criterion Approach (SCA) provides a potential solution for this alternative scenario. By relying exclusively on surrogate endpoint data (and omitting on-trial primary outcome information), the

SCA represents a potential compromise that eliminates the risk of information leakage while showing improved efficiency in the scenario with minor prior-data conflict. However, based on our results this comes at the cost of a significantly higher rate of incorrect AA requests in case of substantial prior-data conflict. Furthermore, it should be noted that implementing a DCA using only historical information for the primary endpoint, without incorporating concurrent data, would offer limited added value. In fact, in such a case, the PPoS would serve merely as a preliminary check, independent on the concurrent trial (as it would be driven only by external information) thus limiting the utility dual-criterion approach itself.

Even though our methodology jointly tests treatment efficacy on two different endpoints, it is inherently uni-variate, meaning that the potential patient level correlation between surrogate and primary endpoints is not explicitly taken into account in the current trial analysis and, therefore, the data collected on the two endpoints are considered independent. However, while this choice is sensible in the setting presented in [22] where the outcome on a surrogate endpoint in phase II is used to inform a future phase III trial, in our case an estimate of the within trial correlation between surrogate and primary endpoints may be obtained at the time of interim analysis using concurrent data. Hence, further work could be done to extend the current statistical model with a bi-variate analysis as discussed in [83].

Moreover, while the methodology presented in the paper is based on probabilities that there is a difference between treatments, the clinical relevance of the observed hazard ratios should also be discussed with the health authorities, considering the clinical context and the patient population.

It is important to emphasize that the methodology proposed in the present work should be interpreted primarily as a supportive framework for sponsors in guiding internal decision-making processes concerning applications for Accelerated Approval (AA) programs. The ultimate decision regarding the granting of AA resides solely with regulatory health authorities and is based on a multitude of considerations, including but not limited to the safety profile of the experimental treatment . Nonetheless, we posit that strengthening the evidentiary criteria for AA could be mutually advantageous. Specifically, such an approach may assist sponsors in mitigating the risk of subsequent withdrawal of approval, while concurrently aiding regulators in preventing the premature commercialization of therapeutics with unproven efficacy. In this context, the proposed Predictive Probability of Success (PPoS) criterion could also serve as a supplementary tool for regulatory agencies in determining the evidentiary requirements at later stages of the development process. For instance, if the PPoS at the interim analysis is found to be particularly high, regulators might consider imposing specific post-marketing commitments or obligations at the time of final analysis. Conversely, if AA is granted on the basis of only a moderately high PPoS, a correspondingly more stringent evidentiary threshold may be warranted for the transition to Full Approval.

While the present work adopts a Bayesian framework, the proposed methodology can, in principle, be extended to a frequentist setting. This would involve performing standard hypothesis tests for efficacy—such as the log-rank test in survival analysis—and replacing the Predictive

Probability of Success with a frequentist analogue, such as *Predictive Power* or *Conditional Power*. In this context, incorporating historical information requires careful methodological choices to preserve statistical validity. Suitable approaches include *test-then-pool* strategies or frequentist hierarchical models, which allow for borrowing while accounting for potential heterogeneity and maintaining control of the type I error rate.

Chapter 4

Including quantitative benefit-risk assessment in seamless phase 2/3 designs with dose selection

4.1 Introduction

Multi-arm multi-stage (MAMS) clinical trials are designs in which multiple active treatments are simultaneously evaluated against a common control group within a single group sequential trial. One advantage of such designs is that the same control patients may be used for all the comparisons, which then results in smaller sample sizes, compared to running separate two-arm trials. Moreover, different kind of adaptations can be made at intermediate points of the trial (referred to as *interim analyses*) based on accumulated data, including (but not limited to) dropping treatments for futility, re-estimating sample sizes (or allocation ratios) and changing the number of future interim looks.

Among the various possible adaptations, the problem of selecting a single treatment from a set of candidate therapies is a well-recognized challenge in drug development [84–87], particularly when the treatments under investigation are initially considered to have comparable potential. Such adaptive designs are especially attractive, as they integrate an exploratory phase during which all candidate treatments are evaluated with a subsequent confirmatory phase, in which only the most promising treatment is advanced to the formal testing stage, where the selected treatment is tested on the primary endpoint of interest only. However particular attention should be paid from a statistical perspective, in order to account properly for the multiplicity arising from the joint evaluation of multiple arms as well as for the adaptive elements. The statistical frameworks to deal with this are either based on the use of the p-value combination within the close-testing procedure [88–91], or on a generalization of the group-sequential design paradigm, thus constructing explicitly a test statistic accounting for the dose multiplicity and the specific selection criteria [10, 92, 11] (these can be viewed as generalization of the Dunnett test [93]).

Selection rules based on the same endpoint tested at final analysis have been vastly explored: in Stallard and Todd [10], the setting where one single dose is selected and taken forward in a multi-stage phase III is considered, in Stallard and Friede [94] the setting where an arbitrary (albeit fixed) number of doses is explored, while in Kelly et al. [92] a rule allowing for the selection of all dose close to the best performing one is proposed. A generalization of the latter approaches allowing for more flexibility in the number of doses to be selected is proposed in Magirr et al. [11]

The MAMS framework can also be incorporated into so-called *seamless designs*, namely trial structures in which distinct phases of drug development are integrated within a single multi-stage study. This approach offers several additional advantages, including a reduction of the "white space" between phases, a decrease in the overall sample size requirements through the accumulation of evidence across stages, and the provision of longer follow-up periods, thereby facilitating earlier access to long-term safety data [91]. In particular, seamless phase II/III designs have been proposed, in which an interim treatment selection is performed during the phase II component of the trial, followed by an efficacy evaluation of the selected treatment in the phase III component. It is important to note that the endpoints used in phase II and phase III are often distinct. For example, in oncology trials, phase II endpoints are typically short-term measures, such as Objective Response Rate (ORR) or Progression-Free Survival (PFS), whereas phase III endpoints generally focus on long-term outcomes, such as Overall Survival (OS). In such settings, basing treatment selection directly on the phase III endpoint is often impractical, as only limited data may be available at the time of the interim analysis. To address this issue, several strategies have been proposed for incorporating early outcome data into the treatment selection process. For instance, Friede et al. [95] considered an interim selection rule based exclusively on early outcomes, Stallard [96] investigated a combined approach using both early and primary endpoint data, while Kunz et al. [97] proposed a data-dependent rule that integrates features of the previous two methods.

If the treatments involved in the trial are different doses of the same drug, however, relying solely on efficacy endpoint data for dose selection may be not acceptable, particularly when there is some uncertainty regarding the safety of the dose considered, and a more comprehensive evaluation of the benefit-risk profile of each dose is advisable. This point has been highlighted in the recent FDA's project OPTIMUS [98], redacted by the Oncology Center of Excellence (OCE), which goal is improving the dose optimization process in oncology drug development. In particular, a more rigorous approach in identifying the optimal dose has been advised, which include a joint assessment of safety and efficacy elements in the dose evaluation process. Anticipating this challenge, Jaki and Hampson [99] proposed a two stage design in which an optimization function including a safety and an efficacy part is used for dose selection at interim, while two independent tests are performed at the time of the final analysis, so that superiority of the selected dose is claimed only if a statistical significant improvement is observed at the final analysis both in safety and efficacy. The optimization function proposed by Jaki and Hampson is

a weighted sum of Z statistics, each of which representing the improvement of a specific dose over the control arm in a specific safety or efficacy endpoint. Although this approach uses all available safety and efficacy information, it may be overly conservative, especially when multiple endpoints are considered due to the numerous univariate tests performed at the final analysis.

The idea of aggregating information coming from different sources to evaluate the overall utility of a drug - especially in the context of dose selection - is the basis of the so called benefit-risk assessment (BRA) procedure. Benefit-risk assessment of a drug consists in balancing its favorable clinical effects versus its undesirable side effects [100], and it's considered a key element to inform decision making in drug development both from sponsor perspective and from health authorities perspective.
According to the European Medicine Agency Benefit-Risk methodology project [101–104], one of the most comprehensive quantitative approaches for BRA is the so called multi-criteria decision analysis [105, 106] (MCDA). The idea behind MCDA is using a utility function to aggregate multiple safety/efficacy criteria in a single univariate score, reflecting the overall benefit-risk profile of the drug, where the relative importance given by the sponsor to each criteria is reflected by the choice of weights to be pre-specified in the utility function. Although in principle different utility functions may be employed, the linear utility function is still the most common and widespread option [107–110], due to its simple implementation. In order to account for the uncertainty of the estimates included in the MCDA criteria, a Bayesian version of the MCDA approach has been proposed [111]. In that, the MCDA utility scores are stochastic quantities, meaning that they are described by probability distributions. A comparison between different drugs within this approach can be made by confronting their posterior probabilities that the MCDA scores are superior to the control one. This approach is known as *probabilistic MCDA* or *stochastic MCDA* [111].

While quantitative benefit-risk balance is largely used after Phase III trials and are becoming more common in the context of regulatory approval process and post-marketing follow-up, however, its use in the design of clinical trials has been intended mainly as a complementary tool for decision making, thus not entering directly in the statistical analysis of the trial.

In this work, we introduce a novel multi-arm seamless two-stage design, in which the dose with the most favorable benefit–risk profile is selected at the first interim analysis (corresponding to the phase II component of the trial), while the efficacy of the selected dose is subsequently evaluated on the primary endpoint at the final analysis (corresponding to the phase III component). The benefit–risk profiles of the candidate doses are assessed using a probabilistic multi-criteria decision analysis (MCDA) framework, which allows for the inclusion of an arbitrary number of safety and efficacy criteria, including both phase II and phase III endpoints, in the construction of the MCDA score. The proposed dose selection procedure shares certain similarities with the approach of Jaki and Hampson [99], as it is based on a trade-off between efficacy and safety. However, in line with Stallard and Todd [10], the hypothesis testing at the final stage relies exclusively on the primary efficacy endpoint. Consequently, the proposed design can be regarded

as positioned between these two approaches, with the advantage of using a well established BRA methodology for treatment selection. The approach of Stallard and Todd [10] and Jaki and Hampson [99] are therefore adopted as comparators in the evaluation of our methodology, along with the approach proposed by Friede et al. [95].

The remaining of the manuscript is structured as follows: in Section 4.2 the trial setting and some backgroud are presented, in Section 4.3 the statistical framework used hereinafter is detailed, in Section 4.4 the methodology proposed is applied in a fictive case study in oncology, while in Section 4.5 an extensive simulation study is performed showing the performance of the proposed approach under different scenarios. Finally, the manuscript closes in Section 4.6 with a discussion.

4.2 Background methodology

4.2.1 Setting

A two stage multi-arm randomized control trial (RCT) is considered, where a number of doses J are compared to a control arm (either placebo or standard of care). The first stage is aimed at selecting one dose among the J included in the trial, while the second stage is aimed at testing efficacy of the selected dose.

Consider $i = 1, \dots, I$ endpoints are monitored during the study for each arm (with $i = 1$ indicating the primary efficacy endpoint), which can be efficacy endpoints if they give information regarding the activity of the dose, or safety endpoints if they are linked to specific safety aspects of the disease. Assume that for all patients the outcomes on the I endpoints follow a parametric distribution, and call $\boldsymbol{\theta_j} = (\theta_{j1}, \dots, \theta_{jI})$ the arm specific vector of parameters of interest for the j-th arm (with $j = 0$ indicating the control arm), which may be the mean responses in case of normally distributed endpoint, the log-hazard in case of survival endpoints or the log-odds in case of binary endpoints. Let us call $\hat{\boldsymbol{\theta}}_j^{\star} = \{\hat{\theta}_{j1}^*, \dots, \hat{\theta}_{jI}^*\}$ the maximum likelihood estimator of $\boldsymbol{\theta}_j$ at stage $\star = \{IA,\ FA\}$, constructed using the data available at the considered stage of the trial.

Let us define the treatment differences parameters $\delta_{ji} = \theta_{ji} - \theta_{0i}$, for $j = 1, \dots, J$. The individual null hypotheses $H_j^0 : \delta_{ji} = 0$ and the respective one-sided alternative hypotheses $H_j^{'} : \delta_{ji} > 0$ with $j = 1, \dots, J$. The objective of the trial in the proposed setting is to test the global null hypothesis $H^0 : \bigcap_{j=1}^{J} H_j^0$, meaning that none of the doses considered is effective on the primary endpoint, versus the alternative hypothesis that at least one dose is superior to the control, i.e. $H' : \bigcup_{j=1}^{J} H_j^{'}$.

The trial is structured as follows: at the time of the interim analysis, the dose j^* is selected based on data available at interim to proceed along with the control arm to the phase III part of the trial, while recruitment is stopped for all the other doses. Then at the time of the final analysis, an appropriate test statistic Z_{j^*1} is constructed using primary endpoint data available for the

selected dose and the control, and the individual null hypothesis $H^0_{j^*}$ is tested at the desired level, so that a rejection is claimed if

$$Z_{j^*1} > \eta \tag{4.1}$$

where η is a critical value, set in order to maintain the false positive rate below a nominal level α. Notice that, considering $H^0 \subset H^0_{j^*}$, then the global null hypothesis H^0 is rejected by extension.

4.2.2 Quantitative Multi-Criteria Decision Analysis (MCDA)

Given the parameters θ_{ji}, indicating the true performance of the j-th dose on the i-th endpoint, the so called *linear MCDA score* [105, 106] for the dose j is defined as follows

$$MCDA_j = \sum_{i=1}^{I} \omega_i u\left(\theta_{ji}\right) \tag{4.2}$$

where ω_i are the relative weights given to each criterion summing up to 1 (intended to reflect the stakeholder opinion regarding the relevance of each endpoint in the decision making), and $u(\cdot)$ is a *partial value function*, i.e. a univariate function which is monotone and bounded in the interval $[0, 1]$. An example of a such function is the so called *linear partial value function*, which is defined as

$$u(\theta_{ji}) = \begin{cases} 0 & \theta_{ji} < \theta_i^L \\ \frac{\theta_{ji} - \theta_i^L}{\theta_i^U - \theta_i^L} & \theta_{ji} \in \left(\theta_i^L\ ,\ \theta_i^U\right) \\ 1 & \theta_{ji} > \theta_i^U \end{cases} \tag{4.3}$$

where θ_i^L and θ_i^U represent respectively the minimum and maximum values for the parameter θ_{ji} which is plausible to observe.
Since θ_{ji} are unknown parameters, one simple way to confront the benefit-risk profiles of the J doses is to compare the estimates $\widehat{MCDA}_j$, constructed by replacing θ_{ji} with its MLE estimates in Equation 4.2.

In order to account for the potential uncertainty in the unknown parameters, another option consists in working in a Bayesian framework which is considering θ_{ji} as random variables following underlying unknown probability distributions. It follows that the MCDA scores are random variables themselves. In order to estimate the latter, first prior distributions $\pi^0_{\theta_{ji}}$ are given to θ_{ji} based on prior available information. Then, once data are observed, the prior distribution are updated to posterior distributions $\pi_{\theta_{ji}}$ and the posterior distribution for each $MCDA_j$ is retrieved from Equation 4.2, where the right-hand side must be seen as a random function of the random variables θ_{ji} (described by their posterior distributions).

The posterior probability that treatment j has a better benefit-risk profile than the control is then defined as

$$p_j = \mathbb{P}\left(\Delta MCDA_j > 0 \mid \mathcal{D}^{IA} \,;\, \pi^0_{\theta_{j1}}, \ldots, \pi^0_{\theta_{jI}}, \pi^0_{\theta_{01}}, \ldots, \pi^0_{\theta_{0I}}\right) \qquad j = 1, \ldots, J \tag{4.4}$$

where $\mathcal{D}^{IA}$ represents the data available at the time of the interim analysis, $\Delta MCDA_j = MCDA_j - MCDA_0$ and the probability is computed with respect to the posterior distributions. A direct comparison of the p_j can be used to determine the treatment with the best benefit-risk profile. This approach is referred in literature as *probabilistic MCDA* [111] and it is increasingly used in regulatory approval and post-marketing follow-up to establish the benefit-risk profile novel drugs [112].

4.3 A novel two-stage design with dose selection based on MCDA

4.3.1 Design

Within the setting proposed in Section 4.2.1, a novel two stage seamless design is detailed herein. Since the outcomes of interest are typically binary or time-to-event (survival) endpoints, for which the parameters of interest, such as the log-odds or log-hazard, are known to be asymptotically normally distributed, let us assume that the maximum likelihood estimators $\hat{\theta}^{\star}_{ji}$ follow a normal distribution

$$\begin{pmatrix} \hat{\theta}^{\star}_{j1} \\ \hat{\theta}^{\star}_{j2} \\ \vdots \\ \hat{\theta}^{\star}_{jI} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \\ \vdots \\ \theta_{jI} \end{pmatrix}, \begin{pmatrix} \sigma^{2,\star}_{j1} & \rho_{12}\sigma^{\star}_{j1}\sigma^{\star}_{j2} & \cdots & \rho_{1I}\sigma^{\star}_{j1}\sigma^{\star}_{jI} \\ \rho_{21}\sigma^{\star}_{j2}\sigma^{\star}_{j1} & \sigma^{2,\star}_{j2} & \cdots & \rho_{2I}\sigma^{\star}_{j2}\sigma^{\star}_{jI} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{I1}\sigma^{\star}_{jI}\sigma^{\star}_{j1} & \rho_{I2}\sigma^{\star}_{jI}\sigma^{\star}_{j2} & \cdots & \sigma^{2,\star}_{jI} \end{pmatrix} \right) \qquad j = 1, \ldots, J \quad \star = IA, FA \tag{4.5}$$

where $\sigma^{\star}_{ji}$ is the stage specific standard error of $\hat{\theta}^{\star}_{ji}$, while ρ_{hk} is the correlation between $\hat{\theta}^{\star}_{jh}$ and $\hat{\theta}^{\star}_{jk}$ (assumed to be known).

At the time of the interim analysis, when the statistics $\hat{\theta}^{IA}_{ji}$ are available for each of the J groups and each of the I endpoints, the *probabilistic MCDA* [111] approach is employed for dose selection, according to the following steps:

1. At the *design* stage, improper normal distributions $\pi^0\left(\theta_{ji}\right)$ are assigned to the parameters of interest and the upper and lower bounds θ^L_i and θ^U_i are pre-specified for the linear partial value function.

2. Posterior distribution for θ_{ji} is found via Bayesian update rule, so that $\pi\left(\theta_{ji} \mid \hat{\theta}_{ji}^{IA}\right) \propto f\left(\hat{\theta}_{ji}^{IA} \mid \theta_{ji}\right) \times \pi^0\left(\theta_{ji}\right)$, and accordingly the posterior distributions for the scores π_{MCDA_j} are found plugging the posterior distribution $\pi\left(\theta_{ji} \mid \hat{\theta}_{ji}^{IA}\right)$ into Equation 4.2. Due to the truncation introduced by the linear partial value function $u(\cdot)$, a Monte Carlo approximation of the distributions for $MCDA_j$ may be needed in this step.

3. The P_j are constructed according to Equation 4.4 and the active dose with largest P_j is selected at the time of the interim analysis, namely $S = \text{argmax}_j P_j$. A futility stopping rule can be incorporated at this stage by imposing the additional requirement that $P_S > \varepsilon$, thus accepting the null hypothesis in case the benefit-risk profile of the selected treatment is particularly poor.

Once the dose S is selected, all other doses are discontinued and accordingly randomization is stopped for these arms. Then at the time of the final analysis $\star = FA$, when the statistics $\hat{\theta}_{j^*1}^{FA}$ and $\hat{\theta}_{01}^{FA}$ is available (remind that $i = 1$ represent the primary efficacy endpoint), the test statistic for the selected dose on the primary endpoint can be constructed as follows

$$Z_{S1}^{FA} = \frac{\hat{\theta}_{S1}^{FA} - \hat{\theta}_{01}^{FA}}{\sqrt{\sigma_{S1}^{2,FA} + \sigma_{01}^{2,FA}}}, \tag{4.6}$$

and the individual null hypothesis H_S^0 is eventually tested accordingly to Equation 4.1.

4.3.2 Type I Error control

To determine the critical value for testing the null hypothesis on the primary endpoint at the final analysis, it is first necessary to derive the corresponding test statistic at the interim stage. Let $\sigma^{2,\dagger}$ denote the inverse of the pre-specified Fisher information targeted for the final analysis, and let t represent the information fraction on the primary endpoint at the interim analysis. The interim test statistic can then be expressed as follows:

$$f_{Z_{S1}^{IA}}(z) = \sum_{j=1}^{J} \int_{-\infty}^{+\infty} f_{S|\hat{\theta}_{01}^{IA},\hat{\theta}_{S1}^{IA}}\left(j \,\middle|\, \hat{\theta}_{01}^{IA} = x\,, \hat{\theta}_{j1}^{IA} = \frac{\sigma^{\dagger} z}{t} + x\right) f_{\hat{\theta}_{j1}^{IA}}\left(\frac{\sigma^{\dagger} z}{t} + x\right) f_{\hat{\theta}_{01}^{IA}}(x)\, dx \tag{4.7}$$

where the formula must be read in this way:

- $f_{S|\hat{\theta}_{01}^{IA},\hat{\theta}_{S1}^{IA}}\left(j \mid \hat{\theta}_{01}^{IA}, \hat{\theta}_{S1}^{IA}\right)$ denotes the probability that treatment j is selected, conditional on the observed estimates $\hat{\theta}_{01}^{IA}$ and $\hat{\theta}_{S1}^{IA}$. While the selection variable S depends on all estimates $\hat{\theta}_{ji}$, we condition explicitly only on $\hat{\theta}_{01}^{IA}$ and $\hat{\theta}_{S1}^{IA}$, as these are the only components influencing the Z-statistic for the selected treatment j.

- $f_{\hat{\theta}^{IA}_{j1}}\left(\frac{\sigma^{\dagger}z}{t}+x\right)$ denotes the probability that the MLE on the primary endpoint for the treatment j is equal to $\frac{\sigma^{\dagger}z}{t}+x$ (which leads to a Z-statistics of z, accordingly to Equation 4.6)

- $f_{\hat{\theta}^{IA}_{01}}(x)$ denotes the probability that the MLE on the primary endpoint for the control arm is equal to x

While the marginal distributions $f_{\hat{\theta}^{IA}_{01}}$ and $f_{\hat{\theta}^{IA}_{j1}}$ are known, as given in Equation 4.5, the conditional distribution $f_{S|\hat{\theta}^{IA}_{01},\hat{\theta}^{IA}_{S1}}$ does not have a closed-form analytical expression. This is due to the complexity introduced by the probabilistic selection criterion and the truncation inherent in the linear partial value function described in Equation 4.3. However, an approximation of the selection rule based on probabilistic MCDA formulated in terms of univariate summary statistics rather than posterior probabilities is possible, and is established in the following theorem 1. A formal proof is provided in Appendix A.

Theorem 1. *Let us consider the interim statistics $\hat{\theta}_{ji}$ as defined in Equation 4.5 (the superscript $\star = IA$ is omitted for sake of notation). Consider the linear partial value function $u(\cdot)$ as defined in Equation 4.3, the MCDA distributions $MCDA_j$ as defined in Equation 4.2, and the posterior probabilities P_j as defined in Equation 4.4. Let us define*

$$\gamma_j = \frac{M_j - M_0}{\sqrt{V_j + V_0}} \qquad j = 1, \dots, J \tag{4.8}$$

where

$$M_j = \sum_{i=1}^{I} \frac{\omega_i \left(\theta_{ji} - \theta_i^L\right)}{\theta_i^U - \theta_i^L} \qquad V_j = \sum_{p=1}^{I}\sum_{q=1}^{I} \frac{\rho_{pq}\omega_p\omega_q\sigma_{jq}\sigma_{jp}}{\left(\mu_p^U - \mu_p^L\right)\left(\mu_q^U - \mu_q^L\right)}. \tag{4.9}$$

and call $\hat{\gamma}_j$ and $\hat{M}_j$ their MLE, obtained substituting θ_{ji} with $\hat{\theta}_{ji}$.
If θ_i^L and θ_i^U are chosen so that the posterior probabilities $\mathbb{P}\left(\theta_i^L < \theta_{ji} < \theta_i^U \,|\hat{\theta}_{ji}\right) \approx 1 \;\; \forall i = 1, \dots, I \;\; \forall j = 0, \dots, J$, then the following holds:

$$\underset{j=1,\dots,J}{\operatorname{argmax}} P_j = \underset{j=1,\dots,J}{\operatorname{argmax}} \hat{\gamma}_j$$

Note that the validity of Theorem 1 relies on the assumption that $\mathbb{P}\left(\theta_i^L < \theta_{ji} < \theta_i^U \mid \hat{\theta}_{ji}\right) \approx 1$ for all $i = 1, \dots, I$ and $j = 0, \dots, J$. We argue that this condition is generally satisfied in practice, as θ_i^L and θ_i^U are defined as boundary values that the parameters θ_{ji} are highly unlikely to exceed. In instances where this assumption is violated, the selection criterion based on $\hat{\gamma}_j$ may still be viewed as an approximation of the P_j-based criterion; the accuracy of this approximation increases as the probability $\mathbb{P}\left(\theta_i^L < \theta_{ji} < \theta_i^U \mid \hat{\theta}_{ji}\right)$ approaches 1.

Let us now consider the dose selection based on $\hat{\gamma}_j$, which is $S = \text{argmax}_{j=1,\dots,J}\,\hat{\gamma}_j$. Since $\hat{M}_j$ are linear combination of $\hat{\theta}_{ji}$ (which are normally distributed), then $\hat{M}_j$ are normally distributed themselves, which makes the analytical derivation of the test statistic possible. In particular, since S depends on $\hat{M}_j$, then the distribution of $S|\hat{\theta}_{j1}, \hat{\theta}_{01}$ can be expanded as:

$$f_{S|\hat{\theta}^{IA}_{j1},\hat{\theta}^{IA}_{01}}(j) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{S|\hat{M}_j,\hat{M}_0}(j) \times f_{\hat{M}_j|\hat{\theta}^{IA}_{j1}}(y) \times f_{\hat{M}_0|\hat{\theta}^{IA}_{01}}(h)\, dydh \tag{4.10}$$

Considering the joint bivariate normal distribution for $(\hat{M}_j, \hat{\theta}^{IA}_{j1})$

$$\begin{pmatrix}\hat{M}_j \\ \hat{\theta}^{IA}_{j1}\end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix}M_j \\ \theta^{IA}_{j1}\end{pmatrix}, \begin{pmatrix}V_j & c_j \\ c_j & \sigma^{2,IA}_{j1}\end{pmatrix}\right), \quad \text{with} \quad c_j = \sum_{i=1}^{I}\omega_i\left[\prod_{k\neq h}\left(\theta^U_k - \theta^L_k\right)\right]\rho_{1i}\sigma^{IA}_{j1}\sigma^{IA}_{ji} \tag{4.11}$$

where c_j is the covariance between $\hat{M}_j$ and $\hat{\theta}^{IA}_j$, it follows from the properties of the multivariate normal distributions that the conditional distribution $\hat{M}_j|\hat{\theta}^{IA}_{j1}$ is again normally distributed

$$\hat{M}_j|\hat{\theta}^{IA}_{j1} \sim N\left(M_j + \frac{c_j}{\sigma^{2,IA}_{j1}}\left(\hat{\theta}^{IA}_{j1} - \theta_{j1}\right), V_j\left(1 - \frac{c_j^2}{\sigma^{2,IA}_{j1}V_j}\right)\right). \tag{4.12}$$

Moreover, since $\hat{\gamma}_k$ are independent conditionally on $\hat{M}_0$, then the conditional distribution of $S|\hat{M}_j, \hat{M}_0$ can be written as

$$\begin{aligned} f_{S|\hat{M}_0,\hat{M}_S}(j^*) &= \prod_{k\neq j^*}\mathbb{P}\left(\hat{M}_k < \hat{M}_0 + \left(\hat{M}_S - \hat{M}_0\right)\sqrt{\frac{V_0+V_k}{V_0+V_S}}\right) = \\ &= \prod_{k\neq j^*}\Phi\left(\frac{\hat{M}_0 + \left(\hat{M}_S - \hat{M}_0\right)\sqrt{\frac{V_0+V_k}{V_0+V_S}} - M_k}{\sqrt{V_k}}\right) \end{aligned} \tag{4.13}$$

The distribution of the Z-statistics at the time of the interim analysis is obtained by substituting Equations 4.12 and 4.13 into Equation 4.10, and subsequently substituting the resulting expression into Equation 4.7.

Exploiting the independent increment of the test statistics Z_{S1}, the conditional distribution of the final test statistics conditional on the interim one is

$$Z^{FA}_{S1}|Z^{IA}_{S1} \sim N\left(tZ_S, 1-t\right). \tag{4.14}$$

Finally, the distribution of the test statistics on the primary endpoint at the final analysis is found by integrating the rejection probability over the distribution of the interim Z statistic

$$\begin{aligned} F_{Z_{S1}^{FA}}(q) &= \int_{-\infty}^{+\infty} F_{Z_{S1}^{FA}}\left(q \mid Z_{S1}^{IA} = z\right) f_{Z_{S1}^{IA}}(z)dz \\ &= \int_{-\infty}^{+\infty} \left[1 - \Phi\left(\frac{q - t_j z}{\sqrt{t_j - 1}}\right)\right] f_{Z_{S1}^{IA}}(z)dz. \end{aligned} \tag{4.15}$$

This result will be applied in the next section to derive appropriate decision boundaries that ensure control of the type I error.

4.3.3 Strong control of type I Error rate

In the design proposed, selection at interim is based on the quantitative assessment of the benefit-risk profile of each dose in the trial, which involves the evaluation of multiple efficacy and safety endpoints. However, only the efficacy on the primary endpoint θ_{S1} is tested for the selected dose at the final analysis, thus the null hypothesis is $\delta_{11} = \delta_{21} = \cdots = \delta_{J1} = 0$ does not involve any specification regarding θ_{ji} for any $i = 2, \ldots, I$, nor for σ_{ji}^2.

In order to control type I error in the *strong sense*, which is under all possible values for δ_{ji} and σ_{ji}^2, with $i = 2, \ldots, I$ and $j = 1, \ldots, J$ it is sufficient to choose η so that the false positive rejection rate is controlled under a *worst case configuration* θ_{ji}^*. The formula for the determination of the critical value is then:

$$\eta = \underset{q}{\operatorname{argmin}} \left| F_{Z_S^{FA}}(q) + \alpha - 1 \right| \qquad \theta_{ji} = \theta_{ji}^* \tag{4.16}$$

The worst case configuration θ_{ji}^* in this setting is the combination of parameters θ_{ji} where, under the null, the test statistic Z_{S1}^{FA} is stochastically maximized, thus also the critical value η is maximized. In this sense, since η is stochastically decreasing with each σ_{ji}^2 and δ_{ji}, it is maximized when $\delta_{ji} = 0$ and σ_{ji}^2 are minimum $\forall i = 2, \ldots, I$ and $\forall j = 1, \ldots, J$ (see Figure C.1 of the Supplementary Material for an illustration).

Notice that if σ_{ji}^2 are independent on θ_{ji}, e.g. in case of normal endpoint with known variance, no minimization is needed as the variance depends only on the expected amount of information at the interim analysis, on the contrary when σ_{ji}^2 depend on θ_{ji} then the *worst case configuration* is found as

$$\theta_{ji}^* = \underset{\theta_{ji}}{\operatorname{argmin}}\, \sigma_{ji}^2\left(\theta_{ji}\right) \qquad i = 2, \ldots, I \quad j = 1, \ldots, J$$

As an example, in case $Y_{jik} \sim \text{Bernoulli}\left(p_{ji}\right)$ and θ_{ji} is defined as the log-odds of p_{ji}, namely $\log\left(p_{ji} \big/ \left(1-p_{ji}\right)\right)$, then the $\hat{\theta}_{ji}$ is asymptotically normally distributed with mean θ_{ji} and variance $\sigma^2_{ji} = 1 \big/ \left(np_{ji}(1-p_{ji})\right)$. It follows that σ^2_{ji} is minimized when $p_{ji} = 0.5$, as a consequence the worst case configuration would be $\theta_{ji} = 0 \; \forall j = 1, \ldots, J$.

In addition it should be noted that, the critical value η increases with the assumed correlations ρ_{1k} for $k \geq 2$ (Figure C.2 of the Supplementary Material), and coincides with the critical values derived by Stallard and Todd [10] when all factors included in the MCDA are perfectly correlated with the primary endpoint, i.e., $\rho_{12}, \ldots, \rho_{1I} = 1$. This is intuitive, since in such a setting the dose-selection procedure is equivalent to that based solely on the primary endpoint. Accordingly, accurate specification of the correlations among endpoints is of fundamental importance. Additional determinants of the critical value η include the information fraction, where larger values of t correspond to larger η, the weights assigned in the MCDA scores (with higher weight on the primary endpoint yielding larger values of η), and the number of experimental dose levels, with larger J leading to larger η. The dependence of η on these design choices, along with the explicit type I error control ensured by Equation 4.16, is summarized in Tables C.5 and C.6 of the Supplementary Material.

4.3.4 Power and Sample size calculation

Power is defined, in the context of multi-arm designs, as the probability of rejecting the global null hypothesis $\delta_{11} = \delta_{21} = \cdots = \delta_{J1} = 0$ when, in fact, $\delta_j > \delta_0 > 0$ for at least one $j \in \{1, \ldots, J\}$. Here, δ_0 denotes the *maximum uninteresting difference*, which represents the largest effect size that is not considered clinically relevant. In multi-arm multi-stage (MAMS) designs involving dose selection, particularly when selection is based on the primary endpoint, it is common practice to evaluate power under the so-called *least favorable configuration* (LFC) [93, 11]. The LFC corresponds to a scenario in which only one dose j is truly effective on the primary endpoint, with an effect size δ^*_1, while all other doses exhibit a moderate effect δ^0_1, lying on the upper boundary of the non-interest region; that is,

$$C^P = \bigcup_{j=1}^{J} C^P_j \qquad\qquad C^P_j = \left\{\theta_{k1} = \theta_{01} + \delta^0_1 \; ; \; k \neq j\right\} \cap \left\{\theta_{j1} = \theta_{01} + \delta^*_1\right\}$$

Notice that this choice of LFC is conservative, in fact, it minimizes the probability to select the best performing dose on the primary endpoint [85], among all the configurations where at least one dose is effective at a target level $\delta^* > \delta^0$.

When the selection is based on MCDA, however, the distribution of S, and so the power itself, does not depend only on $\delta_{11} = \delta_{21} = \cdots = \delta_{J1}$, but also on $\gamma_1, \ldots, \gamma_J$. The idea of *least favorable configuration* can be extended in this context as follows:

$$C^P = \bigcup_{j=1}^{J} C_j^P \qquad C_j^P = \left\{\theta_{k1} = \theta_{01} + \delta_1^0 \,;\, k \neq j\right\} \cap \left\{\gamma_k = \gamma^0 \,;\, k \neq j\right\} \cap \left\{\theta_{j1} = \theta_{01} + \delta_1^*\right\} \cap \left\{\gamma_j = \gamma^*\right\}$$

where γ^* represents a desirable value for the benefit-risk score γ, while γ^0 may represent a non-interesting value for the benefit-risk score γ. It is important to note that, according to Equation 4.8, the benefit-risk score γ_j is a function of both θ_{ji} and the standard errors of the interim estimates, σ_{ji}^{IA}. Consequently, distinct combinations of these parameters may lead to identical values of γ_j. Furthermore, since γ_j represents an aggregated measure without a direct clinical interpretation, its elicitation is inherently challenging. In particular, the standard errors σ_{ji}^{IA} depend on the information fraction available at the interim analysis and are therefore implicitly determined by the trial sample size. For these reasons, the direct specification of meaningful thresholds γ^* and γ^0 is often impractical, especially given the heterogeneity in type and scale of the endpoints entering the MCDA framework. A more pragmatic strategy is to specify clinically interpretable values for δ_i^0 and δ_i^* on the parameter scale, from which the corresponding standard errors σ_{ji}^{IA}, and thus the implied values of γ_j, can be computed at the sample size determination stage via Equation 4.8. This indirect elicitation anchors the design on clinically meaningful quantities while preserving analytical rigor. Assuming equal allocation across all doses, the procedure can be summarized as follows:

- Assumptions are made on the control parameters θ_{0i} for all $i = 1, \ldots, I$. Then, for each endpoint i, a target improvement δ_i^* and a minimal clinically non-relevant improvement δ_i^0 are elicited based on input from the clinical team.
- The common standard deviations $\sigma_{ji}^{\star} = \sigma_i^{\star}$ are determined as a byproduct of the sample size calculation, which is performed to ensure a nominal power of $1 - \beta$, where β denotes the target type II error, under the parameter configuration specified in the first step.
- The values γ^0 and γ^* are retrieved plugging respectively $\theta_{0i} + \delta_i^0$, $\theta_{0i} + \delta_i^*$ and the computed standard deviations σ_{ji}^{IA} in Equation 4.8.

For the second step, sample size can be computed by numerically solving the Equation for the Power, namely

$$\Pi = 1 - \beta = \sum_{j=1}^{J} \Pi^{C_j^P}, \tag{4.17}$$

with respect to $\sigma^{2,\dagger}$, taking the expectation for $\sigma_{ji}^{2,IA}$ and V_j (in case they are not fixed at the time of the interim analysis). In Equation 4.17, $\Pi^{C_j^A}$ is the rejection rate under the j-th alternative configuration, obtained by using Equation 4.15, with the summation in Equation 4.7 restricted

to the event $S = j$ (an analytical expression can be found in the Appendix at the end of the manuscript).

Note that the assumption of equal allocation ratios across doses is necessary in this context. Indeed, if the information available at interim differs across doses, then the same choice of δ_i^* (respectively δ_i^0) would lead to different values of γ^* (respectively γ^0), depending on the specific information available for the most effective dose. In such cases, a conservative approach to elicitation is to assume that the dose with the minimum allocation ratio is also the most effective one. Eliciting under this configuration yields a lower bound for the power. As a general guidance, notice that for a given type II error β, the follwing hold:

1. the larger δ_i^*, the higher the probability to reject the null hypothesis if the dose with that treatment effect on the primary endpoint is selected, the lower the sample size needed to reach a power of $1-\beta$.
2. the higher $\delta_i^* - \delta_i^0$, the higher the probability to select dose j at the time of the interim analysis, the lower the sample size needed to reach a power of $1-\beta$.

4.4 Case Study

4.4.1 Study setting

A seamless phase II/III trial in oncology is considered, where two active doses are tested against a control arm. A total of 450 patients is randomized in 1:1:1 proportion with a constant accrual rate of 20 patients per arm per month, and for each arm considered in the study $I = 4$ different endpoints are evaluated.

Overall Survival (OS) serves as the primary efficacy endpoint (or Phase III endpoint) and is assumed to follow an exponential distribution with hazard parameters denoted by λ_{j1}. The logarithm of the hazard rates, taken with a negative sign, defines the parameters of interest: $\theta_{j1} = \log(\lambda_{j1})$. The estimator $\hat{\theta}_{j1}^{\star}$, where $\star = \text{IA}, \text{FA}$, is assumed to follow an asymptotic normal distribution centered at the true value θ_{j1}, with standard error $\sigma_{j1}^{\star} = \sqrt{1/n_{j1}^{\star}}$. Here, $n_{j1}^{\star}$ represents the number of OS events observed at the conclusion of stage $\star$ in the j-th treatment arm.

Objective Response Rate (ORR) is treated as the secondary efficacy endpoint (or Phase II endpoint) and is modeled using a Bernoulli distribution with true response probabilities denoted by λ_{j2}. The corresponding odds are defined as $O(\lambda_{j2}) = \lambda_{j2}/(1-\lambda_{j2})$, and the parameter of interest is the log-odds, given by $\theta_{j2} = \log(O(\lambda_{j2}))$. The maximum likelihood estimators $\hat{\theta}_{j2}^{\star}$, for $\star = \text{IA}, \text{FA}$, are assumed to follow an asymptotic normal distribution centered at the true value $\theta j2$, with standard error $\sigma_{j2}^{\star} = \sqrt{1/(\lambda_{j2} n_{j2}^{\star}) + 1/((1-\lambda_{j2}) n_{j2}^{\star})}$. Here, $n_{j2}^{\star}$ represents the number of patients evaluable for ORR at the end of stage $\star$ in the j-th treatment arm.

Severe adverse event rate (SevAE) and serious adverse event rate (SerAE) are monitored as safety endpoints. For each patient, these outcomes are assumed to follow a Bernoulli distribution, with parameters λ_{ji} denoting the true probability of experiencing the respective adverse event. The log-odds, defined as $\theta_{ji} = \log\left(\frac{\lambda_{ji}}{1-\lambda_{ji}}\right)$, are taken as the parameters of interest (taken with a negative sign). The maximum likelihood estimators $\hat{\theta}^{\star}_{ji}$, for $\star = \{\text{IA}, \text{FA}\}$, are assumed to follow an asymptotic normal distribution centered at the true value θ_{ji}, with standard error $\sigma^{\star}_{ji} = \sqrt{\frac{1}{\lambda_{ji} n^{\star}_{ji}} + \frac{1}{(1-\lambda_{ji}) n^{\star}_{ji}}}$, where $n^{\star}_{ji}$ is the number of patients evaluable for the corresponding adverse event at the end of stage $\star$ in the j-th treatment arm.

Zero patient-level correlation is assumed between endpoints, which leads accordingly to $\rho_{hk}\ \forall h \in (1, \ldots, 4)$ and $h \neq k$.

4.4.2 Study design

For this study, the control is assumed to have a median OS of 9 months, corresponding to $\theta_{01} = 2.56$, a true ORR of 30%, corresponding to $\theta_{02} = -0.84$, a 30% probability of experiencing a serious adverse event and the same probability to experience a severe adverse event, which corresponds to $\theta_{0i} = 0.84$, with $i = 3, 4$.

The interim analysis for dose selection is planned after a number n of patients overall have been followed up for at least 2 months (which is considered the response assessment time for the three binary endpoints considered, referred to as *delay*).

For the MCDA assessment, a vector of weights $\boldsymbol{\omega} = (0.1, 0.4, 0.25, 0.25)$ is used. This choice is made so that efficacy and safety are equally weighted in the benefit-risk balance. However, considering that the BRA is made at interim (i.e. after the phase II part of the trial) and considering that only few OS data are likely to be observed at that time, a larger weight is assigned to phase II endpoint compared to the phase III endpoint. The two safety criteria are equally weighted. The linear partial value function defined in Equation 4.3 is used in the MCDA, and maximum and minimum values θ_i^U and θ_i^L (which define the expected domain of the random variable θ) are chosen to have a very low chance to observe MLEs more extreme than those values.

The least favorable configuration (LFC) is summarized in Table 4.1. On the phase III endpoint an hazard ratio $\delta_1^* = 0.45$ is targeted (corresponding to a 4 months improvement in OS), while a maximum non-interesting hazard ratio is set to $\delta_1^0 = 0.15$ (corresponding to a 2 moths improvement in OS with respect to baseline). On the phase II endpoint, a log-odds ratio of $\delta_2^* = 1.24$ is targeted (corresponding to a 30% increase in ORR), while a maximum non-interesting log-odds ratio is set to $\delta_2^0 = 0.44$ (corresponding to a 10% increase in ORR with respect to baseline). The two active doses are assumed to be equal to the control on the two safety endpoint, so that $\delta_i^* = \delta_i^0 = 0\ \forall i = 3, 4$.

The required number of events at the final analysis (FA), denoted by d, and the number of evaluable patients at the interim analysis (IA), denoted by n, are jointly determined through a

numerical search for the pair (d,n) that satisfies Equation 4.17. This search is conducted under the least favorable configuration (LFC) specified in Table 4.1, targeting 80% power with an average information fraction on the primary endpoint of $t = 0.2$. In this procedure, the variances of the log-odds for the binary endpoints are specified as $\sigma_{ji}^{2,IA} = \frac{1}{\lambda_{ji} n} + \frac{1}{(1-\lambda_{ji})n}$, while the variances of the log-hazards for the survival endpoint are given by $\sigma_{j1}^{2,IA} = \frac{1}{d_j^{IA}}$, where d_j^{IA}, for $j = 0,1,2$, denote the expected number of events at interim. These are computed to satisfy $d_0^{IA} + d_2^{IA} = td$ under the LFC and the assumed recruitment model. Assuming uniform recruitment and equal allocation, the expected number of events d_j^{IA} admits the analytical expression

$$d_j^{IA} = \frac{N_j^{IA}}{\tau}\left(1 - \int_0^{\tau} e^{\lambda_{j1}(\tau - r)}\, dr\right),$$

where N_j^{IA} denotes the number of patients recruited up to the interim analysis, and τ is the calendar time at which the interim analysis occurs. The latter is determined by

$$\tau = \frac{n}{3 \times \text{accrual rate}} + \text{delay}.$$

Applying this procedure yields an interim analysis after $n = 183$ patients are evaluable for the binary endpoints, and a final analysis triggered once $d = 179$ overall survival events have been observed across the selected dose and the control arm.

The critical value η for testing the primary endpoint at the final analysis is computed exploiting Equation 4.15, with t corresponding to the *observed* information fraction of the selected dose on the primary endpoint. For this computation, the *worst case configuration* is considered according to Section 4.3.4, namely log-hazards $\theta_{01} = \cdots = \theta_{J1}$ are assumed on the primary endpoint, while the log-odds ratios corresponding to the three binary endpoints are set to $\theta_{ji} = 0\ \forall j = 0,1,2$ and $\forall i = 2,3,4$ (corresponding to probability parameters $\lambda_{ji} = 0.5\ \forall j = 0,1,2$ and $\forall i = 2,3,4$).

Table 4.1 Design specification.

	Least Favorable Configuration (LFC)			**MCDA parameters**		
Endpoint	**Control** ($j = 0$)	**dose 1** ($j = 1$)	**dose 2** ($j = 2$)	ω_i	θ_i^L	θ_i^L
$-\log(H_{OS})$	2.31 (med(OS)=7†)	2.56 (med(OS)=9†)	2.76 (med(OS)=11†)	0.1	1.98 (med(OS)=5†)	3.07 (med(OS)=15†)
$\log(O_{ORR})$	-0.84 ($p = 30\%$)	-0.40 ($p = 40\%$)	0.40 ($p = 60\%$)	0.4	-1.38 ($p^{min} = 20\%$)	1.38 ($p^{max} = 70\%$)
$-\log(O_{sevAE})$	0.84 ($p = 30\%$)	0.84 ($p = 30\%$)	0.84 ($p = 30\%$)	0.25	-1.38 ($p^{min} = 20\%$)	1.38 ($p^{max} = 80\%$)
$-\log(O_{serAE})$	0.84 ($p = 30\%$)	0.84 ($p = 30\%$)	0.84 ($p = 30\%$)	0.25	-1.38 ($p^{min} = 20\%$)	1.38 ($p^{max} = 80\%$)

† Overall Survival (OS) is expressed in months.

4.4.3 Analysis

Table 2 shows a summary table of interim statistics for each arm included in the trial on the four endpoints considered.

Table 4.2 Data corresponding to case study.

	$-\log(\boldsymbol{H_{OS}})$		$\log(\boldsymbol{O_{ORR}})$		$-\log(\boldsymbol{O_{sevAE}})$		$-\log(\boldsymbol{O_{serAE}})$		**Probabilistic MCDA**	
Arm	$\hat{\boldsymbol{\theta}}_{j1}^{IA}$	$\boldsymbol{\sigma}_{j1}$	$\hat{\boldsymbol{\theta}}_{j2}^{IA}$	$\boldsymbol{\sigma}_{j2}$	$\hat{\boldsymbol{\theta}}_{j3}^{IA}$	$\boldsymbol{\sigma}_{j3}$	$\hat{\boldsymbol{\theta}}_{j4}^{IA}$	$\boldsymbol{\sigma}_{j4}$	$\boldsymbol{P_j}$	$\boldsymbol{\gamma_j}$
$j=0$ (control)	2.39	0.22	-0.30	0.26	0.82	0.28	0.31	0.26	–	–
$j=1$ (dose 1)	2.72	0.26	-0.50	0.26	1.12	0.30	1.31	0.31	0.92	1.47
$j=2$ (dose 2)	2.82	0.27	0.84	0.27	-0.76	0.27	0.23	0.25	0.75	0.68

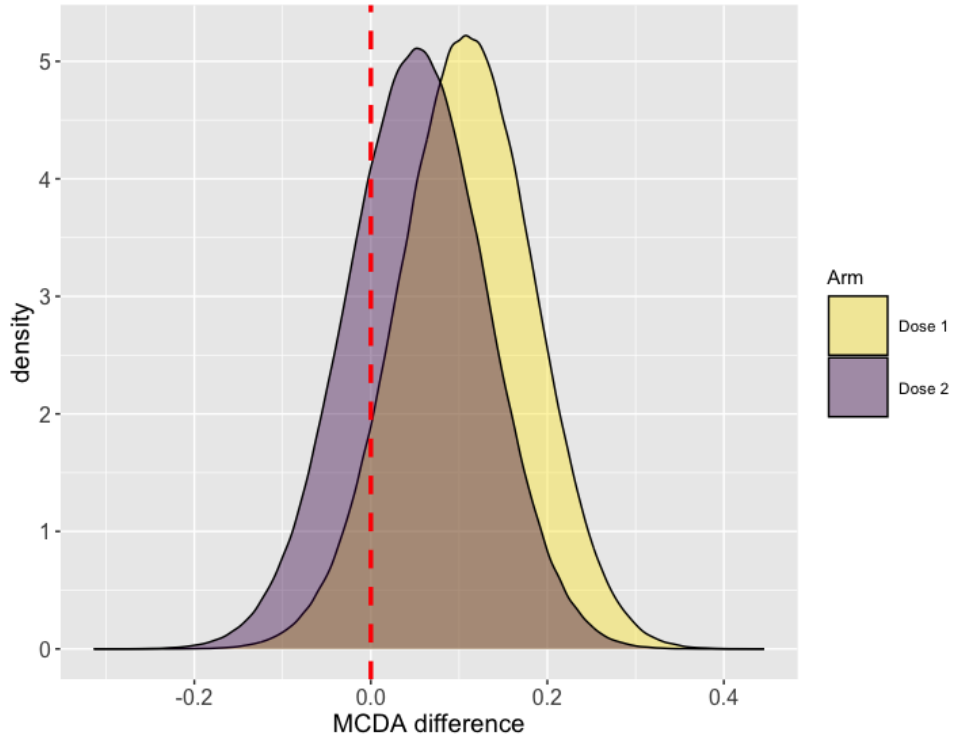


Fig. 4.1 Posterior distribution for the MCDA difference for the two active doses.

Data show that dose 2 is more effective than dose 1 both in improving OS ($\hat{\theta}_{21}^{IA} > \hat{\theta}_{11}^{IA}$) and increasing ORR ($\hat{\theta}_{22}^{IA} < \hat{\theta}_{32}^{IA}$), while on the contrary it performs worse in terms of the two safety endpoints ($\hat{\theta}_{33}^{IA} < \hat{\theta}_{23}^{IA}$ and $\hat{\theta}_{34}^{IA} < \hat{\theta}_{24}^{IA}$).

At the time of the interim analysis, the MCDA distributions for the three doses included in the trial are constructed following the procedure detailed in Section 4.3.1. Left panel of Figure 4.1 shows that the three MCDA distributions have similar variance due to the equal randomization between arms, however, the distributions of the two active doses are shifted towards larger MCDA values, suggesting an overall improvement in benefit-risk balance. The latter is also reflected in Figure 4.1, showing the distributions of the MCDA differences $\Delta MCDA_j$ $j = 1, 2$. Once the posterior distribution for $\Delta MCDA_j$ are available, the posterior probability that the difference in MCDA for the two active doses is greater than zero can be computed, resulting in $P_1 = 0.92$ and $P_2 = 0.76$. As a consequence, dose 1 is taken forward and accordingly randomization is stopped for dose 2.

Note that the same dose would be chosen if selection was based on γ_j as defined in Section 4.3.2.

In fact, plugging the observed interim statistics $\hat{\theta}_{ji}^{IA}$ and σ_{ji}^{IA} into Equation 4.8, we would obtain $\gamma_1 = 1.47$ versus $\gamma_2 = 0.68$.

Considering the selected dose and the control, the interim analysis has been performed after 20% of the pre-planned number of events has been observed (which is $t = 0.203$), therefore the critical value $\eta = 2.027$ is used (expressed on the Z scale). At the time of the final analysis, a standardized test statistic $Z_{S1}^{FA} = 1.62$ is observed for the selected dose, which being lower than the critical value is not compatible with a rejection of the null hypothesis.

Notice that if selection was made based solely on efficacy data (either on the phase II or phase III endpoint), dose 2 would be continued. However, although this would likely bring to a rejection of the null hypothesis at final analysis, due to the poor safety exhibited at the time of the interim analysis, there would be a low chance to obtain a marketing approval by health authorities.

4.5 Simulation study

This section presents an extensive simulation study conducted to evaluate the performance of the proposed methodology. The simulations are carried out within the framework and trial designs introduced in Section 4.4, while varying the characteristics of the active doses to investigate a broad range of potential safety and efficacy profiles. For comparative purposes, alternative designs available in the literature are also included, encompassing different dose selection rules and testing procedures. The evaluation of all approaches is based on appropriately defined operating characteristics (OCs), tailored to the specific requirements of the present context.

4.5.1 Setting

We consider a seamless phase II/III oncology trial comparing two active doses against a control arm, with patients randomized equally (1:1:1) under a constant accrual rate of 20 patients per arm per month. Four endpoints are evaluated: overall survival (OS), modeled with an exponential distribution with hazard rate λ_{j1} and parameterization $\theta_{j1} = -\log(\lambda_{j1})$; Objective Response Rate (ORR), following a Bernoulli distribution with success probability λ_{j2} and corresponding log-odds θ_{j2}; and two safety endpoints, Severe Adverse Events and Serious Adverse Events, both Bernoulli distributed with probabilities $\lambda_{j3}, \lambda_{j4}$ and log-odds θ_{ji} for $i = 3, 4$. Zero correlation is assumed between endpoints.

4.5.2 Design

For this study, the control is assumed to have a median overall survival (OS) of 9 months, corresponding to a log hazard-ratio $\theta_{01} = 2.56$, a true objective response rate (ORR) of 30%, corresponding to a log odds-ratio$\theta_{02} = -0.84$, a 30% probability of experiencing a severe adverse

event, and the same probability to experience a serious adverse event, which corresponds to $\theta_{0i} = 0.84$ for $i = 3,4$. An interim analysis for dose selection is planned after a number n of patients overall have been followed up for at least 2 months (we will refer to this as *delay*), which is the response assessment time for the three binary endpoints considered.

The least favorable configuration (LFC) parameters are the ones illustrated in Table 4.1: for the phase III endpoint a hazard ratio of 0.64 is targeted ($\delta_1^* = 0.45$), corresponding to a 4-month improvement in OS, with a maximum non-interesting hazard ratio of 0.78 ($\delta_1^0 = 0.15$), corresponding to a 2-month improvement in OS. For the phase II endpoint, a log-odds ratio $\delta_2^* = 1.24$ is targeted (corresponding to a 30% increase in ORR), with a maximum non-interesting log-odds ratio $\delta_2^0 = 0.44$ (corresponding to a 10% increase in ORR). For the two active doses, safety endpoints are assumed equal to the control so that $\delta_i^* = \delta_i^0 = 0$ for all $i = 3,4$.

For the multi-criteria decision analysis (MCDA), a vector of weights $\boldsymbol{\omega} = (0.1, 0.4, 0.25, 0.25)$ is used. A linear partial value function is employed, with maximum and minimum values θ_i^U and θ_i^L chosen respectively as the highest and lowest values which is plausible to observe (values are specified in Table 4.1).A non-binding futility stopping rule may be implemented by requiring that the posterior probability of the MCDA score of the selected arm being superior to that of the control exceeds 50%.

The required number of events at the final analysis (sample size), denoted by d, and the number of evaluable patients at the interim analysis, denoted by n, are jointly determined through a numerical search for the pair (d,n) that satisfies Equation 4.17. The present analysis is performed under the *least favorable configuration* (LFC), aiming to achieve a statistical power of 80% with an expected information fraction of $t = 0.2$ for the primary endpoint (note that t depends on the phase II endpoint; therefore it is a random variable rather than a fixed quantity). For the purpose of first evaluation, zero correlation between patient responses is assumed, in that $\rho_{hk} = 0\ \forall h \neq k$ are used in Equation 4.17.

In the sample size calculation procedure, the variances of the log-odds for the binary endpoints are specified as $\sigma_{ji}^{2,IA} = \frac{1}{\lambda_{ji} n} + \frac{1}{(1-\lambda_{ji})n}$, while the variances of the log-hazards for the survival endpoint are given by $\sigma_{j1}^{2,IA} = \frac{1}{d_j^{IA}}$, where d_j^{IA}, for $j = 0,1,2$, denote the expected number of events at interim. These are computed to satisfy $d_0^{IA} + d_2^{IA} = td$ under the LFC and the assumed recruitment model. Assuming uniform recruitment and equal allocation, the expected number of events d_j^{IA} admits the analytical expression

$$d_j^{IA} = \frac{N_j^{IA}}{\tau}\left(1 - \int_0^{\tau} e^{\lambda_{j1}(\tau - r)}\, dr\right) \qquad j = 0,1,2$$

where N_j^{IA} denotes the number of patients recruited up to the interim analysis, and τ is the calendar time at which the interim analysis occurs. The latter is determined by

$$\tau = \frac{n}{3 \times \text{accrual rate}} + \text{delay}.$$

Applying this procedure yields an interim analysis after $n = 183$ patients are evaluable for the binary endpoints, and a final analysis triggered once $d = 179$ overall survival events have been observed across the selected dose and the control arm. An illustration of the numerical search of (n, d) is illustrated in Table C.1 of the Supplementary material.

The critical value η for testing the primary endpoint at the final analysis is determined numerically by solving Equation 4.16, i.e., by identifying the minimal threshold q such that $F_{Z_{S1}^{FA}}(q) > 1-\alpha$ where $F_{Z_{S1}^{FA}}(q)$ is specified in Equation 4.15 and α is fixed at 2.5%. In Equation 4.15, the observed information fraction for the primary endpoint is employed to determine t. The computation is performed under the "worst-case" configuration, wherein $\theta_{01} = \cdots = \theta_{J1}$ for the primary endpoint, and the log-odds ratios for the binary endpoints are set to $\theta_{ji} = 0$ for all $j = 0, 1, 2$ and $i = 2, 3, 4$. This corresponds to probability parameters $\lambda_{ji} = 0.5$ for all j and i.

4.5.3 Scenarios

Using the study described in Section 4.5.1, this section evaluates the operating OCs of the proposed methodology and compares them against some relevant approaches present in literature. Three distinct analyses are conducted: two *Type I Error Analyses*, designed to assess the OCs under a broad range of null configurations (i.e. configurations where $\theta_{01} = \cdots = \theta_{J1}$), and a *Power Analyses*, aimed at evaluating the OCs under selected alternative configurations.

In the conducted simulations, the parameters corresponding to the four endpoints under consideration were varied across three distinct levels. For the two efficacy endpoints, these levels comprised low effect, maximum uninteresting effect, and target effect. For the two safety endpoints, the levels comprised low toxicity, maximum acceptable toxicity, and high toxicity. A schematic representation of the tested levels, along with their qualitative interpretations, is presented in Figure 4.2.

The three analysis performed are the following:

1. **Analysis A: null scenarios with $\boldsymbol{\gamma_1 = \gamma_2}$.** The objective of this analysis is to evaluate scenarios in which the two experimental doses perform equally across all four endpoints considered and, consequently, exhibit the same benefit-risk profile (i.e., $\gamma_1 = \gamma_2$). For this purpose, all parameters associated with dose 1 and dose 2 are varied in parallel across scenarios, except for the overall survival (OS) hazard, which is fixed to that of the control group. Specifically, the ORR parameters are set as $\lambda_{12} = \lambda_{22} \in \{0.3, 0.4, 0.6\}$, the severe adverse event rates are set to $\lambda_{13} = \lambda_{23} \in \{0.3, 0.5, 0.7\}$, and the serious adverse event rates are set to $\lambda_{14} = \lambda_{24} \in \{0.3, 0.5, 0.7\}$, yielding a total of $3 \times 3 \times 3 = 27$ simulated scenarios. Additionally, ten further scenarios are examined by fixing the parameter of dose 1 equal to the control ones, while varying the properties of dose 2 such that in all cases the condition $\gamma_2 = \gamma_1$ is satisfied. An overview of the complete set of scenarios considered is provided in Table C.3 of the Supplementary Material.

2. **Analysis B: null scenarios with $\gamma_1 \neq \gamma_2$.** The objective of this analysis is to evaluate null scenarios in which the two experimental doses are ineffective on the primary endpoint but differ with respect to the remaining three endpoints, thereby leading to distinct benefit-risk profiles (i.e., $\gamma_1 \neq \gamma_2$). Parameters for dose 1 are fixed to those of the control arm, with overall survival (OS) hazard rate $\lambda_{11} = \lambda_{01} = 0.1$, response rate $\lambda_{12} = \lambda_{02} = 0.3$, probability of serious adverse events $\lambda_{13} = \lambda_{03} = 0.3$, and probability of severe adverse events $\lambda_{14} = \lambda_{04} = 0.3$. For dose 2, all parameters except the OS hazard (kept equal to that of the control and dose 1) are varied to generate alternative benefit-risk profiles, specifically $\lambda_{22} \in \{0.3, 0.4, 0.6\}$, $\lambda_{23} \in \{0.3, 0.5, 0.7\}$, and $\lambda_{24} \in \{0.3, 0.5, 0.7\}$. This configuration yields a total of $3 \times 3 \times 3 = 27$ simulated scenarios. A complete overview of the considered scenarios is reported in Table C.2 of the Supplementary Material.

3. **Analysis C: alternative scenarios** The objective of this analysis is to evaluate scenarios in which both experimental doses exhibit efficacy on the primary endpoint, potentially at different levels. Parameters for dose 1 are fixed as $\lambda_{11} = 0.08$, $\lambda_{12} = 0.4$, $\lambda_{13} = 0.3$, and $\lambda_{14} = 0.3$, corresponding to a treatment with a maximal non-interesting effect on the primary endpoint, a moderate Objective Response Rate, and a favorable safety profile. For dose 2, all parameters are varied to explore a range of alternative configurations, specifically $\lambda_{21} \in \{0.08, 0.07\}$, $\lambda_{22} \in \{0.3, 0.6, 0.7\}$, $\lambda_{23} \in \{0.3, 0.5, 0.7\}$, and $\lambda_{24} \in \{0.3, 0.5, 0.7\}$. This setup yields a total of $2 \times 3 \times 3 \times 3 = 54$ distinct simulated scenarios. Independence across endpoints is assumed. A complete overview of the considered scenarios is provided in Table C.4 of the Supplementary Material.

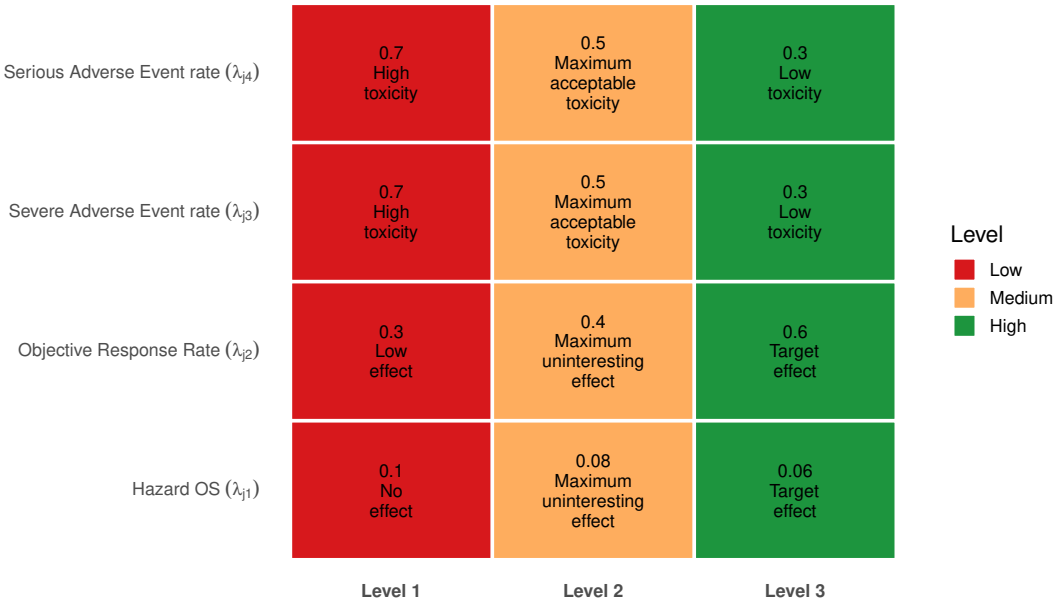


Fig. 4.2 Parameters varied in the simulation study , with corresponding quantitative levels and qualitative interpretation.

For all analyses considered, the sample size found as detailed in Section 4.5.2 was used and for each scenario 20,000 trial replicates were generated under the assumption of zero correlation between endpoints, i.e. $\rho_{hk} = 0 \ \forall h \in \{1, \ldots, 4\}, \forall k \neq h$. Furthermore, in all main analyses, no futility stopping was permitted at the interim analysis; nonetheless, Analysis B was also repeated under a design allowing for non-binding early futility stopping (see Supplementary Material). All data generation and analyses were conducted using the statistical software R [73].

4.5.4 Operating characteristics evaluated (OCs)

The operating characteristics (OCs) evaluated in this study include:

- **False Positive Rejection Rate** (*T1E*): Defined as the joint probability of selecting, at interim analysis, a dose exhibiting no effect on the primary endpoint and subsequently rejecting the null hypothesis for that dose.
- **Probability of Selecting Dose 2** (*%Sel2*): This metric quantifies the likelihood of selecting dose 2 at interim. It serves as an informative measure to assess the capacity of the different considered approaches to make appropriate dose selections during interim analyses. Note that the probability of selecting dose 1 can be obtained by subtracting from 1 the sum of the probability of selecting dose 2 and the probability of early stopping (when such stopping is permitted).
- **Statistical Power**: Defined as the joint probability of selecting the dose with the target treatment effect on the primary endpoint ($\lambda_{j1} = 0.07$, corresponding to a hazard ratio of 0.64) and rejecting the null hypothesis for that dose. This metric provides insight into the efficacy of the methods under consideration in correctly identifying truly effective doses (without accounting for safety).
- **Uninteresting Success Rate** (*Unint. Succ*): Defined as the joint probability of selecting a dose with a positive but *not clinically relevant* treatment effect on the primary endpoint ($0.08 < \lambda_{j1} < 0.1$, corresponding to a hazard ratio between 0.64 and 1) and rejecting the null hypothesis for that dose. This measure is particularly valuable in scenarios where the most effective dose for the primary endpoint is associated with safety concerns, thereby warranting the evaluation of less effective, but safer, dose options.
- **Average Toxicity Levels** (*Tox1* and *Tox2*): These metrics correspond to the averages of the toxicity parameters for dose 1 and dose 2, weighted by the respective probabilities of selecting each dose. Here, *Tox1* refers to average toxicity level related to severe adverse events, and *Tox2* refers to average toxicity level related to serious adverse events. These quantities are critical for evaluating the ability of different approaches to minimize the selection of doses with unfavorable safety profiles at interim analysis. Note that the average

is employed here for convenience; however, it does not represent an actual observable rate but rather an estimate of the expected toxicity level for the selected dose prior to the study.

4.5.5 Competing designs

Stallard & Todd (2003)

In the design proposed by Stallard and Todd [10], the best performing dose on the primary endpoint *only* is selected at the time of the interim analysis, while all the others are discontinued. Considering that the statistics $\hat{\theta}_{j1}$ and $\hat{\theta}_{k1}$ are independent $\forall j \neq k$, then the overall probability to select a dose with interim statistics $\hat{\theta}_{Si} = x$ can be expressed as

$$f_{Z_{S1}^{IA}}(z) = \sum_{j=1}^{J} \frac{1}{\sigma_{j1}^{IA}\sigma_{01}^{IA}} \int_{-\infty}^{+\infty} \prod_{k\neq j} \left[\Phi\left(\frac{x-\theta_{k1}}{\sigma_{k1}^{IA}}\right)\right] \phi\left(\frac{\frac{z\sigma^{\dagger}}{t}+x-\theta_{j1}}{\sigma_{j1}^{IA}}\right) \phi\left(\frac{x-\theta_{01}}{\sigma_{01}^{IA}}\right) dx. \tag{4.18}$$

Given the conditional distribution of the final test statistics conditional on the interim one in Equation 4.14, the cumulative distribution function of the test statistic of the treatment effect on the primary endpoint for the selected dose can be computed using Equation 4.15. The critical value η is found under the null hypothesis by searching over the minimum value of q such that $1 - F_{Z_{S1}^{FA}}(q) < \alpha$, where α is the nominal level for the false positive rate.

It is worth noting that the MCDA-based approach presented in this manuscript can be regarded as a generalization of the method proposed by Stallard and Todd. Specifically, in the MCDA framework the dose selection rule is allowed to incorporate information from multiple endpoints, including those different from the primary efficacy endpoint. The two methods coincide when the weight vector $\boldsymbol{\omega} = (1, 0, 0, \ldots, 0)$ is employed, i.e., when only the primary endpoint is taken into account.

In analyses incorporating a futility stop, the additional constraint $\hat{\theta}_{S1}^{IA} > 0$ is applied at the interim analysis, whereby the trial is discontinued if the control arm demonstrates the most favorable outcome on the primary endpoint.

Jaki & Hampson (2016)

While in the designs considered so far the efficacy of the selected treatment on the primary efficacy endpoint was the ultimate goal of the trial, in the design proposed by Jaki and Hampson [99] the superiority of a dose in *all* endpoints is tested. The null hypothesis considered is accordingly:

$$H^0 = \bigcap_{j=1}^{J} H_j^0 \qquad \text{where} \qquad H_j^0 = \bigcup_{i=1}^{I} \{\delta_{ji} = 0\}$$

Treatment selection at the time of the interim analysis is based on the following optimization function

$$O_j = \sum_{i=1}^{I} w_i Z_{ji}^{IA} \qquad j = 1, \ldots, J \tag{4.19}$$

where w_i are weights indicating the relative importance of each endpoint i in the optimization function. A constraint $\sum_{i=1}^{I} w_i^2 = 1$ is imposed on the weights so that O_j has unit variance. Once the treatment $S = \text{argmax}_j O_j$ is selected, I independent tests are performed at the final analysis and the null hypothesis is rejected if the single null hypotheses H_{Si}^0 are rejected for all i, namely if

$$Z_{Si}^{FA} > \eta_i \qquad \forall i = 1, \ldots I \tag{4.20}$$

Each of the critical values η_i is numerically computed so that the family-wise type I error rate (FWER) is exactly controlled at the nominal level α under the I partial null configurations where $\delta_{jk} = +\infty \; \forall j = 1, \ldots J \;\; \forall k \neq i$, while $\delta_{ji} = 0 \; \forall j = 1, \ldots J$ and $k = i$. Family-wise error rate is defined in this context as the probability to reject H^0 when the selected treatment is not superior to control on at least one endpoint, namely $\exists i^* : \delta_{Si^*} = 0$. While it has been shown that controlling the FWER under these I parameter configurations is a sufficient condition to ensure control under any null configuration, a lower type I error rate (strictly below the nominal level) is expected under the global null hypothesis, owing to the multiplicity of tests required to reject the null hypothesis.

In the present application, the Cox regression model is employed to compute the Z-statistic for the primary endpoint, denoted by $Z_{j1}^{\star}$. For the three binary endpoints considered (ORR, SevAE and SerAE), the corresponding Z-statistics, namely $Z_{j2}^{\star}$, $Z_{j3}^{\star}$, and $Z_{j4}^{\star}$, are derived using the standard maximum likelihood estimators (MLEs) and their associated variances. To enable a fair comparison between this approach and the proposed methodology, the weight vector $\boldsymbol{w} = \sqrt{\boldsymbol{\omega}} = (0.31, 0.63, 0.5, 0.5)$ is used, thereby ensuring that the four endpoints considered in the study are weighted consistently across both approaches. For the final analysis, one-sided hypothesis tests are performed for the two efficacy endpoints, whereas equivalence tests are applied to the two safety endpoints. The equivalence margins for the safety endpoints are defined as 0.2 on the probability scale. In analyses incorporating a futility stop, an additional constraint $O_S > 0$ is imposed at the interim analysis, such that the trial is terminated in advance if the control arm exhibits the most favorable benefit-risk profile.

Friede et al. (2011)

In the design proposed by Friede et al. [95], the null hypothesis $H^0 : \cap_{j=1}^{J} \delta_{j1}$ is considered. Treatment selection at the time of the interim analysis is based solely on the analysis of data on an early outcome (i.e. phase II endpoint, in our setting), so that the treatment with larger Z statistic is selected, namely $S = \text{argmax}\, Z_{j2}^{IA}$. The final testing on the primary endpoint at the final analysis is based on the p-value combination [88] within the closed testing procedure [113] and is summarized as follows:

- At each stage $\star = \{1, 2\}$, all the intersection hypothesis $H_Q : \bigcap_{j \in Q} H_j$ with $Q \subseteq \{1, \ldots, J\}$ are tested at the pre-specified significance level $1 - \alpha$ and the single stage p-values $p_Q^\star$ are computed using *only* data accrued in the $\star$-th stage of the trial. For the subsets Q with cardinality > 1 the maximum between the Z-statistics $\max_j Z_{j1}^\star$ may be used to run a Dunnett-type test, accounting for the correlation due to the common control. Since data for the non-selected treatments may not be available after the first stage of the trial, the Z-statistics corresponding to the stage 2 data are conventionally set to $-\infty$ for all the non-selected treatments. This results in a more conservative tests.

- For each intersection hypothesis considered, the stage-wise p-values are combined using the *weighted inverse normal method* [114] as follows:

$$C\left(p_Q^1, p_Q^2\right) = 1 - \Phi\left[\zeta_1 \Phi^{-1}\left(1 - p_Q^1\right) + \zeta_2 \Phi^{-1}\left(1 - p_Q^2\right)\right], \tag{4.21}$$

  where $\zeta_1^2 + \zeta_2^2 = 1$ and ζ_1 and ζ_2 are typically set proportionally to the square root of the information fraction available at the stage 1 of the trial.

  A test for the hypothesis H_Q is then constructed comparing $C\left(p_Q^1, p_Q^2\right)$ to the nominal level α.

- The closed-tesing procedure is used for the final hypothesis testing, and accordingly the individual null hypothesis for the selected treatment H_{j^*} is rejected if *all* the intersection hypotheses including the selected treatment j^* are rejected at the final analysis, namely

$$\text{Reject } H^0 \iff C\left(p_Q^1, p_Q^2\right) < \alpha \quad \forall Q \ni j^* \tag{4.22}$$

Notice that if p_Q^1 and $p_Q^2 | p_Q^1$ are stochastically no smaller than a uniform $U(0, 1)$ (*p-clud* condition [115]), then a control of the type I error rate is achieved. If p_Q^1 and p_Q^2 are asymptotically uniform under the null hypothesis, then an asymptotic control of the type I error is obtained (asymptotic *p-clud* condition [116]).

In the current application, the Cox regression model is used to compute p-values for the individual null hypotheses, while a Dunnett test is employed to obtain p-values corresponding to the intersection hypotheses. Moreover, since employing a nominal significance level of $\alpha = 0.025$ in

the closed-testing procedure defined in Equation 4.22 leads to an overly conservative test under the assumption of zero correlations [95], an adjustment to $\alpha = 0.053$ is applied herein. In order to ensure independence of the stage-wise p-values, stage 1 p-values (resp. stage 2 p-values) are computed using data from patients enrolled prior to (resp. after) the interim analysis. Consequently, the weights ζ_1 and ζ_2 are defined as the square root of the proportions of patients recruited before and after the interim analysis, respectively. These adjustment ensures that the Type I error rate is controlled exactly at 2.5%, thereby facilitating a fair comparison. In analyses incorporating a futility stop, the additional constraint $\hat{\theta}_{S2}^{IA} > \hat{\theta}_{02}^{IA}$ is applied at the interim analysis, so that the trial is stopped in advance if the control arm exhibits most favorable outcome on the phase II efficacy endpoint.

4.5.6 Results

Results for 12 representative scenarios from Analysis A and Analysis B are reported in Table 4.3, while a summary illustration of the complete set of results is provided in Figure C.3 of the Supplementary Material. The scenarios where selected to span a wide range for benefit-risk scores and different regimens of safety and efficacy.

Across all scenarios of analysis A, for the proposed MCDA-based approach, the probability of selecting dose 1 or dose 2 is consistently close to 50% across all scenarios simulated, reflecting the fact that both doses exhibit identical benefit-risk profiles ($\gamma_1 = \gamma_2$). With respect to type I error, values around the nominal level of 2.5% are observed in all scenarios, with small variations due to numerical error. As both doses have the same probability to be selected, the average toxicity level of the selected dose for a given safety endpoint coincides trivially with average of the dose-specific toxicity rates. It should be noted that this metric lacks direct interpretation in scenarios A1–A3, since the two active doses exhibit identical toxicity rates. Conversely, scenarios A4-A6 demonstrate that identical values of the γ-score may correspond to doses described by distinct parameter configurations. Consequently, doses with relatively large toxicity parameters can be selected with 50% probability (e.g., scenario A6) when elevated toxicity is offset by a high ORR. We advocate, however, that this occurrence is unlikely in practice, as one would generally not expect improvements in OS in the absence of corresponding (or at least non-negligible) gains in ORR.

Under the approach of Stallard & Todd, the selection probability of dose 1 and dose 2 is again approximately 50% across all scenarios, owing to their equal efficacy on the primary endpoint. Type I error is controlled on average close to the nominal level, with no notable variation across scenarios. As before, average toxicity levels for the selected treatment are therefore identical to the average of the dose-specific toxicity rates.

In the approach of Jaki & Hampson, the selection probability for each dose remains around 50% when the experimental arms are identical across all endpoints (and thus yield the same expected O_j), however, it exhibits variations when different configurations of the parameters

Table 4.3 Operating characteristics corresponding to Analyses A and B for 12 scenarios. The parameters used for the control arm, shared across scenarios, are $\lambda_{01} = 0.1$, $\lambda_{02} = 0.3$, $\lambda_{03} = 0.3$, $\lambda_{04} = 0.3$, while parameters for the active doses are specified in the Table. For each scenario, 20,000 datasets were generated.

Scenario	Parameters ($j = 2$)					Approaches	Metrics						
	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	γ_2		%Stop	%Sel2	T1E	Pow	Unint. Succ.	Tox1	Tox2
Equal benefit-risk scores and equal characteristics for the two active doses ($\gamma_1 = \gamma_2$, $\lambda_{1i} = \lambda_{2i}$ $\forall i = 2,3,4$)													
A1	0.1	0.6	0.3	0.3	2.24	MCDA-based	-	49.8	0.024	-	-	0.30	0.30
						Stallard & Todd	-	50.1	0.025	-	-	0.30	0.30
						Jaki & Hampson	-	50.0	0.022	-	-	0.30	0.30
						Friede	-	49.6	0.024	-	-	0.30	0.30
A2	0.1	0.3	0.5	0.3	-0.94	MCDA-based	-	49.4	0.025	-	-	0.50	0.30
						Stallard & Todd	-	49.7	0.026	-	-	0.50	0.30
						Jaki & Hampson	-	49.3	0.000	-	-	0.50	0.30
						Friede	-	49.8	0.025	-	-	0.50	0.30
A3	0.1	0.3	0.5	0.5	1.30	MCDA-based	-	50.2	0.026	-	-	0.50	0.30
						Stallard & Todd	-	50.3	0.027	-	-	0.50	0.30
						Jaki & Hampson	-	50.2	0.000	-	-	0.50	0.30
						Friede	-	49.1	0.025	-	-	0.50	0.30
Different characteristics for the two active doses, leading to the same benefit-risk scores ($\gamma_1 = \gamma_2 = 0.30$, $\lambda_{11} = 0.1$, $\lambda_{12} = 0.3$, $\lambda_{13} = 0.3$, $\lambda_{14} = 0.3$)													
A4	0.1	0.35	0.36	0.32	0.000	MCDA-based	-	49.6	0.024	-	-	0.33	0.31
						Stallard & Todd	-	49.7	0.024	-	-	0.33	0.31
						Jaki & Hampson	-	45.6	0.001	-	-	0.33	0.31
						Friede	-	72.0	0.023	-	-	0.34	0.31
A5	0.1	0.49	0.48	0.42	0.000	MCDA-based	-	49.4	0.024	-	-	0.47	0.43
						Stallard & Todd	-	50.1	0.025	-	-	0.48	0.44
						Jaki & Hampson	-	20.7	0.000	-	-	0.37	0.36
						Friede	-	99.9	0.023	-	-	0.65	0.58
A6†	0.1	0.69	0.65	0.58	0.000	MCDA-based	-	49.8	0.025	-	-	0.39	0.36
						Stallard & Todd	-	49.8	0.025	-	-	0.39	0.36
						Jaki & Hampson	-	36.0	0.000	-	-	0.36	0.34
						Friede	-	98.6	0.025	-	-	0.48	0.42
Different benefit–risk scores in the two doses ($\gamma_1 \neq \gamma_2$, $\lambda_{11} = 0.1$, $\lambda_{12} = 0.3$, $\lambda_{13} = 0.3$, $\lambda_{14} = 0.3$)													
B1	0.1	0.3	0.3	0.3	0.000	MCDA-based	-	49.6	0.024	-	-	0.30	0.30
						Stallard & Todd	-	50.6	0.027	-	-	0.30	0.30
						Jaki & Hampson	-	49.6	0.000	-	-	0.30	0.30
						Friede	-	49.5	0.025	-	-	0.30	0.30
B2†	0.1	0.3	0.5	0.7	-2.85	MCDA-based	-	0.2	0.021	-	-	0.30	0.30
						Stallard & Todd	-	50.7	0.025	-	-	0.40	0.50
						Jaki & Hampson	-	0.000	0.000	-	-	0.30	0.30
						Friede	-	49.2	0.025	-	-	0.40	0.50
B3	0.1	0.4	0.5	0.3	-0.16	MCDA-based	-	43.6	0.024	-	-	0.39	0.30
						Stallard & Todd	-	50.6	0.024	-	-	0.40	0.30
						Jaki & Hampson	-	34.80	0.000	-	-	0.37	0.30
						Friede	-	87.5	0.024	-	-	0.48	0.30
B4†	0.1	0.4	0.7	0.5	-2.10	MCDA-based	-	1.7	0.022	-	-	0.31	0.30
						Stallard & Todd	-	49.9	0.026	-	-	0.50	0.40
						Jaki & Hampson	-	0.5	0.000	-	-	0.30	0.30
						Friede	-	87.5	0.025	-	-	0.65	0.48
B5	0.1	0.6	0.3	0.5	1.32	MCDA-based	-	90.9	0.023	-	-	0.30	0.48
						Stallard & Todd	-	50.0	0.026	-	-	0.30	0.40
						Jaki & Hampson	-	83.4	0.000	-	-	0.30	0.47
						Friede	-	99.9	0.025	-	-	0.30	0.50
B6†	0.1	0.6	0.7	0.7	-1.57	MCDA-based	-	6.0	0.023	-	-	0.32	0.32
						Stallard & Todd	-	50.2	0.024	-	-	0.50	0.50
						Jaki & Hampson	-	1.1	0.000	-	-	0.30	0.30
						Friede	-	99.9	0.025	-	-	0.70	0.70

† Scenario exhibiting unacceptably high toxicity rates

are considered across arms, due to the different criterion used for dose selection. Type I error rates fall strictly below the nominal level due to multiple hypothesis testing at the final analysis and the specific scenarios used to compute the critical values η_i. Particularly, it is higher in scenarios where both doses perform well across all endpoints (e.g. scenario A1), but approaches zero when the selected experimental dose is either ineffective on the phase II efficacy endpoints (e.g. scenarios A2 and A3), or display poor safety in one of the safety endpoint (e.g. scenario A4). The average toxicity of the selected dose is defined as the weighted mean of the toxicity rates associated with the two active doses, where the weights correspond to their respective probabilities of being selected. As before, this metric is not informative in scenarios A1–A3; however, it illustrates that doses with differing characteristics may yield identical values of the γ-score. This, in turn, implies that doses with higher toxicity levels may still be selected with a relatively high probability in unlikely situations where increased toxicity is compensated by pronounced efficacy on the phase II efficacy endpoint (e.g., scenario A6).

Finally, for the approach of Friede et al., the selection probability in dose 1 and dose 2 are approximately balanced at 50% when the two doses share the same ORR (e.g. scenarios A1-A3), while the probability that dose 2 is selected increases when its ORR is greater than dose 1 (e.g. scenarios A4-A6). Type I error is controlled close to the nominal level, with no relevant differences across scenarios. Here, low average toxicity levels of the selected dose are observed when the dose exhibiting the higher objective response rate (ORR) also demonstrates favorable safety properties, whereas high toxicity levels arise when the dose performing best in terms of ORR is associated with safety concerns (e.g., scenario A6). It should be noted, as previously mentioned, that this metric is not meaningful in scenarios A1–A3, where the two active doses share identical characteristics.

Analysis B reveals that when the proposed MCDA-based approach is used, the probability of selecting dose 2 over dose 1 varies with $\gamma_2 - \gamma_1$: it is higher when the benefit-risk score of dose 2 substantially exceeds that of dose 1 (e.g. scenarios B5), approaches 50% when the two doses exhibit comparable benefit-risk profiles (e.g. scenarios B1 and B3), and decreases toward zero when γ_2 is markedly lower than γ_1 (e.g. scenarios B2 and B4). Type I error is strictly controlled below the nominal 2.5% level across all scenarios. This is consistent with the fact that the critical value η is computed under the worst-case scenario, in which all doses share identical parameters θ_{ji} $\forall i = 1,2,3,4$ and $\forall j = 0,1,2$ (or equivalently to λ_{ji} $\forall i = 1,2,3,4$ and $\forall j = 0,1,2$), thereby yielding conservative type I error control under all other scenarios. Notably, dose 1 (which exhibits good safety profile), is generally preferred when dose 2 presents safety concerns (e.g. in scenarios B2, B4 and B6), leading to relatively low average toxicity levels for the selected dose; this occurs because higher toxicity in dose 2 reduces its benefit-risk score γ_2.

Under the approach of Stallard & Todd, the selection probability of dose 1 and dose 2 remains close to 50% across scenarios, reflecting the fact that efficacy on the primary endpoint is identical in the two doses. Type I error is controlled at the nominal level, with no meaningful variation across scenarios. As both doses are equally likely to be selected, the average toxicity levels for

the selected dose in each scenario corresponds approximately to the mean of the toxicity rates of the two doses, which leads to moderately high average toxicity levels for the selected dose in scenarios where dose 2 is associated with safety concerns (e.g. scenarios B2, B4 and B6).

In the approach of Jaki & Hampson, the probability of selecting dose 2 depends on its benefit-risk score relative to the control, $\gamma_2 - \gamma_0$, with higher (resp. lower) values leading to higher (resp. lower) selection probabilities. This arises from the similarity between the optimization function O_j (used by Jaki and Hamoson in the dose selection step) and the benefit-risk score γ_j, as both incorporate all efficacy and safety endpoints. As in Analysis A, the Type I error rates remain strictly below the nominal level, owing to the adjustment for multiple hypothesis testing at the final analysis and the specific scenarios employed in determining the critical values η_i. In particular, since all scenarios considered exhibit either low efficacy or poor safety, the Type I error rate converges to zero across all configurations.

Under the approach of Friede et al., the probability of selecting dose 1 versus dose 2 depends on the difference in the phase II endpoint parameter between the dose with varying characteristics ($j = 2$) and the fixed dose ($j = 1$). When $\lambda_{12} = \lambda_{22} = 0.3$ (e.g. scenarios B1 and B2), the two doses are selected with approximately equal probability (around 50%), whereas the selection probability for dose 2 approaches respectively 88% and 100% when ORR related to dose 2 is larger (e.g. $\lambda_{22} = 0.4$, as in scenarios B3 and B4, or $\lambda_{22} = 0.6$, as in scenarios B5 and B6). Type I error is controlled close to the nominal level, with no meaningful variation across different values of the ORR, in line with the findings of Friede et al. [95] under the assumption of zero correlation. With respect to toxicity, when the dose achieving the highest ORR is also safe (e.g. scenarios B5), the average toxicity of the selected treatment remains low; however, when the dose with the highest ORR is associated with safety concerns (e.g. scenarios B4 and B6), the probability to select a dose with high toxicity rate increases markedly, and accordingly also the average toxicity levels for the selected dose.

An analogous version of the presented table (scenarios B1-B6), based on the probabilistic MCDA approach for interim dose selection within the MCDA-based framework, is provided in Table C.7 of the Supplementary Material. This comparison demonstrates practical equivalence between the two methods, with selection probabilities differing by at most 1.5% and practically no differences in type I error rates. In addition, a further version of the table allowing for a non-binding futility stop is reported in Table C.8 of the Supplementary Material, which shows a general decrease in type I error rates across all scenarios for all the approaches considered.

Results for 12 representative scenarios from Analysis C are reported in Table 4.4, while a summary of the complete set of results is provided in Figure C.4 of the Supplementary Material. The selected scenarios were designed to explore varying levels of efficacy and safety across the four considered endpoints.

Similarly to what was observed in Analysis B, the probability of selecting dose 2 over dose 1 in the proposed MCDA-based approach varies with $\gamma_2 - \gamma_1$: it increases when the benefit-risk

Table 4.4 Operating characteristics corresponding to Analysis C for 12 scenarios. The parameters used for the control arm are $\lambda_{01} = 0.1$, $\lambda_{02} = 0.3$, $\lambda_{03} = 0.3$, $\lambda_{04} = 0.3$, while for dose 1 they are $\lambda_{11} = 0.1$, $\lambda_{12} = 0.4$, $\lambda_{13} = 0.3$, $\lambda_{14} = 0.3$ (leading to $\gamma_1 = 1.08$). For each scenario, 20,000 datasets were generated.

Scenario	Parameters ($j = 2$)					Approaches	Metrics						
	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	γ_2		%Stop	%Sel2	T1E	Pow	Unint. Succ.	Tox1	Tox2
C1	0.06	0.6	0.3	0.3	2.74	MCDA-based	-	95.7	-	0.81	0.02	0.30	0.30
						Stallard & Todd	-	71.3	-	0.61	0.13	0.30	0.30
						Jaki & Hampson	-	94.0	-	0.65	0.01	0.30	0.30
						Friede	-	98.7	-	0.84	0.01	0.30	0.30
C2†	0.06	0.6	0.7	0.7	-1.03	MCDA-based	-	1.7	-	0.02	0.36	0.31	0.31
						Stallard & Todd	-	70.9	-	0.60	0.13	0.58	0.58
						Jaki & Hampson	-	0.2	-	0.00	0.11	0.30	0.30
						Friede	-	98.6	-	0.83	0.01	0.69	0.69
C3†	0.06	0.6	0.3	0.7	-0.86	MCDA-based	-	41.1	-	0.36	0.23	0.30	0.46
						Stallard & Todd	-	70.7	-	0.60	0.13	0.30	0.58
						Jaki & Hampson	-	25.7	-	0.00	0.10	0.30	0.40
						Friede	-	98.7	-	0.84	0.01	0.30	0.69
C4	0.06	0.4	0.3	0.5	0.37	MCDA-based	-	23.5	-	0.20	0.29	0.30	0.35
						Stallard & Todd	-	71.5	-	0.61	0.12	0.30	0.44
						Jaki & Hampson	-	16.9	-	0.00	0.11	0.30	0.33
						Friede	-	49.4	-	0.42	0.23	0.30	0.40
C5†	0.06	0.4	0.7	0.5	-1.53	MCDA-based	-	0.4	-	0.000	0.36	0.30	0.30
						Stallard & Todd	-	71.0	-	0.60	0.13	0.58	0.44
						Jaki & Hampson	-	0.1	-	0.00	0.12	0.30	0.30
						Friede	-	49.8	-	0.44	0.22	0.50	0.40
C6	0.06	0.3	0.3	0.5	0.50	MCDA-based	-	28.2	-	0.24	0.27	0.30	0.30
						Stallard & Todd	-	71.6	-	0.61	0.12	0.30	0.30
						Jaki & Hampson	-	28.6	-	0.01	0.09	0.30	0.30
						Friede	-	11.8	-	0.10	0.39	0.30	0.30
C7	0.08	0.6	0.3	0.3	2.52	MCDA-based	-	93.2	-	-	0.38	0.30	0.30
						Stallard & Todd	-	49.7	-	-	0.41	0.30	0.30
						Jaki & Hampson	-	91.4	-	-	0.30	0.30	0.30
						Friede	-	98.5	-	-	0.42	0.30	0.30
C8†	0.08	0.6	0.7	0.7	-1.27	MCDA-based	-	0.9	-	-	0.36	0.30	0.30
						Stallard & Todd	-	49.9	-	-	0.40	0.50	0.50
						Jaki & Hampson	-	0.1	-	-	0.12	0.30	0.30
						Friede	-	98.6	-	-	0.41	0.69	0.69
C9†	0.08	0.6	0.3	0.7	0.61	MCDA-based	-	32.1	-	-	0.38	0.30	0.43
						Stallard & Todd	-	50.1	-	-	0.40	0.30	0.50
						Jaki & Hampson	-	19.9	-	-	0.10	0.30	0.38
						Friede	-	98.7	-	-	0.41	0.30	0.69
C10	0.08	0.4	0.3	0.5	0.13	MCDA-based	-	17.1	-	-	0.38	0.30	0.33
						Stallard & Todd	-	49.9	-	-	0.40	0.30	0.40
						Jaki & Hampson	-	13.0	-	-	0.11	0.30	0.30
						Friede	-	49.6	-	-	0.40	0.30	0.40
C11†	0.08	0.4	0.7	0.5	-1.79	MCDA-based	-	0.1	-	-	0.36	0.30	0.30
						Stallard & Todd	-	50.3	-	-	0.40	0.50	0.40
						Jaki & Hampson	-	0.000	-	-	0.11	0.30	0.30
						Friede	-	49.8	-	-	0.40	0.50	0.40
C12	0.08	0.3	0.3	0.3	0.29	MCDA-based	-	21.2	-	-	0.37	0.30	0.30
						Stallard & Todd	-	49.6	-	-	0.40	0.30	0.30
						Jaki & Hampson	-	23.0	-	-	0.10	0.30	0.30
						Friede	-	12.0	-	-	0.38	0.30	0.30

† Scenario exhibiting unacceptably high toxicity rates

score of dose 2 substantially exceeds that of dose 1, approaches 50% when the two doses exhibit comparable benefit-risk profiles, and decreases toward zero when γ_2 is substantially lower than γ_1 (see Figure C.4). In scenarios where dose 2 is effective on the primary endpoint at the target level $\lambda_{21} = 0.06$ (scenarios C1-C6), power is closely related to the probability of selecting dose 2 (see Figure C.4). Notably, power approaches 80% under scenario C1 (the least favorable

configuration, LFC) thus justifying the sample size calculation, while progressively decreases as the probability to select dose 2 diminishes. Conversely, the uninteresting success rate exhibits the opposite trend: it is lower when the probability to select dose 1 is low, consistently with the fact that dose 1 represents in our setting the dose with *maximum non interesting* effect. In scenarios where neither of the two active doses attains the target treatment effect on the primary endpoint (scenarios C7-C12), and thus power cannot be assessed, the observed success rate is approximately 37%, suggesting that the proposed approach retains some capacity to detect smaller, albeit non-target, treatment effects. Notably, dose 1 is generally preferred when dose 2 presents safety concerns (e.g., scenario C8), resulting in relatively low average toxicity levels for the selected dose and thereby justifying the reduced power observed in such scenarios. This is because higher toxicity in dose 2 is generally associated with lower values of γ_2, which in turn reduces the probability of selecting that dose. As a drawback, in specific scenarios where dose 2 exhibits a poor benefit–risk profile despite being relatively safe (e.g., scenario C4), for instance due to low efficacy on the phase II endpoint, the resulting power can be substantially reduced. Overall, the MCDA-based approach provides relatively high power when the dose with the target treatment effect on the primary endpoint also exhibits an acceptable safety profile, while it tends to favor less effective doses whenever the best-performing dose displays safety issues. This reduction in power should be considered beneficial, as it generally reflects greater protection against the selection of unsafe doses.

Under the approach of Stallard & Todd, the selection probability of dose 1 versus dose 2 depends solely on the difference in treatment effects between the two doses on the primary endpoint. In particular, it remains close to 71% when $\lambda_{11} = 0.08$ and $\lambda_{21} = 0.06$ (scenarios C1-C6), while it decreases to approximately 50% when $\lambda_{11} = \lambda_{21} = 0.08$ (scenarios C7-C12). As a consequence, power and the uninteresting success rate remain nearly constant in scenarios C1–C6, at about 60% and 13%, respectively. In scenarios C7-C12, where neither active dose achieves the target treatment effect on the primary endpoint, the uninteresting success rate is consistently around 39%. Regarding safety, in scenarios C1-C6 relatively low average toxicity levels are observed when dose 2 (i.e. the dose achieving the target treatment effect) also has a favorable safety profile (e.g. scenarios C1, C4 and C6). Conversely, high average toxicity levels for the selected dose are observed when the best-performing dose on the primary endpoint is associated with increased risks in the safety endpoints (e.g. in scenarios C2 and C5). In scenarios C6-C12, sue to the equal probability to select dose 1 and dose 2, the average toxicity levels for the selected dose are the average of the dose-specific toxicity rates.

Similarly to what was observed in Analysis B, under the approach of Jaki & Hampson the probability of selecting dose 2 depends on its benefit-risk score relative to dose 2, namely $\gamma_2 - \gamma_1$, with higher (resp. lower) values leading to increased (resp. decreased) selection probability. The resulting selection probabilities are broadly similar to those obtained with the MCDA-based approach. In contrast to the latter, however, power in scenarios C1–C6 is markedly lower, reaching its maximum at approximately 65% in configurations where dose 2 is superior (or non inferior, in case of safety endpoints) to control across all endpoints (e.g. scenario C1), but

dropping close to zero in all other scenarios. In scenario C1, the uninteresting success rate is naturally close to zero due to the low probability of selecting dose 1. Notably, even in scenarios where power approaches zero (scenarios C2–C6) and the probability of selecting dose 2 is negligible (hence dose 1 is selected almost exclusively), the uninteresting success rate remains relatively low, thereby highlighting the conservative nature of the approach. In scenarios where neither active dose achieves the target treatment effect on the primary endpoint, the uninteresting success rate is around 29% in scenario C7, where dose 2 has a favorable safety profile, a large ORR, and a moderate treatment effect on the primary endpoint. By contrast, the uninteresting success rate drops to values between 7% and 11% in all other scenarios (C8–C12). As in the MCDA-based approach, dose 2 (the dose potentially associated with safety concerns in our simulations) is selected with non-negligible probability only when it demonstrates a favorable overall benefit-risk profile (e.g. scenarios C1). Conversely, dose 1 (which does not present safety issues) is preferred whenever $\gamma_2 < \gamma_1$. This yields relatively low average toxicity levels for the selected dose across most of the scenarios considered.

Under the approach of Friede et al., the selection probability of dose 1 versus dose 2 depends exclusively on the difference in treatment effects between the two doses on the phase II endpoint. In particular, it remains close to 12% when $\lambda_{22} = 0.3$ (scenarios C6), while it increases to 50% when $\lambda_{22} = \lambda_{12} = 0.4$ (scenarios C4 and C5) and up to 98% when $\lambda_{22} = 0.7$ (scenarios C1, C2 and C3). As a consequence, low power values (around 10%) are observed when the probability of selecting dose 2 is low, whereas substantially higher power values are observed when the probability of selecting dose 2 is high (approximately 84% in scenarios C1-C3). In scenarios where dose 2 exhibits the target treatment effect on the primary endpoint, the uninteresting success rate is inversely related to the probability of selecting this dose, being around 39% when $\lambda_{22} = 0.3$ (scenarios C3–C4) and dropping to zero when $\lambda_{22} \in \{0.4, 0.6\}$. In contrast, in scenarios where neither active dose achieves the target treatment effect on the primary endpoint, the uninteresting success rate varies between 38% and 42%. With respect to toxicity, when the dose achieving the highest ORR also demonstrates a favorable safety profile (e.g. scenario C1), the average toxicity of the selected treatment remains relatively low. Conversely, when the dose with the highest ORR is associated with safety concerns (e.g. scenarios C2 and C6), the average toxicity levels increase markedly.

Overall, the simulation results indicate that the proposed methodology successfully controls the type I error rate at or below the nominal level under all null configurations. As expected, the dose-selection procedure grounded in the MCDA framework exhibits a tendency to prioritize doses characterized by superior benefit-risk profiles. Consequently, this approach generally leads to the discontinuation of doses associated with higher toxicity, yielding lower average toxicity among the selected doses compared with alternative methods in which selection is driven solely by efficacy outcomes (i.e., those of Stallard & Todd and Friede *et al.*), and comparable average toxicity levels relative to the approach proposed by Jaki and Hampson. Unlike the latter, which tends to be overly conservative in terms of power, the MCDA-based approach maintains a high power when the dose corresponding to the target treatment effect on the primary

endpoint also exhibits a good benefit-risk balance. Conversely, power diminishes as the benefit-risk characteristics of the most efficacious dose on the primary endpoint deteriorate. Since the MCDA scores used for interim decision-making represent aggregate measures, it is noteworthy that multiple distinct parameter configurations may lead to identical operating characteristics. Accordingly, *(i)* low power may arise not only from high toxicity rates—which is consistent with the design's intent to avoid selecting excessively toxic doses—but also from scenarios in which poor benefit-risk profiles are driven by inadequate efficacy on non-primary endpoints despite acceptable safety; and *(ii)* doses associated with elevated toxicity may still be selected with high probability if the toxicity is offset by substantial efficacy on the efficacy endpoints. Therefore, comprehensive scenario evaluation is strongly recommended, together with sensitivity analyses for all key design parameters—including the MCDA weights and the upper and lower limits of the linear partial value functions—during the design stage.

4.6 Discussion

In this work we proposed a novel seamless phase II/III design with dose selection, where the dose with the best benefit-risk profile at the end of phase II is continued and tested at the end of phase III versus the control arm. Dose selection is based in our context on the probabilistic MCDA, assuming all the endpoints considered in the benefit-risk analysis are at least approximately normally distributed. An analytical expression for the distribution of the test statistic is derived throughout the manuscript, and a guideline on how to achieve strong control of type I error rate is given.

From a statistical point of view the manuscript offers two main elements of novelty: first, it extend the use of the MCDA within the context of confirmatory trials with dose selction, providing analytical expression for the test statistics and achieving strong control of type I error; second, it demonstrates that under reasonable assumptions, dose selection rule based on the probabilistic MCDA approach is equivalent to a dose selection rule based on the comparison of appropriate univariate statistics.

Simulation results show that including safety elements in the dose selection rule substantially helps in limiting the probability to take forward potentially harmful doses. This feature is particularly appealing, mainly in situations where the safety of the dose involved within the study is not well established before phase II.

The proposed approach is highly flexible as it allows for the inclusion of any kind of internal or external information in the dose selection stage as well as any number of active arms, however a potential limitation is that the pair-wise correlations between each factor included in the MCDA have to be known in advance to compute the decision bounds.

In the present work we focused on a two-stage design, where only one dose is selected at the first interim analysis (coinciding with phase II) and rejection of the null hypothesis is allowed

only at the end of phase III. However, the statistical framework proposed lays the foundations for the development of more complex designs, e.g. allowing the rejection of the null hypothesis at multiple stages and/or letting more than one dose to be taken forward at each stage. The latter considerations draws a path for a potential future research.

Chapter 5

On the interplay between prior weight and variance of the robustification component in Robust Mixture Prior Bayesian Dynamic Borrowing approach

5.1 Introduction

Leveraging historical information in clinical trials is particularly valuable in contexts like rare diseases [117] and pediatric trials [118–120], where recruiting large patient populations is challenging. Bayesian designs are appealing as they allow incorporating available knowledge into prior distributions. However, including external data raises challenges, such as quantifying heterogeneity between external and current data, which can lead to biased estimates and poor operating characteristics if not properly addressed.

Bayesian dynamic borrowing (BDB) sets out to solve such issue by dynamically discounting the use of external information based on a measure of heterogeneity between the prior distribution and the observed data. Several borrowing strategies have been proposed over the years such as Power priors [14, 15], commensurate priors [3] and Robust Mixture Prior (RMP) [121, 17], all of them requiring the specification of a tuning parameter quantifying the amount of borrowing (called *knowledge factor* in an early non clinical reference [121]). A thorough review of the available borrowing methods can be found in Van Rosmalen et al. [12] and Viele et al. [13]. Among them, Robust Mixture Prior (RMP) [16, 17], is acknowledged as one of the most versatile options due to its natural ability of dynamically discounting the amount of borrowed information as the prior-data conflict increases. Examples of practical use of RMP in different contexts of application can be found in literature, e.g. bringing adult information to inform treatment effect on a pediatric trial [120], exploiting expert opinion to inform a prior distribution for a treatment effect [17], borrowing historical information to predict a treatment effect on a primary endpoint

based on a surrogate endpoint [76, 22] or borrowing external control data to discount sample size in the control arm [72].

The idea behind RMP is to construct a prior distribution for the parameter of interest by combining an informative component, derived from external information, and a *robustification* high-variance component in a mixture distribution. The advantage of this approach is that the information contained in the informative component of the mixture impacts the posterior inference in a dynamic way, i.e. mostly in case of agreement between historical and current data, while it is progressively disregarded as the prior-data conflict increases [16]. The RMP approach can be interpreted as a Bayesian Model Averaging (BMA) framework, where the robust component acts as a vague model designed to mitigate the influence of the informative prior in case of data conflict. In this context, the posterior weight assigned to each component is governed by the Bayes factor, which formalizes the relative evidence provided by the trial data in favor of one model over the other.

The robustness of the RMP framework is deeply rooted in the tail behavior of the prior relative to the likelihood. As discussed by [122] and [123], a Bayesian model achieves robustness when the prior's tails are flatter than those of the likelihood, allowing the data to discount the prior information in the presence of extreme conflict. Our joint selection strategy (decreasing n_0) explicitly leverages this principle: by increasing the variance of the robust component, we ensure that the RMP's tails dominate the informative component's influence more effectively as the drift increases.

The main object of investigation of this paper are robust mixtures of normal priors, called normal RMPs, which are vastly used in case of normally distributed (or approximately normally distributed) endpoints. In particular, we will focus on the case in which the informative component of the RMP is a single normal distribution with known mean and variance, and is combined with a robust normal component with higher variance. In this context, three parameters must be specified, namely *i) weight* of the robustification component of the mixture prior, *ii) location* of the robustification component and *iii) variance* of the robustification component. Although it has been shown that all these three factors impact the operating characteristics (see Weru et al. [124]), it is common to focus solely on the selection of the mixture weight related to the informative component (referred to as "mixture weight"), regulating the amount of information to be borrowed. The latter is commonly pre-specified based on the stakeholder degree of confidence in the historical source, while all the other parameters are commonly fixed. For the variance of the robustification component of the mixture it has been argued that extremely large variances should be avoided [17, 124, 125], as they can lead to borrowing of historical information even in case of extreme inconsistency between historical and concurrent data. To avoid this situation, robust weakly informative components have generally been preferred and unit information priors (UIP) [16] have become a common choice. Using weakly informative robustification components, however, has some drawbacks, in particular *i)* it is sensitive to the choice of the location of the robustification component [124], and *ii)* it causes an inflation of type I error rate in case of the

major inconsistency between historical and current data.

In this work, we demonstrate that the borrowing properties of the RMP are defined by the *joint* specification of prior weight and variance of the robustification component and these two parameters should be chosen together. We theoretically demonstrate that RMP with high-variance robustification components is a viable choice, provided a jointly optimized selection of prior weight and variance of the robustification component. We argue that this approach is advantageous as *i)* it practically makes the choice of the location of the robustification component impactless and *ii)* it effectively prevents from the asymptotic inflation of the type I error rate, which arises - in the case of weakly informative robustification components - when major inconsistency between historical and current data is observed.

The manuscript is organized as follows: Sections 5.2–5.6 focus on the normal setting. Specifically, Section 5.2 introduces the RMP model and its application in the normal setting; Section 5.3 presents the motivation for this work; Section 5.4 details the theoretical findings for the normal setting; Section 5.5 provides a proof-of-concept analysis highlighting the key benefits of the proposed methodology; and Section 5.6 outlines a novel procedure for hyper-parameter selection. Section 5.7 discusses the extension to the binary case with the Beta RMP, while Section 5.8 presents the extension to scenarios in which the informative component of the RMP is itself a mixture. Finally, Section 5.9 concludes with a discussion.

5.2 Methodology

5.2.1 Setting

Bayesian Design of a Randomized Control Trial (RCT)

Consider a randomized controlled trial (RCT) evaluating a novel treatment against placebo or standard of care. Let X_t and X_c denote the normally distributed mean treatment and control responses with unknown means θ_t and θ_c, and known variances $\sigma_t^2 = s^2/n_t$ and $\sigma_c^2 = s^2/n_c$, where s is the common variance of individual responses and n_j $(j = t, c)$ the arm-specific sample sizes.

The treatment effect $\delta = \theta_t - \theta_c$ is the parameter of interest, with $H_0 : \delta = 0$ tested against $H_A : \delta > 0$. Priors $\pi_t(\cdot)$ and $\pi_c(\cdot)$ are specified for θ_t and θ_c.

Trial success is declared when the posterior probability of a positive treatment effect exceeds a prespecified threshold:

$$\mathbb{P}_{\pi_c,\pi_t}\big(\delta > 0 \mid x_c, x_t\big) > 1 - \eta, \tag{5.1}$$

where x_c and x_t are observed mean responses. The threshold $1 - \eta$ represents the required posterior evidence for efficacy; with smaller η values imply more stringent criteria.

Frequentist and Bayesian Operating Characteristics

The *type I error rate*, the probability of rejecting H_0 when $\delta = 0$, is computed by integrating the success condition over the data likelihoods:

$$\alpha(H) = \iint_{\mathbb{R}^2} \mathbb{1}\left\{\mathbb{P}_{\pi_c,\pi_t}(\delta > 0 | x_c, x_t) > 1 - \eta\right\} f_{X_c}(x_c|\theta_c = H) f_{X_t}(x_t|\theta_t = H)\, dx_c\, dx_t, \tag{5.2}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and f_{X_c}, f_{X_t} denote the sampling distributions. *Power* is obtained analogously under $\theta_t = H + \delta^*$ and $\theta_c = H$, for a target effect $\delta^* > 0$.

type I error rate and power are frequentist quantities, as they condition on fixed parameter values. To assess Bayesian designs more comprehensively, Best et al. [80] proposed averaging α over a *design prior* Π_c, namely:

$$\alpha_{\text{avg}}^{\Pi_c} = \int_{\mathbb{R}} \alpha(t)\, \Pi_c(t)\, dt. \tag{5.3}$$

A *design prior* is the prior distribution used during the *planning* of the trial to reflect plausible values for the parameters, which allows evaluation of Bayesian operating characteristics such as average Type I error and power. It is not necessarily the same as the prior used in the *analysis* of the trial, which represents the formal beliefs applied to the data once observed. The design prior is primarily a tool for trial design and simulation, whereas the analysis prior is used for inference and decision-making.

Metrics for the posterior estimation

Besides testing, performance is evaluated through estimation metrics. The posterior median $\hat{\delta}$ serves as point estimate, and bias, variance, and mean squared error (MSE) quantify its accuracy (see Supplementary Material for formulas).

5.2.2 Robust Mixture Prior

Let $\pi_{\text{inf}}(\cdot)$ be an informative prior for θ_c. The *Robust Mixture Prior (RMP)* combines this with a weakly informative or non-informative robustification component $\pi_{\text{rob}}(\cdot)$:

$$\pi_c(\theta_c) = \omega\, \pi_{\text{inf}}(\theta_c) + (1 - \omega)\, \pi_{\text{rob}}(\theta_c), \tag{5.4}$$

where $\omega \in [0, 1]$ is the prior weight on the informative component. The robustification term downweights historical information when inconsistent with current data.

After observing x_c, the posterior is again a mixture:

$$g(\theta_c|x_c) = \tilde{\omega}\, g_{\text{inf}}(\theta_c|x_c) + (1 - \tilde{\omega})\, g_{\text{rob}}(\theta_c|x_c), \tag{5.5}$$

where each component posterior is $g_\star(\theta_c|x_c) = f(x_c|\theta_c)\pi_\star(\theta_c)/f(x_c|\pi_\star)$, with $\star \in \{\text{inf}, \text{rob}\}$. The updated weight depends on x_c via the formula

$$\tilde{\omega}(x_c) = \frac{\omega f(x_c|\pi_{\text{inf}})}{\omega f(x_c|\pi_{\text{inf}}) + (1-\omega) f(x_c|\pi_{\text{rob}})}. \tag{5.6}$$

A proof of Equation (5.5) and (5.6) is in the Supplementary Material.

Equation (5.6) can be expressed equivalently in terms of odds as

$$\tilde{\Omega}(x_c) = \Omega \frac{f(x_c|\pi_{\text{inf}})}{f(x_c|\pi_{\text{rob}})}, \tag{5.7}$$

with $\Omega = \omega/(1-\omega)$ and $\tilde{\Omega} = \tilde{\omega}/(1-\tilde{\omega})$. It can be noticed that weights (and odds) adjust borrowing dynamically according to the data's compatibility with prior information, namely increases when the observed response x_c is compatible with the informative component of the mixture while decreases otherwise.

Note that in Equations (5.6) and (5.7), posterior weights and posterior odds are well-defined functions of the observed mean response, conditional on the specified RMP for θ_c. For simplicity, this dependence will be implicitly understood in subsequent sections and explicitly stated only when necessary.

5.2.3 Normal Robust Mixture Prior

When both mixture components are Normal,

$$\pi_{\text{inf}}(\theta_c) = \mathcal{N}(\mu_{\text{inf}}, \sigma^2_{\text{inf}}), \quad \pi_{\text{rob}}(\theta_c) = \mathcal{N}(\mu_{\text{rob}}, \sigma^2_{\text{rob}} = s^2/n_0),$$

the conjugacy ensures that the posterior remains a Normal mixture with updated parameters. Moreover, the corresponding prior predictive distributions are also Normal:

$$f(x_c|\pi_\star) = \frac{1}{\sqrt{2\pi v^2_\star}} \exp\left[-\frac{(x_c - \mu_\star)^2}{2v^2_\star}\right], \quad v^2_\star = \sigma^2_\star + \sigma^2_c, \tag{5.8}$$

for $\star \in \{\text{inf}, \text{rob}\}$. As a consequence, letting $R = v_{\text{rob}}/v_{\text{inf}}$, then Equation (5.7) becomes

$$\tilde{\Omega}(x_c) = \beta(\omega, \sigma^2_{\text{rob}}) \exp\left\{-\frac{d^2}{2v^2_{\text{inf}}} + \frac{(x_c - \mu_{\text{rob}})^2}{2R^2 v^2_{\text{inf}}}\right\}. \tag{5.9}$$

In the latter, $\beta\left(\omega, \sigma^2_{\text{rob}}\right) = \Omega \times R$, while d represents the realization of the random variable $X_c - \mu_{\text{inf}} \sim \mathcal{N}\left(D, \sigma^2_c\right)$, with mean D representing the true *drift* parameter (also referred to as *prior-data conflict* hereinafter), indicating the level of inconsistency between concurrent data and historical information provided in the informative component of the RMP. Note that defining

the function $\beta(\cdot)$ will become useful later on.

Equation (5.9) shows that the posterior odds $\tilde{\Omega}$ depend on the choice of Ω (which is a deterministic function of the prior weight ω), the location parameter of the robustification component μ_{rob} and the variance of the robustification component σ^2_{rob}.

Notice that, since the robustification component must be less informative than the informative one, $R > 1$ (often $R \gg 1$ when π_{rob} is nearly non-informative).

5.3 Motivation for the work

5.3.1 Background

Robust Mixture Priors (RMPs) are widely applied in randomized controlled trials (RCTs) to borrow information for the control arm [80, 72, 126]. Several approaches exist for specifying the mixture weight ω [52, 51], yet the selection of hyperparameters for the robustification component has received limited attention.

Large variances for the robustification prior are often adopted to represent minimal prior knowledge; however, such weakly informative choices may retain excessive influence of the informative component even under strong prior–data conflict—an effect known as *Lindley's paradox* [17, 124, 125]. Schmidli et al. [16] proposed mitigating this through a *unit-information prior* (UIP), namely a distribution which effective sample size (ESS)[127] is equal to 1.

While practical and commonly used, this approach introduces two main challenges: *(i)* the prespecification of the robustification mean μ_{rob}, which strongly affects posterior inference [124]; and *(ii)* the asymptotic inflation of the Type I error in the presence of substantial discrepancies between the historical and current control data[124, 80]. Here, the term *asymptotic inflation* refers to the progressive increase in the Type I error rate as the drift parameter D increases, such that the Type I error approaches 1 as $D \to +\infty$.

The following case study illustrates these issues in Normal RMPs within hybrid-control RCTs, providing the basis for the theoretical developments in Section 5.4.

5.3.2 Illustrative Trial

Consider a two-arm RCT comparing treatment and control (placebo or standard of care). Individual outcomes in both arms follow normal distributions with unit variance ($s = 1$), as a consequence the mean responses in the two arms are:

$$X_t \sim \mathcal{N}(\theta_t, n_t^{-1}), \quad X_c \sim \mathcal{N}(\theta_c, n_c^{-1}).$$

The trial allocates $n_t = 150$ patients to treatment and $n_c = 50$ to control (3:1 ratio). Trial success is defined by Equation (5.1) with $\eta = 0.05$.

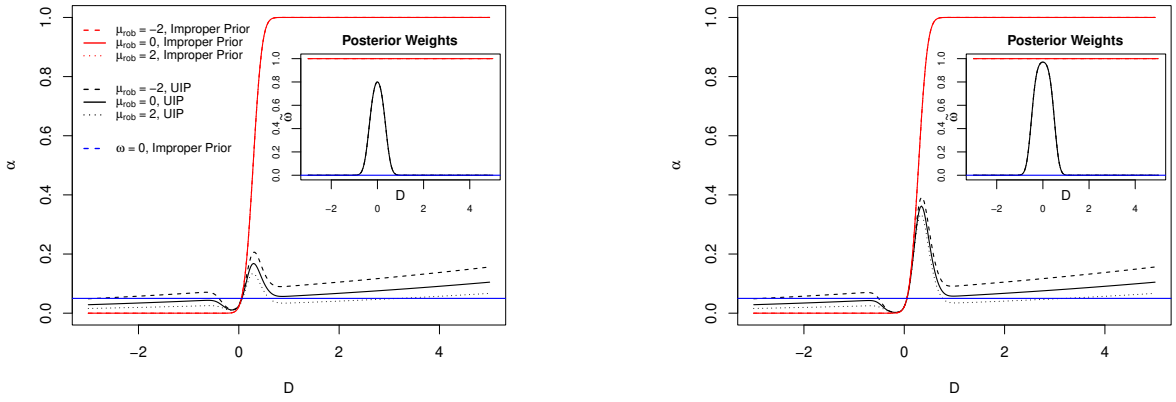
No prior information is available for θ_t, so a non-informative prior $\theta_t \sim \mathcal{N}(\mu_{\text{rob}}, n_0^{-1})$ is used. For θ_c, an informative prior $\mathcal{N}(\mu_{\text{inf}}, n_{\text{inf}}^{-1})$ with effective sample size $n_{\text{inf}} = 100$ and mean $\mu_{\text{inf}} = 0$ is combined with a non-informative prior $\mathcal{N}(\mu_{\text{rob}}, n_0^{-1})$ through an RMP with weight ω.

Performance metrics include the type I error rate (Equation 5.2), power (for target $\delta^* = 0.31$), and the average posterior weight $\tilde{\omega}$, obtained by integrating Equation (5.6) over the data likelihood.

Different RMP configurations are examined, considering mixture weights $\omega \in \{0.5, 0.9\}$ to represent, respectively, moderate and strong confidence in the historical information. Six sub-scenarios are defined by varying the hyperparameters of the robustification component. Specifically, the location parameter is set to $\mu_{\text{rob}} \in \{-2, 0, 2\}$, while the variance takes values $\sigma^2_{\text{rob}} \in \{1, 10^{100}\}$, the former corresponding to a unit-information prior and the latter approximating an improper prior. A reference setting with $\omega = 0$ and $\sigma^2_{\text{rob}} = 10^{100}$ represents a standard non-informative Bayesian design. Performance metrics are assessed across a range of *drift* values D.

5.3.3 Analysis

Figure 5.1 depicts the Type I error rate as a function of the drift parameter D for $\omega = 0.5$ (panel(a)) and $\omega = 0.9$ (panel (b)), illustrating the behavior under various combinations of hyper-parameters μ_{rob} and σ^2_{rob} for the robustification component of the RMP.



(a) Type I Error with $\omega = 0.5$

(b) Type I Error with $\omega = 0.9$

Fig. 5.1 Type I error $\alpha(D)$ under different choices of parameters for the RMP. Red curves: improper prior distributions ($\sigma^2_{\text{rob}} = 10^{100}$). Black curves: unit-information prior ($\sigma^2_{\text{rob}} = 1$). Different choices of μ_{rob} are denoted with different line types. Panel (a): analysis with prior mixture weight $\omega = 0.5$. Panel (b): analysis with prior mixture weight $\omega = 0.9$.

The benefit of using RMP in combination with UIP robustification component is evident in case of minor prior-data conflict ($D \approx 0$), where a reduction in type I error is observed with respect to the baseline nominal level. Indeed, in this region of the parameter space, the high similarity between the informative component of the mixture and the data likelihood results in high borrowing ($\tilde{\omega} \gg 0$) and accordingly in an improvement of the frequentist operating characteristics.

As the prior-data conflict grows, the borrowing extent decreases due to the progressively lower similarity between the informative component of the RMP and the data likelihood, however, when an intermediate level of drift is observed, still a certain degree of "undesirable" borrowing is observed ($\tilde{\omega} > 0$), resulting in a biased estimation of the control parameter. As a results, an increase in type I error is observed in case of (moderate) positive drifts, whilst a further deflation of the type I error is observed in case of (moderate) negative drifts.
In case of major prior-data conflict, no borrowing is observed ($\tilde{\omega} \approx 0$), meaning that posterior inference is uniquely driven by the robustification component. Whilst in this region of the parameter space it is reasonable to expect that posterior inference is driven by data, leading therefore to a constant type I error approaching the nominal level, an asymptotic increase (resp. decrease) of the latter is observed converging to 1 (resp. 0) when $D \to +\infty$ (resp. $D \to -\infty$). The investigation of this counterintuitive behavior represents a first point of interest and will be addressed in the next sections.

Figure 5.1 shows that this asymptotic behavior of the RMP is the same across the different choices of the location of the robustification component in RMP, when UIP is used as robustification component. However, in this case the choice of μ_{rob} significantly impacts the type I error behavior, which is uniformly increased in the whole parameter space as μ_{rob} increases, while is uniformly decreased as μ_{rob} increases. The sensitivity of the operating characteristics to the choice of μ_{rob} is a second point of interest which will be discussed in this manuscript.

When using a RMP with improper robustification component, high borrowing is observed at any level of prior-data conflict, with $\tilde{\omega} = 1$ across all the parameter space. The latter situation is referred in literature as *Lindley's paradox* [17, 124, 125]. As a result, type I error is almost constant at $\alpha = 0$ for $D < -0.2$, it increases steeply to $\alpha = 1$ for $-0.2 \leq D \leq 0.5$ and remains at this level for larger values of D. In this case, the impact of μ_{rob} on the type I error is practically null, as it can be noticed by the overlapping red curves.

5.3.4 Research questions

In section 5.3.4 we have shown that there are some issues related to the use RMP in the context of hybrid control RCT. These are:

1. The asymptotic inflation of type I error for large positive values of prior-data conflict, when weakly informative robustification components are employed.

2. The sensitivity of the operating characteristics to the choice of μ_{rob}, when weakly informative robustification components are employed.

3. The apparent failure in discounting information borrowing as the prior-data conflict increases, when large variance robustification components are used (Lindley's paradox).

In the next sections the cause of these issues will be theoretically investigated, and a solution to all of them will be proposed.

5.4 Analytical results

In the following sections, the issues related to the current use of the RMP framework illustrated in Section 5.3.4 will be theoretically addressed. In particular, it will be proven that RMPs with large variance robustification components are able to effectively mitigate all these problems without incurring in Lindley's paradox.

5.4.1 Asymptotic inflation of type I error

The cause of the asymptotic type I error rate inflation, along with the conditions under which the latter is prevented are investigated in Theorem 2. In particular, it is proven that type I error rate inflation occurs when an upwards bias is induced by the robustification component π_{rob} of the RMP on the posterior mean for the treatment difference. For a fixed value of the mixture weight ω, this bias is inversely proportional to the variance of the robustification component σ^2_{rob}, and in particular it is null if the latter diverges to $+\infty$ at least as fast as the drift parameter D. Under this condition, an asymptotic control of the type I error rate is achieved, thus making the choice of large variance robustification components in RMPs particularly attractive.

Theorem 2. *Consider a RCT where mean control and treatment responses are normal* $X_c \sim \mathcal{N}\left(\theta_c, \sigma_c^2\right)$, $X_t \sim \mathcal{N}\left(\theta_t, \sigma_t^2\right)$, *and assume* $\sigma_t^2 = K\sigma_c^2$ *(where* K^{-1} *is the randomization ratio, assumed > 1). Assume a RMP* $\pi_c(\theta_c) = \omega\pi_{inf}(\theta_c) + (1-\omega)\pi_{rob}(\theta_c)$ *is used for the control parameter, where* $\pi_{inf}(\theta_c)$ *and* $\pi_{rob}(\theta_c)$ *are the PDF of normally distributed random variables with parameters* μ_{inf}, σ^2_{inf} *and* μ_{rob}, σ^2_{rob} *respectively; while a normal prior distribution* $\theta_t \sim \mathcal{N}\left(\mu_t, \sigma^2_{rob}\right)$ *is given to the treatment parameter. Consider the type I error rate* $\alpha(\cdot)$ *as defined in Equation (5.2), corresponding to the null hypothesis* $H_0 : \theta_c = \theta_t = D + \mu_{inf}$, *where* $D = \theta_c - \mu_{inf}$ *is the drift parameter. Then the following hold:*

$$\lim_{D\to+\infty} \alpha\left(D + \mu_{inf}\right) = \eta \iff \lim_{D\to+\infty} \frac{D}{\sigma^2_{rob}} = 0$$

A formal proof of Theorem 2 can be found in the supplementary material. A numerical validation of this result is shown in Section 5.5, while a practical use of the latter in parameter selection can be found in Section 5.6.

5.4.2 The impact of the selection of μ_{rob}

The robustification component of the mixture acts to *robustly* model the tails of the informative component's prior distribution. Ideally, it represents a lack of prior knowledge, thereby hindering precise elicitation of its location parameter μ_{rob}. This choice, however, may significantly impact the posterior inference, as demonstrated by Weru et al. [124].
Theorem 3 investigates the condition under which the choice of μ_{rob} becomes impact-less in the posterior inference, showing that employing robustification components with large variances effectively prevents from bias stemming from the chosen location, enabling then the use of any convenient value for μ_{rob}.

Theorem 3. *Consider a normal random variable modeling the mean control response* $X_c \sim \mathcal{N}\left(\theta_c, \sigma_c^2\right)$*, and assume two distinct RMPs are used for the underlying parameter* θ_c*, namely*

$$\pi_c^{(1)}(\theta_c) = \omega\pi_{inf}(\theta_c) + (1-\omega)\pi_{rob}^{(1)}(\theta_c) \qquad \pi_c^{(2)}(\theta_c) = \omega\pi_{inf}(\theta_c) + (1-\omega)\pi_{rob}^{(2)}(\theta_c)$$

where $\pi_{inf}(\theta_c)$ *and* $\pi_{rob}^{(i)}(\theta_c)$ *are the PDF of normally distributed random variables with parameters* μ_{inf}*,* σ_{inf}^2 *and* $\mu_{rob}^{(i)}$*,* σ_{rob}^2 *respectively with* $i \in \{1,2\}$*.*
Consider the posterior distributions $g(\theta_c|x_c, \pi_c^{(1)})$ *and* $g(\theta_c|x_c, \pi_c^{(2)})$*, then*

$$\lim_{\sigma_{rob}^2 \to +\infty} g(\theta_c|x_c, \pi_c^{(1)}) = \lim_{\sigma_{rob}^2 \to +\infty} g(\theta_c|x_c, \pi_c^{(2)}) \qquad \forall x_c \in \mathbb{R}$$

A formal proof of Theorem 3 can be found in the supplementary material. A numerical validation of this result is presented in Section 5.5, while a practical use of the latter in parameter selection is proposed in Section 5.6.

5.4.3 The Lindley's paradox

The phenomenon termed "Lindley's paradox" within the context of robust mixture priors (RMPs) describes the counterintuitive situation where full borrowing (defined as $\tilde{\omega} = 1$) occurs despite significant prior-data conflict. Literature suggests this arises when the RMP's robustification component is improper [17, 124, 125]. This occurs because the prior predictive distribution for the robustification component, shown in Equation (5.8), becomes improper ($R \to +\infty$), leading to $\tilde{\omega} = 1$ for all observed control responses x_c according to Equation (5.9). In Theorem 4 we show that this behavior is due to the hidden underlying assumption that the mixture weight ω is

fixed and independent on the choice of σ^2_{rob}. We find that relaxing this assumption, effectively prevents from the occurring of Lindley's paradox.

Theorem 4. *Consider a normal random variable $X_c \sim \mathcal{N}\left(\theta_c, \sigma_c^2\right)$, and assume a RMP is used for the parameter θ_c, namely $\pi_c(\theta_c) = \omega\pi_{inf}(\theta_c) + (1-\omega)\pi_{rob}(\theta_c)$, where $\pi_{inf}(\theta_c)$ and $\pi_{rob}(\theta_c)$ are the PDF of normally distributed random variables with parameters μ_{inf}, σ^2_{inf} and μ_{rob}, σ^2_{rob} respectively. The following hold:*

1. *if $\Omega < +\infty$, then*

$$\lim_{\sigma^2_{rob}\to+\infty} \tilde{\omega}\left(x_c, \pi_{inf}(\theta_c), \pi_{rob}(\theta_c), \omega\right) = 1 \qquad \forall x_c \in (-\infty, +\infty)$$

2. *if $\Omega \sim O(R)$ for $\sigma^2_{rob} \to +\infty$, then*

$$\lim_{\sigma^2_{rob}\to+\infty} \tilde{\omega}\left(x_c, \pi_{inf}(\theta_c), \pi_{rob}(\theta_c), \omega\right) \neq 1 \qquad \forall x_c \in (-\infty, +\infty)$$

The preceding theorem demonstrates that Lindley's paradox arises, as $\sigma^2_{\text{rob}} \to +\infty$, when the prior weight ω (or prior odds Ω) is fixed independently of σ^2_{rob}. Conversely, if ω and σ^2_{rob} are jointly selected such that the prior odds Ω are of the same order of magnitude as R - as $\sigma^2_{\text{rob}} \to +\infty$ - then Lindley's paradox is avoided. The latter holds because as $\sigma^2_{\text{rob}} \to \infty$, the posterior odds $\tilde{\Omega}$ can be written following Equation (T2.1) as

$$\tilde{\Omega}(x_c; \omega, \sigma^2_{\text{rob}}) = \beta(\omega, \sigma^2_{\text{rob}}) \times \exp\left[-\frac{d^2}{2v^2_{\text{inf}}}\right], \tag{5.10}$$

where the influence of the RMP on the posterior odds is entirely captured by the function $\beta(\omega, \sigma^2_{\text{rob}})$. As a consequence, all combinations of ω and σ^2_{rob} yielding $\beta(\omega, \sigma^2_{\text{rob}}) = \beta^*$ share the same "borrowing profile", resulting in identical posterior odds (and thus, posterior weights) for any observed value x_c.

The parameter β^* governs the RMP's flexibility in borrowing information across the x_c space, determining the rate at which posterior weights decrease with increasing prior-data conflict. Specifically, it represents the posterior odds when no drift is observed, quantifying the maximum borrowing achievable by the RMP. Therefore, β^* will be referred to as the *borrowing strength*. It is important to note that while these pairs yield identical posterior weights, posterior inference for θ_c could differ in principle across RMPs due to variations in $g_{\text{rob}}(\theta_c | x_c, \pi_{\text{rob}})$, resulting from differing choices of μ_{rob} and σ^2_{rob}. However, as $\sigma^2_{\text{rob}} \to \infty$, the robust posterior becomes independent of μ_{rob}, leading to similar inference for θ_c across all pairs across the entire control response parameter space.

Note that the asymptotic approximation of posterior odds in Equation (5.10) is valid only when $R \gg 1$ ($v_{\text{rob}} \gg v_{\text{inf}}$), a reasonable assumption given the robustification component of the RMP is specifically designed for robustification.

5.5 Practical considerations

Using the same trial design considered in Section 5.3.2, in the following sections we will focus on the validation of the use of the RMPs with large variance robustification components in the context of unbalanced RCT with hybrid control arms. In Section 5.5.1 we show that large variance robustification components can be employed without incurring in Lindley's paradox, if the parameters ω and σ^2_{rob} are jointly selected. In Section 5.5.2 we show that RMPs with large variance robustification components effectively prevent from asymptotic type I error inflation. In Section 5.5.3 we show that frequentist operating characteristics of RMPs with large variance robustification components are independent on the choice of μ_{rob}.

5.5.1 Overcoming Lindley's paradox

In Section 5.4.3 it has been proven that different pairs $(\omega, \sigma^2_{\text{rob}})$ may induce the same posterior weights distribution on the control response space. The latter is illustrated in Figure 5.2.

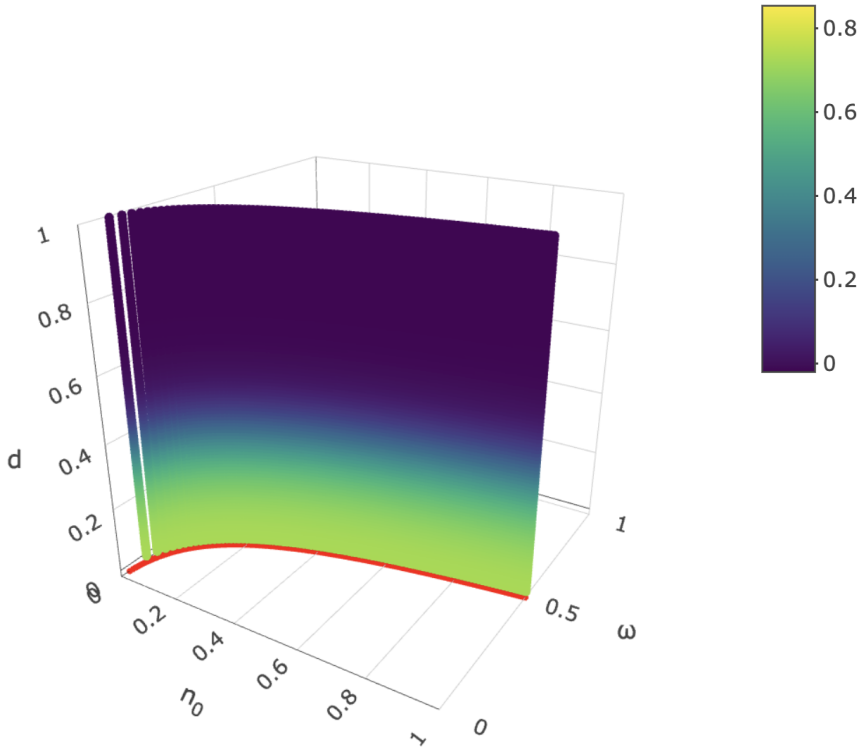


Fig. 5.2 Posterior weight $\tilde{\omega}$ as a function of effecive sample size of the robust component n_0, prior weight ω and observed control response x_c. The red curve in the (n_0, ω) represents all RMPs with $\beta^* = 5.83$.

Figure 5.2 presents a three-dimensional representation with parameters ω and $n_0 = \sigma^{-2}_{\text{rob}}$ on the horizontal axes and the observed control response x_c on the vertical axis. The red curve embedded

in the (ω, n_0) plane delineates the set of parameter pairs (ω, n_0) satisfying $\beta(\omega, n_0) = 5.83$, each representing a distinct RMP. Notice that this value has been specifically selected so to include the pair $\omega = 0.5$, $n_0 = 1$, so that

$$\beta^* = \beta(0.5, 1) = \frac{0.5}{1-0.5} \times \sqrt{\frac{1+1/50}{1/100+1/50}}. \tag{5.11}$$

The figure was generated by varying the effective sample size of the robust component over the interval $(0.01, 1)$ with a step of 0.01. For each value, the prior weight ω was determined to satisfy Equation 5.11, and the posterior odds were computed for each pair (ω, n_0) using Equation 5.7. The posterior weights were then obtained using the formula $\Omega = 1/(1+\Omega)$. The vertical colored lines in the figure depict the posterior weights $\tilde{\omega}$ as a function of x_c for all RMPs considered along the red curve, the yellow color indicating a posterior weight of 1 (full borrowing) and the blue color indicating a posterior weight of 0 (no borrowing).

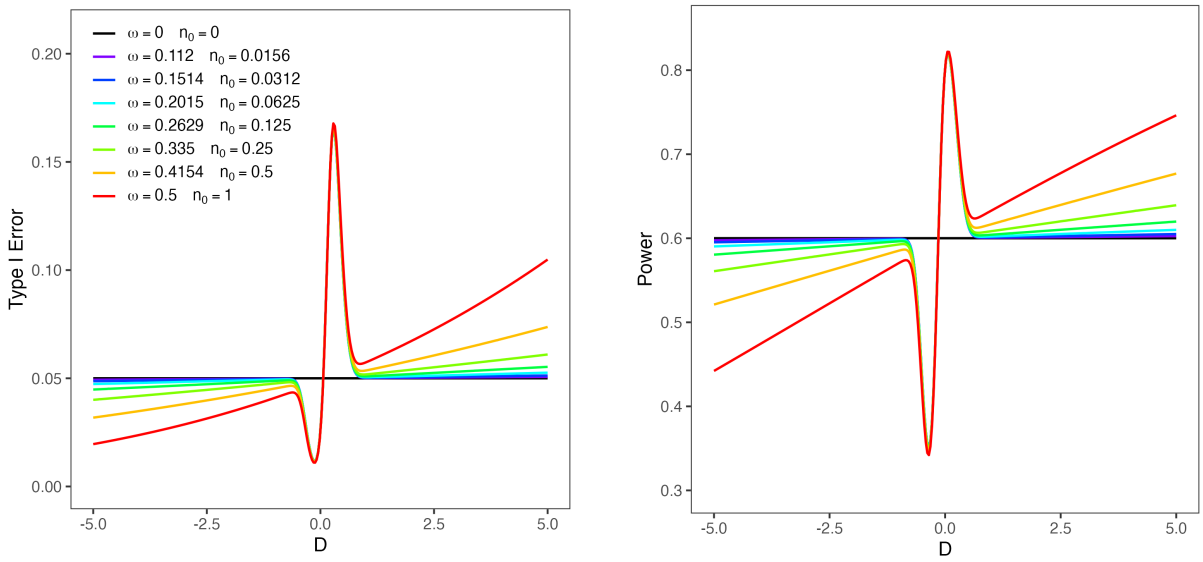
The vertical lines originating from each point on the red curve exhibit a continuous color gradient along the x_c axis, indicating that the posterior weights $\tilde{\omega}$, as a function of the control response x_c, depend solely on the chosen value of β^*. Consequently, all pairs (ω, n_0) yielding the same β^* correspond to identical posterior weight profiles.

These observations suggest that Lindley's paradox is effectively mitigated by a joint selection of ω and σ^2_{rob}. Specifically, the posterior weight profile characteristic of any RMP with a weakly informative robustification component (e.g, UIP) can be replicated using robustification components with arbitrarily large variance. Further visualizations of posterior weights under varying β^* values are provided in the supplementary materials.

5.5.2 Overcoming asymptotic type I error inflation

While the preceding analysis demonstrates that a set of RMPs share a common posterior weight profile $\tilde{\omega}$, this does not guarantee identical posterior inferences on the control parameter θ_c. Posterior inference is influenced not only by posterior weights but also by the posterior distributions of the individual RMP components, which are functions of their hyper-parameters.

In this section, an analysis of the frequentist operating characteristics is conducted, with specific attention to the problem of asymptotic type I error rate inflation. In addition, the link between the latter and the posterior inference metrics (bias, variance and MSE) is discussed.



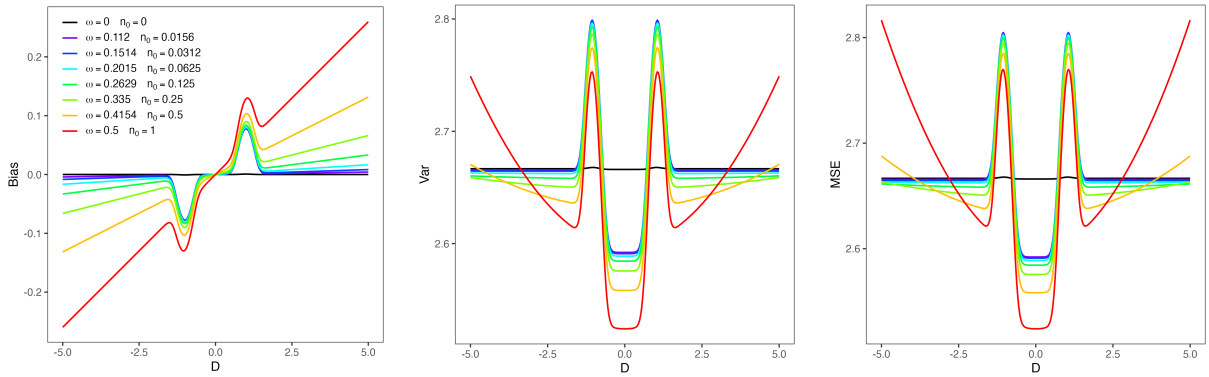
(a) Type I Error

(b) Power

Fig. 5.3 Panel (a): type I error. Panel (b): power under $\delta^* = 0.31$. Colors represent different couples of (ω, n_0), corresponding to $\beta = 5.83$.

This application considers eight distinct RMPs, generated by varying the effective sample size of the control parameter, n_0, across the set $\{(\frac{1}{2})^k | k = 0, \ldots, 7\}$, and the prior mixture weight, ω, across the set $\{0, 0.5, 0.415, 0.335, 0.263, 0.201, 0.151, 0.112\}$. All considered pairs, excluding the first (representing an improper prior), belong to the level set $\beta(n_0, \omega) = 5.83$, thus exhibiting the shared posterior weight profile discussed in Section 5.5.1. For each RMP, type I error rate (Figure 5.3a) and power (Figure 5.3b), are assessed, with power calculated for a treatment difference of $\delta^* = 0.31$. Posterior inference is evaluated using bias (Figure 5.4a), variance (Figure 5.4b), and mean squared error (MSE) (Figure 5.4c).

For small to moderate prior-data conflicts, the power (Figure 5.3b) and type I error rate (Figure 5.3a) curves overlap for all RMPs. This occurs because both variances and bias are comparable in these regions. Consequently, the posterior distributions of the treatment difference δ are similar across pairs, centered near $\delta = 0$ (for type I error rate) and $\delta = \delta^*$ (for power). This results in highly similar null hypothesis rejection rates for all RMPs.



(a) Bias (b) Variance (c) MSE

Fig. 5.4 Panel (a): bias. Panel (b): variance. Panel (c): mean squared error. Colors represent different couples of (ω, n_0), all corresponding to $\beta^* = 0.171$.

Conversely, significant differences among the pairs emerge under large prior-data conflicts, where RMPs with weakly informative robustification components exhibit inflation (deflation) of both type I error rate and power for large positive (negative) drifts. However, this effect is attenuated for RMPs with less informative robustification components, practically disappearing when $n_0 < (\frac{1}{2})^6$. In these regions, substantial differences in bias among the RMPs impact type I error rate and power, which deviate considerably from their nominal levels for RMPs with more informative robustification components, while remaining near their nominal values for RMPs with less informative robustification components.

ω	n_0	α_{max}	$\alpha(50)$	$\alpha_{avg}^{\text{VAG}}$	$\alpha_{avg}^{\text{INF}}$	$\alpha_{avg}^{\text{RMP}}$	Pow(0)	Sweet spot width
0	10^{-100}	0.0500	0.0500	0.0500	0.0500	0.0500	0.600	0.000
0.500	1.000	0.168	0.9914	0.2955	0.0394	0.0492	0.803	0.207
0.415	0.500	0.167	0.6478	0.1522	0.0397	0.0496	0.803	0.206
0.335	0.250	0.166	0.2643	0.0785	0.0399	0.0498	0.802	0.207
0.263	0.125	0.166	0.1278	0.0574	0.0399	0.0499	0.802	0.207
0.201	0.062	0.166	0.0822	0.0520	0.0400	0.0499	0.802	0.207
0.151	0.031	0.165	0.0645	0.0507	0.0400	0.0500	0.802	0.207
0.112	0.016	0.165	0.0569	0.0503	0.0400	0.0500	0.802	0.207

Table 5.1 Maximum type I error (α_{max}), average type I error (α_{avg}), power gain under no data-conflict Pow(0) and width of the sweet spot for different couples of (ω, n_0), all corresponding to $\beta^* = 5.83$.

Table 5.1 summarizes key characteristics of the observed curves. These include the maximum type I error rate inflation, α_{max}, constrained to the interval $-5 < D < 5$ (a plausible response range); the power gain, Pow(0), when the informative component of the RMP perfectly matches the control data; the type I error rate under extreme drift, $\alpha(50)$; the average type I error rate across different design priors (an improper prior, the informative component of the RMP, and the RMP itself); and the width of the "sweet spot" region [13]. The "sweet spot" is defined as the interval of D values where type I error rate and Power are respectively below and above

their nominal levels (5% and 60% in this application). All considered (ω, n_0) pairs demonstrate comparable performance in terms of maximum type I error rate, α_{max}, power gain, $\text{Pow}(0)$, and sweet spot width. However, a significant difference emerges when examining $\alpha(50)$. This value is notably higher for RMPs with weakly informative robustification components (approaching 100% for the UIP), progressively decreasing towards 5% as the informativeness of the robustification component increases.

Averaging type I error rate across an improper prior distribution reveals a marked inflation for RMPs with weakly informative robustification components, as consequence of the asymptotic type I error rate increase discussed previously. The type I error rate decrease observed for negative drifts does not fully compensate for the inflation because the range of increase (from 5% to 100%) is considerably larger than the range of decrease (from 5% to 0%), leading to a greater weighting of the inflation in the averaging process.

Conversely, minimal differences are observed among pairs when averaging type I error rate across more informative priors, such as the informative component of the RMP or the RMP itself. These priors are concentrated around regions of small drifts, where all RMPs have practically identical type I error rate curves. The type I error rate reduction exhibited by all RMPs in this region keeps the average type I error rate controlled at the nominal level (in the strong sense, when using the informative component or the RMP as the design prior).

It is worth noticing that the restriction of the drift range to $D \in [-5, 5]$ for the evaluation of α_{max} is intentionally chosen to focus on the region where the posterior weight $\tilde{\omega}$ is non-negligible, meaning where active borrowing occurs. In this zone, RMPs characterized by the same β^* but different (ω, n_0) pairs exhibit nearly identical Type I error profiles and maximum inflation values. While this focus might seem to limit the frequentist perspective, the frequentist validity is preserved by the asymptotic properties established in Theorem 1: for larger drifts, the Type I error rate is guaranteed to revert to the nominal level η, ensuring that the maximum inflation observed within $D \in [-5, 5]$ is indeed the global peak of concern for the practitioner.

It is also crucial to highlight that for any desired level of borrowing strength β^*, increasing the variance of the robustification component (i.e., decreasing n_0) while jointly adjusting ω provides a systematic way to attenuate Type I error inflation (leading to a control of the latter to the limit of $n_0 \to 0$. While it is true that increasing β^* with a fixed n_0 would exacerbate the inflation, our results demonstrate that the joint selection (ω, n_0) allows the practitioner to maintain the same degree of information borrowing while significantly enhancing robustness.

In summary, RMPs with high-variance robustification components achieve comparable performance to those with weakly informative robustification components, while simultaneously mitigating type I error rate inflation. This results in average type I error rate remaining below the nominal level when the RMP or its informative component are used as design priors (as demonstrated in Best et al. [80]), but also controlled just slightly above the nominal level when

improper priors are used; thus guaranteeing an higher overall protection to incorrect rejections of the the null hypothesis.

5.5.3 Overcoming biases due to the specification of μ_{rob}

Figure 5.5 investigate the influence of robustification component location on the type I error rate within the Robust Mixture Prior (RMP). For each of the first six (ω, n_0) pairs analyzed in Figure 5.3 and Table 5.1, five type I error rate and power curves (as functions of the drift parameter D) are presented, corresponding to variations in the robustification component location parameter, μ_{rob}, across the set $\{-2, -1, 0, 1, 2\}$.

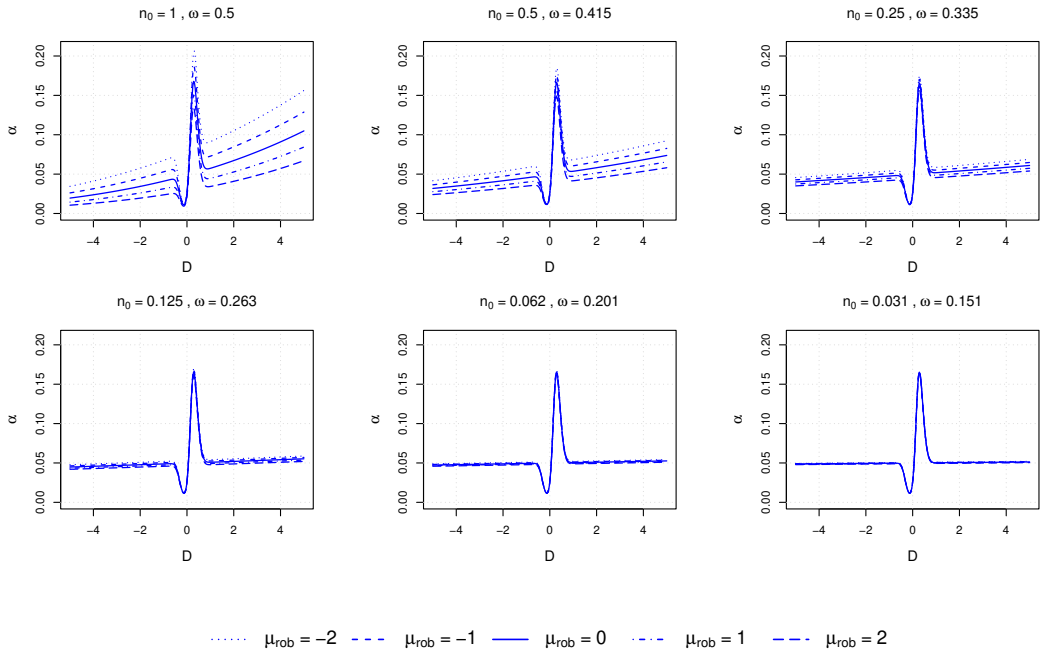


Fig. 5.5 For each panel representing a different couples of (ω, n_0), type I error as a function of the prior-data conflict D is displayed for five different values of the location of the robustification component μ_{rob}.

The figures demonstrates that for large n_0 values (e.g., UIP), operating characteristics exhibit high sensitivity to the location parameter μ_{rob}. Consistently with what shown in Section 5.3, increasing μ_{rob} uniformly inflates both Type I error curve, while decreasing μ_{rob} has the opposite effect. Conversely, as n_0 decreases (and accordingly σ^2_{rob} increases), the impact of μ_{rob} on posterior inference diminishes, as evidenced by the substantial overlap of the Type I error curves when $n_0 = 0.031$.

5.6 Hyper-parameters elicitation

5.6.1 On the interpretation of the prior weight

The use of normal RMPs in practice necessitates the pre-specification of hyper-parameters: the robustification component location μ_{rob}, the robustification component variance σ^2_{rob}, and the mixture weight ω. Current practice often prioritizes default values for the two former parameters, centering the robustification component at the informative component mean ($\mu_{\text{rob}} = \mu_{\text{inf}}$) and selecting a unit-information robust variance [16]. The mixture weight ω is then normally determined based on stakeholder or experts confidence in the data supporting the informative component.
This elicitation is typically driven by questions like *"how much is the probability that historical data are relevant in the current setting?"* or *"how much confidence do you have in historical data being representative of the current data?"*. For instance, high confidence (or high probability) might lead to $\omega = 0.9$, whereas low confidence might lead to $\omega = 0.3$.

While straightforward to communicate, this interpretation may disregard the crucial interplay between ω and σ^2_{rob}, significantly influencing RMP performance as it only concerns one parameter of the RMP, while it is argued above that they should be chosen in accordance with the variance of the robustification component. Furthermore, implies that the current choice of ω is unrelated to the choice of the robustification component. In fact, following the results above, we argue that the interpretation (and as a result the elicitation) of the weight should come together with the choice of the robustification component.

We have proven in Section 5.4.3 that the *borrowing strength* β^* is the key parameter influencing the borrowing profile of the RMP. This suggests that an equivalent prior degree of confidence in historical data should correspond to a lower ω for RMPs with a larger robustification component variance and a higher ω for RMPs with a smaller robustification component variance. As a consequence, we posit that ω should be viewed as a *relative* confidence measure between the informative model π_{inf} and the robust model π_{rob}, which specification should then depend on how informative the robustification component itself is.

Given the suggested interpretation of ω, we propose the following procedure for its elicitation.

5.6.2 An approach for hyper-parameters elicitation

A four-step elicitation approach is proposed:

1. Standard deviation of the robustification component of the RMP σ_{rob} is set to a large value. A possible option is setting it to $\sigma_{\text{rob}} = 1000 \times s$, where s represents the standard deviation of the considered endpoint (note that even higher values can be used, but as demonstrated above they will have no impact on the inference).

2. The location of the robustification component μ_{rob} is set equal to the location of the informative component μ_{inf}.

3. Clinicians are asked to determine an "equipoise drift" value d^*, representing the potentially observed control response that would induce maximum uncertainty regarding the relevance of historical data. Prompting questions could be: *"At what control response value would you be 50% confident that the historical component is relevant for the current trial and 50% that it is not?"* or *"At what control response value would you suspect a systematic difference between historical and concurrent control data?"*.

4. Once specified σ_{rob} and d^*, the prior odds Ω is obtained such that $\tilde{\Omega}(d^* + \mu_{\text{inf}}) = 1$ (or equivalently $\tilde{\omega} = 0.5$), inverting equation (5.9) as follows:

$$\Omega = \frac{R}{\exp\left\{-\frac{d^{*2}}{2v_{\text{inf}}^2} + \frac{(x_c - \mu_{\text{rob}})^2}{2R^2 v_{\text{inf}}^2}\right\}} \tag{5.12}$$

   and accordingly the prior weight is retrieved as $\omega = \frac{\Omega}{1+\Omega}$.

It is important to acknowledge that the elicitation of the equipoise drift d^* cannot be performed in a vacuum because it is inherently tied to the expected variances of both the historical and concurrent data. Specifically, the scale and clinical meaning of d^* are dictated by the predictive standard deviation $v_{inf} = \sqrt{\sigma_{inf}^2 + \sigma_c^2}$, which directly depends on the sample sizes n_{inf} and n_c. Therefore, to prevent confounding the location of the drift with the precision of the trial, the prompting questions must explicitly contextualize these uncertainties for the experts.

In practice, this means clinicians should be provided with the expected standard errors or visual aids, such as plots of the prior predictive distributions under the historical model. The prompting question can then be refined as: *"Given that our historical data has a standard error of σ_{inf} and the current control arm will have a standard error of σ_c, at what observed difference d^* would you suspect a systematic conflict?"* Alternatively, to simplify elicitation, d^* could be elicited on a standardized scale (e.g., as a multiple of the predictive standard deviation v_{inf}) rather than on the raw response scale.

Our hyper-parameter selection routine combines the benefits of RMPs with large variance robustification components and expert interaction. Moreover, while elicitation of the mixture weight ω poses challenges due to its complex interpretability, elicitation on the drift scale offers straightforward interpretation, thus justifying the approach.

5.7 Beta-Binomial case

5.7.1 Beta Robust Mixture Prior

Let us now consider the setting in which a RCT is performed with a binary outcome so that the total number of responses is $X_c \sim \text{Bin}(\theta_c, n_c)$, where n_c is the number of patients allocated to the control arm and $\theta_c \in (0,1)$ represents the response parameter on the probability scale.

The Robust Mixture Prior in this case can be chosen as a mixture of two Beta distribution, namely $\text{Beta}(a_{\text{inf}}, b_{\text{inf}})$ for the informative component and $\text{Beta}(a_{\text{rob}}, b_{\text{rob}})$ for the robustification component. Then the prior predictive density of the data is a Beta-Binomial, namely

$$f(x_c|\pi_\star) = \binom{n_c}{x_c} \frac{B(a_\star + x_c, b_\star + n_c - x_c)}{B(a_\star, b_\star)} \qquad \star = \{\text{inf}, \text{rob}\} \tag{5.13}$$

where $x_c \in (0, n_c)$ is the observed number of responders in the control arm and $B(\cdot)$ represents the Beta function. Working out with the Gamma function expression of the Beta function, it follows that the odds update of Equation (5.7) can be expressed in this case as

$$\Omega(x_c) = \beta(\omega, a_{\text{rob}}, b_{\text{rob}}) \times \frac{B(a_{\text{inf}} + x_c, b_{\text{inf}} + n_c - x_c)}{B(a_{\text{rob}} + x_c, b_{\text{rob}} + n_c - x_c)\, B(a_{\text{inf}}, b_{\text{inf}})}, \tag{5.14}$$

where the function $\beta(\omega, a_{\text{rob}}, b_{\text{rob}})$ can be expressed as

$$\beta(\omega, a_{\text{rob}}, b_{\text{rob}}) = \Omega \cdot B(a_{\text{rob}}, b_{\text{rob}}) \tag{5.15}$$

Note that although a_{rob} and b_{rob} may differ, setting them equal and small is a reasonable choice when aiming to represent limited prior knowledge. In common practice, specifications such as $\text{Beta}(1,1)$ or $\text{Beta}(0.5, 0.5)$ (Jeffreys prior) are typically employed for this purpose.

5.7.2 The Lindley's paradox in the Beta-Binomial case

Similarly to the normal case, also in the Beta-Binomial case the phenomenon of the Lindley's paradox occurs when a large variance distributions is used as a robust component of the RMP. Specifically, this happens - for a fixed ω - when the parameter of the Beta distribution related to the robust component approaches 0, because $\Gamma(0^+) \to +\infty$ and accordingly following Equation (5.15) the posterior odds goes to $+\infty$ and accordingly the posterior weights ω goes to 1. Similarly to what done in the normal case in Theorem 4, in Theorem 5 we show that this behavior is due to the hidden underlying assumption that the mixture weight ω is fixed and independent on the choice of a_{rob} and b_{rob}. We find that relaxing this assumption, effectively prevents from the occurring of Lindley's paradox.

Theorem 5. *Consider a binomial random variable $X_c \sim Bin(\theta_c, n_c)$, and assume a RMP is used for the parameter θ_c, namely $\pi_c(\theta_c) = \omega\pi_{inf}(\theta_c) + (1-\omega)\pi_{rob}(\theta_c)$, where $\pi_{inf}(\theta_c)$ and $\pi_{rob}(\theta_c)$ are the PDF of Beta distributed random variables with parameters a_{inf}, b_{inf} and $a_{rob} = b_{rob} = \varepsilon$, respectively. The following hold:*

1. *if $\Omega < +\infty$, then*

$$\lim_{\varepsilon \to 0} \tilde{\omega}\left(x_c, \pi_{inf}(\theta_c), \pi_{rob}(\theta_c), \omega\right) = 1 \qquad \forall x_c \in (0, n_c)$$

2. *if $\Omega \sim O(\varepsilon)$ for $\varepsilon \to 0$, then*

$$\lim_{\varepsilon \to 0} \tilde{\omega}\left(x_c, \pi_{inf}(\theta_c), \pi_{rob}(\theta_c), \omega\right) \neq 1 \qquad \forall x_c \in (0, n_c)$$

A formal proof of Theorem 4 can be found in the Supplementary material.
The preceding theorem demonstrates that Lindley's paradox arises, as the parameters of the robust component of the RMP approaches zero, when the prior weight ω (or prior odds Ω) is fixed independently of the parameters of the robust component. Conversely, if ω and $a_{\text{rob}} = b_{\text{rob}} = \varepsilon$ are jointly selected such that the prior odds Ω remain of the same order of magnitude as the parameters of the robust component, namely $\Omega \sim O(\varepsilon)$, then Lindley's paradox is avoided.

This occurs because, as $\varepsilon \to 0$, the posterior odds $\tilde{\Omega}$ can be expressed following Equations (5.14) and (5.15) as

$$\tilde{\Omega}(x_c; \omega, \varepsilon) = \beta(\omega, \varepsilon) \times \frac{B\left(a_{\text{inf}} + x_c, b_{\text{inf}} + n_c - x_c\right)}{B\left(x_c, n_c - x_c\right) B\left(a_{\text{inf}}, b_{\text{inf}}\right)}, \tag{5.16}$$

where the influence of the RMP on the posterior odds is entirely captured by the function $\beta(\omega, \varepsilon)$ defined in Equation (5.15). It follows that, similarly to what shown in the normal case, all combinations of ω and ε yielding the same $\beta(\omega, \varepsilon) = \beta^*$ share the same "borrowing profile", resulting in identical posterior odds and posterior weights $\tilde{\omega}$ for any observed number of responders x_c.

The parameter β^* governs the RMP's flexibility in borrowing information across the x_c space, determining the rate at which posterior weights decrease in the presence of prior-data conflict.

It is important to note that, while these pairs (ω, ε) yield identical posterior weights, posterior inference for θ_c could in principle differ across RMPs due to variations in the robust posterior component $g_{\text{rob}}(\theta_c | x_c, \pi_{\text{rob}})$ arising from different choices of ε. However, as $\varepsilon \to 0$, the posterior distribution related to the robust component of the RMP tends to lose its dependence on the prior parameters, thus leading to similar inference for θ_c across all such pairs.

5.7.3 Practical Considerations

In the Supplementary Material, the results presented in Section 7 are validated through a numerical investigation. Specifically, we considered a randomized controlled trial (RCT) in

which $n_c = 100$ patients are assigned to the control arm, while $n_t = 200$ patients are allocated to the treatment arm. The number of responses in each arm follows a binomial distribution $X_* \sim \text{Bin}(\theta_*, n_*)$, $* = \{c, t\}$.

A Jeffreys prior, $\text{Beta}(0.5, 0.5)$, is used for the treatment parameter θ_t, whereas various robust mixture priors (RMPs) are explored as prior distributions for the control parameter θ_c. The informative component of the RMP is fixed to $\text{Beta}(50, 50)$, reflecting a prior knowledge on the control parameter being close to $\theta_c = 0.5$. The success rule is the same expressed in Equation (5.1), where δ represents the log odds ratio corresponding to the two parameters, namely $\delta = \log\left(\frac{\theta_t(1-\theta_c)}{\theta_c(1-\theta_t)}\right)$.

Analogously to the normal case, Figure D.1 in the Supplementary Material illustrates how the posterior weights vary as a function of the observed number of responses in the control arm, when the prior weight ω and the parameters of the robust component of the RMP, $a_{\text{rob}} = b_{\text{rob}}$, are jointly chosen to satisfy the condition $\beta^* = 12.56$. Notice that this value has been arbitrarily selected so to include the pair $\omega = 0.8$, $a_{\text{rob}} = b_{\text{rob}} = 0.5$, so that $\beta^* = \beta(0.8, 0.5) = \frac{0.8}{1-0.8} \cdot B(0.5, 0.5)$.

The figure shows that, for all parameter pairs satisfying $\beta^* = 12.56$, the variation of the posterior weights $\tilde{\omega}$ with respect to the number of control responses x_c is closely aligned. This indicates that all such RMPs exhibit the same borrowing profile, and particularly that borrowing is possible even when a_{rob} and b_{rob} are very small, thus confirming that the Lindley's paradox can be effectively avoided provided a joint selection of the pair $(\omega, a_{\text{rob}} = b_{\text{rob}})$.

This behavior is further confirmed by examining the type I error rate and power plots in Figure 5.6, as well as the bias, variance, and mean squared error plots in Figure D.2.

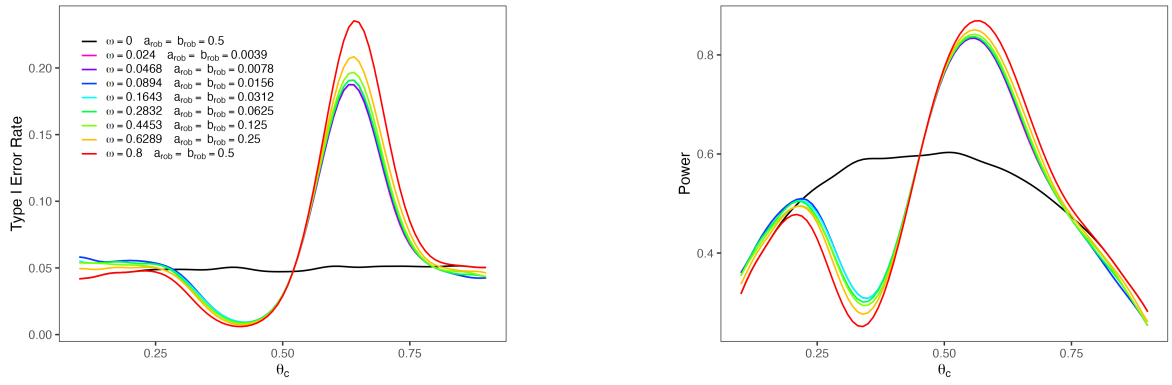


Fig. 5.6 Panel (a): Type I error rate

; Panel (b): power under a target log-odds ratio $\delta^* = 0.47$, both evaluated in the Beta–Binomial setting. Colors indicate different pairs of $(\omega, a_{\text{rob}} = b_{\text{rob}})$ corresponding to $\beta^* = 12.56$.

In these figures, eight pairs $(\omega, a_{\text{rob}} = b_{\text{rob}})$ satisfying $\beta^* = 12.56$ are shown, and the operating characteristics corresponding to different RMPs are displayed across the true control parameter $\theta_c \in (0.1, 0.9)$. In particular, the curves corresponding to different pairs $(\omega, a_{\text{rob}} = b_{\text{rob}})$ follow very similar trends across the θ_c range. A near-complete overlap is observed for pairs with

$a_{\text{rob}} = b_{\text{rob}} < 0.1$ across the parameter space, while some deviations occur in regions of moderate prior-data conflict, i.e., when more informative Beta priors are employed as the robust component of the RMP. For instance, using a Beta$(0.5, 0.5)$ prior produces similar OCs in regions of minor drift, but the maximum type I error increases noticeably (approximately 6% higher) relative to RMPs with weaker robust components, due to higher bias in regions of intermediate conflict.

Consistent with the normal case, we conclude that employing quasi non-informative Beta distributions as the robust component in the Beta RMP is feasible without inducing Lindley's paradox, provided that the prior weight ω and the parameters of the robust component are jointly selected. Moreover, using weakly informative robust components mitigates bias in regions of the parameter space where type I error inflation is most pronounced, thus offering greater protection against potential inflation arising from moderate drift between concurrent and historical data.

Finally, it is noteworthy that, in the Beta-Binomial setting, asymptotic type I error inflation is not a concern, as the extent of prior-data conflict is inherently bounded by the domain of the parameter θ_c.

5.8 Extension to a Mixture Informative component

The proposed framework can be generalized to scenarios where the informative component of the Robust Mixture Prior (RMP) is itself a finite mixture of distributions, such as Normals or Betas. This extension is highly relevant in practice, as the informative component is frequently derived from a meta-analytic predictive (MAP) prior. Because MAP priors typically lack a closed-form density, they are routinely and accurately approximated using finite mixtures, making this adaptation both practical and computationally advantageous.

Let the informative component of the RMP be expressed as

$$\pi_{\text{inf}}(\theta_c) = \sum_{k=1}^{K} \xi_k \, \pi_{\text{inf}}^{(k)}, \tag{5.17}$$

where $\sum_{k=1}^{K} \xi_k = 1$. Denote by ω the weight assigned to the informative component and by $1-\omega$ the weight assigned to the robust component. The overall RMP can then be represented as a mixture of $K+1$ components:

$$\pi_c(\theta_c) = \sum_{k=1}^{K} \omega \, \xi_k \, \pi_{\text{inf}}^{(k)} + (1-\omega) \, \pi_{\text{rob}}. \tag{5.18}$$

Define $\eta_k = \omega \xi_k$ for $k = 1, \ldots, K$ and $\eta_{K+1} = 1 - \omega$. Let $\Omega_k = \eta_k / (1 - \eta_k)$ denote the odds associated with the k-th component of the RMP. An extension of Equation 5.7 to this setting,

expressed in terms of the reciprocal of the odds rather than the odds themselves (for convenience), can be written as

$$\tilde{\Omega}_h^{-1}(x_c) = \sum_{\substack{k=1 \\ k\neq h}}^{K} \frac{\xi_k \, f\left(x_c \mid \pi_{\text{inf}}^{(k)}\right)}{\xi_h \, f\left(x_c \mid \pi_{\text{inf}}^{(h)}\right)} + \frac{1}{\xi_h} \Omega_{K+1}^{-1} \frac{f(x_c \mid \pi_{\text{rob}})}{f\left(x_c \mid \pi_{\text{inf}}^{(h)}\right)} \qquad h = 1, \ldots, K \tag{5.19}$$

and the posterior weight related to the robust component can be retrieved as $\tilde{\eta}_{K+1} = 1 - \sum_{k=1}^{K} \tilde{\eta}_k$. Note that Equation (5.19) reduces to Equation (5.7) when $K = 1$.

It is worth noting that the first summation term in the above expression does not depend on the prior weights assigned to the informative and non-informative components, but only on the fixed weights ξ_k associated with each element of the informative part of the RMP. Moreover, it is independent of the specification of the robust component of the RMP. The reciprocal of the second term, in contrast, coincides with Equation 5.7, rescaled by a component-specific factor ξ_h. Consequently, the asymptotic decomposition derived in the previous sections (for both the continuous and binary cases) remains valid, and the proposed methodology can be seamlessly extended to the mixture-based framework.

5.9 Discussion

Robust Mixture Priors (RMPs) are a prominent dynamic borrowing approach used to incorporate historical control data in the analysis of a current randomized trial. However, specifying parameters for the RMP components, particularly the robustification component and mixture weights, presents a challenge, as these parameters strongly influence posterior inferences. While improper normal distributions may seem intuitive for the robustification component, their use has been discouraged due to the potential for Lindley's paradox, prompting a preference for weakly informative priors. Employing the unit-information prior (UIP) [16] has become common; nevertheless, this choice remains somewhat arbitrary and context-dependent [125]. Specifically, concerns have been raised regarding the UIP's potential over-informativeness in trials with limited sample sizes [124], as well as the theoretical unbounded type I error rate in unbalanced trials using UIP [80].

In this article, we demonstrate, for both normal and binary endpoints, that jointly eliciting the mixture weight and the hyperparameters of the robustification component within a Robust Mixture Prior (RMP) framework effectively mitigates Lindley's paradox, even when using arbitrarily large variances.

This approach offers several practical advantages. In the normal case, it practically eliminates the impact of the location of the robustification component and prevents asymptotic type I error rate inflation in unbalanced trials, which is a critical regulatory consideration. While asymptotic inflation does not occur in balanced trials, these scenarios are of limited practical interest, as the

main goal of borrowing is to reduce sample size on the control arm.

For binary endpoints, asymptotic type I error inflation does not occur due to the natural bounds of the probability parameter (0 to 1). Nevertheless, employing a large-variance robustification component (i.e., a Beta distribution with parameters approaching 0) has been shown to reduce the maximum type I error inflation compared to the commonly used Jeffreys prior.

Beyond the inferential advantages, the proposed approach offers a significant computational benefit. While the implementation of Robust Mixture Priors frequently relies on MCMC-based algorithms, such methods often encounter numerical instability and convergence issues when dealing with diffuse components (i.e., very small n_0), as the estimation of Bayes factors becomes increasingly unreliable in high-variance settings. In contrast, our framework relies on exact analytical expressions for the posterior weights. This mathematical tractability not only ensures numerical stability even in the limit of extreme robustification ($n_0 \to 0$) but also eliminates the simulation errors inherent in MCMC.

We illustrate these properties through a proof-of-concept case study. Additionally, we propose a novel routine for selecting hyperparameters that combines a large-variance robustification component with an expert opinion-driven prior weight, ω.

We further extend the methodology to the setting where the informative component of the RMP itself is a mixture of normal distributions, enhancing the flexibility of the approach.

Importantly, the insights derived from this work are general and extend to any framework employing a Robust Mixture Prior (RMP). The demonstrated interplay between the prior weight ω and the robustification component π_{rob} is not limited to the specific implementation proposed here but is also relevant to other approaches that rely on RMPs, including those based on empirical Bayes formulations such as the EB-rMAP [52] and the SAM prior [51]. Consequently, our findings provide a unifying perspective that can inform the specification and calibration of RMP-based borrowing mechanisms across diverse methodological frameworks.

Although the mathematical results could, in principle, be extended to one-arm trials where borrowing is performed on the treatment effect scale, exploring this application is beyond the scope of the current study. We leave the investigation of one-arm trial extensions and the evaluation of whether similar advantages hold in practice as future work.

Chapter 6

Conclusions and future work

6.1 Conclusions

In this dissertation, we have investigated how the comprehensive use of information originating from multiple sources, both internal and external to the trial, can enhance the efficiency of randomized controlled trials.

In Chapter 2, we demonstrated that leveraging a well-documented historical relationship between a surrogate short-term endpoint and a primary long-term endpoint of interest can effectively enhance a futility stopping rule based on the predictive probability of success (PPoS), provided that the surrogacy relationship between the surrogate and the primary efficacy endpoint is well established.

In Chapter 3, we showed that the same PPoS criterion can be employed to strengthen the evidence derived from a surrogate endpoint in randomized controlled trials that allow for interim evaluation in support of accelerated approval. Furthermore, we demonstrated that the inclusion of historical information within the Bayesian framework can enhance trial efficiency, while the inflation of Type I error rates can be mitigated through the use of dynamic borrowing approaches.

In Chapter 4, we demonstrated that the multi-criteria decision analysis (MCDA) framework can be integrated within a seamless Phase II/III design to balance efficacy and safety considerations in the treatment selection process at interim, thereby helping to prevent potentially harmful treatments from advancing to Phase III.

Finally, in Chapter 5, we addressed the selection of hyper-parameters in the context of the Robust Mixture Prior (RMP) Bayesian dynamic borrowing approach. We showed that the operating characteristics of RMPs are governed by the joint specification of these parameters and demonstrated that adopting robust components with large variance can improve Type I error control, provided that an appropriate joint elicitation of hyper-parameters is conducted at the design stage.

6.2 Future work

Building on the work presented in this dissertation, several potential directions for future research are outlined below.

In Chapter 2, the surrogate prior introduced by Saint-Hilary *et al.* was employed to establish a futility stopping rule grounded in the *Predictive Probability of Success* (PPoS). Future research directions may involve expanding this methodological framework to incorporate a broader spectrum of adaptive design modifications, such as interim sample size re-estimation procedures. Noteworthy foundations for such extensions include the approaches developed by Brown *et al.* [128], Spiegelhalter *et al.* [71], Lecoutre *et al.* [129], and Lee and Zelen [130], wherein sample size determination is predicated on the predictive probability of trial success. Alternative strategies for sample size calculation may consider the length of the posterior credibility interval, as discussed in Pham-Gia and Turkkan [131], Joseph *et al.* [132], and Pezeshk [133], or may involve optimizing a utility function within a decision-theoretic framework as proposed by Stallard [134], Claxton *et al.* [135], Sahu and Smith [136], and Berry *et al.* [137]. Furthermore, the current Bayesian dynamic borrowing approach utilizes a two-component mixture prior, consisting of a weakly informative component alongside the surrogate prior as the informative component. Future investigations could explore alternative prior configurations, such as a three-component mixture prior that introduces an additional informative element incorporating evidence from earlier clinical development phases (e.g., Phase II trials). This extension has the potential to enhance both the flexibility and efficiency of the borrowing mechanism by enabling differential weighting of prior information sources based on their relevance and reliability.

Future developments of Chapter 3 could explore the extension of the proposed dual-criterion framework toward multi-endpoint adaptive designs, in which several surrogate and clinical outcomes are evaluated jointly. Such an extension may enhance both the efficiency of the design and the accuracy of interim decision-making when multiple biomarkers or intermediate outcomes are available. Nevertheless, it would introduce the methodological challenge of modeling potentially complex dependencies among endpoints. Another promising direction of research concerns the application of the proposed methodology within *platform trials* [138] or *basket trial* [139, 140] structures, where multiple experimental treatments or disease subtypes are investigated simultaneously under a unified statistical framework. Embedding the dual-criterion approach in these trial architectures could promote consistent accelerated approval decisions across correlated therapeutic settings and may enable dynamic information borrowing across study arms [141]. Finally, the integration of causal inference methodologies, for instance within the potential outcomes framework [142, 143], could strengthen the assessment of surrogacy and causal validity between short-term and long-term endpoints. This would allow a more explicit quantification of how treatment effects propagate from surrogate to primary outcomes, thereby improving both the interpretability and the regulatory credibility of accelerated approval requests.

Throughout this dissertation, the concept of historical borrowing from external control data has been thoroughly investigated, with particular focus on the Robust Mixture Prior (RMP) methodology within the Bayesian Dynamic Borrowing framework. Typically, the selection of design parameters in this context relies on sensitivity analyses or expert elicitation, as discussed in Chapter 4. However, in confirmatory clinical trial settings, where controlling the type I error rate is paramount, such parameter choices must also address the potential inflation of type I error arising from prior-data conflicts. In line with this principle, Calderazzo et al. [144] proposed a comprehensive framework for borrowing historical information that caps the type I error at a pre-specified level by employing a data-dependent, adaptive adjustment of the Bayesian test decision thresholds. Within the framework of Robust Mixture Priors, an important avenue for future research would be to establish analytical relationships between the prior hyper-parameters and the resulting type I error rates. Such developments would enable a fully objective and transparent selection of hyper-parameters, thereby enhancing the rigor and reproducibility of historical borrowing methodologies in clinical trial design.

In several of the contexts explored throughout this dissertation, particularly in Chapters 2 and 3, we focused on the modeling of multiple clinical endpoints (e.g., progression-free survival [PFS] and overall survival [OS]). For the sake of methodological simplicity, the dependency between these endpoints at the patient level was not explicitly incorporated into the proposed models. However, in certain applications, explicitly modeling the correlation between endpoints through a joint modeling framework may enhance both inference and interpretability. A promising avenue for future research involves the development and application of methods that account for patient-level dependencies across outcomes—whether binary, time-to-event, or count data. In the time-to-event setting, several approaches have been proposed. For instance, Link [145] and Emoto and Matthews [146] introduce joint modeling techniques that rely on restricted joint distributions via semi-parametric or parametric formulations. Alternatively, Déjardin et al. [147] employ a multi-state modeling framework for joint inference. Additional approaches situated within the semi-competing risks framework can be found in the works of Fine, Jiang, and Chappell [148], Wang [149], and Peng and Fine [150] or in Fu et al, [151] in the Bayesian context.

References

[1] Philip Pallmann, Andrew W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Lora Flight, Lisa V. Hampson, Jac D. Holmes, Adrian P. Mander, Matthew R. Sydes, Susana S. Villar, James M.S. Wason, Christopher J. Weir, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1):29, 2018.

[2] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman andHall/CRC, 1999.

[3] Stuart J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29:175–188, 3 1976.

[4] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979.

[5] David L. Demets and K. K. Gordon Lan. Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13(13-14):1341–1352, 1994.

[6] Max Halperin, K.K. Gordon Lan, James H. Ware, Norman J. Johnson, and David L. DeMets. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials*, 3(4):311–323, 1982.

[7] David J. Spiegelhalter, Laurence S. Freedman, and Patrick R. Blackburn. Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, 7(1):8–17, 1986.

[8] Mauro Gasparini, Lilla Di Scala, Frank Bretz, and Amy Racine-Poon. Predictive probability of success in clinical drug development. *Epidemiology Biostatistics and Public Health*, 10, 03 2013.

[9] Anthony O'Hagan, John W. Stevens, and Michael J. Campbell. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201, 2005.

[10] N. Stallard and S. Todd. Sequential designs for phase iii clinical trials incorporating treatment selection. *Statistics in Medicine*, 22(5):689–703, 2003.

[11] Dominic Magirr, Thomas Jaki, and John Whitehead. A generalized dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*, 99(2):494–501, 2012.

[12] Joost van Rosmalen, David Dejardin, Yvette van Norden, Bob Löwenberg, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27:3167–3182, 10 2018.

[13] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G. Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13:41–54, 1 2014.

[14] Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67:1047–56, 9 2011.

[15] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in medicine*, 34:3724–49, 12 2015.

[16] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70:1023–1032, 12 2014.

[17] Timothy Mutsvari, Dominique Tytgat, and Rosalind Walley. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 15:28–36, 1 2016.

[18] Kaspar Rufibach, Paul Jordan, and Markus Abt. 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(2):191–201, 2014.

[19] H.C. Bucher, G.H. Guyatt, D.J. Cook, A. Holbrook, F.A. McAlister, and Evidence-Based Medicine Working Group. Users' guides to the medical literature: Xix. applying clinical trial results a. how to use an article measuring the effect of an intervention on surrogate end points. *JAMA*, 282(8):771–778, 1999.

[20] H. Poad, S. Khan, L. Wheaton, A. Thomas, M. Sweeting, and S. Bujkiewicz. The validity of surrogate endpoints in sub groups of metastatic colorectal cancer patients defined by treatment class and kras status. *Cancers (Basel)*, 14(21):5391, 2022.

[21] Marc Buyse, Tomasz Burzykowski, Kevin Carroll, Stefan Michiels, Daniel J. Sargent, Langdon L. Miller, Gary L. Elfring, Jean-Pierre Pignon, and Pascal Piedbois. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology*, 25(33):5218–5224, 2007.

[22] G. Saint-Hilary, V. Barboux, M. Pannaux, M. Gasparini, V. Robert, and G. Mastrantonio. Predictive probability of success using surrogate endpoints. *Statistics in Medicine*, 38(10):1753–1774, 2019.

[23] Hui Quan, Zhi Xu, Jun Luo, Guillaume Paux, Meehyung Cho, and Xun Chen. Utilization of treatment effect on a surrogate endpoint for planning a study to evaluate treatment effect on a final endpoint. *Pharmaceutical Statistics*, 22(4):633–649, 2023.

[24] Z. Zhang, Y. Lin, and J. Liu. Probability of study success (prss) evaluation based on multiple endpoints in late phase oncology drug development. *Statistics in Biopharmaceutical Research*, 15:675–688, 2022.

[25] Sylwia Bujkiewicz, John R. Thompson, Richard D. Riley, and Keith R. Abrams. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Statistics in Medicine*, 35(7):1063–1089, 2016.

[26] KyungMann Kim and Anastasios A. Tsiatis. Independent increments in group sequential tests: a review. *SORT: Statistics and Operations Research Transactions*, 44(2):223–264, 2020.

[27] Michael J. Daniels and Michael D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16(17):1965–1982, 1997.

[28] Institut de Recherches Internationales Servier. A randomised, open-label, multi-centre, two-arm phase 3 study comparing futuximab/modotuximab in combination with trifluridine/tipiracil to trifluridine/tipiracil single agent with a safety lead-in part in participants with kras/nras and braf wild type metastatic colorectal cancer previously treated with standard treatment and anti-egfr therapy. Servier Clinical Trials, 2024. ClinicalTrials.gov Identifier: https://clinicaltrials.gov/ct2/show/NCT05223673.

[29] D. Arnold, G. W. Prager, A. Quintela, A. Stein, S. Moreno Vera, N. Mounedji, and J. Taieb. Beyond second-line therapy in patients with metastatic colorectal cancer: A systematic review. *Annals of Oncology*, 29(4):835–856, 2018.

[30] Derek J Jonker, Chris J O'Callaghan, Christos S Karapetis, John R Zalcberg, Dongsheng Tu, Heather-Jane Au, Scott R Berry, Marianne Krahn, Timothy Price, R John Simes, Niall C Tebbutt, Guy van Hazel, Rafal Wierzbicki, Christiane Langer, and Malcolm J Moore. Cetuximab for the treatment of colorectal cancer. *The New England journal of medicine*, 357:2040–8, 11 2007. DOI: 10.1056/NEJMoa071834.

[31] Lillian L Siu, Jeremy D Shapiro, Derek J Jonker, Chris S Karapetis, John R Zalcberg, John Simes, Felix Couture, Malcolm J Moore, Timothy J Price, Jehan Siddiqui, Louise M Nott, Danielle Charpentier, Winston Liauw, Michael B Sawyer, Michael Jefford, Nadine M Magoski, Andrew Haydon, Ian Walters, Jolie Ringash, Dongsheng Tu, and Chris J O'Callaghan. Phase iii randomized, placebo-controlled study of cetuximab plus brivanib alaninate versus cetuximab plus placebo in patients with metastatic, chemotherapy-refractory, wild-type k-ras colorectal carcinoma: the ncic clinical trials group and agitg co.20 trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 31:2477–84, 7 2013. DOI: 10.1200/JCO.2012.46.0543.

[32] Timothy J Price, Marc Peeters, Tae Won Kim, Jin Li, Stefano Cascinu, Paul Ruff, Atilli Satya Suresh, Anne Thomas, Sergei Tjulandin, Kathy Zhang, Swaminathan Murugappan, and Roger Sidhu. Panitumumab versus cetuximab in patients with chemotherapy-refractory wild-type kras exon 2 metastatic colorectal cancer (aspecct): a randomised, multicentre, open-label, non-inferiority phase 3 study. *The Lancet. Oncology*, 15:569–79, 5 2014. DOI: 10.1016/S1470-2045(14)70118-4.

[33] Francesco Sclafani, Tae Y Kim, David Cunningham, Tae W Kim, Josep Tabernero, Hans J Schmoll, Jae K Roh, Sun Y Kim, Young S Park, Tormod K Guren, Eliza Hawkes, Steven J Clarke, David Ferry, Jan-Erik Frödin, Mark Ayers, Michael Nebozhyn, Clare Peckitt, Andrey Loboda, David J Mauro, and David J Watkins. A randomized phase ii/iii study of dalotuzumab in combination with cetuximab and irinotecan in chemorefractory, kras wild-type, metastatic colorectal cancer. *Journal of the National Cancer Institute*, 107:djv258, 12 2015. DOI: 10.1093/jnci/djv258.

[34] E. Van Cutsem, T. Yoshino, H.-J. Lenz, S. Lonardi, A. Falcone, M.L. Limon, M.P. Saunders, A. Sobrero, E. Maiello, Y.S. Park, R. Ferreiro Monteagudo, Y.S. Hong, J. Tomasek, H. Taniguchi, F. Ciardiello, J. Hocke, Z. Oum'hamed, S. Vlassak, M. Studeny, and J. Tabernero. gastrointestinal tumours, colorectal nintedanib plus best supportive care (bsc) versus placebo plus bsc for the treatment of patients (pts) with colorectal cancer (crc) refractory to standard therapies: Results of the phase iii lume-colon 1 study. *Annals of Oncology*, 27:vi559, 10 2016. DOI: 10.1093/annonc/mdw435.12.

[35] Jianming Xu, Tae Won Kim, Lin Shen, Virote Sriuranpong, Hongming Pan, Ruihua Xu, Weijian Guo, Sae-Won Han, Tianshu Liu, Young Suk Park, Chunmei Shi, Yuxian Bai, Feng Bi, Joong Bae Ahn, Shukui Qin, Qi Li, Changping Wu, Dong Ma, Donghu Lin,

and Jin Li. Results of a randomized, double-blind, placebo-controlled, phase iii trial of trifluridine/tipiracil (tas-102) monotherapy in asian patients with previously treated metastatic colorectal cancer: The terra study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 36:350–358, 2 2018. DOI: 10.1200/JCO.2017.74.3245.

[36] Mario M Leitao, Usha S Kreaden, Vincent Laudone, Bernard J Park, Emmanouil P Pappou, John W Davis, David C Rice, George J Chang, Emma C Rossi, April E Hebert, April Slee, and Mithat Gonen. The recourse study: Long-term oncologic outcomes associated with robotically assisted minimally invasive procedures for endometrial, cervical, colorectal, lung, or prostate cancer: A systematic review and meta-analysis. *Annals of surgery*, 277:387–396, 3 2023. DOI: 10.1097/SLA.0000000000005698.

[37] Axel Grothey, Eric Van Cutsem, Alberto Sobrero, Salvatore Siena, Alfredo Falcone, Marc Ychou, Yves Humblet, Olivier Bouché, Laurent Mineur, Carlo Barone, Antoine Adenis, Josep Tabernero, Takayuki Yoshino, Heinz-Josef Lenz, Richard M Goldberg, Daniel J Sargent, Frank Cihon, Lisa Cupit, Andrea Wagner, Dirk Laurent, and CORRECT Study Group. Regorafenib monotherapy for previously treated metastatic colorectal cancer (correct): an international, multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet (London, England)*, 381:303–12, 1 2013. DOI: 10.1016/S0140-6736(12)61900-X.

[38] Jin Li, Shukui Qin, Ruihua Xu, Thomas C C Yau, Brigette Ma, Hongming Pan, Jianming Xu, Yuxian Bai, Yihebali Chi, Liwei Wang, Kun-Huei Yeh, Feng Bi, Ying Cheng, Anh Tuan Le, Jen-Kou Lin, Tianshu Liu, Dong Ma, Christian Kappeler, Joachim Kalmus, Tae Won Kim, and CONCUR Investigators. Regorafenib plus best supportive care versus placebo plus best supportive care in asian patients with previously treated metastatic colorectal cancer (concur): a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet. Oncology*, 16:619–29, 6 2015. DOI: 10.1016/S1470-2045(15)70156-7.

[39] T Yoshino, J M Cleary, E Van Cutsem, R J Mayer, A Ohtsu, E Shinozaki, A Falcone, K Yamazaki, T Nishina, R Garcia-Carbonero, Y Komatsu, H Baba, G Argilés, A Tsuji, A Sobrero, K Yamaguchi, M Peeters, K Muro, A Zaniboni, N Sugimoto, Y Shimada, Y Tsuji, H S Hochster, T Moriwaki, B Tran, T Esaki, C Hamada, T Tanase, F Benedetti, L Makris, F Yamashita, and H-J Lenz. Neutropenia and survival outcomes in metastatic colorectal cancer patients treated with trifluridine/tipiracil in the recourse and j003 trials. *Annals of oncology : official journal of the European Society for Medical Oncology*, 31:88–95, 1 2020. DOI: 10.1016/j.annonc.2019.10.005.

[40] Per Pfeiffer, Hafdan Sorbye, Camilla Qvortrup, Mia Karlberg, Christian Kersten, Kirsten Vistisen, Birgitta Lindh, Jon Kroll Bjerregaard, and Bengt Glimelius. Maintenance therapy with cetuximab every second week in the first-line treatment of metastatic colorectal cancer: The nordic-7.5 study by the nordic colorectal cancer biomodulation group. *Clinical colorectal cancer*, 14:170–6, 9 2015. DOI: 10.1016/j.clcc.2015.03.002.

[41] Jin Li, Shukui Qin, Rui-Hua Xu, Lin Shen, Jianming Xu, Yuxian Bai, Lei Yang, Yanhong Deng, Zhen-Dong Chen, Haijun Zhong, Hongming Pan, Weijian Guo, Yongqian Shu, Ying Yuan, Jianfeng Zhou, Nong Xu, Tianshu Liu, Dong Ma, Changping Wu, Ying Cheng, Donghui Chen, Wei Li, Sanyuan Sun, Zhuang Yu, Peiguo Cao, Haihui Chen, Jiejun Wang, Shubin Wang, Hongbing Wang, Songhua Fan, Ye Hua, and Weiguo Su. Effect of fruquintinib vs placebo on overall survival in patients with previously treated metastatic colorectal cancer: The fresco randomized clinical trial. *JAMA*, 319:2486–2496, 6 2018. DOI: 10.1001/jama.2018.7855.

[42] Cathy Eng, Tae Won Kim, Johanna Bendell, Guillem Argilés, Niall C Tebbutt, Maria Di Bartolomeo, Alfredo Falcone, Marwan Fakih, Mark Kozloff, Neil H Segal, Alberto

Sobrero, Yibing Yan, Ilsung Chang, Anne Uyei, Louise Roberts, Fortunato Ciardiello, and IMblaze370 Investigators. Atezolizumab with or without cobimetinib versus regorafenib in previously treated metastatic colorectal cancer (imblaze370): a multicentre, open-label, phase 3, randomised, controlled trial. *The Lancet. Oncology*, 20:849–861, 6 2019. DOI: 10.1016/S1470-2045(19)30027-0.

[43] R Core Team. R: A language and environment for statistical computing, 2023. Accessed: 2023-10-19.

[44] Stan Development Team. Rstan: The r interface to stan, 2023. Accessed: 2023-10-19.

[45] Yun Li and Jeremy M. G. Taylor. Predicting treatment effects using biomarker data in a meta-analysis of clinical trials. *Statistics in Medicine*, 29(18):1875–1889, 2010.

[46] Jeremy Oakley and Tony O'Hagan. Shelf: The sheffield elicitation framework, 2016. Accessed: 2023-10-19.

[47] Nigel Dallow, Nicky Best, and Timothy H. Montague. Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*, 17(4):301–316, 2018.

[48] Lisa V. Hampson, John Whitehead, Despina Eleftheriou, and Paul Brogan. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33(24):4186–4201, October 2014.

[49] Lisa V. Hampson, Björn Bornkamp, Björn Holzhauer, Joseph Kahn, Markus R. Lange, Wen-Lin Luo, Giovanni Della Cioppa, Kelvin Stott, and Steffen Ballerstedt. Improving the assessment of the probability of success in late stage drug development. *Pharmaceutical Statistics*, 21(2):439–459, 2022.

[50] M. Moatti, S. Zohar, T. Facon, P. Moreau, J. Y. Mary, and S. Chevret. Modeling of experts' divergent prior beliefs for a sequential phase iii clinical trial. *Clinical Trials*, 10(4):505–514, 2013.

[51] Peng Yang, Yuansong Zhao, Lei Nie, Jonathon Vallejo, and Ying Yuan. Sam: Self-adapting mixture prior to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79(4):2857–2868, 2023.

[52] Zhen Zhang, Yifan Lin, and Jun Liu. Probability of study success (prss) evaluation based on multiple endpoints in late phase oncology drug development. *Statistics in Biopharmaceutical Research*, 15(4):675–688, 2023.

[53] Richard Simon. Optimal two-stage designs for phase ii clinical trials. *Controlled Clinical Trials*, 10(1):1–10, 1989.

[54] Sin-Ho Jung, Taiyeong Lee, KyungMann Kim, and Stephen L. George. Admissible two-stage designs for phase ii cancer clinical trials. *Statistics in Medicine*, 23(4):561–569, 2004.

[55] P. Gallo, L. Mao, and V. H. Shih. Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics*, 24(5):976–993, 2014.

[56] Dong Xi, Paul Gallo, and David Ohlssen. On the optimal timing of futility interim analyses. *Statistics in Biopharmaceutical Research*, 9(3):293–301, 2017.

[57] Xieran Li, Carolin Herrmann, and Geraldine Rauch. Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint. *BMC Medical Research Methodology*, 20(1):274, 2020.

[58] Svenja Schüler, Meinhard Kieser, and Geraldine Rauch. Choice of futility boundaries for group sequential designs with two endpoints. *BMC Medical Research Methodology*, 17(1):119, 2017.

[59] Xiaofei Wang and Stephen L. George. Futility monitoring for randomized clinical trials with non-proportional hazards: An optimal conditional power approach. *Clinical Trials*, 20(6):603–612, 2023.

[60] Frédéric Fiteni, Virginie Westeel, and Franck Bonnetain. Surrogate endpoints for overall survival in lung cancer trials: A review. *Expert Review of Anticancer Therapy*, 17(5):447–454, 2017.

[61] L. A. Gharzai, R. Jiang, D. Wallington, et al. Intermediate clinical endpoints for surrogacy in localised prostate cancer: An aggregate meta-analysis. *The Lancet Oncology*, 22(3):402–410, 2021.

[62] Marc Buyse, Geert Molenberghs, Xavier Paoletti, Koji Oba, Ariel Alonso, Wim Van der Elst, and Tomasz Burzykowski. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, 58(1):104–132, 2016.

[63] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977. DOI: 10.2307/2335684.

[64] S S Emerson and T R Fleming. Symmetric group sequential test designs. *Biometrics*, 45:905–23, 9 1989. DOI: 10.2307/2531692.

[65] FDA. *Accelerated Approval.*, 2023.

[66] FDA. *Clinical trial considerations to support Accelerated Approval of oncology therapeutics: guidance for industry.*, 2023.

[67] Lisa Belin, Aidan Tan, Yann De Rycke, and Agnès Dechartres. Progression-free survival as a surrogate for overall survival in oncology trials: a methodological systematic review. *British journal of cancer*, 122:1707–1714, 5 2020. DOI: 10.1038/s41416-020-0805-y.

[68] FDA. *Ongoing | Cancer Accelerated Approvals*, 2023.

[69] FDA. *Withdrawn | Cancer Accelerated Approvals*, 2023.

[70] FDA. *Verified Clinical Benefit | Cancer Accelerated Approvals*, 2023.

[71] David J. Spiegelhalter, Laurence S. Freedman, and Mahesh K. B. Parmar. Applying bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, 12(15-16):1501–1511, 1993.

[72] Satrajit Roychoudhury and Beat Neuenschwander. Bayesian leveraging of historical control data for a clinical trial with time-to-event endpoint. *Statistics in Medicine*, 39:984–995, 3 2020.

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[74] Gernot Wassmer and Friedrich Pahlke. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*, 2023. R package version 3.3.4.

[75] Martyn Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2023. R package version 4-14.

[76] Ronan Fougeray, Loïck Vidot, Marco Ratta, Zhaoyang Teng, Donia Skanji, and Gaëlle Saint-Hilary. Futility interim analysis based on probability of success using a surrogate endpoint. *Pharmaceutical Statistics*, 7 2024.

[77] Luca Pozzi, Heinz Schmidli, and David I. Ohlssen. A bayesian hierarchical surrogate outcome model for multiple sclerosis. *Pharmaceutical Statistics*, 15:341–348, 7 2016. DOI: 10.1002/pst.1749.

[78] FDA. *Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products*, 2020.

[79] FDA. *Adaptive Designs for Clinical Trials of Drugs and Biologics*, 2019.

[80] Nicky Best, Maxine Ajimi, Beat Neuenschwander, Gaëlle Saint-Hilary, and Simon Wandel. Beyond the classical type i error: Bayesian metrics for bayesian designs using informative priors. *Statistics in Biopharmaceutical Research*, 17:183–196, 4 2025.

[81] Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian analysis*, 7:639–674, 8 2012. DOI: 10.1214/12-BA722.

[82] Liyun Jiang, Lei Nie, and Ying Yuan. Elastic priors to dynamically borrow information from historical data in clinical trials. *Biometrics*, 79:49–60, 3 2023. DOI: 10.1111/biom.13551.

[83] Zhen Zhang, Yong Lin, and Jingyi Liu. Probability of study success (prss) evaluation based on multiple endpoints in late phase oncology drug development. *Statistics in Biopharmaceutical Research*, 15:675–688, 7 2023. DOI: 10.1080/19466315.2022.2120532.

[84] Edward Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35:174–180, 3 1964.

[85] Peter F. Thall, Richard Simon, and Susan S. Ellenberg. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75:303, 6 1988.

[86] P F Thall, R Simon, and S S Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*, 45:537–47, 6 1989.

[87] Daniel J. Schaid, Sam Wieand, and Terry M. Therneau. Optimal two-stage screening designs for survival comparisons. *Biometrika*, 77:507, 9 1990.

[88] P. Bauer and K. Kohne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50:1029, 12 1994.

[89] P Bauer and M Kieser. Combining different phases in the development of medical treatments within a single trial. *Statistics in medicine*, 18:1833–48, 7 1999.

[90] Martin Posch, Franz Koenig, Michael Branson, Werner Brannath, Cornelia Dunger-Baldauf, and Peter Bauer. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*, 24:3697–3714, 12 2005.

[91] Frank Bretz, Heinz Schmidli, Franz König, Amy Racine, and Willi Maurer. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, 48:623–634, 8 2006.

[92] Patrick J. Kelly, Nigel Stallard, and Susan Todd. An adaptive group sequential design for phase ii/iii clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, 15:641–658, 7 2005.

[93] Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50:1096, 12 1955.

[94] Nigel Stallard and Tim Friede. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*, 27:6209–6227, 12 2008.

[95] T. Friede, N. Parsons, N. Stallard, S. Todd, E. Valdes Marquez, J. Chataway, and R. Nicholas. Designing a seamless phase ii/iii clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine*, 30:1528–1540, 6 2011.

[96] Nigel Stallard. A confirmatory seamless phase ii/iii clinical trial design incorporating short-term endpoint information. *Statistics in Medicine*, 29:959–971, 4 2010.

[97] Cornelia Ursula Kunz, Tim Friede, Nick Parsons, Susan Todd, and Nigel Stallard. Data-driven treatment selection for seamless phase ii/iii trials incorporating early-outcome data. *Pharmaceutical Statistics*, 13:238–246, 7 2014.

[98] *Optimizing the dosage of human prescription drugs and biological products for the treatment of oncologic diseases: Guidance for industry. Oncology Center of Excellence (OCE), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). U.S. Department of Health and Human Services, Food and Drug Administration (2024).*

[99] Thomas Jaki and Lisa V. Hampson. Designing multi-arm multi-stage clinical trials using a risk-benefit criterion for treatment selection. *Statistics in Medicine*, 35:522–533, 2 2016.

[100] Erica Aranha Suzumura, Bruna de Oliveira Ascef, Fernando Henrique de Albuquerque Maia, Aline Frossard Ribeiro Bortoluzzi, Sidney Marcel Domingues, Natalia Santos Farias, Franciele Cordeiro Gabriel, Beate Jahn, Uwe Siebert, and Patricia Coelho de Soarez. Methodological guidelines and publications of benefit-risk assessment for health technology assessment: a scoping review. *BMJ open*, 14:e086603, 6 2024.

[101] *European Medical Agency (EMA). Benefit-risk methodology project. Work package 1 Report: description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network.*

[102] *European Medical Agency (EMA). Benefit-risk methodology project. Work package 2 report: applicability of current tools and processes for regulatory benefit-risk assessment.*

[103] *European Medical Agency (EMA). Benefit-risk methodology project. Work package 3 report: Field tests.*

[104] *European Medical Agency (EMA). Benefit-risk methodology project. Work package 4 report: benefit-risk tools and processes.*

[105] Filip Mussen, Sam Salek, and Stuart Walker. A quantitative approach to benefit-risk assessment of medicines – part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiology and Drug Safety*, 16:S2–S15, 7 2007.

[106] Kevin Marsh, Tereza Lanitis, David Neasham, Panagiotis Orfanos, and Jaime Caro. Assessing the value of healthcare interventions using multi-criteria decision analysis: A review of the literature. *PharmacoEconomics*, 32:345–365, 4 2014.

[107] Tommi Tervonen, Gert van Valkenhoef, Erik Buskens, Hans L. Hillege, and Douwe Postmus. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30:1419–1428, 5 2011.

[108] Lydie Marcelon, Thomas Verstraeten, Geraldine Dominiak-Felden, and François Simondon. Quantitative benefit–risk assessment by mcda of the quadrivalent hpv vaccine for preventing anal cancer in males. *Expert Review of Vaccines*, 15:139–148, 1 2016.

[109] Richard Nixon, Christoph Dierig, Shahrul Mt-Isa, Isabelle Stöckert, Thaison Tong, Silvia Kuhls, Gemma Hodgson, John Pears, Ed Waddingham, Kimberley Hockley, and Andrew Thomson. A case study using the proact-url and brat frameworks for structured benefit risk assessment. *Biometrical Journal*, 58:8–27, 1 2016.

[110] Gaelle Saint-Hilary, Stephanie Cadour, Veronique Robert, and Mauro Gasparini. A simple way to unify multicriteria decision analysis (mcda) and stochastic multicriteria acceptability analysis (smaa) using a dirichlet distribution in benefit–risk assessment. *Biometrical Journal*, 59:567–578, 5 2017.

[111] Henk Broekhuizen, Catharina G. M. Groothuis-Oudshoorn, A. Brett Hauber, Jeroen P. Jansen, and Maarten J. IJzerman. Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis. *BMC Medical Informatics and Decision Making*, 15:102, 12 2015.

[112] J.W. Valle, G.K. Abou-Alfa, R.K. Kelley, M.A. Lowery, R.T. Shroff, Y. Bian, G. Saint-Hilary, H. Liu, Z. Teng, Z. Hua, C. Gliser, A. Vogel, and M.M. Javle. Quantitative benefit–risk assessment of data from the phase iii claridhy study of ivosidenib versus placebo in patients with midh1 cholangiocarcinoma. *ESMO Gastrointestinal Oncology*, 8:100159, 2025.

[113] Tim Friede and Nigel Stallard. A comparison of methods for adaptive treatment selection. *Biometrical Journal*, 50:767–781, 10 2008.

[114] Walter Lehmacher and Gernot Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55:1286–1290, 12 1999.

[115] Werner Brannath, Martin Posch, and Peter Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97:236–244, 3 2002.

[116] Werner Brannath, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28:1445–1463, 5 2009.

[117] Marc Dunoyer. Accelerating access to treatments for rare diseases. *Nature Reviews Drug Discovery*, 10(7):475–476, July 2011.

[118] Julia Dunne, William J. Rodriguez, M. Dianne Murphy, B. Nhi Beasley, Gilbert J. Burckart, Jane D. Filie, Linda L. Lewis, Hari C. Sachs, Philip H. Sheridan, Peter Starke, and Lynne P. Yao. Extrapolation of Adult Data and Other Data in Pediatric Drug-Development Programs. *Pediatrics*, 128(5):e1242–e1249, November 2011.

[119] David A Schoenfeld, Hui Zheng, and Dianne M Finkelstein. Bayesian design using adult data to augment pediatric trials. *Clinical Trials*, 6(4):297–304, August 2009.

[120] Christian Röver, Simon Wandel, and Tim Friede. Model averaging for robust extrapolation in evidence synthesis. *Statistics in Medicine*, 38:674–694, 2 2019.

[121] Andre Kleyner, Shrikar Bhagath, Mauro Gasparini, Jeffrey Robinson, and Mark Bender. Bayesian techniques to reduce the sample size in automotive electronics attribute testing. *Microelectronics and Reliability*, 37(6):879–883, 1997.

[122] Anthony O'Hagan and Jonathan Forster. *Kendall's Advanced Theory of Statistics*, volume 2B. Arnold, London, 2010.

[123] Bernd Holzhauer. Methods for using aggregate historical control data in meta-analyses of clinical trials with time-to-event endpoints. *Statistics in Biopharmaceutical Research*, 12(1):107–116, 2020.

[124] Vivienn Weru, Annette Kopp-Schneider, Manuel Wiesenfarth, Sebastian Weber, and Silvia Calderazzo. Information borrowing in bayesian clinical trials: choice of tuning parameters for the robust mixture prior. 12 2024.

[125] Andrea Callegaro, Nicholas Galwey, and Juan J Abellan. Historical controls in clinical trials: a note on linking pocock's model with the robust mixture priors. *Biostatistics*, 24:443–448, 4 2023.

[126] Andrea Callegaro, Naveen Karkada, Emmanuel Aris, and Toufik Zahaf. Vaccine clinical trials with dynamic borrowing of historical controls: Two retrospective studies. *Pharmaceutical Statistics*, 22:475–491, 5 2023.

[127] Satoshi Morita, Peter F. Thall, and Peter Müller. Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602, 2008.

[128] Barry W. Brown, Jay Herson, E. Neely Atkinson, and M. Elizabeth Rozell. Projection from previous studies: A bayesian and frequentist compromise. *Controlled Clinical Trials*, 8(1):29–44, 1987.

[129] Bruno Lecoutre. Two useful distributions for bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, 79(1):93–105, 1999.

[130] Susan J. Lee and Marvin Zelen. Clinical trials and sample size considerations: Another perspective. *Statistical Science*, 15(2):95–110, 2000.

[131] T. Pham-Gia and N. Turkkan. Sample size determination in bayesian analysis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(4):389–397, 1992.

[132] Lawrence Joseph, David B. Wolfson, and Roxanne du Berger. Some comments on bayesian sample size determination. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(2):167–171, 1995.

[133] Hamed Pezeshk. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research*, 12(6):489–504, 2003.

[134] Nigel Stallard. Sample size determination for phase ii clinical trials based on bayesian decision theory. *Biometrics*, 54(1):279–294, 1998.

[135] Karl Claxton, Larry F. Lacey, and Stephen G. Walker. Selecting treatments: A decision theoretic approach. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 163(2):211–225, 2000.

[136] S. K. Sahu and T. M.F. Smith. A bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(2):235–253, 2006.

[137] Scott M. Berry, Bradley P. Carlin, J. Jack Lee, and Peter Müller. *Bayesian Adaptive Methods for Clinical Trials*, volume 38 of *Chapman & Hall/CRC Biostatistics Series*. CRC Press, Boca Raton, FL, 2010.

[138] Scott M. Berry, John T. Connor, and Robert J. Lewis. The platform trial: An efficient strategy for evaluating multiple treatments. *JAMA*, 313(16):1619–1620, 2015.

[139] Amanda J. Redig Pasi A. Jänne. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *JCO Precision Oncology*, 2015:1–5, 2015.

[140] Janet Woodcock and Lisa M. LaVange. Innovative trial designs and master protocols: Building a framework for efficient drug development. *Clinical Pharmacology & Therapeutics*, 103(1):25–28, 2018.

[141] Libby Daniells, Pavel Mozgunov, Alun Bedding, and Thomas Jaki. A comparison of bayesian information borrowing methods in basket trials and a novel proposal of modified exchangeability-nonexchangeability method. *Statistics in Medicine*, 42(24):4392–4417, 2023.

[142] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[143] Donald B. Rubin. Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):467–488, 2006.

[144] Silvia Calderazzo, Manuel Wiesenfarth, Vivienn Weru, and Annette Kopp-Schneider. Principled type i error rate inflation in two-arm clinical trial designs with external control information borrowing. *arXiv preprint arXiv:2508.16348*, 2025.

[145] W. A. Link. A model for informative censoring. *Journal of the American Statistical Association*, 84(407):749–752, 1989.

[146] S. E. Emoto and P. C. Matthews. A weibull model for dependent censoring. *The Annals of Statistics*, 18(4):1556–1577, 1990.

[147] David Dejardin, Emmanuel Lesaffre, and Geert Verbeke. Joint modeling of progression-free survival and death in advanced cancer clinical trials. *Statistics in Medicine*, 29(16):1724–1734, 2010.

[148] J.P. Fine, H. Jiang, and R. Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, 2001.

[149] Weijing Wang. Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):257–273, 2003.

[150] Limin Peng and Jason P. Fine. Regression modeling of semicompeting risks data. *Biometrics*, 63(1):96–108, 2007.

[151] Haoda Fu, Yanping Wang, Jingyi Liu, Pandurang M. Kulkarni, and Allen S. Melemed. Joint modeling of progression-free survival and overall survival by a bayesian normal induced copula estimation model. *Statistics in Medicine*, 32(2):240–254, 2013.

Appendix A

Supplementary Material - Chapter 2

PPoS derivation for the design employed

Let us consider a design with 3 interim analyses – namely IA0 and IA1 for futility and IA2 both for futility and efficacy – before the final analysis FA. Let suppose we are interested in calculating the PPoS of the trial after data at IA0 are collected, the trial success can be achieved in two situations:

- The trial does not stop for futility at IA0 and IA1 and stops for efficacy at IA2.
- The trial does not stop for futility at IA0 and IA1, does not stop either for futility nor efficacy at IA2 and achieves statistical significance at FA.

The PPoS (expressed for sake of simplicity as the complement of the rejection region) can therefore be derived as follows:

$$
\begin{aligned}
\text{PPoS} = 1 - \Big[& P\left(\hat{\theta}_{IA1} > u_{IA1} \middle| \hat{\theta}_{IA0}\right) + P\left(\hat{\theta}_{IA1} < u_{IA1} \cap \hat{\theta}_{IA2} > u_{IA2} \middle| \hat{\theta}_{IA0}\right) + \\
& + P\left(\hat{\theta}_{IA1} < u_{IA1} \cap l_{IA2} < \hat{\theta}_{IA2} < u_{IA2} \cap \hat{\theta}_{F} > u_{F} \middle| \hat{\theta}_{IA0}\right) \Big] = \\
= 1 - \Big[& P\left(\hat{\theta}_{IA1} > u_{IA1} \middle| \hat{\theta}_{IA0}\right) + \int_{-\infty}^{u_1} P\left(\hat{\theta}_{IA2} > u_{IA2} \middle| \hat{\theta}_{IA1}\right) P\left(\hat{\theta}_{IA1} = y \middle| \hat{\theta}_{IA0}\right) dy + \\
& + \int_{l_2}^{u_2} \int_{-\infty}^{u_1} P\left(\hat{\theta}_{F} > u_{F} \middle| \hat{\theta}_{IA2}\right) P\left(\hat{\theta}_{IA2} = z \middle| \hat{\theta}_{IA1}\right) P\left(\hat{\theta}_{IA1} = y \middle| \hat{\theta}_{IA0}\right) dy dz \Big] =
\end{aligned}
$$

$$
\begin{aligned}
=1-\Big[&P\left(t_{IA0}\hat{\theta}_{IA0}+t_{IA1-IA0}\hat{\theta}_{IA1-IA0}<u_{IA1}\middle|\hat{\theta}_{IA0}\right)+\\
&+\int_{-\infty}^{u_1}P\left(t_{IA1}\hat{\theta}_{IA1}+t_{IA2-IA1}\hat{\theta}_{IA2-IA1}<u_{IA2}\middle|\hat{\theta}_{IA1}\right)P\left(\hat{\theta}_{IA1}=y\middle|\hat{\theta}_{IA0}\right)dy+\\
&+\int_{l_2}^{u_2}\int_{-\infty}^{u_1}P\left(t_{IA2}\hat{\theta}_{IA2}+t_{F-IA2}\hat{\theta}_{F-IA2}<s\middle|\hat{\theta}_{IA2}\right)P\left(\hat{\theta}_{IA2}=z\middle|\hat{\theta}_{IA1}\right)P\left(\hat{\theta}_{IA1}=y\middle|\hat{\theta}_{IA0}\right)dydz\Big]
\end{aligned}
$$

$$
\begin{aligned}
=1-\Bigg[&P\left(\hat{\theta}_{IA1-IA0}<\frac{u_{IA1}-t_{IA0}\hat{\theta}_{IA0}}{t_{IA1-IA0}}\middle|\hat{\theta}_{IA0}\right)+\\
&+\int_{u_1}^{+\infty}P\left(\hat{\theta}_{IA2-IA1}<\frac{u_{IA2}-t_{IA1}\hat{\theta}_{IA1}}{t_{IA2-IA1}}\middle|\hat{\theta}_{IA1}\right)P\left(\hat{\theta}_{IA1}=y\middle|\hat{\theta}_{IA0}\right)dy+\\
&+\int_{l_2}^{u_2}\int_{-\infty}^{u_1}P\left(\hat{\theta}_{F-IA2}<\frac{s-t_{IA2}\hat{\theta}_{IA2}}{t_{F-IA2}}\middle|\hat{\theta}_{IA2}\right)P\left(\hat{\theta}_{IA2}=z\middle|\hat{\theta}_{IA1}\right)P\left(\hat{\theta}_{IA1}=y\middle|\hat{\theta}_{IA0}\right)dydz\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
=1-\Bigg[&\int_{-\infty}^{\frac{u_1-t_{IA0}\hat{\theta}_{IA0}}{t_{IA1-IA0}}}h^S_{\hat{\theta}_{IA1-IA0}}(t)dt+\\
&+\int_{-\infty}^{\frac{u_2-t_{IA1}\hat{\theta}_{IA1}}{t_{IA2-IA1}}}\int_{u_1}^{+\infty}h^S_{\hat{\theta}_{IA2-IA1}}(t)h^S_{\hat{\theta}_{IA1-IA0}}(x)dxdt+\\
&+\int_{-\infty}^{\frac{s-t_{IA2}\hat{\theta}_{IA2}}{t_{F-IA2}}}\int_{\frac{l_2-t_{IA1}x}{t_{IA2-IA1}}}^{\frac{u_2-t_{IA1}x}{t_{IA2-IA1}}}\int_{u_1}^{+\infty}h^S_{\hat{\theta}_{F-IA2}}(t)h^S_{\hat{\theta}_{IA2-IA1}}(y)h^S_{\hat{\theta}_{IA1-IA0}}(x)dxdydt\Bigg]
\end{aligned}
$$

Results varying IF of IA0

In figure S1 and S2 the comparison between information fractions (IF) IF=0.089 and IF=0.2 is made in terms of probability to continue after the early interim.

The two designs lead to a similar decrease trend in terms of Go probability, but it is shown that – due to the reduced variance of the predictive distribution - having a later early interim increases the performances of all the 3 scenarios, reducing accordingly the probability to make incorrect decision at IA0. However, it may have a non-negligible impact in terms of resources or number of patients enrolled in the trial, since the futility decision is made later

If the treatment is effective (Scenario 1,2,3) the benefit of using a surrogate prior (SURR design) is increased up to 5% in case of moderate prior data conflict and around 1% in case of minor prior data conflict, while the drop in case of large prior-data conflict is decreased up to 20%. On the other hand, using a vague prior (VAGUE design) with IA0 at IF=0.2 increases the probability to make a correct decision up to 6% not depending on the prior-data conflict level.

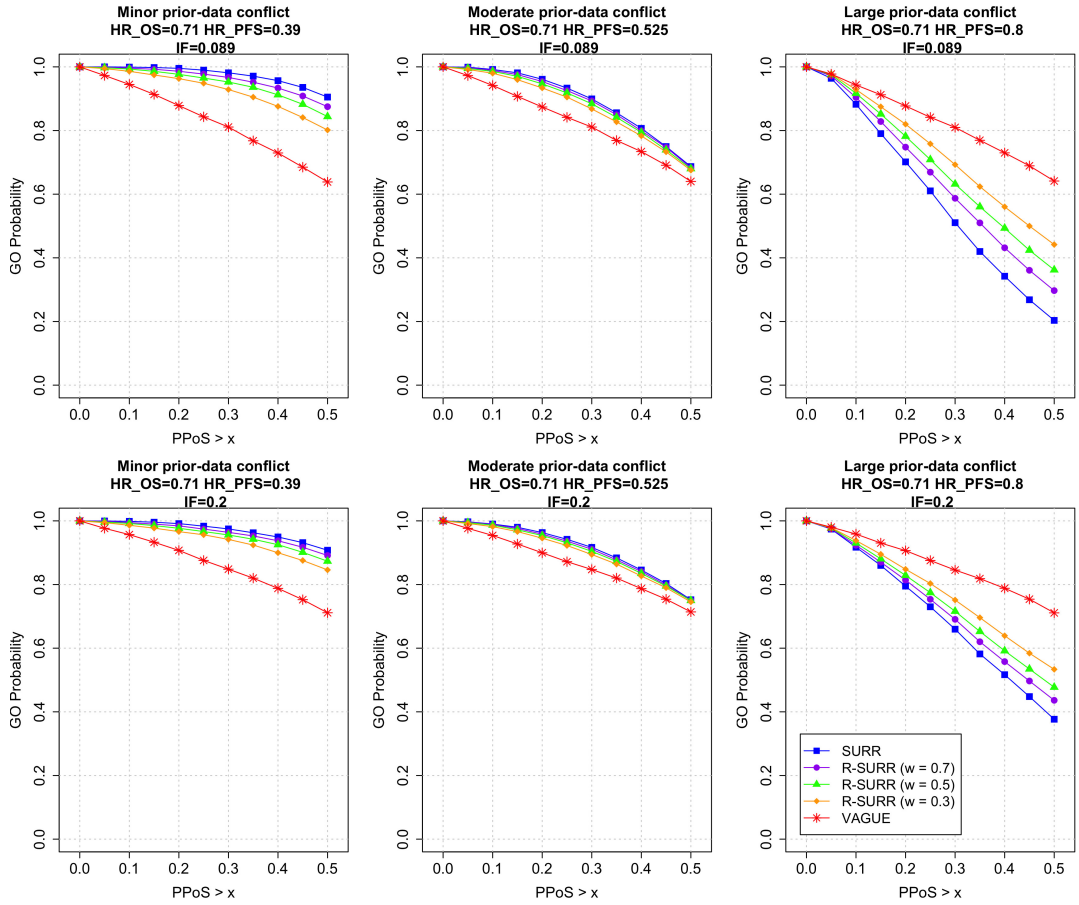


Fig. A.1 Probability to continue after IA0 for effective treatments (comparison between IF=0.089 and 0.2)

Similarly, if the treatment is ineffective (scenario 4,5,6) there is an increase of the probability of making a correct decision at IA0 of up to 6%, 15% and 20% for the surrogate prior (respectively in case of minor, moderate and large prior-data conflict) and up to 7% for the vague prior not depending on the discrepancy between historical and current data.

In terms of study power (Figure S3), the study power using a later IA0 is increased by up to 1%, 7% and 18% (in case of minor, moderate and large prior-data conflict, respectively) for SURR design and up to 7% for VAGUE design, not depending on the prior data-conflict level.

Figure S4, on the other hand, shows that a later early interim analyses is not beneficial in terms of type I error reduction. Indeed it leads to an increase of up to 11%, 6% and 3% (respectively in case of minor, moderate and large prior-data conflict) for surrogate prior and up to 5% for the vague prior. All the reasoning described above remain valid also for R-SURR design which behaviour is between VAGUE design and SURR design.

In summary, we have shown that the timing of the early interim analyses has an impact on the study behaviour: while an earlier interim permits to detect ineffective treatments early (lower type I error), a later early interim permits to reach a higher study power in all prior-data conflict levels. The timing of the interim analysis should be chosen based on the operating characteristics,

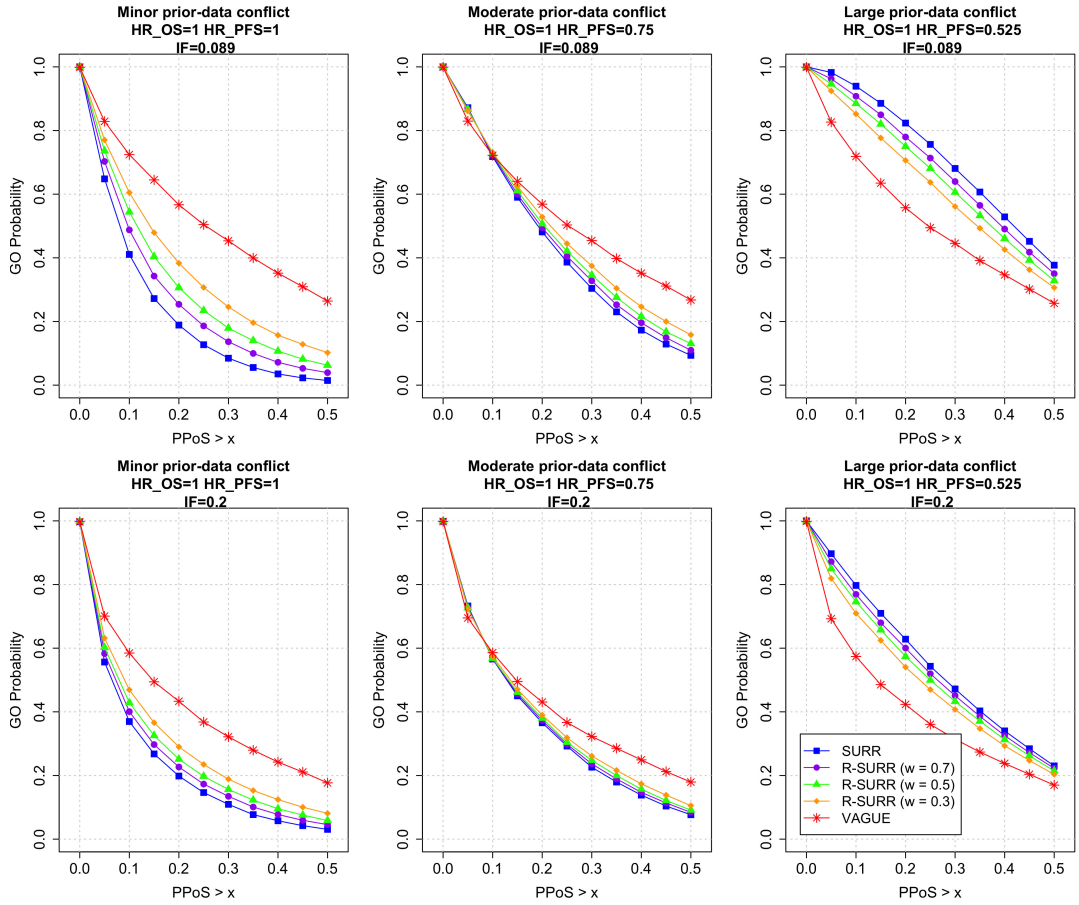


Fig. A.2 Probability to continue after IA0 for ineffective treatments (comparison between IF=0.089 and 0.2)

but also based on the number of patients enrolled in the study at this time, and other trial-specific strategic reasons.

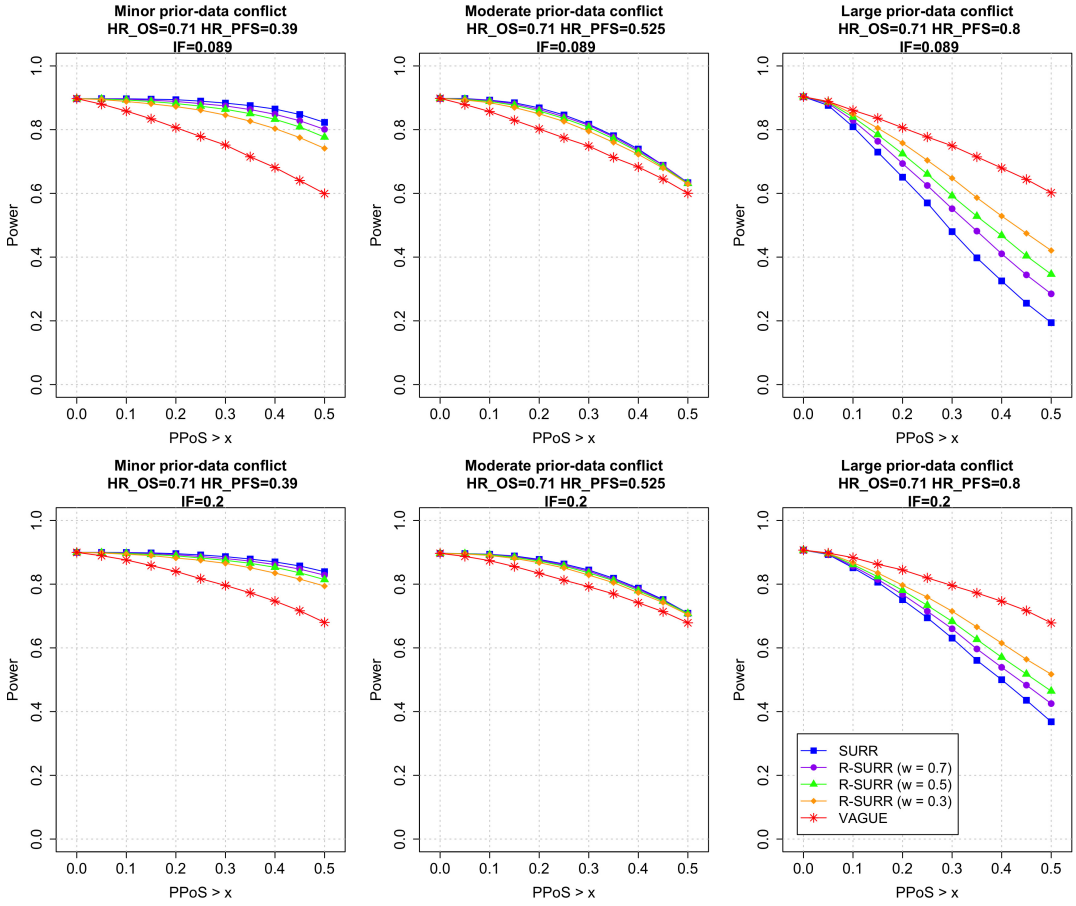


Fig. A.3 Power (comparison between IF=0.089 and 0.2)

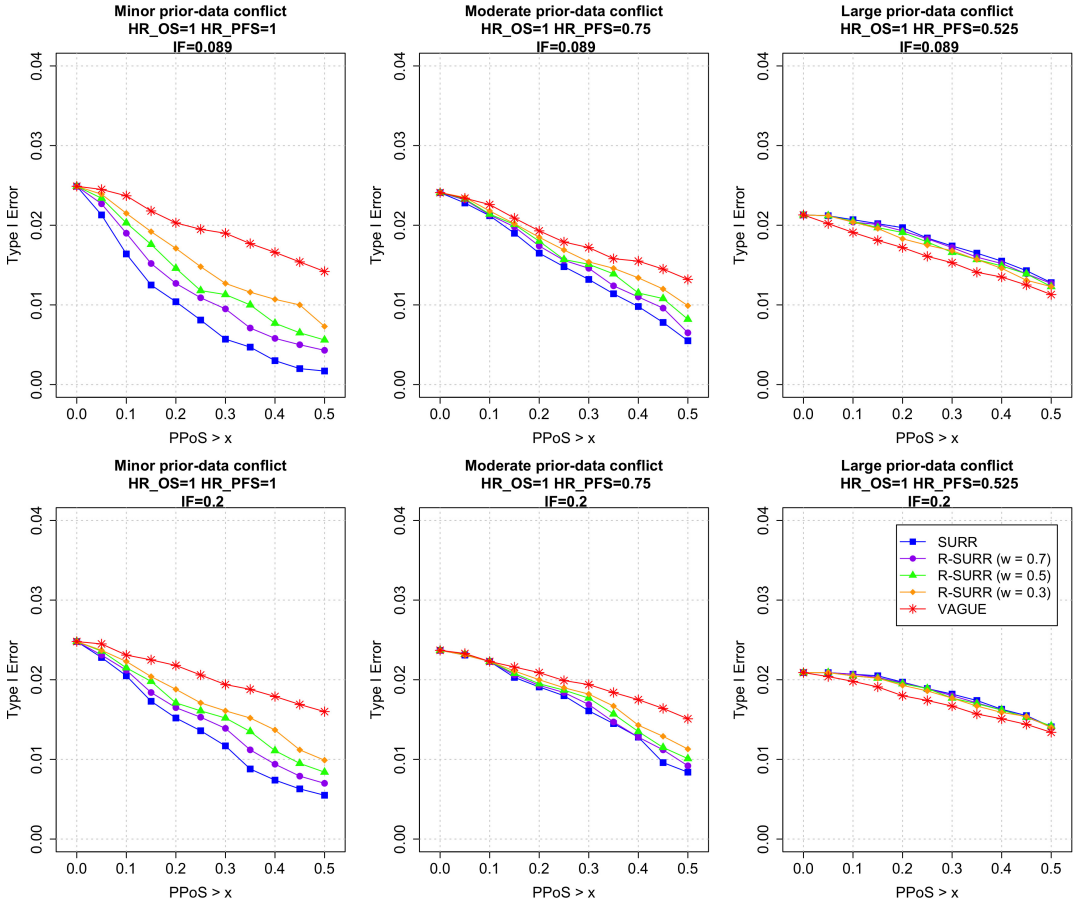


Fig. A.4 Type I error (comparison between IF=0.089 and 0.2)

Additional informations on the meta-analysis

Table A.1 Clinical trials included in the Meta-Analytic model: point estimates and 95% confidence intervals for the surrogate and primary endpoints.

Clinical Trials Included in the Meta-Analytic Model	HR(OS) [95% CI]	HR(PFS) [95% CI]
TERRA [35]	0.79 [0.62; 0.99]	0.43 [0.34; 0.54]
PFEIFFER [40]	0.55 [0.32; 0.94]	0.45 [0.29; 0.72]
RECOURSE [36]	0.68 [0.58; 0.81]	0.48 [0.41; 0.57]
UPDATE.J003 [39]	0.63 [0.45; 0.87]	0.41 [0.28; 0.59]
CORRECT [37]	0.77 [0.64; 0.94]	0.49 [0.42; 0.58]
CONCUR [38]	0.55 [0.40; 0.77]	0.31 [0.22; 0.44]
VAN.CUTSEM.15 [34]	1.00 [0.82; 1.22]	0.54 [0.44; 0.66]
JONKER.13 [30]	0.77 [0.64; 0.92]	0.68 [0.57; 0.80]
SIU.18 [31]	0.88 [0.74; 1.03]	0.72 [0.62; 0.84]
PRICE.22 [32]	0.97 [0.84; 1.11]	1.00 [0.88; 1.14]
SCLAFFANI [33]	1.41 [0.99; 2.00]	1.33 [0.98; 1.83]
VANCUTSEM [34]	1.01 [0.86; 1.19]	0.58 [0.49; 0.69]
FRESCO [41]	0.65 [0.51; 0.83]	0.26 [0.21; 0.34]
IMBLAZE370.COMBI [42]	1.00 [0.73; 1.38]	1.25 [0.94; 1.65]
IMBLAZE370.ATEZO [42]	1.19 [0.83; 1.71]	1.39 [1.00; 1.94]

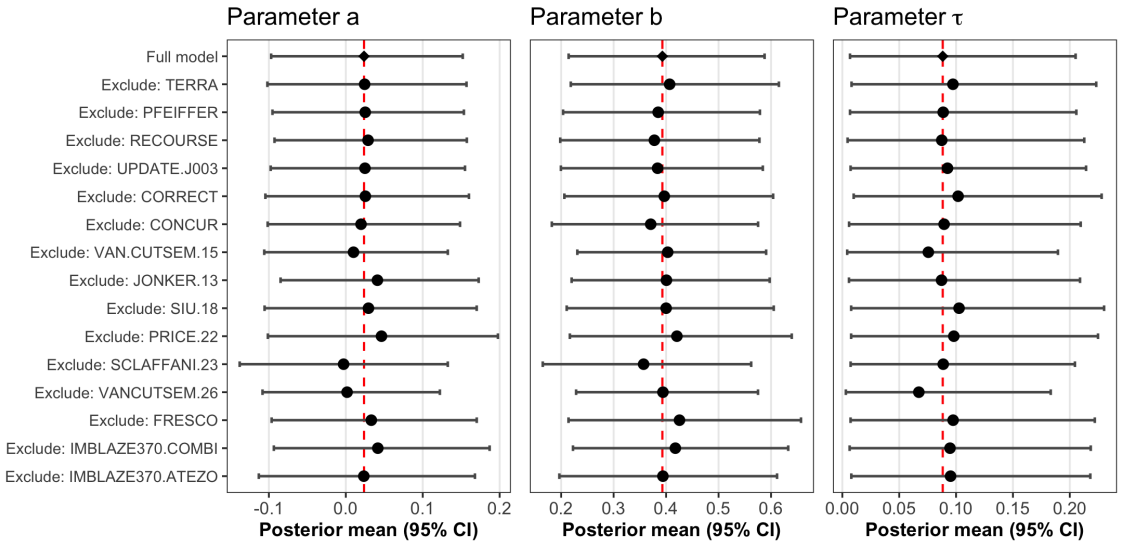


Fig. A.5 Leave-one-out sensitivity analysis for the coefficients a, b and τ of the meta-analytic regression line.

Gain in posterior ESS at IA0

Table A.2 Average gain in effective Sample Size (ESS) of the posterior distribution for the primary endpoint at the early interim analysis with respect to the vague prior analysis

Scenario	Mixture weight (w)				
	$w = 1$ (Surrogate Prior)	$w = 0.7$	$w = 0.5$	$w = 0.3$	$w = 0$ (Vague Prior)
S1	1416	1143	942	671	0
S2	1365	1093	897	635	0
S3	1396	1068	860	599	0
S4	1379	1108	910	645	0
S5	1369	1100	905	643	0
S6	1386	1077	875	614	0

Appendix B

Supplementary Material - Chapter 3

Supplementary Tables

Table B.1 Clinical trials included in the Meta-Analytic model

Clinical Trials Included in the Meta-Analytic Model	**HR(OS) [95% CI]**	**HR(PFS) [95% CI]**
TERRA [35]	0.79 [0.62; 0.99]	0.43 [0.34; 0.54]
PFEIFFER [40]	0.55 [0.32; 0.94]	0.45 [0.29; 0.72]
RECOURSE [36]	0.68 [0.58; 0.81]	0.48 [0.41; 0.57]
UPDATE.J003 [39]	0.63 [0.45; 0.87]	0.41 [0.28; 0.59]
CORRECT [37]	0.77 [0.64; 0.94]	0.49 [0.42; 0.58]
CONCUR [38]	0.55 [0.40; 0.77]	0.31 [0.22; 0.44]
VAN.CUTSEM.15 [34]	1.00 [0.82; 1.22]	0.54 [0.44; 0.66]
JONKER.13 [30]	0.77 [0.64; 0.92]	0.68 [0.57; 0.80]
SIU.18 [31]	0.88 [0.74; 1.03]	0.72 [0.62; 0.84]
PRICE.22 [32]	0.97 [0.84; 1.11]	1.00 [0.88; 1.14]
SCLAFFANI [33]	1.41 [0.99; 2.00]	1.33 [0.98; 1.83]
VANCUTSEM [34]	1.01 [0.86; 1.19]	0.58 [0.49; 0.69]
FRESCO [41]	0.65 [0.51; 0.83]	0.26 [0.21; 0.34]
IMBLAZE370.COMBI [42]	1.00 [0.73; 1.38]	1.25 [0.94; 1.65]
IMBLAZE370.ATEZO [42]	1.19 [0.83; 1.71]	1.39 [1.00; 1.94]

Table B.2 Comparison between single-criterion approach (SCA) and dual-criterion approach (DCA). Patient level correlation between PFS and OS is set to 0.45.

Scenario	Accelerated Approval Rate		Confirmation Rate		Full Approval Rate		Global type I Error Rate	
	SCA	DCA (no borrow)	SCA	DCA (no borrow)	SCA	DCA (no borrow)	SCA	DCA (no borrow)
A0 LOW	100	43.4	90.4	98.4	90.4	90.4	–	–
A1 LOW	40.1	25.0	95.0	98.8	90.4	90.4	–	–
N0 LOW	1.7	0.0	–	–	1.1	1.1	2.8	1.1
N1 LOW	96.1	1.7	–	–	1.1	1.1	96.1	2.6
A0	100	45.1	90.7	98.2	90.7	90.7	–	–
A1	44.0	27.1	96.1	99.3	90.7	90.7	–	–
N0	1.3	0.1	–	–	1.0	1.0	2.3	1.1
N1	97.0	1.4	–	–	1.0	1.0	97.0	2.2
A0 HIGH	100	47.0	90.9	98.7	90.9	90.9	–	–
A1 HIGH	46.8	28.7	94.7	99.7	90.9	90.9	–	–
N0 HIGH	1.5	0.3	–	–	1.2	1.2	2.7	1.5
N1 HIGH	98.6	1.5	–	–	1.2	1.2	98.6	2.4

Table B.3 Comparison between the Dual-Criterion Approach without historical borrowing (*no borrow*) and with historical borrowing (*borrow*).Patient level correlation between PFS and OS is set to 0.45.

Scenario	Accelerated Approval Rate		Confirmation Rate		Full Approval Rate		Global type I Error Rate	
	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)	DCA (no borrow)	DCA (borrow)
A0 LOW	43.4	64.5	98.4	97.4	90.4	90.4	–	–
A1 LOW	25.0	20.8	98.8	98.6	90.4	90.4	–	–
N0 LOW	0.0	0.0	–	–	1.1	1.1	2.8	1.1
N1 LOW	1.7	1.3	–	–	1.1	1.1	2.6	2.2
A0	45.1	73.1	98.2	97.1	90.7	90.7	–	–
A1	27.1	25.6	99.3	99.6	90.7	90.7	–	–
N0	0.1	0.0	–	–	1.0	1.0	1.1	1.0
N1	1.4	1.6	–	–	1.0	1.0	2.2	2.4
A0 HIGH	47.0	77.9	98.7	96.8	90.9	90.9	–	–
A1 HIGH	28.7	29.2	97.7	99.7	90.9	90.9	–	–
N0 HIGH	0.3	0.0	–	–	1.2	1.2	1.5	1.2
N1 HIGH	1.5	2.1	–	–	1.2	1.2	2.4	3.0

Supplementary Figures

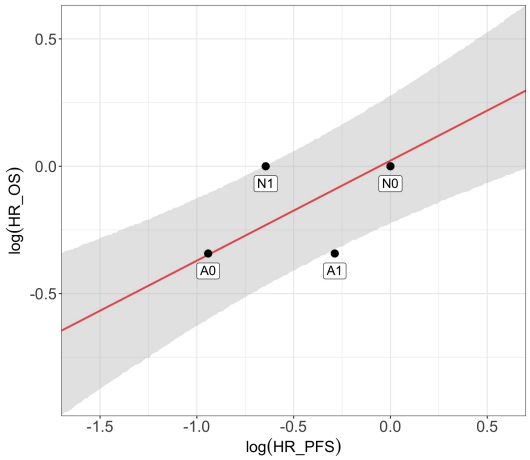


Fig. B.1 Graphical representation of the considered scenarios with respect to the meta-analytic regression line. Scenarios close to the red line (A0 and N0) are in accordance with the historical information, scenarios far from the red line (A1, N1) are in conflict with historical information.

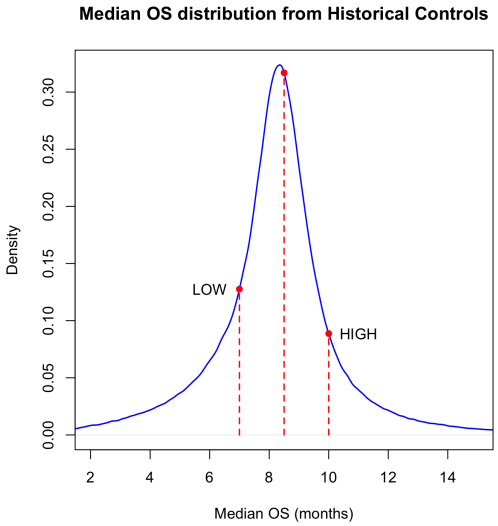


Fig. B.2 Graphical representation of the considered scenarios with respect to the meta-analytic distribution of the median OS from the historical control informations. Scenarios referred with "LOW" present a lower median OS with respect to the one expected from the historical MAP while scenarios referred with "HIGH" present an higher median OS with respect to the one expected from the historical MAP. Scenarios with no label do not present relevant drift in median OS with respect to the historical MAP.

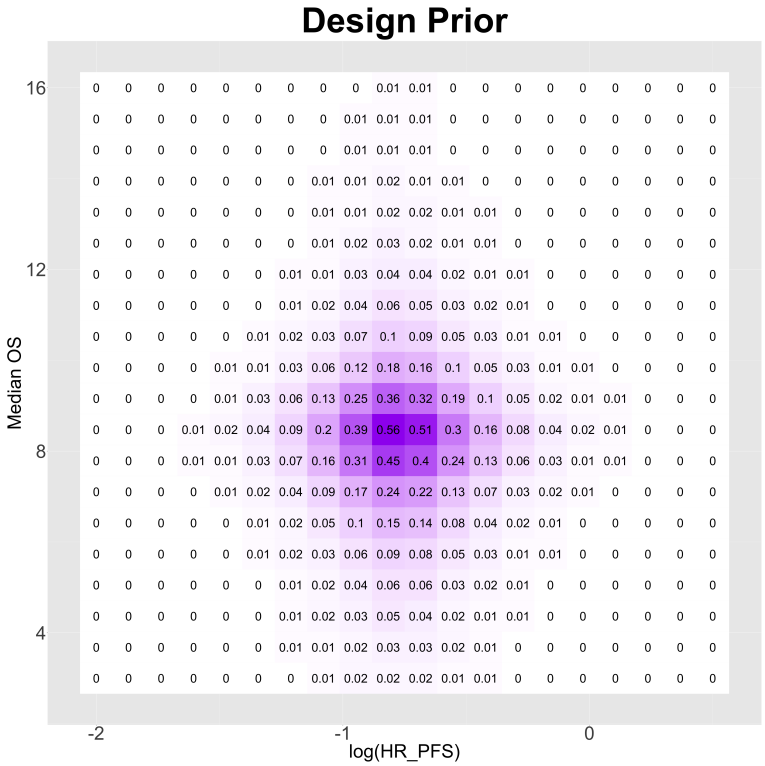


Fig. B.3 Graphical representation of the bi-variate Design prior density on the simulation grid.

(a) (b)

Fig. B.4 Maximum Accelerated Approval Power max(AA-Pow) computed for different pairs of the prior mixture weights (w_h, w_s) in the set $\mathcal{W} = (0.1, 0.3, 0.5, 0.7, 0.9)$

Appendix C

Supplementary Material - Chapter 4

Proof of Theorem 1

Proof. If θ_i^L and θ_i^U are chosen so that the posterior probabilities $\mathbb{P}\left(\theta_i^L < \theta_{ji} < \theta_i^U \,|\hat{\theta}_{ji}\right) \approx 1 \;\; \forall i = 1,\ldots,I \;\; \forall j = 0,\ldots,J$, then the linear partial value function $u(\cdot)$ takes the following form

$$u(\theta_{ji}) = \frac{\theta_{ji} - \theta_i^L}{\theta_i^U - \theta_i^L}$$

which due to the normality of $\hat{\theta}_{ji}$, is normally distributed with mean $\frac{\theta_{ji}-\theta_i^L}{\theta_i^U-\theta_i^L}$ and variance $\frac{\sigma_{ji}^2}{(\theta_i^U-\theta_i^L)^2}$. The distributions of the MCDA scores are then linear combinations of normally distributed random variables, thus are normally distributed themselves $MCDA_j \sim N\left(M_j\,,\,V_j\right)$, with mean and variance given by the following expressions

$$M_j = \sum_{i=1}^{I} \frac{\omega_i\left(\theta_{ji} - \theta_i^L\right)}{\theta_i^U - \theta_i^L} \qquad V_j = \sum_{p=1}^{I}\sum_{q=1}^{I} \frac{\rho_{pq}\omega_p\omega_q\sigma_{jq}\sigma_{jp}}{\left(\mu_p^U - \mu_p^L\right)\left(\mu_q^U - \mu_q^L\right)}. \tag{C.1}$$

Accordingly, $\Delta MCDA_j = MCDA_j - MCDA_0$ follows a normal distribution $N\left(M_j - M_j\,,\,V_j + V_j\right)$, and from Equation 4.4 follows that:

$$P_j = \mathbb{P}\left(\Delta MCDA_j > 0\right) = \Phi\left(\frac{M_j - M_0}{\sqrt{V_j + V_0}}\right) = \Phi\left(\gamma_j\right)$$

The statement follows from the fact that the normal CDF $\Phi(\cdot)$ is monotonically increasing. □

Extended expression for γ_j

$$\hat{\gamma}_j = \frac{\sum_{i=1}^{I} \omega_i \left[\prod_{k\neq i}\left(\theta_k^U - \theta_k^L\right)\right]\left(\hat{\theta}_{ji} - \hat{\theta}_{0i}\right)}{\sqrt{\sum_{p=1}^{I}\sum_{q=1}^{I}\rho_{pq}\omega_p\omega_q\left[\prod_{k\neq p}\left(\theta_k^U - \theta_k^L\right)\right]\left[\prod_{k\neq q}\left(\theta_k^U - \theta_k^L\right)\right]\left(\sigma_{jq}\sigma_{jp} + \sigma_{0q}\sigma_{0p}\right)}},$$

Analytical formula for the Power

$$\Pi^{C_j^P} = \frac{1}{K} \times \iiiint \prod_{k\neq j} \Phi\left(\frac{y + (h-y)\sqrt{\frac{V_0+V_j}{V_0+V_k}} - M_k^0}{\sqrt{V_k^0}}\right)\left[1 - \Phi\left(\frac{\eta - t_j z}{\sqrt{1-t_j}}\right)\right]$$
$$\phi\left(\frac{h - M_j - \frac{c_j}{\sigma_{j1}^2}\left(z\sqrt{\sigma_{j1}^2 + \sigma_{01}^2} + x - \theta_{j1}\right)}{\sqrt{\left(V_j\left(1 - \frac{c_j^2}{\sigma_{j1}^2 V_j}\right)\right)}}\right)\phi\left(\frac{y - M_0 - \frac{c_0}{\sigma_{01}^2}(x - \theta_{01})}{\sqrt{\left(V_0\left(1 - \frac{c_0^2}{\sigma_{01}^2 V_0}\right)\right)}}\right)$$
$$\phi\left(\frac{\frac{\sigma^\dagger z}{t} + x - \theta_{j1}}{\sigma_{j1}^{IA}}\right)\phi\left(\frac{x - \theta_{01}}{\sigma_{01}^{IA}}\right) dx dy dh dz$$

where

$$K = \sigma_{j1}^{IA}\sigma_{01}^{IA}\sqrt{V_0 V_j\left(1 - \frac{c_0^2}{\sigma_{01}^2 V_0}\right)\left(1 - \frac{c_j^2}{\sigma_{j1}^2 V_j}\right)} \qquad \text{and} \qquad M_k^0 = M_0 + \gamma^0\sqrt{V_k + V_0}$$

$\mathbb{E}[t]$	delay	accrual rate	γ_1	γ_2	n	d	$\mathbb{E}$[Pow]
0.2	2	60	1.04	2.65	171	166	0.771
0.2	2	60	1.05	2.67	174	169	0.779
0.2	2	60	1.06	2.70	177	172	0.786
0.2	2	60	1.07	2.72	180	176	0.796
0.2	2	60	1.08	2.74	183	179	0.803
0.2	2	60	1.09	2.77	186	182	0.810
0.2	2	60	1.10	2.79	189	186	0.818

Table C.1 Numerical investigation of the number of evaluable patients for ORR (n) and the number of OS events (d) in the considered setting.

Table C.2 Simulated scenarios for the Type I Error Analysis assuming dose 1 and dose 2 share the same benefit-risk profile ($\gamma_1 = \gamma_2$). For the control arm $j = 0$, the parameters are fixed to $\lambda_{01} = 0.10$, $\lambda_{02} = 0.30$, $\lambda_{03} = 0.30$, $\lambda_{04} = 0.30$.

Equal benefit-risk scores and equal characteristics for the two active doses ($\gamma_1 = \gamma_2$, $\lambda_{1i} = \lambda_{2i}$ $\forall i = 2,3,4$)					
Scenario	λ_{21}	λ_{22}	λ_{23}	λ_{24}	$\gamma_1 = \gamma_2$
1	0.10 (fixed)	0.30	0.30	0.30	0.000
2		0.40	0.30	0.30	0.79
3		0.60	0.30	0.30	2.24
4		0.30	0.50	0.30	-0.94
5		0.40	0.50	0.30	-0.16
6		0.60	0.50	0.30	1.30
7		0.30	0.70	0.30	-1.86
8		0.40	0.70	0.30	-1.10
9		0.60	0.70	0.30	0.35
10		0.30	0.30	0.50	-0.94
11		0.40	0.30	0.50	-0.16
12		0.60	0.30	0.50	2.08
13		0.30	0.50	0.50	1.30
14		0.40	0.50	0.50	-1.89
15		0.60	0.50	0.50	-1.12
16		0.30	0.70	0.50	0.35
17		0.40	0.70	0.50	-2.81
18		0.60	0.70	0.50	-0.60
19		0.30	0.30	0.70	-1.86
20		0.40	0.30	0.70	-1.10
21		0.60	0.30	0.70	0.35
22		0.30	0.50	0.70	-2.81
23		0.40	0.50	0.70	-2.06
24		0.60	0.50	0.70	0.60
25		0.30	0.70	0.70	-3.72
26		0.40	0.70	0.70	-2.99
27		0.60	0.70	0.70	-1.54
Different characteristics for the two active doses, leading to the same benefit-risk scores ($\gamma_1 = \gamma_2 = 0.00$, $\lambda_{11} = 0.10$, $\lambda_{12} = 0.30$, $\lambda_{13} = 0.30$, $\lambda_{14} = 0.30$)					
Scenario	λ_{21}	λ_{22}	λ_{23}	λ_{24}	$\gamma_1 = \gamma_2$
28	0.10 (fixed)	0.35	0.36	0.32	0.00
29		0.49	0.32	0.36	0.00
30		0.49	0.48	0.42	0.00
31		0.49	0.42	0.48	0.00
32		0.69	0.65	0.68	0.00
33		0.69	0.68	0.65	0.00
34		0.56	0.70	0.31	0.00
35		0.56	0.31	0.70	0.00
36		0.48	0.56	0.33	0.00
37		0.48	0.33	0.56	0.00
38		0.63	0.44	0.68	0.00
39		0.63	0.68	0.44	0.00
40		0.56	0.31	0.70	0.00
41		0.56	0.70	0.31	0.00

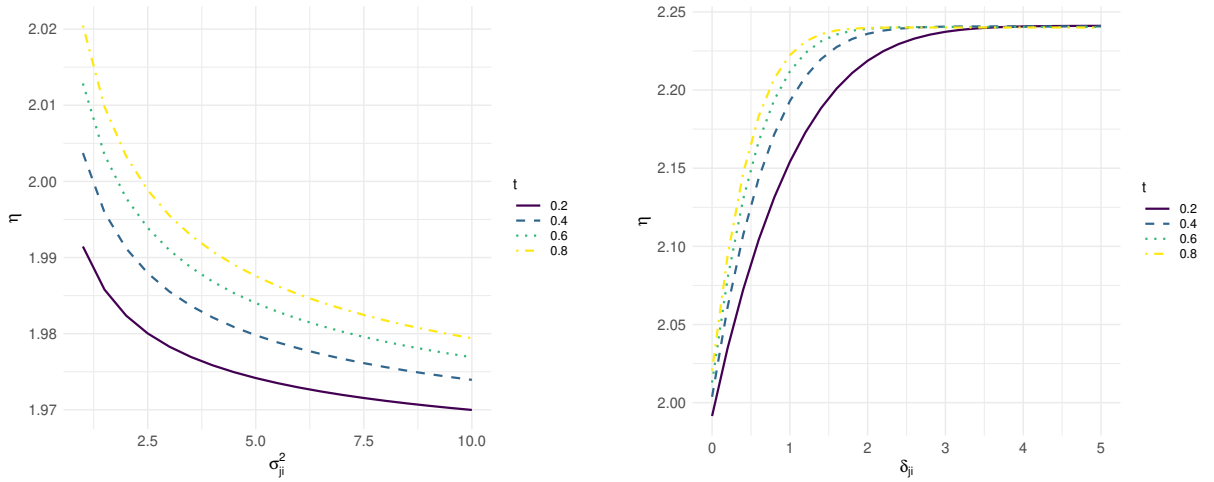
Table C.3 Simulated scenarios for the Type I Error Analysis assuming dose 1 and dose 2 have different benefit-risk profiles ($\gamma_1 \neq \gamma_2$). For the control arm $j = 0$, the parameters are fixed to $\lambda_{01} = 0.10$, $\lambda_{02} = 0.30$, $\lambda_{03} = 0.30$, $\lambda_{04} = 0.30$.

Scenario	**Dose $j = 1$**					**Dose $j = 2$**				
	λ_{11}	λ_{12}	λ_{13}	λ_{14}	γ_1	λ_{21}	λ_{22}	λ_{23}	λ_{24}	γ_2
1							0.30	0.30	0.30	0.000
2							0.40	0.30	0.30	0.79
3							0.60	0.30	0.30	2.24
4							0.30	0.50	0.30	-0.94
5							0.40	0.50	0.30	-0.16
6							0.60	0.50	0.30	1.30
7							0.30	0.70	0.30	-1.86
8							0.40	0.70	0.30	-1.10
9							0.60	0.70	0.30	0.35
10							0.30	0.30	0.50	-0.94
11							0.40	0.30	0.50	-0.16
12							0.60	0.30	0.50	2.08
13							0.30	0.50	0.50	1.30
14	0.10 (fixed)	0.30 (fixed)	0.30 (fixed)	0.30 (fixed)	0 (fixed)	0.10 (fixed)	0.40	0.50	0.50	-1.89
15							0.60	0.50	0.50	-1.12
16							0.30	0.70	0.50	0.35
17							0.40	0.70	0.50	-2.81
18							0.60	0.70	0.50	-0.60
19							0.30	0.30	0.70	-1.86
20							0.40	0.30	0.70	-1.10
21							0.60	0.30	0.70	0.35
22							0.30	0.50	0.70	-2.81
23							0.40	0.50	0.70	-2.06
24							0.60	0.50	0.70	0.60
25							0.30	0.70	0.70	-3.72
26							0.40	0.70	0.70	-2.99
27							0.60	0.70	0.70	-1.54

Nota: For the control arm $j = 0$, the parameters are fixed to $\lambda_{01} = 0.1$, $\lambda_{02} = 0.3$, $\lambda_{03} = 0.3$, $\lambda_{04} = 0.3$.

Table C.4 Simulated scenarios for the Power Analysis. For the control arm $j = 0$, the parameters are fixed to $\lambda_{01} = 0.10$, $\lambda_{02} = 0.30$, $\lambda_{03} = 0.30$, $\lambda_{04} = 0.30$.

Scenario	**Dose $j = 1$**					**Dose $j = 2$**				
	λ_{11}	λ_{12}	λ_{13}	λ_{14}	γ_1	λ_{21}	λ_{22}	λ_{23}	λ_{24}	γ_2
1						0.08	0.3	0.3	0.3	0.285
2						0.06	0.3	0.3	0.3	0.50
3						0.08	0.4	0.3	0.3	1.07
4						0.06	0.4	0.3	0.3	1.30
5						0.08	0.6	0.3	0.3	2.52
6						0.06	0.6	0.3	0.3	2.74
7						0.08	0.3	0.5	0.3	-0.66
8						0.06	0.3	0.5	0.3	-0.43
9						0.08	0.4	0.5	0.3	0.13
10						0.06	0.4	0.5	0.3	0.36
11						0.08	0.6	0.5	0.3	1.59
12						0.06	0.6	0.5	0.3	-1.81
13						0.08	0.3	0.7	0.3	-1.58
14						0.06	0.3	0.7	0.3	-1.36
15						0.08	0.4	0.7	0.3	-0.82
16						0.06	0.4	0.7	0.3	-0.59
17						0.08	0.6	0.7	0.3	0.63
18						0.06	0.6	0.7	0.3	0.85
19						0.08	0.3	0.3	0.5	-0.66
20						0.06	0.3	0.3	0.5	-0.43
21						0.08	0.4	0.3	0.5	0.13
22						0.06	0.4	0.3	0.5	0.36
23						0.08	0.6	0.3	0.5	1.59
24						0.06	0.6	0.3	0.5	1.81
25						0.08	0.3	0.5	0.5	-1.61
26						0.06	0.3	0.5	0.5	-1.38
27	0.08	0.3	0.3	0.3	0.28	0.08	0.4	0.5	0.5	-0.83
28						0.06	0.4	0.5	0.5	-0.60
29						0.08	0.6	0.5	0.5	0.63
30						0.06	0.6	0.5	0.5	0.87
31						0.08	0.3	0.7	0.5	-2.53
32						0.06	0.3	0.7	0.5	-2.31
33						0.08	0.4	0.7	0.5	-1.78
34						0.06	0.4	0.7	0.5	-1.55
35						0.08	0.6	0.7	0.5	-0.32
36						0.06	0.6	0.7	0.5	0.09
37						0.08	0.3	0.3	0.7	-1.58
38						0.06	0.3	0.3	0.7	-1.36
39						0.08	0.4	0.3	0.7	-0.82
40						0.06	0.4	0.3	0.7	-0.59
41						0.08	0.6	0.3	0.7	0.63
42						0.06	0.6	0.3	0.7	0.85
43						0.08	0.3	0.5	0.7	-2.53
44						0.06	0.3	0.5	0.7	-2.31
45						0.08	0.4	0.5	0.7	-1.78
46						0.06	0.4	0.5	0.7	-1.55
47						0.08	0.6	0.5	0.7	-0.32
48						0.06	0.6	0.5	0.7	-0.09
49						0.08	0.3	0.7	0.7	-3.44
50						0.06	0.3	0.7	0.7	-3.22
51						0.08	0.4	0.7	0.7	-2.71
52						0.06	0.4	0.7	0.7	-2.48
53						0.08	0.6	0.7	0.7	-1.26
54						0.06	0.6	0.7	0.7	-1.04



(a) (b)

Fig. C.1 Panel (a): critical value η as a function of the variances σ^2_{ji} (considered equal $\forall\ (j,i)$. We considered $\hat{\theta}_{ji} \sim N(0,\sigma^2_{ji})\ \forall j = 1,\ldots,J$. Panel (b) critical value η as a function of the treatment differences δ_{2i}. We considered $\hat{\theta}_{ji} \sim N(0,1)\ \forall i,\ \forall j = 0,\ldots,J-1$, while for arm J we consider $\hat{\theta}_{Ji} \sim N(\delta,1)\ \forall i$. In both cases $\theta^U_i = 3$ and $\theta^L_i = -3\ \forall i = 1,\ldots,4$, and the vector of weights $\boldsymbol{\omega} = (0.1, 0.4, 0.25, 0.25)$

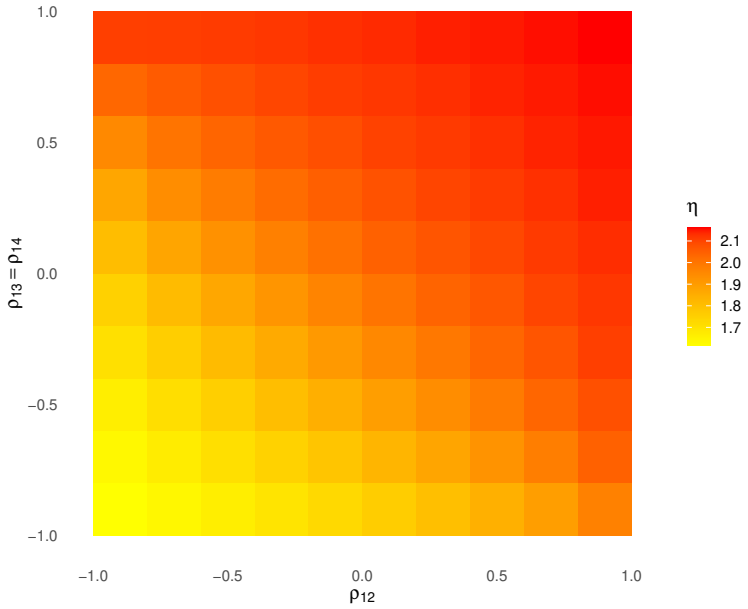


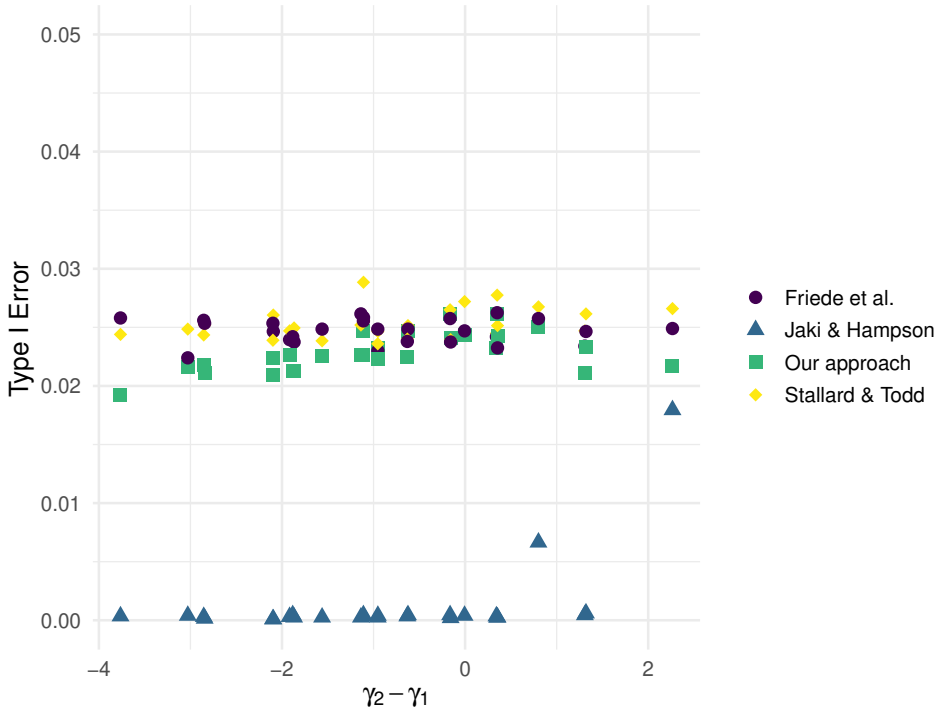
Fig. C.2 Critical value η for different levels of correlations. The correlations not displayed are fixed to $\rho_{hk} = \rho_{h1}\rho_{k1}$ in order to have a positive definite covariance matrix. We considered $\hat{\theta}_{ji} \sim N(0,1)$ $\forall j = 0,\ldots,J$, $\theta^U_i = 3$ and $\theta^L_i = -3\ \forall i = 1,\ldots,4$, and the vector of weights $\boldsymbol{\omega} = (0.1, 0.4, 0.25, 0.25)$

Table C.5 Type I error under various choices of weights ω and information fractions t. Results are obtained generating 10^6 trials, considering $\hat{\theta}_{ji} \sim N(0,1)\ \forall j = 1,\dots,J$, $\theta_i^U = 3$, $\theta_i^L = -3\ \forall i = 1,\dots,4$ and $\rho_{hk} = 0$ $\forall h = 1,\dots,J$, $k \neq h$.

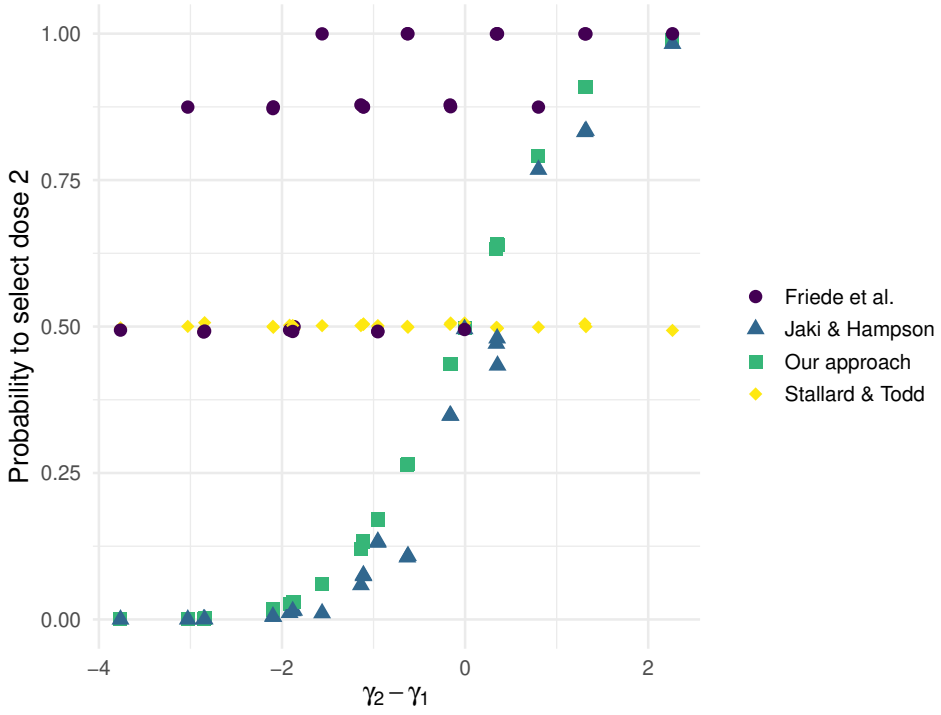
		$J=2$		$J=3$	
$(\omega_1,\omega_2,\omega_3,\omega_4)$	t	η	T1E	η	T1E
(0.7, 0.1, 0.1, 0.1)	0.1	2.068	0.0251	2.123	0.0248
	0.3	2.129	0.0250	2.218	0.0250
	0.5	2.162	0.0249	2.271	0.0253
	0.7	2.185	0.0250	2.306	0.0252
	0.9	2.201	0.0249	2.331	0.0250
(0.4, 0.2, 0.2, 0.2)	0.1	2.046	0.0249	2.090	0.0251
	0.3	2.098	0.0249	2.171	0.0250
	0.5	2.129	0.0253	2.218	0.0250
	0.7	2.151	0.0252	2.252	0.0252
	0.9	2.168	0.0252	2.279	0.0252
(0.25, 0.25, 0.25, 0.25)	0.1	2.019	0.0249	2.049	0.0250
	0.3	2.057	0.0252	2.107	0.0248
	0.5	2.081	0.0251	2.144	0.0254
	0.7	2.099	0.0252	2.172	0.0249
	0.9	2.114	0.0250	2.194	0.0252
(0.1, 0.4, 0.25, 0.25)	0.1	1.982	0.0248	1.994	0.0249
	0.3	1.998	0.0249	2.018	0.0253
	0.5	2.009	0.0250	2.033	0.0252
	0.7	2.017	0.0250	2.046	0.0251
	0.9	2.024	0.0251	2.057	0.0249

Table C.6 Type I error under various choices of weights ω and information fractions t. Results are obtained generating 10^6 trials, considering $\hat{\theta}_{ji} \sim N(0,1)\ \forall j = 1,\dots,J$, $\theta_i^U = 3$ and $\theta_i^L = -3\ \forall i = 1,\dots,4$. Correlations $\rho_{12} = \rho_{21} = 0.3$ are assumed, while all the other correlations are set to 0.

		$J=2$		$J=3$	
$(\omega_1,\omega_2,\omega_3,\omega_4)$	t	η	T1E	η	T1E
(0.7, 0.1, 0.1, 0.1)	0.1	2.068	0.0250	2.124	0.0250
	0.3	2.129	0.0250	2.218	0.0249
	0.5	2.163	0.0249	2.271	0.0249
	0.7	2.185	0.0249	2.307	0.0250
	0.9	2.201	0.0250	2.332	0.0252
(0.4, 0.2, 0.2, 0.2)	0.1	2.051	0.0252	2.098	0.0252
	0.3	2.106	0.0249	2.181	0.0252
	0.5	2.137	0.0250	2.231	0.0250
	0.7	2.159	0.0251	2.265	0.0249
	0.9	2.176	0.0253	2.292	0.0249
(0.25, 0.25, 0.25, 0.25)	0.1	2.031	0.0250	2.067	0.0251
	0.3	2.075	0.0249	2.134	0.0251
	0.5	2.102	0.0251	2.176	0.0253
	0.7	2.122	0.0252	2.207	0.0250
	0.9	2.138	0.0252	2.232	0.0250
(0.1, 0.4, 0.25, 0.25)	0.1	2.007	0.0253	2.030	0.0247
	0.3	2.038	0.0249	2.077	0.0251
	0.5	2.055	0.0252	2.108	0.0250
	0.7	2.073	0.0247	2.131	0.0250
	0.9	2.085	0.0252	2.150	0.0251



(a)



(b)

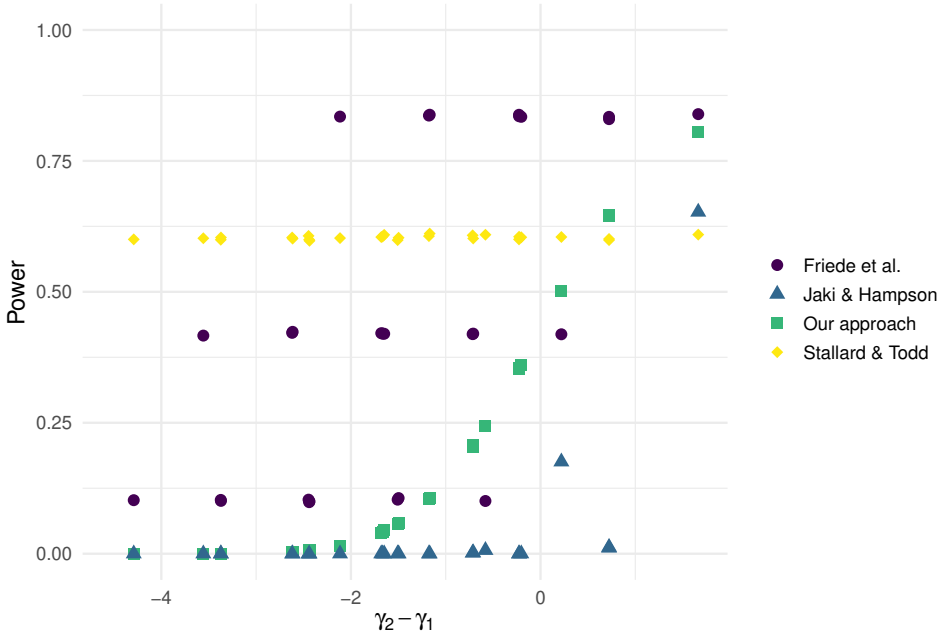
Fig. C.3 Summary results for the 27 scenarios simulated in the type I Error analysis. Panel (C.3a): type I error for all the competing approaches at varying levels of γ_2. Panel (C.3b): probability to select treatment 2 at varying levels of γ_2.

Table C.7 Analysis of the same data as in Table 3 (scenario B1-B6), using the probabilistic MCDA criterion for dose selection. The parameters used for the control arm and dose 1, shared across scenarios, are $\lambda_{01} = \lambda_{11} = 0.1$, $\lambda_{02} = \lambda_{12} = 0.3$, $\lambda_{03} = \lambda_{13} = 0.3$, $\lambda_{04} = \lambda_{14} = 0.3$

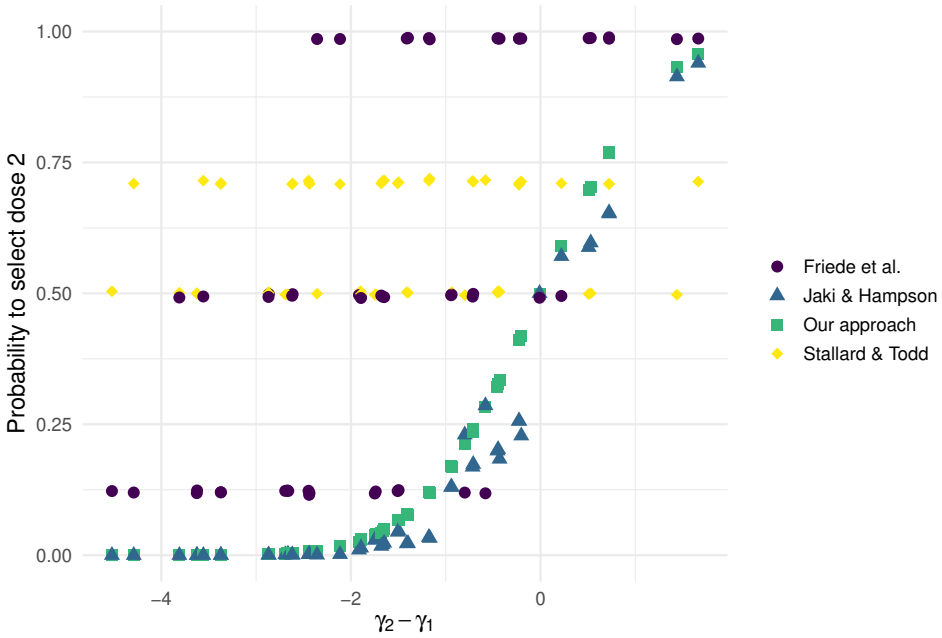
Scenario	**Parameters ($j = 2$)**					**Approaches**	**Metrics**						
	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	γ_2		**%Stop**	**%Sel2**	**T1E**	**Pow**	**Unint. Succ.**	**Tox1**	**Tox2**
B1	0.1	0.3	0.3	0.3	0.000	MCDA-based	-	49.6	0.025	-	-	0.30	0.30
						Stallard & Todd	-	50.6	0.027	-	-	0.30	0.30
						Jaki & Hampson	-	49.6	0.000	-	-	0.30	0.30
						Friede	-	49.5	0.025	-	-	0.30	0.30
B2	0.1	0.3	0.5	0.7	-2.85	MCDA-based	-	0.1	0.021	-	-	0.30	0.30
						Stallard & Todd	-	50.7	0.025	-	-	0.40	0.50
						Jaki & Hampson	-	0.000	0.000	-	-	0.30	0.30
						Friede	-	49.2	0.025	-	-	0.40	0.50
B3	0.1	0.4	0.5	0.3	-0.16	MCDA-based	-	43.1	0.024	-	-	0.39	0.30
						Stallard & Todd	-	50.6	0.024	-	-	0.40	0.30
						Jaki & Hampson	-	34.80	0.000	-	-	0.37	0.30
						Friede	-	87.5	0.024	-	-	0.48	0.30
B4	0.1	0.4	0.7	0.5	-2.10	MCDA-based	-	1.4	0.022	-	-	0.31	0.30
						Stallard & Todd	-	49.9	0.026	-	-	0.50	0.40
						Jaki & Hampson	-	0.5	0.000	-	-	0.30	0.30
						Friede	-	87.5	0.025	-	-	0.65	0.48
B5	0.1	0.6	0.3	0.5	-3.77	MCDA-based	-	92.0	0.023	-	-	0.30	0.48
						Stallard & Todd	-	50.0	0.026	-	-	0.30	0.40
						Jaki & Hampson	-	83.4	0.000	-	-	0.30	0.47
						Friede	-	99.9	0.025	-	-	0.30	0.50
B6	0.1	0.6	0.7	0.7	1.32	MCDA-based	-	5.4	0.022	-	-	0.32	0.32
						Stallard & Todd	-	50.2	0.024	-	-	0.50	0.50
						Jaki & Hampson	-	1.1	0.000	-	-	0.30	0.30
						Friede	-	99.9	0.025	-	-	0.70	0.70

Table C.8 Analysis of the same data as in Table 3 (scenario B1-B6), with the only difference being the possibility to stop early the study for futility. The futility stop rules used are $\max_j \gamma_j < 0$ (MCDA based approach), $\max_j Z_{j1}^{IA} < 0$ with $j \geq 1$ (Stallard & Todd), $\max_j O_j < 0$ with $j \geq 1$ (Jaki & Hampson) and $\max_j Z_{j2}^{IA} < 0$ with $j \geq 1$ (Friede et al.). The parameters used for the control arm and dose 1, shared across scenarios, are $\lambda_{01} = \lambda_{11} = 0.1$, $\lambda_{02} = \lambda_{12} = 0.3$, $\lambda_{03} = \lambda_{13} = 0.3$, $\lambda_{04} = \lambda_{14} = 0.3$

Scenario	**Parameters ($j = 2$)**					**Approaches**	**Metrics**						
	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	γ_2		**%Stop**	**%Sel2**	**T1E**	**Pow**	**Unint. Succ.**	**Tox1**	**Tox2**
B1	0.1	0.3	0.3	0.3	0.000	MCDA-based	33.0	33.2	0.019	-	-	0.30	0.30
						Stallard & Todd	33.4	33.5	0.025	-	-	0.30	0.30
						Jaki & Hampson	33.0	33.1	0.000	-	-	0.30	0.30
						Friede	33.5	32.9	0.016	-	-	0.30	0.30
B2	0.1	0.3	0.5	0.7	-2.85	MCDA-based	49.8	0.000	0.010	-	-	0.30	0.30
						Stallard & Todd	33.2	33.7	0.017	-	-	0.40	0.50
						Jaki & Hampson	49.6	0.000	0.000	-	-	0.30	0.30
						Friede	34.5	32.1	0.014	-	-	0.40	0.50
B3	0.1	0.4	0.5	0.3	-0.16	MCDA-based	36.4	27.1	0.017	-	-	0.39	0.30
						Stallard & Todd	33.3	33.6	0.022	-	-	0.40	0.30
						Jaki & Hampson	40.3	19.0	0.000	-	-	0.37	0.30
						Friede	10.6	79.3	0.021	-	-	0.48	0.30
B4	0.1	0.4	0.7	0.5	-2.10	MCDA-based	49.9	0.3	0.011	-	-	0.31	0.30
						Stallard & Todd	33.9	32.7	0.019	-	-	0.50	0.40
						Jaki & Hampson	50.1	0.000	0.000	-	-	0.30	0.30
						Friede	10.3	79.4	0.020	-	-	0.65	0.48
B5	0.1	0.6	0.3	0.5	-3.77	MCDA-based	7.9	84.6	0.020	-	-	0.30	0.48
						Stallard & Todd	33.6	33.3	0.021	-	-	0.30	0.40
						Jaki & Hampson	12.4	74.0	0.001	-	-	0.30	0.47
						Friede	0.000	99.9	0.023	-	-	0.30	0.50
B6	0.1	0.6	0.7	0.7	1.32	MCDA-based	49.3	1.5	0.009	-	-	0.32	0.32
						Stallard & Todd	32.8	33.8	0.015	-	-	0.50	0.50
						Jaki & Hampson	50.1	0.000	0.000	-	-	0.30	0.30
						Friede	0.000	99.9	0.021	-	-	0.70	0.70



(a)



(b)

Fig. C.4 Summary results for the 54 scenarios simulated in the Power analysis. Panel (C.4a): power for all the competing approaches at varying levels of $\gamma_2 - \gamma_1$. Panel (C.4b): probability to select treatment 2 at varying levels of $\gamma_2 - \gamma_1$.

Appendix D

Supplementary Material - Chapter 5

Theoretical results

Proof of Theorem 1

Consider a RCT where mean control and treatment responses are normal $X_c \sim \mathcal{N}(\theta_c, \sigma_c^2)$, $X_t \sim \mathcal{N}(\theta_t, \sigma_t^2)$, and assume $\sigma_t^2 = K\sigma_c^2$ (where K^{-1} is the randomization ratio, assumed > 1). Assume a RMP $\pi_c(\theta_c) = \omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}(\theta_c)$ is used for the control parameter, where $\pi_{\text{inf}}(\theta_c)$ and $\pi_{\text{rob}}(\theta_c)$ are the PDF of normally distributed random variables with parameters μ_{inf}, σ_{inf}^2 and μ_{rob}, σ_{rob}^2 respectively; while a normal prior distribution $\theta_t \sim \mathcal{N}(\mu_t, \sigma_{\text{rob}}^2)$ is given to the treatment parameter. Consider the type I error rate $\alpha(\cdot)$ as defined in Equation (5.2), corresponding to the null hypothesis $H_0 : \theta_c = \theta_t = D + \mu_{\text{inf}}$, where $D = \theta_c - \mu_{\text{inf}}$ is the drift parameter. Consider moreover the following path $\left(D, \sigma_{\text{rob}}^2(D)\right)$. Then the following hold:

$$\lim_{D\to+\infty} \alpha(D + \mu_{\text{inf}}) = \eta \iff \lim_{D\to+\infty} \frac{D}{\sigma_{\text{rob}}^2(D)} = 0$$

Proof. Consider the following change of variable: $H = D + \mu_{\text{inf}}$, so that the thesis of the theorem becomes:

$$\lim_{H\to+\infty} \alpha(H) = \eta \iff \lim_{H\to+\infty} \frac{H}{\sigma_{\text{rob}}^2(H)} = 0\,.$$

It should be noted that, in the latter, the dependence of σ_{rob}^2 on the rescaled drift H implies that these two are jointly varied along the functional path $\left(H, \sigma_{\text{rob}}^2(H)\right)$; for the sake of conciseness, this dependence is implicitly assumed throughout the remainder of the proof. Since under the null hypotheses $\theta_c = \theta_t = H$ control and treatment responses are respectively $X_c \sim \mathcal{N}(H, \sigma_c^2)$ and $X_t \sim \mathcal{N}(H, \sigma_t^2)$, then the observed mean responses can be expressed as $X_c = H + \Delta_c$, where $\Delta_c \sim \mathcal{N}(0, \sigma_c^2)$ and $X_t = H + \Delta_t$, where $\Delta_t \sim \mathcal{N}(0, \sigma_t^2)$.

It follows from Equation (5.9) that

$$\lim_{H\to+\infty}\tilde{\Omega}(X_c)=\lim_{H\to+\infty}\tilde{\Omega}(H+\Delta_c)=\lim_{H\to+\infty}\tilde{\Omega}(H)=0 \implies \lim_{H\to+\infty}\tilde{\omega}(X_c)=0$$

where the second equality holds since $\Delta_c \sim o(H)$ for $H \to +\infty$.
As a consequence Equation (5.5) reduces to

$$\lim_{H\to+\infty} g(\theta_c \mid x_c,\pi_{\text{inf}},\pi_{\text{rob}})=\lim_{H\to+\infty} g_{\text{rob}}(\theta_c|x_c,\pi_{\text{rob}})$$

where $g_{\text{rob}}(\cdot|x_c,\pi_{\text{rob}})$ is the PDF of a normal distribution $\mathcal{N}\left(\mu_{\text{c}}^{\text{post}},\sigma_{\text{c}}^{2,\text{post}}\right)$, with

$$\mu_{\text{c}}^{\text{post}}=\frac{\sigma_{\text{rob}}^2 x_c+\sigma_c^2\mu_{\text{rob}}}{\sigma_c^2+\sigma_{\text{rob}}^2}=\frac{\sigma_{\text{rob}}^2 H+\sigma_{\text{rob}}^2\Delta_c+\sigma_c^2\mu_{\text{rob}}}{\sigma_c^2+\sigma_{\text{rob}}^2} \qquad \sigma_{\text{c}}^{2,\text{post}}=\frac{\sigma_c^2\sigma_{\text{rob}}^2}{\sigma_c^2+\sigma_{\text{rob}}^2} \tag{T1.1}$$

Using the same argument the posterior distribution for θ_t is $\mathcal{N}\left(\mu_t^{\text{post}},\sigma_t^{2,\text{post}}\right)$; with

$$\mu_t^{\text{post}}=\frac{\sigma_{\text{rob},t}^2 x_t+K\sigma_c^2\mu_t}{K\sigma_c^2+\sigma_{\text{rob},t}^2}=\frac{\sigma_{\text{rob},t}^2 H+\sigma_{\text{rob},t}^2\Delta_t+K\sigma_c^2\mu_t}{K\sigma_c^2+\sigma_{\text{rob},t}^2} \qquad \sigma_t^{2,\text{post}}=\frac{K\sigma_c^2\sigma_{\text{rob},t}^2}{K\sigma_c^2+\sigma_{\text{rob},t}^2} \tag{T1.2}$$

Since the posterior densities for θ_c and θ_t are normally distributed, then the posterior probability for the mean treatment difference parameter is normal itself, i.e. $\delta^{\text{post}} \sim \mathcal{N}\left(\mu_t^{\text{post}}-\mu_c^{\text{post}},\sigma_t^{2,\text{post}}+\sigma_c^{2,\text{post}}\right)$. Notice that while the variance of the latter distribution is a fixed quantity, as it does not depend on H; the mean is a random variable depending on Δ_c and Δ_t.
Let us prove the two implications of the Theorem separately.

$\boxed{\implies}$ Let us proceed by contradiction. If $\lim_{H\to+\infty}\frac{H}{\sigma_{\text{rob}}^2}=+\infty$, then exploiting the equalities in T1.1 and T1.2, and ignoring negligible terms it holds that:

$$\lim_{H\to+\infty}\mu_t^{\text{post}}-\mu_c^{\text{post}}=\lim_{H\to+\infty}\frac{H(1-K)\sigma_{\text{rob}}^2\sigma_c^2}{(K\sigma_c^2+\sigma_{\text{rob}}^2)(\sigma_c^2+\sigma_{\text{rob}}^2)}=+\infty \qquad \forall x_c,x_t\in\mathbb{R}$$

and from Equation (5.1) follows that

$$\lim_{H\to+\infty}\mathbb{P}(\delta>0\mid x_c,x_t)=\Phi(+\infty)=1>1-\eta \quad \forall x_c,x_t\in\mathbb{R}$$

meaning that success is achieved with probability 1 as $H \to +\infty$, and accordingly

$$\lim_{H\to+\infty}\mathbb{1}\{\mathbb{P}(\delta>0\mid x_c,x_t)\}=\mathbb{1}\{(-\infty,+\infty)\times(-\infty,+\infty))\}$$

Type I error $\alpha(D+\mu_{\text{inf}})$ is easily obtained by integrating the success over the likelihood

$$\begin{aligned}\lim_{H\to+\infty}\alpha(H) &= \lim_{H\to+\infty}\iint_{\mathbb{R}^2}\mathbb{1}\left\{\mathbb{P}(\delta>0\mid x_c,x_t)>\eta\right\} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t\\ &= \iint_{\mathbb{R}^2}\lim_{H\to+\infty}\mathbb{1}\left\{\mathbb{P}(\delta>0\mid x_c,x_t)>\eta\right\} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t\\ &= \iint_{\mathbb{R}^2} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t = 1\end{aligned}$$

$\boxed{\Longleftarrow}$ If $\lim_{H\to+\infty}\frac{H}{\sigma^2_{\text{rob}}}=C$ with $C\neq+\infty$, then exploiting the equalities in T1.1 and T1.2, and ignoring negligible terms it holds that:

$$\lim_{H\to+\infty}\mu_t^{\text{post}}-\mu_c^{\text{post}} = x_t-x_c+C(1-K)\sigma_c^2 \qquad \lim_{H\to+\infty}\sigma_c^{2,\text{post}}=\sigma_c^2 \qquad \lim_{H\to+\infty}\sigma_t^{2,\text{post}}=\sigma_t^2$$

and from Equation (5.1) follows that

$$\lim_{H\to+\infty}\mathbb{P}(\delta>0\mid x_c,x_t)>1-\eta \iff \frac{x_t-x_c+C(1-K)\sigma_c^2}{\sqrt{\sigma_t^2+\sigma_c^2}}>z_\eta$$

where z_η is the η quantile of a standard normal distribution.
The limit of the type I error for $H\to+\infty$ is:

$$\begin{aligned}\lim_{H\to+\infty}\alpha(H) &= \lim_{H\to+\infty}\iint_{\mathbb{R}^2}\mathbb{1}\left\{\mathbb{P}(\delta>0\mid x_c,x_t)>\eta\right\} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t\\ &= \iint_{\mathbb{R}^2}\mathbb{1}\left\{\mathbb{P}(\delta>0\mid x_c,x_t)>\eta\right\} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t\\ &= \iint_{\mathbb{R}^2}\mathbb{1}\left\{\frac{x_t-x_c+C(1-K)\sigma_c^2}{\sqrt{\sigma_t^2+\sigma_c^2}}>z_\eta\right\} f_{X_c}(x_c|\theta_c=H) f_{X_t}(x_t|\theta_t=H)\,dx_c\,dx_t\\ &= \int_{z_\eta\sqrt{\sigma_t^2+\sigma_c^2}-C(1-K)\sigma_c^2}^{+\infty} f_{X_t-X_c}(\xi)d\xi =\\ &= \int_{z_\eta\sqrt{\sigma_t^2+\sigma_c^2}-C(1-K)\sigma_c^2}^{z_\eta\sqrt{\sigma_t^2+\sigma_c^2}} f_{X_t-X_c}(\xi)d\xi + \int_{z_\eta\sqrt{\sigma_t^2+\sigma_c^2}}^{+\infty} f_{X_t-X_c}(\xi)d\xi =\\ &= 1-\Phi^{-1}(z_\eta)+\varepsilon(C) =\\ &= \eta+\varepsilon(C)\geq\eta\end{aligned}$$

where $\xi=x_t-x_c$ and we remind that $X_t-X_c\sim\mathcal{N}\left(0,\sigma_t^2+\sigma_c^2\right)$. It can be noticed that when $C>0$, then $\alpha(H)$ converges to a value greater than the nominal level η and lower than 1; while an exact control at the nominal level η is achieved when $C=0$; from which the result follows. □

Proof of Theorem 2

Consider a normal random variable modeling the mean control response $X_c \sim \mathcal{N}\left(\theta_c, \sigma_c^2\right)$, and assume two distinct RMPs are used for the underlying parameter θ_c, namely

$$\pi_c^{(1)}(\theta_c) = \omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}^{(1)}(\theta_c) \qquad \pi_c^{(2)}(\theta_c) = \omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}^{(2)}(\theta_c)$$

where $\pi_{\text{inf}}(\theta_c)$ and $\pi_{\text{rob}}^{(i)}(\theta_c)$ are the PDF of normally distributed random variables with parameters μ_{inf}, σ_{inf}^2 and $\mu_{\text{rob}}^{(i)}$, σ_{rob}^2 respectively with $i \in \{1,2\}$.
Consider the posterior distributions $g(\theta_c|x_c, \pi_c^{(1)})$ and $g(\theta_c|x_c, \pi_c^{(2)})$, then

$$\lim_{\sigma_{\text{rob}}^2 \to +\infty} g(\theta_c|x_c, \pi_c^{(1)}) = \lim_{\sigma_{\text{rob}}^2 \to +\infty} g(\theta_c|x_c, \pi_c^{(2)}) \qquad \forall x_c \in \mathbb{R}$$

Proof. The two RMPs for θ_c differ only for the the locations of their robustification components, which impact the posterior weights $\tilde{\omega}$ and the posterior corresponding to the robustification component $g_{\text{rob}}(\theta_c|x_c, \pi_{\text{rob}}^{(i)})$. In the following, the argument will be proven by working independently on these two objects.
Recalling that $R = \frac{v_{\text{rob}}}{v_{\text{inf}}}$, $\quad v_{\text{rob}}^2 = \sigma_{\text{rob}}^2 + \sigma_c^2$, $\quad v_{\text{inf}}^2 = \sigma_{\text{inf}}^2 + \sigma_c^2$, it holds that, for $\sigma_{\text{rob}}^2 \to +\infty$, then $R(\sigma_{\text{rob}}^2) \to +\infty$. As a consequence, given Equation (5.7), it holds that for $\sigma_{\text{rob}}^2 \to +\infty$, then

$$\frac{1}{R^2}\frac{(x_c - \mu_{\text{rob}})^2}{2v_{\text{inf}}^2} \sim o\left(\frac{d^2}{2v_{\text{inf}}^2}\right) \qquad \Longrightarrow \qquad \tilde{\Omega} \sim \Omega R \exp\left\{\frac{d^2}{2v_{\text{inf}}^2}\right\}. \tag{T2.1}$$

The latter is independent on $\mu_{\text{rob}}^{(i)}$; as a consequence

$$\lim_{\sigma_{\text{rob}}^2 \to +\infty} \tilde{\omega}(x_c; \pi_{\text{inf}}, \pi_{\text{rob}}^{(1)}, \omega) = \lim_{\sigma_{\text{rob}}^2 \to +\infty} \tilde{\omega}(x_c; \pi_{\text{inf}}, \pi_{\text{rob}}^{(2)}, \omega) \qquad \forall x_c \in \mathbb{R} \tag{T2.2}$$

Moreover, the posterior distribution $g_{\text{rob}}(\theta_c|x_c, \pi_{\text{rob}}^{(i)})$ corresponding to each robustification component is normal with parameters $\mu_{\text{rob}}^{(i),\text{post}}$ and $\sigma_{\text{rob}}^{2,\text{post}}$, with

$$\mu_{\text{rob}}^{(i),\text{post}} = \frac{\sigma_{\text{rob}}^2 x_c + \sigma_c^2 \mu_{\text{rob}}^{(i)}}{\sigma_c^2 + \sigma_{\text{rob}}^2} \qquad \sigma_{\text{c}}^{2,\text{post}} = \frac{\sigma_c^2 \sigma_{\text{rob}}^2}{\sigma_c^2 + \sigma_{\text{rob}}^2}$$

Notice that the variance, which is the same in the two RMPs, does not depend on $\mu_{\text{rob}}^{(i)}$, moreover for the mean we have that for $\sigma_{\text{rob}}^2 \to +\infty$, then

$$\mu_{\text{rob}}^{(i),\text{post}} \sim \frac{\sigma_{\text{rob}}^2 x_c}{\sigma_c^2 + \sigma_{\text{rob}}^2}$$

which is independent on $\mu_{\text{rob}}^{(i)}$. It follows that

$$\lim_{\sigma_{\text{rob}}^2 \to +\infty} g_{\text{rob}}(\theta_c | x_c, \pi_{\text{rob}}^{(1)}) = \lim_{\sigma_{\text{rob}}^2 \to +\infty} g_{\text{rob}}(\theta_c | x_c, \pi_{\text{rob}}^{(2)}) \tag{T2.3}$$

The argument follows from Equation (T2.2) and T2.3. □

Proof of Theorem 3

Consider a normal random variable $X_c \sim \mathcal{N}\left(\theta_c, \sigma_c^2\right)$, and assume a RMP is used for the parameter θ_c, namely $\pi_c(\theta_c) = \omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}(\theta_c)$, where $\pi_{\text{inf}}(\theta_c)$ and $\pi_{\text{rob}}(\theta_c)$ are the PDF of normally distributed random variables with parameters μ_{inf}, σ_{inf}^2 and μ_{rob}, σ_{rob}^2 respectively. The following hold:

1. if $\Omega < +\infty$, then

$$\lim_{\sigma_{\text{rob}}^2 \to +\infty} \tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) = 1 \qquad \forall x_c \in (-\infty, +\infty)$$

   Proof. From the asymptotic equivalence in T2.1, considering that $\Omega < +\infty$ and considering that $R \to +\infty$ for $\sigma_{\text{rob}}^2 \to +\infty$, then the argument follows. □

2. if $\Omega \sim O(1/R)$ for $\sigma_{\text{rob}}^2 \to +\infty$, then

$$\lim_{\sigma_{\text{rob}}^2 \to +\infty} \tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) \neq 1 \qquad \forall x_c \in (-\infty, +\infty)$$

   Proof. From the asymptotic equivalence in T2.1, considering that $\Omega \sim O(1/R) \Rightarrow \beta(\omega, R) < +\infty$ for $\sigma_{\text{rob}}^2 \to +\infty$, then the argument follows. □

Proof of Theorem 4

Consider a binomial random variable $X_c \sim \text{Bin}(\theta_c, n_c)$, and assume a RMP is used for the parameter θ_c, namely $\pi_c(\theta_c) = \omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}(\theta_c)$, where $\pi_{\text{inf}}(\theta_c)$ and $\pi_{\text{rob}}(\theta_c)$ are the PDF of Beta distributed random variables with parameters a_{inf}, b_{inf} and $a_{\text{rob}} = b_{\text{rob}} = \varepsilon$, respectively. The following hold:

1. if $\Omega < +\infty$, then

$$\lim_{\varepsilon \to 0} \tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) = 1 \qquad \forall x_c \in (0, n_c)$$

   Proof. From Equation (5.15), and expressing the Beta function using the Gamma functions $B(x,y) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, the posterior odds under the Robust Mixture Prior (RMP)

in the Beta-Binomial setting can be written as

$$\Omega(x_c) = \beta(\omega, a_{\text{rob}}, b_{\text{rob}}) \times \frac{\Gamma(x_c + a_{\text{inf}})\Gamma(n_c - x_c + b_{\text{inf}})\Gamma(a_{\text{inf}} + b_{\text{inf}})}{\Gamma(n_c + a_{\text{inf}} + b_{\text{inf}})\Gamma(a_{\text{inf}})\Gamma(b_{\text{inf}})} \times \frac{\Gamma(n_c + a_{\text{rob}} + b_{\text{rob}})}{\Gamma(x_c + a_{\text{rob}})\Gamma(n_c - x_c + b_{\text{rob}})},$$

where

$$\beta(\omega, a_{\text{rob}}, b_{\text{rob}}) = \frac{\omega}{1-\omega} \cdot \frac{\Gamma(a_{\text{rob}})\Gamma(b_{\text{rob}})}{\Gamma(a_{\text{rob}} + b_{\text{rob}})}.$$

Under the assumptions of the theorem $a_{\text{rob}} = b_{\text{rob}} = \varepsilon$ with $\varepsilon \to 0^+$, and using the well-known asymptotic expansion $\Gamma(\varepsilon) \sim 1/\varepsilon$ as $\varepsilon \to 0^+$, and the fact that $\Gamma(x_c + \varepsilon) \to \Gamma(x_c)$ for $x_c > 0$, we obtain

$$\Gamma(a_{\text{rob}})\Gamma(b_{\text{rob}}) \sim \frac{1}{\varepsilon^2}, \quad \Gamma(a_{\text{rob}} + b_{\text{rob}}) = \Gamma(2\varepsilon) \sim \frac{1}{2\varepsilon},$$

and $\Gamma(n_c + a_{\text{rob}} + b_{\text{rob}}) \sim \Gamma(n_c)$.

Substituting these limits into the definition of $\beta(\omega, a_{\text{rob}}, b_{\text{rob}})$ gives

$$\beta(\omega, a_{\text{rob}}, b_{\text{rob}}) \sim \frac{\omega}{1-\omega} \cdot \frac{2}{\varepsilon} \to +\infty \quad \text{as} \quad \varepsilon \to 0$$

The remaining multiplicative factor in the expression for $\tilde{\Omega}(x_c)$,

$$C(x_c, n_c) = \frac{B(a_{\text{inf}} + x_c, b_{\text{inf}} + n_c - x_c)}{B(x_c, n_c - x_c)\, B(a_{\text{inf}}, b_{\text{inf}})},$$

is finite and positive for all $x_c \in (0, n_c)$. Therefore,

$$\tilde{\Omega}(x_c) = \beta(\omega, a_{\text{rob}}, b_{\text{rob}}) \cdot C(x_c, n_c) \to +\infty \quad \text{as } \varepsilon \to 0^+.$$

Finally, the posterior weight of the informative component is

$$\tilde{\omega}(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega) = \frac{\tilde{\Omega}(x_c)}{1 + \tilde{\Omega}(x_c)}.$$

Since $\tilde{\Omega}(x_c) \to +\infty$, it follows that

$$\lim_{\varepsilon \to 0^+} \tilde{\omega}(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega) = 1, \quad \forall x_c \in (0, n_c).$$

□

2. if $\Omega \sim O(\varepsilon)$ for $\varepsilon \to 0$, then

$$\lim_{\varepsilon \to 0} \tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) \neq 1 \qquad \forall x_c \in (0, n_c)$$

Proof. Assume again that $a_{\text{rob}} = b_{\text{rob}} = \varepsilon$ with $\varepsilon \to 0^+$. In Point 1, we observed that as $\varepsilon \to 0^+$, $\Gamma(\varepsilon) \sim 1/\varepsilon$ and $\Gamma(2\varepsilon) \sim 1/(2\varepsilon)$, so that $\beta(\omega, \varepsilon, \varepsilon)$ diverges as $O(1/\varepsilon)$. This divergence was responsible for $\Omega(x_c) \to +\infty$, leading to $\tilde{\omega} \to 1$.

Here, we relax the assumption of a fixed ω and instead assume that $\Omega(x_c)$ satisfies the asymptotic scaling

$$\Omega \sim O(\varepsilon) \quad \text{as } \varepsilon \to 0^+,$$

This means that $\Omega(x_c)$ and ε are of the same order of magnitude, i.e.

$$\frac{\Omega}{\varepsilon} \to K,$$

for some finite, positive constant $K > 0$.

It follows that as $\varepsilon \to 0^+$,

$$\begin{aligned}\tilde{\Omega}(x_c) &= \beta(\omega, \varepsilon, \varepsilon) \cdot C(x_c, n_c) \\ &= K \cdot C(x_c, n_c) = \tilde{K} < +\infty\end{aligned}$$

Substituting this asymptotic behavior into the expression for the posterior weight,

$$\tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) = \frac{\tilde{\Omega}(x_c)}{1 + \tilde{\Omega}(x_c)},$$

we obtain that as $\varepsilon \to 0^+$,

$$\lim_{\varepsilon \to 0^+} \tilde{\omega}\left(x_c, \pi_{\text{inf}}(\theta_c), \pi_{\text{rob}}(\theta_c), \omega\right) = \frac{\tilde{K}}{1 + \tilde{K}} < 1, \quad \forall x_c \in (0, n_c).$$

□

Proof of Equations (5.5) and (5.6)

$$g\left(\theta_c|x_c,\pi_c\right) = \frac{\left[\omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}(\theta_c)\right] f\left(x_c|\theta_c\right)}{\int_{-\infty}^{+\infty}\left[\omega\pi_{\text{inf}}(\theta_c) + (1-\omega)\pi_{\text{rob}}(\theta_c)\right] f\left(x_c|\theta_c\right) d\theta_c} =$$

$$= \frac{\omega\pi_{\text{inf}}(\theta_c) f\left(x_c|\theta_c\right) + (1-\omega)\pi_{\text{rob}}(\theta_c) f\left(x_c|\theta_c\right)}{\omega\int_{-\infty}^{+\infty}\pi_{\text{inf}}(\theta_c) f\left(x_c|\theta_c\right) d\theta_c + (1-\omega)\int_{-\infty}^{+\infty}\pi_{\text{rob}}(\theta_c) f\left(x_c|\theta_c\right) d\theta_c} =$$

$$= \frac{\omega\pi_{\text{inf}}(\theta_c) f\left(x_c|\theta_c\right) + (1-\omega)\pi_{\text{rob}}(\theta_c) f\left(x_c|\theta_c\right)}{\omega f\left(x_c|\pi_{\text{inf}}\right) + (1-\omega) f\left(x_c|\pi_{\text{rob}}\right)} =$$

$$= \frac{\omega\pi_{\text{inf}}(\theta_c) f\left(x_c|\theta_c\right)}{\omega f\left(x_c|\pi_{\text{inf}}\right) + (1-\omega) f\left(x_c|\pi_{\text{rob}}\right)} + \frac{(1-\omega)\pi_{\text{rob}}(\theta_c) f\left(x_c|\theta_c\right)}{\omega f\left(x_c|\pi_{\text{inf}}\right) + (1-\omega) f\left(x_c|\pi_{\text{rob}}\right)} =$$

$$= \frac{f\left(x_c|\theta_c\right)\pi_{\text{inf}}\left(\theta_c\right)}{f\left(x_c|\pi_{\text{inf}}\right)} \times \frac{\omega f\left(x_c|\pi_{\text{inf}}\right)}{\omega f\left(x_c|\pi_{\text{inf}}\right) + (1-\omega) f\left(x_c|\pi_{\text{rob}}\right)} +$$

$$+ \frac{f\left(x_c|\theta_c\right)\pi_{\text{rob}}\left(\theta_c\right)}{f\left(x_c|\pi_{\text{rob}}\right)} \times \frac{(1-\omega) f\left(x_c|\pi_{\text{rob}}\right)}{\omega f\left(x_c|\pi_{\text{inf}}\right) + (1-\omega) f\left(x_c|\pi_{\text{rob}}\right)} .$$

Formulas for the metrics used in posterior inference

Bias is defined as:

$$b(\hat{\delta}) = \mathbb{E}\left[\hat{\delta} - \delta\right] = \iint_{\mathbb{R}^2}\left(\hat{\delta} - \delta\right) f_{X_c}(x_c) f_{X_t}(x_t)\, dx_c\, dx_t ,$$

Variance is defined as:

$$Var(\hat{\delta}) = \mathbb{E}\left[\left(\hat{\delta} - \mathbb{E}\left[\delta\right]\right)^2\right] = \iint_{\mathbb{R}^2}\left(\hat{\delta} - \mathbb{E}\left[\delta\right]\right)^2 f_{X_c}(x_c) f_{X_t}(x_t)\, dx_c\, dx_t$$

Mean Squared Error (MSE) is defined as:

$$MSE(\hat{\delta}) = \mathbb{E}\left[\left(\hat{\delta} - \delta\right)^2\right] = \iint_{\mathbb{R}^2}\left(\hat{\delta} - \delta\right)^2 f_{X_c}(x_c) f_{X_t}(x_t)\, dx_c\, dx_t$$

Additional plots for the beta-binomial case

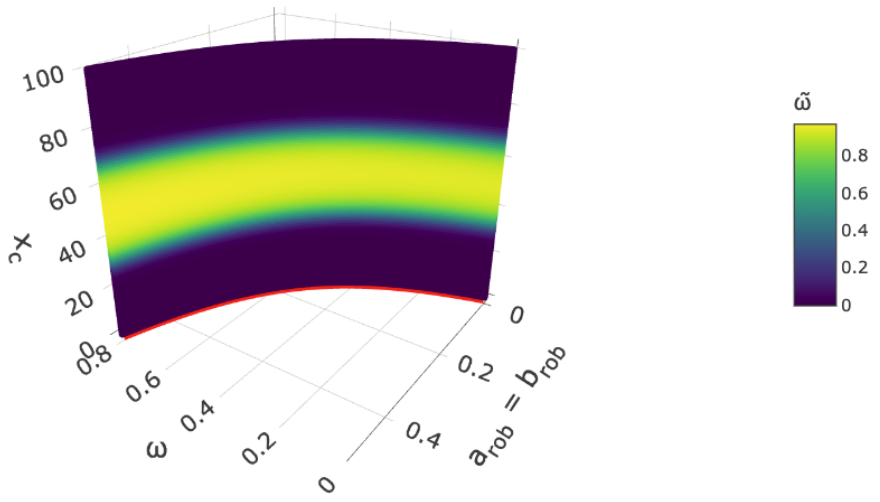


Fig. D.1 Posterior weight $\tilde{\omega}$ as a function of $a_{\text{rob}} = a_{\text{rob}}$, ω and x_c. The red curve in the horizontal plane represents all RMPs with $\beta^* = 12.56$.

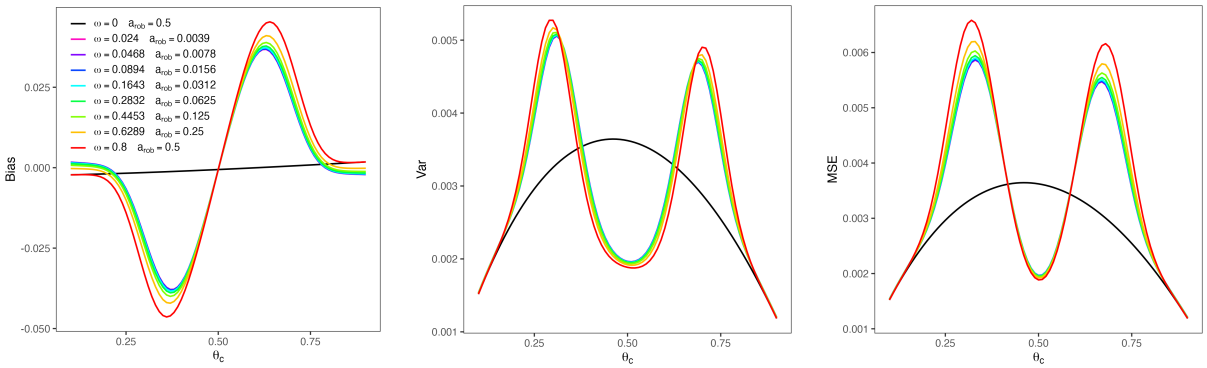


Fig. D.2 Panel (a): bias; Panel (b): variance; Panel (c): mean squared error in the Beta–Binomial setting, all computed using the posterior mean of the treatment effect parameter δ as the point estimate. Colors indicate different combinations of $(\omega, a_{\text{rob}} = b_{\text{rob}})$, each corresponding to $\beta^* = 12.56$.