Doctoral Dissertation
Doctoral Program in Pure and Applied Mathematics (36^{th}cycle)

Statistical methods for using available evidence to support decision-making

By

Fulvio Di Stefano

******

Supervisors:
Prof. Mauro Gasparini, Supervisor
Dr. Gaelle Saint-Hilary, Co-Supervisor

Doctoral Examination Committee:
Prof. Stefania Gubbiotti, Referee, Università degli studi di Roma "La Sapienza"
Prof. Moreno Ursino, Referee, Institut national de la santé et de la recherche médicale

Politecnico di Torino
2024

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.

Fulvio Di Stefano
2024

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

I would like to dedicate this thesis to my beloved grandmother Vincenza, the first person to encourage me to the study of mathematics.

Acknowledgements

I would like to thank all the people which, in one way on another, have contributed to making this work possible.

Firstly, I express my sincerest gratitude to Professor Mauro Gasparini, for his guidance and inspiration over these years. His unwavering support and encouragement has pushed me to always aim high and strive for excellence.

I would also like to extend my thanks to Doctor Gaelle Saint-Hilary, who has been my supervisor. She has guided me with great rigor and discipline, that I will not forget in my working career.

I am grateful to all my co-authors, which have contributed their valuable insights and expertise to our work.

My heartfelt gratitude goes to my mother Francesca, for being the rock upon which I stood and always there for me through thick and thin. Her endless love and hard work have not gone unnoticed.

I am forever indebted to my girlfriend Cristina, who has provided unwavering love and support throughout my journey, sharing the ups and downs with me.

Lastly, I would like to express my sincerest gratitude to my family and friends, who have always been there to provide emotional support during the critical moments of my journey.

Abstract

This work focuses on the development of statistical methods which permit to make use of available evidence to support decision-making. In particular, it deepens three research areas: the incorporation of historical data in early phase clinical trials, a novel method to perform adaptive screening in a certain sub-population and the comparison of different estimation methods in adaptive designs with time-to-event endpoints. Four methodologies are presented. The first one regards the incorporation of healthy volunteer data on receptor occupancy in a phase II proof of concept trial. The second one regards an analysis on the incorporation of preclinical animal toxicological data in a phase I trial. The third one, motivated by a case-study on a COVID-19 screening in a university community, regards a novel methodology to test adaptively whether a certain subpopulation proportion follows the same time evolution as the general population proportion. The last one is a comparison of different estimation methods to account for selection bias in adaptive enrichment designs with time-to-event endpoints.

These methodologies are valuable quantitative tools to include available evidence to support decision-making. They have strong theoretical foundations and have been tested in real life case studies. Moreover, they can potentially be applied to a variety of other problems and provide useful tools that can help to make more accurate and informed decisions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Recent developments of statistical methods

As the name suggests, Statistics was born as the "science of the state". Therefore, its original intent was to deal with questions of general administration, such as counting armaments, resources and people. Its main mathematical developments were launched in the 16th and 17th century by the work of Cardano, Fermat and Pascal in probability, mainly applied to the theory of gambling, and were prosecuted by the great names of Mathematics in the following centuries, like Laplace, Bernoulli and Gauss, just to name a few. The foundation of modern Statistics in the 20th century traces back to the work of Pearson and Fisher, who are often referred as the fathers of modern Statistics. Most of their research and, in general, of the 20th century Statistics is focused on what today is called frequentist statistics. However, by the end of the century, the work of the English reverend Thomas Bayes, which dated back three hundred years, has been rediscovered and Bayesian statistics has been renewed. It aims at combining some sort of previous knowledge with data from the current process, obtaining an estimation which is able to take both information sources into account. In the early days, Bayesian statistics was used just for simple and special cases, where analytical solution of the calculations were available. However, thanks to modern computational capacity, Bayesian complex and time-consuming computations have been made feasible and every year new and modern updates permit to improve existing methods.

Notwithstanding its humble origins, Statistics today is widely used to support decision-making in many different fields, using methods from both a Bayesian and frequentist background. New niches of research are continuously carved in this frame by researchers, who then explore and analyse the newfangled crafted possibilities.

Novel studies, novel ideas and novel methodologies have been the engine of innovation over the history of humankind, from Pythagoras to Fisher, and permitted the world to advance. However, sometimes it may happen that these novelties collide with the ongoing regulations. The regulatory framework is constructed to be conservative and encourage well understood and tested methods. This is especially true in healthcare regulations, where protecting, curing and, as the name suggests, taking care of people are its foundations and main goals. Statistics researchers have to pursue their main activity with this in mind and find novel methodologies that permit to reduce patient burden and make more informed decisions. Then, they have to demonstrate to the regulators the safety and superiority of those methods. This was especially true during the last four years, a time in which the world has experienced a terrible historic event which led to concentrate all of its strength on a possible resolution. Indeed, the Covid-19 pandemic, a trauma for the whole humankind, has been an extremely good example of human cooperation and interaction. Statisticians have been of fundamental importance in the critical context of handling the emergency. The newspapers all over the world were full of statistical indexes of contagion, of prediction, of vaccine and treatment efficacy. Statistics was back to its original role, in some sense, serving each state of the world with its scientific rigour. However, in this case, the discoveries of the great statisticians of the past were applied, reminding the true role of researchers in society: explore, analyse and craft novel methodologies, ideas and possibilities, hoping that their discoveries can be of support for posterity.

1.2 Bayesian statistics in healthcare regulatory approval

Most documents released by regulatory authorities deal with statistical concepts, like p-values, significance tests and confidence intervals, related to classical frequentist approaches. On the other hand, Bayesian statistics is focused on the concept of prior distribution, which is updated via the observed data to obtain a posterior distribution. This posterior distribution is used for statistical inference and, in turn, may become a prior distribution for a subsequent model. Very few documents mentioned Bayesian approaches, until, in 1998, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E-9 “Guidance on Statistical Principles for Clinical Trials” [2] for pharmaceutical products has been released. In its content, the guidance explicitly mentions that, due to the predominance of frequentist approaches in the statistical community, it largely refers to terms and concepts related to a frequentist framework. However, it clarifies that other approaches, explicitly mentioning the Bayesian one, can be considered for analysis, when the reasons for their use is clear and the results and analyses are sufficiently robust. This first guidance was pioneering the use of Bayesian analyses in the regulatory framework.

The United States of America Food and Drug Administration (FDA) started a discussion in the same year to consider Bayesian submission for regulatory approval of medical devices. This discussion ended in the publication of the FDA "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" [3], in 2010. This is the first FDA released guidance which refers to Bayesian statistics in its title. The rationale behind the publication of this guidance specifically for medical devices is twofold. The first reason is that medical devices usually evolve and, therefore, the previous version of the device may be very similar to the following version. The updates of medical devices are usually released after a few years and contain lot of information to be used as prior data for the newer device. The second reason is that their mechanism of action is often local and well understood, unlike what happens in drug development. The guideline outlines the importance of quantitative data over subjective data in the definition of the prior. The latter may cause controversies when it comes to their evaluation. However, it is clearly stated that the use of Bayesian models may accelerate drug development and decrease burden for patients, permitting to use all available evidence to make more informed decisions. The reason why this guideline was published at this time is that the technological advance in computational capacity and speed had made possible to define and test complex and realistic Bayesian models. Still, software and technologies are only a part of the process of planning a Bayesian design. At first, specific statistical and computational expertise have to be gained by the organization planning the trial. Subsequently, the prior definition, the amount of information obtained from the trial and the model to combine the two have to be defined. These require extensive model planning and model building. At last, after the trial has been conducted, calculations have to be double checked, using for example a different software, and it needs to be reminded that Bayesian and frequentist approaches may differ in their conclusion. Therefore, the use of Bayesian models requires at least the same diligence in constructing the design as the one of a classical clinical trial. The objectives, endpoints, conditions, population and statistical analysis of the trial have to be clearly defined and a sound planning and execution is fundamental, as it is the adjustment for covariates, which may affect the results, and the choice of appropriate controls.

Another point of interest raised when dealing with Bayesian design is exchangeability. In the guideline, it is stated that "units (patients or trials) are considered exchangeable if the probability of observing any particular set of outcomes on those units is invariant to any reordering of the units". While for the patients in a single trial this is a common assumption, things are a little different when dealing with trials themselves. These are usually conducted in different points in time and, therefore, changes in the general population or general methods to test or treat the patients and their condition may have arisen, causing the trials not to be fully exchangeable. Bayesian models have to take into account these possibilities and are required to be robust enough to handle this situation. Additionally, other considerations in the design of a Bayesian trials regard the sample size. While for classical frequentist designs the sample size is usually fixed, for Bayesian trials (and more recent frequentist designs), stopping criteria for efficacy, significance, clinical relevance or other pre-specified hypothesis are used, which entail variable sample sizes. FDA recommends to pre-specify at least a minimum and maximum sample, in order to comply with economic, ethic and regulatory considerations and to reach adequate operating characteristics of the trial. In fact, operating characteristics like type I error, type II error and power are fundamental for drug regulation and need to be assessed both in the planning of the Bayesian trial and in its subsequent analysis, such to evaluate the performance of the design under various conditions. Assessing the operating characteristics of a Bayesian design is an important step in determining the validity and reliability of the trial results, as it helps to ensure that the design is well-suited for the specific research question being addressed. To sum up, the guideline suggests a list of additional information to enclose in a Bayesian design, which includes: prior information, success criteria, trial sample size determination rationale, operating characteristics, prior probability of the study claim, posterior effective sample size and the code used for the study. Alongside, a sensitivity analysis, which involves investigations into various factors that could potentially affect the trial results, such as deviations from distributional assumptions or alternative prior distribution, is also recommended.

In 2016, the FDA released a second guidance focused on the use of Bayesian models in efficacy of medical devices assessment in pediatric population, named "Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices" [4]. Conducting clinical trials in the pediatrics is a challenging activity due to several factors, like limited patient availability, differences in treatment response compared to adults, ethical considerations, and logistical difficulties, such as coordinating with multiple caregivers or obtaining informed consent from parents or guardians. Despite, it is important to ensure that new treatments are safe and effective for this very delicate population. The aim of the document is to increase the availability of safe and effective pediatric devices by providing guidance on using existing clinical data to demonstrate safety and effectiveness in pre-market approval applications, de novo requests, and humanitarian device exemptions. It also outlines when it is appropriate to use existing clinical data to support pediatric device indications and labeling, the approach of the FDA in determining if extrapolation is appropriate and to what extent data can be leveraged, describing statistical methods that can be used to leverage data in a way that improves precision for pediatric inferences. There are two main differences with the previous guideline. The first one is that it is clearly stated that other sources of data can be either other pediatric or adult studies and that the data can be borrowed both for the control and the experimental arm. The second one is that it contains in appendix details on a specific three layer hierarchical modelling structure for borrowing information. Bayesian borrowing methods are considered useful for including all available evidence in the trial and of help in performing faster and more precise trials, reducing patient burden.

1.3 The use of Bayesian models in early phase drug development

While the aforementioned guidelines focus on the needs and recommendations from regulatory authorities for drug approval, the same methods may be used for the early phase assessment of drug efficacy. The use of statistical techniques that borrow information from historical data has become more common in recent years, thanks to the availability of statistical software and the development of platforms that provide access to historical studies. This approach has the potential to greatly speed up the drug development process, as it allows researchers to leverage existing data rather than starting from scratch in every new study.

When using historical data in early phase there are at least three main aspects to be taken into account. First, it is important to ensure that there is a clear rationale for using the data. This means that the data should be relevant to the current study and should be able to provide valuable insights into the safety and effectiveness of the drug. In addition to having a clear rationale for using historical data, it is also important to ensure that the data meets predetermined quality standards. This means that the data should be accurate, reliable, and relevant to the current study. It is also important to consider the potential biases or limitations of the data, as these could affect the results of the study. Finally, it is important to ensure that there is enough information to weigh the risks and benefits of using historical data. For example, if the data comes from a small or unrepresentative sample, it may not be reliable enough to support the conclusions of the study. Similarly, if the risks of using the data outweigh the potential benefits, it may not be advisable to use the data. Therefore, borrowing information from historical studies can be a useful tool, but it is not without its limitations, since the data may not be directly applicable to the current study, or there may be differences in the study populations or treatment regimens that make it difficult to directly incorporate results from previous trials.

One important milestone in the use of historical data can be the incorporation of data from healthy volunteers in phase II studies in patients. Methodologies of this type can help in assessing the efficacy and clinical relevance of a new compound, since phase II studies are typically the first time that a drug is tested in a small group of patients and are designed to evaluate the effectiveness and safety of the drug in a more controlled setting. A second, but not less important, milestone can be the incorporation of preclinical animal data in phase I studies. This can provide valuable information about the safety and effectiveness of a drug. Preclinical animal studies, usually conducted before a drug is tested in humans, can help to identify any potential safety concerns regarding the dosing and potential toxicities related to the use of the new compound. In both cases, the data used for borrowing may come from external datasets, as well as internal studies. In any case, it is important that internal drug assessments using the incorporation of historical data via Bayesian models meet the same standards seen in Section 1.2 as regulatory standards. This means that the internal protocols should contain at least the same information that would be included in a submission to the regulatory authority, and the analyses and studies should be performed with the same diligence as if a submission were the next step.

Despite the challenges and limitations of using historical data, the possibilities and new horizons opened by these approaches are significant. In particular, they have the potential to greatly benefit the development of drugs for target therapies or rare diseases, where there may be limited data available, alongside their current use in pediatrics. However, it is important to carefully consider the risks and benefits of using historical data, and to ensure that the data meets predetermined quality standards. As such, the use of historical data in early phase drug development should be approached with caution, but also with an eye towards the potential benefits it can provide. In conclusion, the use of statistical techniques that borrow information from historical data can be a useful tool in the early phases of drug development. It is important to ensure that there is a clear rationale for using the data, that the data meets predetermined quality standards, and that there is enough information to weigh the risks and benefits of using the data. By carefully considering these factors, researchers can take advantage of the potential benefits of historical data while minimizing the risks. In the future, it is likely that the use of historical data will become more common and will help to accelerate the drug development process, not only for rare diseases or pediatric populations.

1.4 Adaptive design trials

In recent years, novel statistical methodologies have been proposed to enhance drug development. Adaptive design trials have been constructed to allow for pre-planned modifications during the trial, such as refining sample size, stopping the trial or specific doses for lack of efficacy, stopping the trial for success, reshuffling patients among treatment arms, selecting populations more likely to benefit from the treatment [5]. These designs are mainly regulated by the FDA's Guidance for Industry on Adaptive Designs for Clinical Trials of Drugs and Biologics [6], issued in 2019. The guidance emphasizes the benefits of utilizing adaptive designs in clinical trials while underscoring the importance of adhering to key principles for regulatory approval.

Adaptive designs offer several advantages, including the ability to adapt the trial based on interim analyses, enabling informed decision-making. These designs enhance various aspects, such as the statistical efficiency of trials, particularly in terms of power, potentially reducing the number of patients receiving ineffective treatments and providing ethical advantages. Additionally, adaptive designs facilitate a better understanding of a drug efficacy, especially if it proves more effective in specific subgroups within the overall population. Lastly, adaptive designs may be more appealing to sponsors and patients since they allow for modifications to randomization, making them more useful to these stakeholders.

On the other hand, the drawbacks of adaptive designs are also highlighted in the guidance. The primary disadvantage is the necessity for specific analytical methods to prevent erroneous calculations or biased estimates. Additionally, some of the benefits of adaptive designs may be counterbalanced by trade-offs in other areas, such as a lower minimum sample size that may be offset by a higher maximum one. In terms of trial conduct, adaptive designs introduce logistical challenges to maintain trial integrity. Lastly, the final results of an adaptive design may differ from the initial findings, thereby increasing the complexity of interpreting the results.

Hence, when preparing for an adaptive design trial, various considerations must be addressed. Firstly, it is crucial to control the likelihood of erroneous conclusions and calculations resulting from the nature of the design. Consequently, reliable estimations of the true treatment and control effects are necessary, which directly impact the determination of the drug's efficacy. Careful trial planning becomes highly significant, encompassing factors such as the number and timing of interim analyses, the type of adaptation, the selection of statistical inferential methods, and the specific algorithms governing adaptation decisions. Prespecifying these aspects allows for consistent results during the simulation and determination of the design operating characteristics, thereby facilitating interpretability. Ultimately, ensuring trial integrity is of utmost importance in order to obtain reliable results.

The statistical literature encompasses numerous studies and analyses to enable a more informed use of these novel techniques. In this scenario, the statistical community plays a pivotal role in offering methods that enable the use of available evidence to enhance decision-making processes. Statisticians play a central role in constantly uncovering and expanding new areas of research, enabling humankind to enhance their current circumstances and continually progress in the most appropriate way possible.

1.5 The support of the statistical community during the Covid-19 pandemic

The COVID-19 pandemic has had a major impact on people lives globally. Since the end of 2019, when the first news about a new virus in the Wuhan market were coming from, what seemed, a far and distant China, a long way has come. Unfortunately, Italy has been the first European country to be heavily affected by the pandemic, being struck by a bolt from the blue. The news of the first lockdown in Italy in March 2020 spread all over the world, seeming an incredible and extraordinary event. Unfortunately, it was not an isolated case. Already in those first confused days fighting against the new disease, the role played by statisticians was considered crucial. This was the first time when the majority of the world population started to hear those statistical indexes of diffusion, which would become part of the common vocabulary of those days. One after another, the SARS-CoV-2 virus spread to all European countries, and then to the world, showing what was statistically clear, i.e. that borders cannot stop a disease, but just slow it. As fast as it was possible, medical research started its processes and looked for a possible cure and vaccine for the virus.

During summer 2020, while the first long steps in medical research were taken, a partial easing of restrictions was put in place by governments. When winter was coming, without a proper cure or vaccination, but already with a working Covid-19 screening infrastructure, Italy has seen the rise of differentiated lockdowns. Different regions were put under different restrictions according to the local state of the pandemic and health services in that precise time. Again, statistical indexes of contagion and of hospital occupancy played a crucial role in political decisions and various models were used to contain the virus, not only in Italy but on a global basis. In this context, also private and public organizations, like universities, started their own screening campaigns to try to intercept and avoid outbreaks.

The first good news since the start of the pandemic came by the end of the year 2020, with the publication of the first two studies on vaccine efficacy [7, 8] which showed protection against Covid-19 and gave a concrete hope for a possible end of the emergency. In the following month, these were flanked by two other interim analyses [9, 10] which showed the efficacy of as many vaccines. Humankind had shown an incredible spirit of cooperation which lead the research on vaccine to be as fast as to break all records.

Therefore, the year 2021 started with the organization of the vaccination campaigns in many countries and the effects of the vaccination were rapidly seen. Taking this into account, the statistical community started to update the models used to support political decisions with the ultimate developments of the pandemic, like the vaccination rate or the re-infection rate. During the whole summer 2021, while the vaccination campaign proceeded expeditiously, statisticians were always at the centre of the public attention, given their importance in that specific historical moment. In the following months, from winter to spring, the restrictions based on differentiated lockdown were still in place in Italy but, thankfully, the high vaccination rate avoided the worst.

During 2022, a continuous easing of the restriction has been experienced, caused by the modified and less severe symptoms of the illness, rather than a slowdown of the spread of the virus. Notwithstanding, the vaccination campaigns were still ongoing, flanked by the ones on common flu, and statisticians were still analysing data, trying to keep the spread of the virus under control and identify possible sudden modifications of its behaviour. The pandemic seemed to be going towards an end, with very few restrictions still in place.

Today, the impact of the pandemic is still profound and clear in people's mind throughout the world. In this context, the crucial role played by statisticians in understanding and combating the disease has been highlighted. From analyzing data and developing models to support decision-making, to designing and analyzing clinical trials and other research studies, the statistical community has been instrumental in advancing the understanding of the disease and finding ways to deal with it. As the pandemic continues to evolve, the work of statisticians will remain vital in helping to control and eventually overcome the emergency.

1.6 Content of this report

There are many statistical methods that can be used for including available evidence to support decision-making, also in the topic discussed in this introduction. In particular, this report focuses on three research axes: the incorporation of historical data in early phase (Chapter 2 and 3), a novel method to perform adaptive screening in a certain sub-population (Chapter 4) and a comparison of estimation methods in adaptive designs (Chapter 5).

The chapters in this report are self-contained and correspond to articles published in a statistical journal, except for one which is ongoing work. The preambles for each chapter provide some context about the publication.

Specifically, Chapter 2 presents a novel methodology to incorporate available evidence from pharmacokinetics data on a new compound in a phase II trial. A physiology based pharmacokinetic model, which has been tuned on healthy volunteers, is used to obtain prior information on patients in an upcoming phase II trial, taking into account the peculiar differences due to the illness.

In Chapter 3, the incorporation of available evidence from pre-clinical animal data in a phase I oncology trial is discussed. Methods are based on the Bayesian logistic regression model and a comparison of different approaches and dose-escalation criteria is described.

In Chapter 4, a novel methodology to test adaptively whether a sub-population proportion follows the same time evolution as the population proportion is presented. The motivating case study is the COVID-19 screening in a university community, taking into account the time evolution of the pandemic in the whole country.

Chapter 5 discusses a comparison of estimation methods to account for selection bias in adaptive enrichment designs with time-to-event endpoint. The different methods are analysed applied to a case study in heart failure.

Chapter 6 provides concluding remarks.

Chapter 2

Incorporation of healthy volunteers data into a phase II proof-of-concept trial

Background

This chapter is published as:

Fulvio Di Stefano, Christelle Rodrigues, Stephanie Galtier, Sandrine Guilleminot, Veronique Robert, Mauro Gasparini and Gaelle Saint-Hilary. *Incorporation of healthy volunteers data on receptor occupancy into a phase II proof-of-concept trial using a Bayesian dynamic borrowing design.* Biometrical Journal, 2023. DOI: 10.1002/bimj.202200305.

2.1 Introduction

Extrapolation of relevant information from existing research has emerged as a focal point within the field of statistical pharmaceutical investigation. The evolving landscape of clinical trials occasionally yields reduced sample sizes during the testing phases, as noted in Bradley et al. (2012) [11]. In specific situations, such as when dealing with targeted therapies and personalized approaches, performing extensive studies may not be a viable or realistic option. The consequence of smaller sample sizes is a predictable reduction in precision. Fortunately, there exist innovative solutions to face this challenge, including the leveraging of historical data.

This chapter introduces an approach that integrates data extrapolated from a minimal physiology-based pharmacokinetic (mPBPK) model observed in healthy volunteers into the design of clinical trials involving patients. This strategy is a component of the broader Model-based Drug Development framework, a field with an abundant literature on optimizing trial designs and making well-informed decisions [12].

The extrapolation of data from healthy volunteers to patients is a well-explored subject in the pharmacometrics literature [13]. Previous research within this domain has predominantly focused on extrapolating data from adults to pediatric populations [14, 15]. Additionally, some researchers have examined bridging models to identify disparities between healthy volunteers and patients [16, 17]. Within the statistical literature, several publications have focused on the extrapolation from pharmacokinetic models to clinical data [18], and increasing attention has been given to the integration of historical data [19–21]. The principal benefit of these approaches lies in their capacity to estimate treatment effects with precision and increase study power, all while controlling, or even diminishing [22], type I errors in cases of consistency between historical and concurrent data. However, in instances of conflict between prior and current data, type I errors might be exacerbated. Bayesian dynamic borrowing (BDB) techniques have been designed to address this challenge [23, 24] and offer the advantage of down-weighting historical data during analysis when they differ from the study data [22, 25–27]. These techniques have also undergone evaluation in the context of platform trials [28].

In the subsequent sections, we present a motivating case study and outline the primary steps of our proposed approach in Section 2.2. Section 2.3 presents the methodology, encompassing the mPBPK model, the extrapolation process for obtaining informative priors, and the BDB design. The proposed approach is then applied to an immuno-inflammation case study in Section 2.4, where we present comprehensive operating characteristics for various designs to determine the most suitable one. Additionally, we conduct a hypothetical analysis as if study data were available, providing insight into the prospective final analysis. Finally, Section 2.5 serves as a concluding discussion.

2.2 Motivating example

The motivation for this study stemmed from a real-life case in immuno-inflammation, where there was a need to leverage historical receptor occupancy (RO) data from a phase I study involving healthy volunteers. The intention was to use this historical information to inform the design of a phase II proof-of-concept (PoC) study in patients. The rationale behind this approach was driven by ethical considerations and practical constraints imposed by the rarity of the disease, which limited the available sample size for the phase II investigation. To construct a robust understanding of the drug's behavior, RO data from healthy volunteers, in conjunction with pharmacokinetic data, were used to develop a mPBPK model for the drug at the conclusion of the phase I study. RO, defined as the fraction of receptors occupied by a specific drug out of the total receptor population, plays a pivotal role in elucidating or confirming the mechanism of action of certain medications [29]. It enables the quantification and characterization of the drug's binding profile to the target [30]. RO serves as a valuable pharmacodynamic biomarker for characterising the relationship between drug exposure and response, especially when coupled with a pharmacokinetic profile. Its evaluation in phase I studies aids in forecasting treatment efficacy and facilitates dose selection in subsequent investigations. In the context of this study, the model developed in healthy volunteers was extended to patients, accounting for both the variability in population parameters and the peculiar differences attributable to the pathological condition.

Simultaneously, a phase II PoC study was considered, designed as a randomized, double-blinded, two-arm trial assessing both efficacy and safety. This study employed a multiple dosing regimen spanning 13 weeks and enrolled 45 patients, with 30 allocated to the treatment group and 15 to the control group. The clinical endpoint was continuous, with negative values signifying improved efficacy, denoting an enhancement in the pathological condition compared to baseline. A reduction of 3 units in the clinical endpoint was considered clinically relevant. The clinical endpoint is assumed to follow a normal distribution with a reported standard deviation of 6 from the literature. Furthermore, a linear relationship between RO and the clinical endpoint at the patient level had been estimated from a prior, internal, and unpublished study.

To design the study, BDB methods, which will be expounded upon in the next sections, were employed to integrate prior evidence from the phase I data on RO obtained from healthy volunteers. The primary objectives were to enhance the study's statistical power while keeping the sample size manageable and to enable informed decision-making at the conclusion of the PoC study by incorporating all available evidence from both current and past investigations.

Regrettably, specific details regarding the disease in question cannot be disclosed due to confidentiality constraints. Nonetheless, the scientific inquiry and simulations performed here are presented in an anonymised manner to offer valuable insights for potential future applications. Although the decision was ultimately made not to pursue this particular approach, the research presented here lays out a general framework that can be applied to various disease areas where RO is expected to be linked to drug efficacy. This includes, but is not limited to, domains like monoclonal therapeutic antibodies in oncology [31, 32] or certain neurological disorders [33]. In other applications, these methods could also facilitate the substitution of RO with a biomarker closely tied to the desired clinical endpoint, potentially resulting in a stronger correlation and/or reduced variability, depending on the available previous data or literature in the context of the project.

2.3 Methods

In the upcoming section, we will present the approaches utilized to model RO in both healthy volunteers and patients, outlining the methods used to integrate historical data into a BDB design, and elucidating the techniques employed for extrapolating from phase I to phase II data.

2.3.1 mPBPK model

Physiology-based pharmacokinetic (PBPK) models are mathematical frameworks employed to predict how a drug will behave in terms of exposure and response under various dosing regimens within a specific target population. In contrast to traditional pharmacokinetic models, PBPK models rely on established anatomical, physiological, physical, and chemical principles for parameterization. These models often feature a larger number of compartments, with multiple differential equations capturing the drug's dynamics across these compartments. To streamline the implementation of PBPK models, mPBPK models were developed. These mPBPK models simplify the complexities of PBPK models by grouping tissues with similar kinetic properties into two categories based on their endothelial structure: "tight" and "leaky" compartments. For an in-depth exploration of PBPK and mPBPK models, please refer to Jones at al. (2013) [32] and Cao and Jusko (2014) [34], respectively.

In the context of the upcoming case-study, the subsequent differential equation elucidates the behavior of RO within the leaky compartment, responsible for generating the relevant secretions. A comprehensive depiction of the full mPBPK model employed in this case study can be found in the Supplementary Material and is generally represented as a function $f(\cdot)$.

$$\frac{\partial C_{leaky}}{\partial t} = f\left(C_{leaky}, C_{free}, V_{max}, K_m, V_{leaky}\right)$$
$$RO = \frac{C_{leaky}}{C_{leaky} + K_m}.$$

Table 2.1 Key parameters of the mPBPK model. V_{max} and K_m are individualized parameters derived from prior investigations involving healthy volunteers. V_{leaky} is a constant volume applicable to all individuals. Additional information regarding the model and these parameters can be located in the Supplementary Material.

Parameter	**Description**
C_{leaky}	Concentration of the drug in the leaky compartment
C_P	Concentration of the drug in the plasma
V_{max}	Maximum binding capacity in the binding site
K_m	Concentration of the free (not bound) drug
V_{leaky}	Volume of the distribution of the drug in the leaky compartment

The main parameters used in the model are shown in Table 2.1.

The unique nature of the disease mandates a modification in the mPBPK model for patients. Specifically, the maximum binding capacity at the binding site (V_{max}) for patients must be scaled by an individual factor denoted as λ. This scaling factor is assumed to follow to a log-normal distribution, and its parameters are estimated based on data from external studies. This adjustment is considered the sole difference between the mPBPK model applied to healthy volunteers and that applied to patients.

2.3.2 Bayesian dynamic borrowing design

The upcoming phase II PoC study is designed as a BDB design, aiming to incorporate historical data to enhance the evidential basis for informed decision-making.

Let θ_T and θ_C represent the true mean values of the clinical endpoint within the treatment and control arms of the new study, respectively. Furthermore, let π_T and π_C denote the informative continuous prior distributions corresponding to θ_T and θ_C, respectively, which encapsulate the knowledge derived from the clinical endpoint data in the phase I study (as discussed in Section 2.3.3). To address potential conflicts between the prior information and the concurrent data, these distributions are robustified by adopting a mixture prior approach. This approach combines them with two less informative distributions, denoted as π_T^V and π_C^V, which exhibit the same mean but substantially larger variance. The robustified distributions are derived as:

$$\begin{aligned}\pi_T^R &= w\pi_T + (1-w)\pi_T^V \\ \pi_C^R &= w\pi_C + (1-w)\pi_C^V\end{aligned}$$

The distributions π_T^R and π_C^R will serve as the prior distributions for θ_T and θ_C in the phase II PoC study.

The prior weight, denoted as w, represents the level of confidence, prior to observing the phase II data, in the relevance of the information extrapolated from phase I to the phase II study. For simplicity, the same prior weight is applied to both the treatment and control arms, under the assumption that the extrapolated data holds equal relevance for both arms. However, if there are scientific justifications, distinct prior weights could be considered for each arm. The choice of the prior weight(s) necessitates careful deliberation, typically striking a balance between the preconceived expectations regarding the extrapolated data relevance and the frequentist operational characteristics. A comprehensive examination of the operational characteristics of the adopted methodology is furnished in the ensuing case-study to facilitate the selection of design parameters, including the prior weight assigned to the extrapolated data. Additionally, a sensitivity analysis known as a tipping point analysis [35] is presented to illustrate the process of assessing the robustness of the outcomes following the data collection phase.

2.3.3 Extrapolation

The extrapolation from phase I to phase II data involves two sequential steps: firstly, utilizing data from healthy volunteers to predict RO in patients, and secondly, employing the predicted RO values in patients to estimate the clinical endpoint. The methodology for both treatment arms is essentially identical; however, for simplicity, we present the methodology specifically for the treatment arm, denoted as T.

Let N_T represent the number of patients in the treatment arm, and γ_i denote the logit transformation of the true RO value for patient $i = 1,...,N_T$, defined as $\gamma_i = \text{logit}(\text{RO}_i) = \log(\text{RO}_i/(1-\text{RO}_i))$. Furthermore, let θ_i signify the true treatment effect on the clinical endpoint for each patient $i = 1,...,N_T$. Consequently, the true mean of the clinical endpoint in the treatment arm is $\theta_T = \frac{1}{N_T}\sum_{i=1}^{N_T}\theta_i$. The association between the logit transformation of RO and the clinical endpoint for each patient is assumed to have been previously estimated, either through existing literature or prior studies. While this assumption holds to a certain extent if the post-phase I dose selection relies on RO results, in practical scenarios, the estimation of this relationship may be uncertain and may need to be substituted with clinical assumptions.

Following a similar approach to Saint-Hilary et al. (2018) [36], a Bayesian regression model is applied, where $\theta_i \sim N(a_i + b_i\gamma_i, \tau_i^2)$ for $i = 1,\ldots,N_T$, with $a_i \sim N(\mu_a, \sigma_a^2)$, $b_i \sim N(\mu_b, \sigma_b^2)$, and $\tau_i \sim HN(\frac{1}{\mu_\tau})$.

The prior distribution π_T for θ_T is approximated using the following methodology:

- Generate K sets of $\hat{\gamma}_{i,k}$ values for each patient $i = 1,...,N_T$ and for each simulation $k = 1,...,K$ using the mPBPK model.
- Sample values for $a_{i,k}$, $b_{i,k}$, and $\tau_{i,k}$ from the estimated distributions of a_i, b_i, and τ_i for each patient $i = 1,...,N_T$ and each simulation $k = 1,...,K$.
- Sample $\hat{\theta}_{i,k}$ values for each patient $i = 1,...,N_T$ and each simulation $k = 1,...,K$ from normal distribution $N(a_{i,k} + b_{i,k}\hat{\gamma}_{i,k},\ \tau_{i,k}^2)$.

- For each patient $i = 1, ..., N_T$, select a single value $\hat{\theta}_{i,k^*}$ from the set of $\hat{\theta}_{i,k}$ values, where k^* represents a fictive trial.
- Calculate the mean effect on the clinical endpoint for the entire treatment arm in the trial, denoted as $\hat{\theta}_{T,k^*} = \frac{1}{N_T} \sum_{i=1}^{N_T} \hat{\theta}_{i,k^*}$.
- Reiterate the previous two steps a large number of times to approximate the distributions of the mean clinical endpoint in the trial.

This procedure yields estimates of the mean effect on the clinical endpoint for the treatment arm, denoted as $\hat{\theta}_{T,k^*}$, which are utilized to approximate the prior distribution π_T. The same steps are replicated to obtain an approximation of the prior distribution π_C for the control arm.

2.4 Case-study in immuno-inflammation

The methodologies proposed in this paper are now applied to a practical case study. The analyses involving the mPBPK model were conducted using Simulx Version 2020R1, while the BDB design was implemented using R Version 4.0.5, along with the RBesT package Version 1.6.1. It's important to note that, as previously mentioned, the data used in this section are simulated due to confidentiality constraints.

In this study, a dual criterion for success is considered. First, it involves achieving a statistically significant reduction in the clinical endpoint within the treatment group compared to the control group at a one-sided 10% significance level. This is translated into a posterior probability condition as $\mathrm{P}[(\theta_T - \theta_C) < 0] > 0.9$. Second, there is an emphasis on ensuring a reasonably high level of confidence that the reduction in the clinical endpoint in the treatment group is at least 3 points greater than that in the control group, expressed in terms of posterior probability as $\mathrm{P}[(\theta_T - \theta_C) < -3] > 0.5$. These criteria were selected to strike a balance between feasibility and risk, with the second criterion chosen to mitigate the risk of making incorrect decisions [37]. As outlined in Chuang-Stein and Kirby (2017) [38], the rationale for a "Go" decision under this approach is to have at least a 50% confidence that the true treatment effect exceeds a predefined clinical threshold while maintaining sufficient precision to be confident that the true treatment effect is greater than zero.

In a frequentist framework, which is equivalent to a Bayesian design assuming an implicit improper conjugate Normal prior with zero precision, the utilization of these two criteria results in an overall type I error rate of approximately 5.2% and a statistical power of 50% in scenarios where the true difference between treatment and control is just 3 points. This implies that the study, without employing any borrowing techniques, is designed to have a false positive rate of approximately 5% when there is no true difference between treatment and control. Additionally, the study exhibits a power of less than 14% for detecting true differences of less than 1 point between treatment and control, while it boasts a power of around 70% when the genuine difference between treatment and control reaches 4 points.

2.4.1 mPKPB model and BDB design

The distribution of the factor λ, which is used to adjust the maximum binding capacity (V_{max}) in patients compared to healthy volunteers, is estimated based on external studies as follows: $\log(\lambda) \sim N(1.297, 0.604^2)$. This distribution has a median of approximately 3.659, indicating that V_{max} tends to be substantially higher in patients compared to healthy volunteers.

Following the procedures outlined in Section 2.3.3, we simulated a total of 1000 clinical trials to derive the prior distribution for each arm. The choice of this number was made after assessing that further increasing the number of simulations did not significantly enhance the overall precision of the results. It's worth noting that the required number of iterations may vary in different contexts, and it should be determined on a case-by-case basis.

The replicates of ROs at the end of week 13 were generated using Simulix. The histograms of the logit-transformed ROs, denoted as $\hat{\gamma}_{i,k}$ for $i = 1, ..., N_T$ (or N_C for the control) and $k = 1, ..., K$, are provided in the Supplementary Material. The relationship between RO and the clinical endpoint for each patient had been established in a previous (internal and unpublished) study, with the parameters set as follows: $\theta_i \sim N(a_i + b_i\gamma, \tau_i^2)$, $a_i \sim N(-2.893, 0.029^2)$, $b_i \sim N(-0.181, 0.003^2)$, and $\tau_i \sim HN(\frac{1}{5.040})$ for $i = 1, ..., N_T$ (or N_C for the control).

The two informative prior distributions obtained are as follows:

$$\pi_T \sim N(-3.786, 1.148^2)$$
$$\pi_C \sim N(-0.018, 1.595^2)$$

To robustify these informative priors, vague priors π_T^V and π_C^V were introduced. These vague priors have the same mean as the informative priors but a standard deviation of 6, which corresponds to the sampling standard deviation of the clinical endpoint observed in external studies. Therefore, the vague prior components are equivalent to the information equivalent of "one patient" in each arm.

Table 2.2 presents key characteristics of various mixtures of the treatment and control priors for different values of the prior weight assigned to the informative component (w). These characteristics include the mean, standard deviation, and effective sample size (ESS) as calculated using the ELIR method [39]. The informative priors collectively contribute an ESS of 41 patients, with the treatment prior having an ESS of 27 patients and the control prior having an ESS of 14 patients, both slightly lower than the number of patients enrolled in their respective arms. As the prior weight on the vague component increases, the ESS decreases until it reaches an ESS of 2 patients when the phase I information is entirely disregarded ($w = 1$), as expected based on the construction of the priors.

Table 2.2 Characteristics of various priors corresponding to different weights (w) allocated to the informative component. A weight of 1 signifies the full incorporation of information without any robustification, whereas a weight of 0 implies exclusive reliance on the vague priors. ESS stands for Effective Sample Size, as calculated through the ELIR method.

Weight (w)	**Treatment Mean**	**Treatment Standard Deviation**	**Treatment ESS (ELIR)**	**Control Mean**	**Control Standard Deviation**	**Control ESS (ELIR)**	**Total ESS (ELIR)**
1	-3.786	1.148	27	-0.018	1.595	14	41
0.8	-3.786	2.873	18	-0.018	3.039	9	27
0.65	-3.786	3.668	13	-0.018	3.775	7	20
0.5	-3.786	4.320	9	-0.018	4.390	5	14
0	-3.786	6	1	-0.018	6	1	2

2.4.2 Operating Characteristics

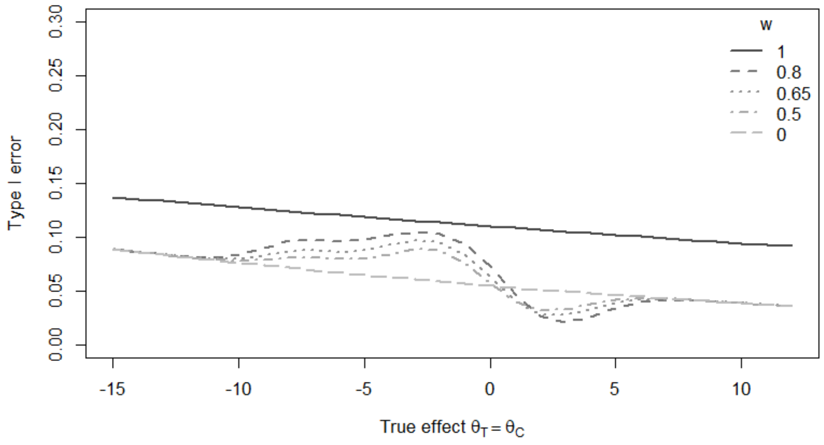


Fig. 2.1 The figure depicts the relationship between the Type I error and the effect size in both arms, assuming equal effect sizes, while considering 30 patients in the treatment arm, 15 patients in the control arm, and varying prior weights (w) assigned to the informative component.

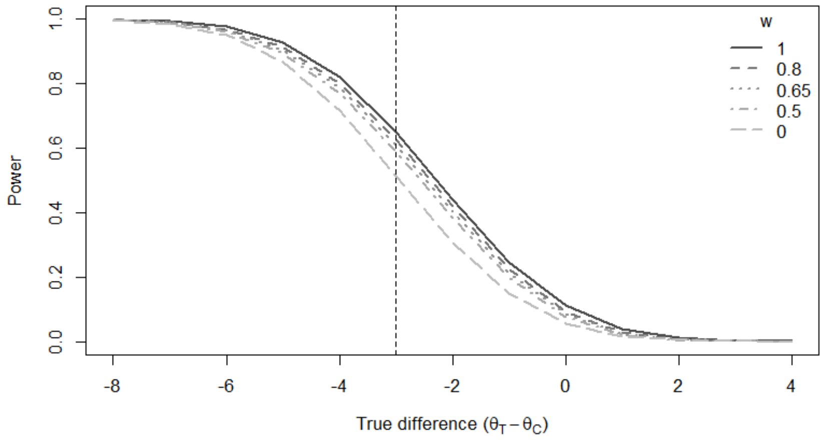


Fig. 2.2 The figure displays the power within the BDB design as a function of the discrepancy between the effects in each arm. This analysis assumes a control effect of $\theta_C = -1$, involves 30 patients in the treatment arm, 15 patients in the control arm, and various prior weights (w) assigned to the informative component. It's worth noting that a smaller difference between the treatment and control effects indicates a more favorable treatment outcome. The vertical dotted line signifies the minimum clinically significant difference of -3.

Table 2.3 Summary of the operating characteristics for BDB designs with 30 patients in the treatment arm, 15 patients in the control arm, varying prior weights (w) on the informative component, and design without borrowing. The plausible for the treatment and control effects is specified as [-7.3, 4.9].

Design	**Type I error when $\theta_T = \theta_C = -1$**	**Maximum type I error over the plausible range** (value at which occurs)	**Range of values where type I error is greater than 10%** (probability under π_T and π_C)	**Power when** $\theta_T = -4$ and $\theta_C = -1$
BDB with w=1	11.2%	12.3% (-7.3)	[-7.3, 4.9] (99.8%)	64.9%
BDB with w=0.8	9.3%	10.5% (-2.8)	[-4.4,-1.6] (10.7%)	62.8%
BDB with w=0.65	8.4%	9.7% (-2.6)	-	61.0%
BDB with w=0.5	7.6%	8.9% (-2.6)	-	59.1%
BDB with w=0	5.7%	7% (-7.3)	-	51.2%
Frequentist	5.2%	5.2% (all)	-	50%

Operating characteristics are presented to assess the performance of the BDB design and facilitate the selection of an appropriate weight for the informative component (w). Figure 2.1, Figure 2.2, and Table 2.3 offer insights into the type I error and power for various w values, taking into account the phase II PoC study's parameters. Plausible ranges for the treatment and control effects were established using the 0.1% percentile of π_T as the lower bound and the 99.9% percentile of π_C as the upper bound. These percentiles define a range of plausible values spanning from -7.3 to 4.9, according to prior knowledge.

Figure 2.1 illustrates the type I error across different assumptions regarding the true effect on the clinical endpoint, assuming it is identical for both the treatment and control arms. The type I error rates remain below 10% when the weight on the informative component is zero ($w = 0$). However, they increase as the true effect decreases due to the unbalanced design. The vague priors used for both arms contain the same amount of information (equivalent to one subject's observation). Therefore, the prior has a greater impact on the control arm because there are more patients in the treatment arm to counterbalance it. Consequently, when the true effect is lower than the effect extrapolated from RO, the posterior estimate for the control arm decreases to a lesser extent than that of the treatment arm. This leads to an increased posterior estimate of the difference between treatments and an increased type I error. Conversely, when all prior information is borrowed ($w = 1$), there is a substantial increase in the type I error, exceeding 10% for all values and increasing as the true effect decreases, again due to the unbalanced nature of the design. It's important to note that this scenario, where $w = 1$ and the effective sample size (ESS) of the borrowed data nearly matches the phase II study's sample size, is likely to be excessive and is presented for illustrative purposes only. However, for intermediate values of w, the increases in type I error are smaller and less pronounced. The type I error surpasses 10% within the plausible range only when $w = 0.8$ and the true effect falls within $[-4.4, -1.6]$, reaching a peak of 10.5% when the true effect is -2.8. The probability that both the treatment and control effects fall within the range $[-4.4, -1.6]$ is only 10.7% based on π_T and π_C. For true values lower than -4.4, the type I error remains between 10% and the type I error for the BDB design when the informative prior is completely ignored ($w = 0$). However, for true values higher than -1.6, the type I error decreases rapidly, remaining below the type I error with $w = 0$ for nearly all positive values of the true effect.

Figure 2.2 portrays the power as a function of the difference between treatment and control group effects, assuming a control effect of -1. The vertical dashed line signifies the second component of the dual success criterion: a mean reduction in clinical endpoint in the treatment group of 3 points or more than the control group. In this context, the design without borrowing has a power of 50%, by design. Conversely, the BDB designs exhibit increased power with higher weights: 59.1% ($w = 0.5$), 61% ($w = 0.65$), 62.8% ($w = 0.8$), and 64.9% ($w = 1$). Notably, the BDB design with $w = 0$ already displays greater power (51.2%) compared to the design without borrowing due to the incorporation of information equivalent to one observation in each arm via the vague priors.

Table 2.3 summarizes the operating characteristics of the design without borrowing and the BDB designs, encompassing type I error, maximum type I error, the range where the type I error exceeds 10%, and power. These results illustrate that integrating extrapolated RO data and employing a BDB design can elevate the study's power while curbing the inflation of the type I error in most scenarios. This approach effectively augments the sample size, equivalent to adding 14 to 41 patients according to the ELIR method, which is especially valuable in early-phase PoC studies with limited sample sizes.

It can be observed from the figure in Supplementary Material that γ in the treatment arm exhibits bimodal behavior. This arises because some patients may experience a faster or slower decay in RO, influenced by individual characteristics. An additional analysis, detailed in the Supplementary Material, addresses this behavior through an extension of the mixture prior. This extension incorporates an extra informative component to better capture such a distribution. The resulting model and operating characteristics closely resemble those presented here.

Furthermore, the Supplementary Material provides additional operating characteristics. Heatmaps reveal that both pointwise type I error and maximum type I error tend to increase with higher weights and smaller sample sizes. Conversely, power tends to rise with increased weights and diminished sample sizes. Separate heatmaps are generated for the first success criterion (statistical significance) and the second success criterion (clinical relevance) only. The results indicate that the second success criterion, which requires a clinically significant difference between treatment and control arms ($\mathrm{P}[(\theta_T - \theta_C) < -3] > 0.5$), primarily influences the overall operating characteristics of the designs.

While the choice of the weight for the informative component should be evaluated on a case-by-case basis depending on project-specific considerations, let's assume, for now, that a weight of $w = 0.8$ is deemed appropriate. According to Table 2.3, this weight results in a type I error of around 10% and a power of 62.8%. With this weight, the long-term operating characteristics of the BDB design, accounting for potential deviations between the true treatment/control effects and the prior means, are presented. The findings indicate that the average posterior weight on the prior distribution increases when the prior and true effects align closely, leading to a narrower posterior credibility interval (CrI) compared to a design without borrowing in cases of prior data consistency. However, as deviations increase, less weight is assigned to the RO-informed evidence, resulting in reduced precision gains. Additional results, including bias and illustrative outcomes under selected data scenarios, are presented to provide a comprehensive understanding of the BDB designs.

2.4.3 Fictive Analysis

To prepare for the final analysis of the trial, simulated and analyzed fictive results are presented in Table 2.4. The 80% credible intervals (CrI) are reported to ensure alignment with the first success criterion, which corresponds to a one-sided type I error of 10%. In the simulated data, the observed treatment mean is -4 with an 80% CrI of [-6.1, -1.9] in a cohort of $N_T = 30$ patients, while the observed control mean is -1 with an 80% CrI of [-4, 2] in a cohort of $N_C = 15$ patients.

The analysis of the design without borrowing is presented in the bottom row, revealing an observed treatment difference of -3 with an 80% CrI that encompasses zero. Consequently, neither of the success criteria is met. In contrast, the Bayesian analysis in the top row combines evidence from the current study data and the extrapolated data from RO. The posterior treatment difference is -3.5, slightly influenced by the informative priors, and exhibits better precision than the analysis of the design without borrowing. The 80% CrI does not include zero, and both success criteria are satisfied.

Sensitivity analyses, such as the tipping point analysis [35] illustrated in Figure 2.3, are crucial for comprehending and evaluating the robustness of the findings. By exploring the impact of different weights on the conclusions, these analyses offer valuable insights into the design's sensitivity to the prior belief regarding the relevance of the extrapolated RO data to the new study's effects. In this particular case, the analysis demonstrates that the conclusions of the BDB design remain robust even with variations in the weight, as the first success criterion is only met for weights greater than $w = 0.1$, underscoring the design's reliability.

A hypothetical scenario involving a false positive is presented in the Supplementary Material to illustrate that incorrect decisions can be averted through appropriate analysis and the application of sensitivity assessments.

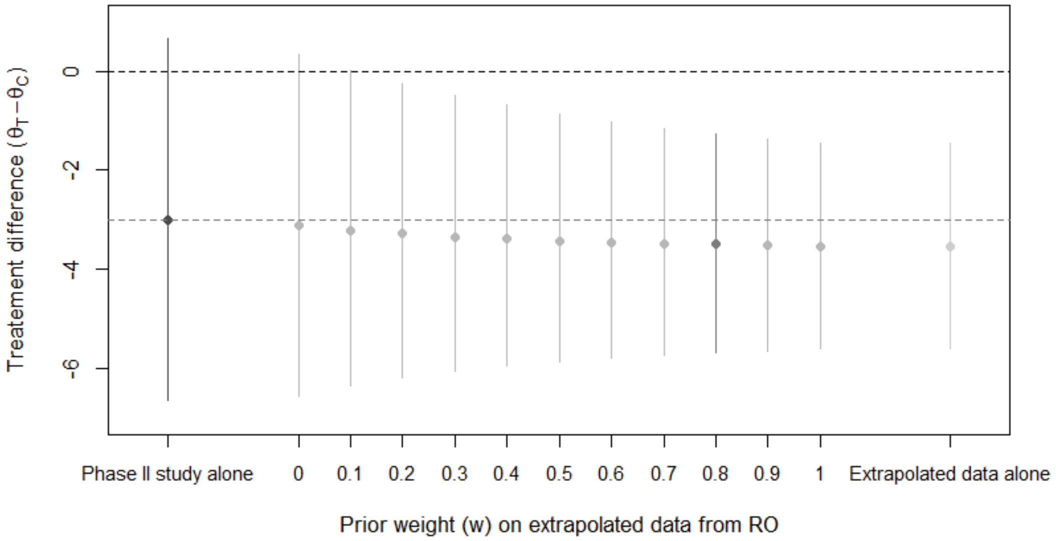


Fig. 2.3 Sensitivity analysis conducted using hypothetical but realistic data, displaying the posterior mean and 80% credible interval (CrI) for the estimated treatment difference in relation to the prior weight. The two dashed lines on the graph represent the two success criteria thresholds: $(\mathrm{P}[(\theta_T - \theta_C) < 0] > 0.9)$ and $(\mathrm{P}[(\theta_T - \theta_C) < -3] > 0.5)$.

Table 2.4 Summary of the primary analysis on the treatment difference, treatment and control response, utilizing hypothetical but realistic data. The lower row showcases the simulated observed data from a design without borrowing, clearly illustrating a failure to meet the success criteria. In contrast, the upper row presents the outcomes achieved by combining fictive observed data and informative components using a BDB design with a weight of $w = 0.8$ demonstrating the fulfillment of the success criteria.

Evidence Source	**Treatment difference [80%CrI]**	**Treatment effect [80%CrI]**	**Control effect [80%CrI]**
Phase I + phase II	-3.5 [-5.7;-1.3]	-3.9 [-5.1;-2.6]	-0.4 [-2.2;1.4]
Phase II only (frequentist)	-3 [-6.6;0.6]	-4 [-6.1;-1.9]	-1 [-4;2]

2.5 Discussion

This work introduces a methodology for integrating data gathered in a phase I study on RO in healthy volunteers as prior information for a phase II PoC study in patients. The results and comprehensive assessment of the design demonstrate that incorporating extrapolated RO data through a BDB design leads to enhanced study power while maintaining the type I error at acceptable levels. These findings support previous studies employing historical data borrowing [22, 25–27]. Importantly, they also facilitate the evaluation of the influence of the prior weight choice on the outcomes, enabling more informed decisions rather than relying solely on external beliefs [36]. The fictive analyses presented here illustrate how results can be presented and their robustness assessed when real study data are available. Furthermore, they illustrate how additional evidence from RO data can reduce the risk of leaving phase II results in a "consider zone" [40], where decisions about drug development continuity are unclear. Additional analyses detailed in the Supplementary Material further underscore the value of such assessments. Incorporating RO data significantly contributes to the study, potentially doubling the effective sample size (ESS). Nevertheless, the analysis indicates that using an informative prior with $w = 1$ is unlikely, given its impact on operating characteristics. In more practical scenarios, integrating RO data leads to sample size increments ranging from one-third to two-thirds of the study's total sample size.

While specific details about the disease motivating this work remain confidential, the statistical methodology presented here has broader applicability in other drug development scenarios where drug efficacy is expected to correlate with some RO. This approach can be a valuable tool for optimizing the design and analysis of such trials. It's important to emphasize that this method is intended for use in early-phase studies and should not replace the need for later-stage randomized trials to confirm drug effects.

The 50% power observed in our example may raise concerns for some readers. However, this is inherent in the utilization of a dual-criterion design based on both statistical significance and clinical relevance [41, 42]. In such designs, trial success is determined not only by achieving statistical significance but also by surpassing a clinically meaningful threshold for the treatment effect estimate. The power calculated at this threshold value is approximately 50% because, if the true parameter equals the threshold, there is an equal probability that the effect estimate will fall on either side of it.

The proposed methodology has some limitations, primarily linked to the extrapolations from healthy volunteers to patients and from RO to the clinical endpoint, which require external evidence. Indeed, while the BDB design has several advantages over traditional designs, potential issues should be considered. One such concern is that the prior evidence used may not represent the current study population accurately or could be based on flawed or incomplete information. This can lead to biased estimates, especially if the relationship between RO and the clinical endpoint is incorrectly estimated, potentially propagating errors or biases to the current study. However, mPBPK models are well-established in the literature [43] and regulatory frameworks [44, 45], and their predictive performance can be validated on independent datasets [46].

Additionally, we assume that the relationship between RO and the clinical endpoint has been previously estimated based on the literature or past studies. While this assumption should be somewhat accurate when the dose is intended to be chosen based on RO results, the estimation of this relationship may be impractical in practice and replaced by clinical assumptions. We strongly discourage selecting doses based on RO without a reliable empirical estimation of the relationship between RO and the clinical endpoint, acknowledging that this assumption is the primary limitation of our proposed methodology.

Before progressing to human trials, a translational framework should have demonstrated a clear connection between drug target exposure, the desired pharmacodynamic effect (biomarker), and model efficacy [11]. RO serves as a marker indicating the direction and magnitude of treatment activity, offering insights into whether the drug is achieving the desired effect. However, there may be other, more proximate biomarkers that are closer to the clinical endpoint, such as proof of principle (POP) biomarkers or Proof of Concept (PoC) biomarkers [11]. Employing these biomarkers alongside RO could further enhance confidence in the relationship with the clinical endpoint. Surrogate markers measuring the pharmacological treatment effect [47] or available pre-clinical information can also be valuable for guiding dose selection. In the example presented, the PoC study assesses a well-established clinical endpoint in the target indication, enriched with information borrowed from a surrogate marker. If RO were found to be an inadequate predictor of the clinical response, the risk of misguided borrowing from RO data is mitigated by the BDB design, which discards prior data in cases of conflicts with new data. Furthermore, the impact of the strength of the relationship between RO and the clinical endpoint on design properties could be further evaluated, as demonstrated for probabilities of success [36].

It's essential to note that the number of simulated patients in the fictive trial affects the precision of the prior distributions. If this number exceeds the number of enrolled patients (N), the prior distributions may have an effective sample size larger than N. To address this concern, downweighting the prior could be considered by increasing the variance of all components by the same factor, as demonstrated in recent studies [48]. This approach allows new data to have sufficient weight when the prior has a high effective sample size, ensuring that the posterior distribution appropriately reflects the agreement or conflict between prior beliefs and current data. Therefore, in our proposed methodology, simulating a number of patients equal to N enables the consideration of prior-data conflicts while allowing the current data to guide decisions. The downweighting approach can be applied if the prior has a high effective sample size.

The proposed methodology builds upon the approach proposed in Saint-Hilary et al. (2018) [36] by extending it to incorporate healthy volunteer RO data. In this work, the focus is on using historical data from a phase I study in healthy volunteers to inform the design and analysis of a phase II PoC study in patients. One natural extension of this methodology is to consider a seamless Phase I/II design within a Bayesian framework, where data from the Phase I stage inform the analysis of the Phase II stage. As another potential extension, a longitudinal analysis could explore the behavior of RO over time and its relationship with the clinical endpoint. Additionally, sensitivity analyses regarding the factor λ, which relates RO in healthy volunteers to patients, could be performed under different assumptions about its distribution to assess its impact on the BDB design's operating characteristics. Lastly, the current study assumes a fixed, independent distribution for a_i, b_i, and τ_i in all patients $i = 1, ..., N$. However, considering different distributions based on patient characteristics could be explored, with the joint distribution estimated accordingly [36]. Moreover, alternative relationships between the logit of RO and the clinical endpoint could be investigated, such as linear improvements in the clinical endpoint for the logit of RO above a minimum level or piece-wise relationships.

In conclusion, the proposed methodology is expected to be a valuable tool for supporting decision-making in early phases, where the number of patients is limited. It has the potential for broader application in various contexts and with different biomarkers or activity criteria, demonstrating the possibility of improving the efficiency of PoC trials by leveraging historical information in drug development.

Chapter 3

Incorporation of pre-clinical animal toxicology data in phase I trials

Background

This chapter presents a discussion on an ongoing work.

3.1 Introduction

Nowadays, preclinical animal data in oncology studies are mainly used for clinical consideration, such as pharmacological safety and general toxicology, and for the determination of the pharmacokinetic and pharmacodynamic profile of a new drug in view of the first in human study [49]. Also, data collected in animal studies, such as toxicity events, are used to determine the maximum safe starting dose for the first-in-human trial, according to the current regulation based on allometric scaling and some additional safety factors [1]. The allometric scaling principle is based on proportionality between doses and dimension of the animals in terms of body weight or body surface area [50], and is widely used in the field.

The animal data contain lots of information that could be used more formally in the design via borrowing methods and may help in determining the safety profile of new compounds in a context where, usually, the available sample size for the trial is limited. In this sense, several approaches have been proposed to incorporate historical data into new studies [26, 22] and some studies have already made use of historical controls in exploratory, or even confirmatory, trials [25, 51, 48]. Among them, the meta-analytic predictive (MAP) approach [23, 24, 52] and the power prior approach [53–55] are some of the most widely used. Both are Bayesian approaches, accounting for the variation across the different sources of evidence and assigning some 'weight' on their contribution to the combined evidence. The fundamental difference between the two methods relies on how they treat a possible inconsistency of the new study data with the historical data, so-called prior-data conflict [56] or drift [22]. This inconsistency could bias the assessment of the estimated effects, leading to incorrect decisions (e.g. pursue the development of ineffective therapies, or incorrect early stop). The power prior discounts the historical data by elevating their likelihood to a certain power. The robustified version of the MAP approach consists in a hierarchical model with a prior defined as a mixture between the distribution obtained from historical data with a certain vague prior, defined as a distribution with high variance, inducing a discounting of historical data in case of drift. In both methods, one concern that is often raised is how to determine the appropriate parameters [23, 24, 52, 53, 55, 57, 58], in particular the prior confidence given on the historical data. These choices should rely on clinical judgments regarding the relevance of the historical data to the current trial and operating characteristics to assess their impact on the model and the conclusions.

The objective of this chapter is to explore the integration of pre-clinical animal data into Phase I oncology trials using the MAP and the power prior approaches, employing a Bayesian logistic regression model (BLRM) [59]. In the following, a motivating case study is introduced in section 3.2. The different methodologies are presented in Section 3.3. A discussion and some concluding remarks are provided in Section 3.4.

3.2 Motivating case study in oncology

This work is motivated by a case study in oncology. The objective is to evaluate the safety profile of a new drug and to determine its maximum tolerated dose (MTD). The dose escalation procedure of the study is governed by the BLRM. This model, used to estimate the dose-response relationship, is widely used and has demonstrated good operating characteristics. It permits to allocate more patients to the maximum tolerated dose (MTD) and to identify it with higher probability with respect to other models [60, 61]. BLRM is also very flexible and permits to use different dose-escalation procedures. A maximum of 30 patients will be enrolled in the study, divided in cohorts of 3. The following doses are planned for testing: 25, 50, 100, 200, 400, 800, 1400 mg. The starting dose is fixed at 50mg, selected according to the International Conference Harmonization (ICH) S9 guidelines for choosing a starting dose for a first-in-human trial conducted in patients with cancer [49]. Cohorts of patients will be treated with the drug until the MTD(s) are identified. After each cohort of patients is completed, the dose recommendation by the model is based on the probability that the true dose limiting toxicity (DLT) rate for each dose lies in one of the following categories: [0,16%] under-dosing, (16%,33%) targeted toxicity, [33%,100%] excessive toxicity. The choice of the thresholds has been carefully determined and is a fairly common choice in this kind of trials. For extended safety, it is prescribed that doses for the next cohort will not be more than doubled, independently from the escalation criteria used. Dose escalation will continue until the following conditions are met: all patients have been treated or all doses are declared overtoxic. It can be noted that the dose recommended by the model at any stage of the trial is based on the entire history of all available DLT information from previous cohorts, as opposed to only the number of DLTs observed in the last group of patients. The MTD(s) will be chosen, at the end of the trial, as the dose(s) with the highest probability of targeted toxicity. The final recommended dose for the phase II studies will be based on the MTD(s) estimated by the model or, if no DLTs are experienced, a recommended dose for expansion will be determined.

Given the small sample size available for the phase I study and the availability of pre-clinical toxicological data from rats and monkeys, there is interest for the incorporation of the evidence from these pre-clinical data to make dose-escalation decisions and final recommendations. Different methodologies for the incorporation of toxicological animal data in the phase I dose-escalation study are described in the present work. The robust MAP approach (following existing methodology [62, 63]) and the power prior approach are presented to show how robust is the informative prior they build from animal data. A novel methodology is hereby introduced regarding the determination of parameters for both approaches, utilizing external data as a basis. Finally, stopping rules that can be added to the models to avoid overdosing are shown, using two different approaches: the escalation with overdose control (EWOC) [64] criterion and one of its most recent extensions, the unit probability mass (UPM)-based add-on rule [65].

3.3 Methods

Considering a phase I dose escalation trial in oncology. Pre-clinical data, namely D_0, consist of a certain number of studies I, which have been previously conducted on certain animal species. A single animal species has been treated with the target drug on different dose levels $\mathscr{D}_i = \{d_{i1},...,d_{iJ_i}; d_{i1} < ... < d_{iJ_i}\}$ in each study $i = 1,...,I$. Moreover, only one species was tested in each study. A certain number of DLT r_{ij} are observed over a certain number of tested animals n_{ij} for each dose $j = 1,...,J_i$.

3.3.1 Bayesian Logistic Regression Model

Zheng et al. [62] define the following model to incorporate preclinical animal data into oncology studies, based on the BLRM. The number of toxicities r_{ij} are supposed to follow a binomial distribution with probability of toxicity p_{ij}, dependent on each dose tested $j = 1,...,J_i$ in study $i = 1,...,I$:

$$\begin{aligned} r_{ij}|p_{ij},n_{ij} &\sim Binomial(p_{ij},n_{ij}) \\ \text{logit}(p_{ij}) &= \theta_{1i} + exp(\theta_{2i})\log(\delta_i d_{ij}/d_{Ref}) \end{aligned} \tag{3.1}$$

In formula (3.1), d_{Ref} is a given reference dose. δ_i represent allometric scaling factors. However, the δ_i are not treated as fixed values, but as lognormal random variables, to account for the inherent uncertainty of these factors. The parameters of the lognormal distributions are defined to be consistent with the FDA recommendation on allometric scaling [1]. The median value corresponds to the reference value in the FDA guidelines, while the 2.5th and 97.5th percentile correspond to the associated working range. In Table 3.1 the distributions and the FDA reference values can be found, while the methodology to obtain the parameters of the distributions is treated in detail in Zheng et al. [62]. θ_i are the parameters of the BLRM model for the species considered in study $i = 1,...,I$. These are defined as:

Table 3.1 Log-normal prior parameters $LN(\lambda, \nu^2)$ for species-specific allometric translational factors, using body surface area (BSA) and body weight (BW) reference and working range values from the FDA guidelines [1].

	BW (kg)			HED (mg/kg)		HED (mg/m2)	
Species	**Reference**	**Working range**	**BSA** (m^2)	λ	ν	λ	ν
Mouse	0.02	(0.011, 0.034)	0.007	-2.562	0.298	1.050	0.283
Hamster	0.08	(0.047, 0.157)	0.016	-2.002	0.302	1.609	0.287
Rat	0.15	(0.080, 0.270)	0.025	-1.820	0.323	1.792	0.309
Ferret	0.30	(0.160, 0.540)	0.043	-1.669	0.323	1.943	0.309
Guinea pig	0.40	(0.208, 0.700)	0.050	-1.532	0.315	2.079	0.301
Rabbit	1.80	(0.900, 3.000)	0.150	-1.127	0.290	2.485	0.274
Dog	10	(5, 17)	0.500	-0.616	0.301	2.996	0.286
Monkeys	3	(1.400, 4.900)	0.250	-1.127	0.273	2.485	0.256
Marmoset	0.35	(0.140, 0.720)	0.060	-1.848	0.401	1.764	0.389
Squirrel monkey	0.60	(0.290, 0.970)	0.090	-1.715	0.269	1.897	0.252
Baboon	12	(7, 23)	0.600	-0.616	0.306	2.996	0.291
Micro-pig	20	(10, 33)	0.740	-0.315	0.284	3.297	0.268
Mini-pig	40	(25, 64)	1.140	-0.054	0.258	3.558	0.240

$$\begin{gathered}\theta_i|\mu_i,\Psi \sim BVN(\mu_i,\Psi)\\ \mu_i=\begin{pmatrix}\mu_{1i}\\ \mu_{2i}\end{pmatrix}\text{ and }\Psi=\begin{pmatrix}\tau_1^2 & \rho\tau_1\tau_2\\ \rho\tau_1\tau_2 & \tau_2^2\end{pmatrix}\\ \tau_1\sim HN(\sigma_1),\ \tau_2\sim HN(\sigma_2),\ \rho\sim U(-1,1)\end{gathered}\tag{3.2}$$

where μ_i represents a species/study-specific mean vector and Ψ is a variance matrix representative of the between-trial variability. $HN(\sigma)$ represents the distribution $N(0,\sigma^2)$ truncated to cover the positive values, while $U[a,b]$ represents the uniform distribution between a and b. In Zheng et al. model one additional "layer" is added to (3.2), to account for another source of variance representative of the between-species variability. However, since in our motivating example (and often in the reality), only one species is tested in each study, this additional layer is excluded from this analysis.

Meta Analytic Predictive Approach

For the first-in-human study (indexed by i^*) the robust meta-analytic predictive approach [23, 24] is described. The first part of the model reflects equation (3.1):

$$\begin{aligned} r_{i^*j}|p_{i^*j},n_{i^*j} &\sim Binomial(p_{i^*j},n_{i^*j}), \text{ for } j=1,...J_{i^*} \\ \text{logit}(p_{i^*j}) &= \theta_{1i^*}+exp(\theta_{2i^*})\log(d_{i^*j}/d_{Ref}) \end{aligned} \quad (3.3)$$

However, no allometric factor is used, since it is 1 for the human study. Then:

$$\begin{aligned} \theta_{i^*} &= \sum_{i=1}^{I} w_i\pi_i + w_r\pi_r \\ \pi_i &\sim BVN(\mu_i,\Psi) \\ \pi_r &\sim BVN(m_r,R_r) \\ m_r = \begin{pmatrix} m_{r1} \\ m_{r2} \end{pmatrix} &\text{ and } R_r = \begin{pmatrix} \sigma_{r1}^2 & 0 \\ 0 & \sigma_{r2}^2 \end{pmatrix} \end{aligned} \quad (3.4)$$

where m_{r1}, m_{r2}, σ_{r1} and σ_{r2} are fixed values.

The π_i distributions are informative distributions which carry the information on the animal data in study $i=1,...,I$. According to this model, the contribution of the studies to the first-in-human trial depends on the expected prior probability of exchangeability w_i for $i=1,...,I$. This is the prior belief that the behaviour of the drug in humans is comparable to the one of the species present in the studies. On the other hand, a certain prior probability w_r is placed on a robust vague prior π_r, which accounts for the fact that the behaviour of the drug in human may differ completely from the one in animals. This distribution has large variance, which permits to discount prior data in case of conflict between the animal and the current data. In the end, the prior based on animal data for the first-in-human trial is determined from the mixture of these distributions, each with its own weight, implying that $(\sum_{i=1}^{I} w_i)+w_r=1$.

Power Prior Approach

The main idea of the power prior approach [53–55] is to add variability to (and therefore robustify) the animal data by elevating their likelihood to a given power $0 \leq \alpha_i \leq 1$ for each study $i=1,...,I$. The power prior approach can be implemented in the previous model by leaving equations (3.1), (3.2) and (3.3) unchanged. Then, considering a vague prior distribution $\pi_r(\theta_{i^*})$ on the parameters for the first-in-human study, equation (3.4) is replaced by:

$$\begin{aligned}\pi(\theta_{i^*}|D_0,\alpha) &\propto \prod_{i=1}^{I}\mathscr{L}(\theta_i|D_{0i})^{\alpha_i}\pi_r(\theta_{i^*})\\ \pi_r(\theta_{i^*}) &\sim BVN(m_r,R_r)\end{aligned}\tag{3.5}$$

As for the weight in the MAP prior approach, the determination of each exponent requires some special attention. Some approaches suggest to make use of the same data D_0 to determine the exponent [55, 66, 67], since the use of a random exponent can lead to an intractable analytical calculation and highly computationally intensive model [57, 58]. However, in this case, we obtain a data-dependent prior that may lead to difficulties in interpreting the Bayesian model [68].

3.3.2 Determination of the parameters of the model

A novel methodology to derive the parameters for the aformentioned approaches is presented. It is based on a survey conducted by Olson et al. [69] on the concordance of toxicity studies in animals and humans which reports the number of concordant and non-concordant studies for animal species, alongside the pre-clinical concordance by therapeutic class for all animal data. The proportion of reported concordant studies for rats (R) is $p_R = \frac{86}{86+75}$, where 86 are concordant and 75 are non-concordant, and for monkeys (M) is $p_M = \frac{41}{41+17}$, where 41 are concordant and 17 are non-concordant. The general pre-clinical concordance from animal data in anti-cancer therapies is $w_A = 0.84$.

Using this independent source, it can be calculated the prior weight for the rat data in a MAP model as $w_R = \frac{p_R}{p_R+p_M}w_A$ and the weight for the monkey data as $w_M = \frac{p_M}{p_R+p_M}w_A$. It results that the weight for the robust component is equal to $w_r = 1-(\sum_{i=1}^{I} w_i) = 1-w_A$ and is interpreted as the prior probability of non-concordance between pre-clinical animal data and human data in anti-cancer therapies. The exponents for a power prior approach can be set to $\alpha_R = p_R w_A$ and $\alpha_M = p_M w_A$ for rat and monkey data, respectively, without the need to further re-scale them.

3.3.3 Dose escalation procedure and overdose control

The BLRM suggests as the next dose to be tested the one which has the highest probability of target toxicity. This may result in an overly aggressive escalation [59]. Therefore, an overdose control rule is commonly integrated into the BLRM to prevent overdosing.

In the following, two different procedures for overdose control are described. The first is the dose escalation with overdose control (EWOC) [64] criterion:

$$d_{sel} = \max\{d_{i^*j} \in \mathscr{D}_{i^*} : \mathrm{P}(p_{i^*j} > 0.33|D) \leq 0.25\} \tag{3.6}$$

where d_{sel} is the selected dose for the next cohort and D is the data collected up to the current point in the phase I trial. This choice is widely used in clinical trials with BLRM and is also suggested in Neuenschwander et al. [59].

Critics argue that this criterion, which is constructed taking into account just the overdose probability, may result in an excessively conservative escalation [70]. For this reason, Zhang et al. [65] propose to integrate the EWOC with the following UPM-based add-on rule:

$$\mathscr{U}_{i^*j}(\text{under}) > g(r_j)\mathscr{U}_{i^*j+1}(\text{over}) \tag{3.7}$$

where $\mathscr{U}_j(\text{under})$ is the probability of underdosing in dose d_{i^*j} divided by the length of the underdosing interval (0.16 in this case), $\mathscr{U}_j(\text{over})$ is the probability of overdosing in dose d_{i^*j} divided by the length of the overdosing interval (0.67) and r_j is the ratio d_{i^*j+1}/d_{i^*j}. In this second rule, the impact of the width of the interval on the overdosing and underdosing probabilities is taken into account. With the same observed data, the wider the interval, the higher the probability of falling into it. Therefore, dividing the probabilities by the width of the corresponding intervals helps in addressing this issue and yields to the the unit probability mass (UPM) used in the mTPI design [71]. Moreover, the ratio of the two consecutive doses r_j is also taken into account, giving the possibility to better control the procedure from this point of view.

Practically, after each cohort, the rule (3.7) will be checked and, if met, will imply escalation to dose d_{i^*j+1}. Otherwise, the dose escalation procedure will be conducted according to rule (3.6). According to Zhang et al. [65], this method assigns more patients to the MTD and allows better accuracy in its identification.

3.4 Discussion

The primary contrast between the power prior and robust MAP approaches is that the former employs static borrowing, with pre-defined exponents, while the latter employs dynamic borrowing. As a consequence, the power prior is unable to accommodate unforeseen disparities between historical and current data, and it is necessary to maintain the same level of confidence in the historical data regardless of the present outcome [22, 72]. The analysis of other extensions of the power prior, such as the modified power prior, is possible. However, these methods are known to require extensive computing [55, 57, 58], have the potential to excessively attenuate the influence of historical data [22], and may require highly informative distributions for the exponential parameters [72]. Therefore, in scenarios where there is a conflict between prior and current data, the robust MAP approach is expected to outperform the power prior approach. In contrast, the power prior is anticipated to exhibit lower variance compared to the robust MAP approach [26]. This is because the robustification process substantially increases the overall variance of the distribution and the quantity of information contained in the tails. As a result, if there is agreement between the historical and current data, the power prior might outperform the robust MAP approach. However, if there is high variability between animal studies, the contrary is expected. This is because the power prior is not designed to handle potential high variability between historical trials, as it is not based on a hierarchical model and lacks appropriate safeguards against such a situation [26].

When it comes to overdose control, the BLRM with EWOC is considered to be a highly safe option, that is unlikely to select a toxic dose as the MTD [70]. Additionally, this approach is known to have a low risk of overdosing a significant portion of patients due to its explicit assessment of overdosing control through the use of EWOC [59, 70]. On the contrary, the UPM-based add-on rule, although less intuitively interpretable than classical probabilities, considers both underdosing and overdosing. As a result, it is expected to achieve greater accuracy in identifying the MTD compared to EWOC. It is likely to enroll more patients at the MTD, increase the number of patients at higher doses, and have a higher DLT rate [65]. In the safest scenarios, the add-on rule is anticipated to facilitate faster dose escalation [59] while delivering comparable results to the EWOC criterion in a scenario where all doses are toxic [65]. For this reason, the requirement to at most double the dose provides a guiding constraint for more aggressive escalation without significantly increasing the risk of harm to patients [65].

The novel methodology to define the parameters for the robust MAP and power prior approach is expected to help the study designers to choose the parameters of the models. Its main advantage relies on the possibility to use external data to define the parameters for the model, which can have an easy interpretation for communication with the clinical teams. It permits to compare the two models here presented on equal terms, which means with the same given prior probability of concordance between the animal data and the current data. The calculation of operating characteristics, as presented in chapter 2, are still needed to assess the behaviour of the designs. Expert knowledge elicitation [73] is another option to determine the parameters. In the case of the power prior, other methods have been proposed in the literature to select the prior exponents, but these methods require extensive numerical calculations or the use of current data in defining a parameter that should be selected "a priori" [58]. Additionally, the creators of the power prior method also note that the use of a fixed exponent is more easily interpretable [55]. Therefore, the methodology presented in this study can be valuable in the presence of appropriate data from external sources, such as the literature or internal studies that are distinct from the ones used in the model.

One of the limitations of the current study is the reliance on allometric scaling to convert animal doses to human doses. While allometric scaling is appropriate for some drugs, it may not be appropriate for others [50]. However, the use of a random allometric parameter helps to account for the uncertainty in such cases and evaluate the suitability of the scaling method [62]. Moreover, it should be noted that the methods presented in this work rely on the availability of multiple pre-clinical trials conducted on different animal species. However, if only a single pre-clinical trial is available, hierarchical models can still be employed by setting a conservative prior distribution to the between-trial standard deviation [23], which will remain unchanged in the posterior. On the other hand, the power prior models can still be applied without modification to the methodology. In addition, it is important to note that phase I MTD trials with a small number of patients (20-40), particularly if the dose-toxicity curve is relatively flat and the MTD is at a high dose, may identify a dose as the MTD despite the possibility of high uncertainty in this conclusion [59].

As regards the study conducted by Olson et al., it is worth noting that no restrictions on the time frame for submitting qualifying data sets were imposed, and that the inclusive years of data collection for the entire database are not disclosed, making it unclear whether the unevenness of study designs over time may have affected the database analysis results. Furthermore, the study did not consider the "false positive" and "true negative" outcomes to assess the predictive value of prospective preclinical toxicity biomarker signals in identifying human toxicities [69]. Finally, no attempt to optimize the parameters for the compared methods has been made and, instead, the parameters are fixed based on existing literature. It is possible that different parameters for the borrowing mechanism could lead to better performance of some methods over others.

Chapter 4

Adaptive screening of a sub-population

Background

This chapter is published as:

Fulvio Di Stefano and Mauro Gasparini. *Adaptive Covid-19 screening of a sub-population.* In *Studies in Theoretical and Applied Statistics*. Springer, 2023. DOI: 10.1007/978-3-031-16609-9_8.

4.1 Introduction

In the context of the ongoing pandemic, certain working, studying, or social communities require intensified screening to promptly detect outbreaks within their members, as mentioned in [74]. Our specific case study focuses on Politecnico di Torino (POLITO), a public university, where a screening process is planned to prevent clusters among students attending in-person classes during the first semester of the academic year 2021-2022, spanning from 27 September 2021 to 14 January 2022. The objective of this study is to compare methods that dynamically and repeatedly test whether a particular sub-population remains similar to the general population it is a subset of with regards to a binary characteristic, like infection status (infected/not infected), that changes over time.

The literature extensively addresses outbreak detection, with various studies conducted on this subject [75, 76]. Existing methods like Tukey's fences [77] and other static outlier detection techniques [78] aim to identify the presence of excessive characteristics in random samples. Attribute control charts, as seen in [79], apply similar concepts to time series data. This work goes beyond those models by introducing forecasting techniques, creating a comprehensive methodology that combines characteristic forecasting with subpopulation anomaly detection. By integrating the forecasting aspect, the methodology combines the strengths of control charts, which effectively identify sudden and substantial deviations in the characteristic of interest [76], with a forecasting technique that considers data variability. The developed methodology is applied to the case study at POLITO, where it obtains adaptively varying thresholds for COVID-19 screening. These thresholds consider the time evolution of the pandemic across the entire country, resulting in alert thresholds that are not fixed but can adjust to the predicted future progression of the pandemic.

4.2 Methods

4.2.1 Modeling time evolving proportions

Consider a general population of approximately constant size N_P, where P_t represents the proportion of individuals carrying a characteristic of interest for all time points $t \geq 0$. The value of P_t is uncertain and follows an unknown stochastic process, indicating that the number of individuals with the characteristic varies over time. Within this population, there exists a specific well-defined subpopulation of interest with a size denoted by N_S, where $N_S \leq N_P$. The proportion of individuals carrying the characteristic of interest in this subpopulation at each time point $t \geq 0$ is represented by p_t. This value, p_t, follows its own distinct stochastic process, reflecting the dynamic nature of the proportion of interest within the subpopulation. If the subpopulation is conformal to the general population it belongs to, meaning the subpopulation and the general population are homogeneous at all possible scales of observations, then the proportion p_t of the characteristic of interest within the subpopulation should be approximately equal to P_t within the general population. However, it is possible for the characteristic of interest to evolve differently in the subpopulation compared to the general population. At a specific time point t_0, we have estimates of P_t for past time steps $t \leq t_0$, denoted as $\hat{P}_t$ for all $t \leq t_0$. These estimates are based on samples of varying sizes. However, this work disregards the additional uncertainty arising from these sample sizes for reasons that will be explained in detail later on. The objective of this methodology is to predict P_{t_0+1} based on estimates of P_t for all $t \leq t_0$, and subsequently conduct a statistical test to determine if the subpopulation proportion p_{t_0+1} is significantly higher than P_{t_0+1}. The focus here is on one-sided upper tests because we are specifically concerned with the possibility of an excessively high proportion p_t compared to P_t. This concern is exemplified by the POLITO case study, where detecting evidence of an excessive proportion within the subpopulation would lead to the implementation of stricter measures, such as confinement or distance learning, to control the situation.

4.2.2 Forecasting using ARMA models

With the available observed time series $\hat{P}_t$ for all $t \leq t_0$, one of the main objectives is to predict the future value P_{t_0+1}. There are two key reasons for this: firstly, to establish a reference value that adapts with the progression of the underlying characteristic of interest in the general population, and secondly, to incorporate all the gathered information up to the current point into the next prediction.

Prominent methods for predicting future values of a time series when the underlying process is unknown are ARMA models (an excellent introductory resource can be found in [80]), widely popular in the econometric literature. ARMA models have been extended in various ways and can be easily adapted to various time series. While we do not claim that ARMA models are universally adequate for all predictions, we propose employing them as a practical tool to update a population reference value.

Each ARMA model is characterized by two order parameters denoted as (p,q). These parameters need to be "identified" through an empirical model selection process based on available data, and several unknown regression parameters must be estimated from the data as well. We can apply a generic ARMA(p,q) model to our variable of interest, P_t, after an initial logarithmic transformation. This choice has been motivated by the inherent positivity of P_t, which is bounded above by 1 and has typically a small value much closer to 0. In addition, the logarithmic transformation permits to take into account the inherent etheroschedasticity of the model, allowing a more accurate prediction. An alternative choice might have been the logit transformation, which is, however, almost equivalent for values close to 0. The model is constructed as follows:

$$\log(P_t) = K + a_1 \log(P_{t-1}) + ... + a_p \log(P_{t-p}) + \varepsilon_t + b_1 \varepsilon_{t-1} + ... + b_q \varepsilon_{t-q}$$

where K represents the underlying mean of the process, the ε_t are the error terms. The model involves coefficients a_i for $i = 1, ..., p$, which are associated with the auto-regressive part, and coefficients b_j for $j = 1, ..., q$, which are associated with the moving average part. The error terms ε_t are assumed to be independent and identically distributed, following a normal distribution with mean 0 and variance σ^2.

The literature extensively covers the estimation of the model's order and coefficients [80, 81], which is beyond the scope of this work. However, it is worth mentioning briefly that the values of p and q can be selected to minimize the Bayesian information criterion (BIC) [81, 82].

In our specific setup, the estimates $\hat{P}_t$ for all $t \leq t_0$ can be utilized as proxies for P_t to estimate the model's parameters. After determining the order (p, q) of the model and estimating the parameters $\hat{a}_i$ and $\hat{b}_j$ for $i = 1, ..., p$ and $j = 1, ..., q$, as well as $\hat{K}$ and $\hat{\sigma}^2$, and obtaining the realizations of the errors $\hat{\varepsilon}_t$ for all $t \leq t_0$, we can proceed with the forecasting. Taking into account all available information up to t_0, the prediction for the next time-step can be calculated as follows:

$$\log(\tilde{P}_{t_0+1}) = \hat{K} + \hat{a}_1 \log(\hat{P}_{t_0}) + ... + \hat{a}_p \log(\hat{P}_{t_0-p}) + \hat{b}_1 \hat{\varepsilon}_{t_0} + ... + \hat{b}_q \hat{\varepsilon}_{t_0-q}$$

This quantity can be considered to be approximately normally distributed with a variance of σ^2, which is estimated as $\hat{\sigma}^2$ for practical purposes.

4.2.3 Detecting excessive presence of the characteristic of interest in the subpopulation

After forecasting for the next time period in the entire population, we can make inferences regarding the presence, particularly potential excessive presence, of the characteristic of interest in the subpopulation. To accomplish this, we consider a random sample of n_S individuals out of the total N_S individuals in the subpopulation.

Let X_t represent the number of individuals who carry the characteristic of interest among the tested individuals (n_S) at time t. The distribution of X_t can be modeled using a hypergeometric distribution with a discrete density given by:

$$\text{Prob}_{p_t}(X_t = x) = \frac{\binom{N_S p_t}{x}\binom{N_S(1-p_t)}{n_S - x}}{\binom{N_S}{n_S}}$$

which, given $N_S >> n_S$, can be approximated by a Binomial distribution:

$$\text{Prob}_{p_t}(X_t = x) = \binom{n_S}{x} p_t^x (1 - p_t)^{n_S - x}$$

where p_t is an unknown parameter that changes over time. We aim to formally test the following system of hypotheses at a significance level of $1 - \alpha$:

$$\begin{cases} H_0 : p_{t_0+1} = P_{t_0+1} \\ H_A : p_{t_0+1} > P_{t_0+1} \end{cases}$$

Based on the methods discussed in the previous section, at time t_0, we have a forecast $\tilde{P}_{t_0+1}$ for the next period along with an estimate $\hat{\sigma}$ of its uncertainty. With this information, we have various options for further steps to take.

Method 1: direct thresholding. Utilize the normal approximation obtained from the ARMA model to derive an explicit threshold. τ_{1,t_0+1} and use the decision rule

$$\text{“Reject } H_0 \text{ if } X_{t_0+1} > \tau_{1,t_0+1} := n_S \exp(\log(\tilde{P}_{t_0+1}) + z_{1-\alpha}\hat{\sigma})\text{”},$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the normal distribution.

Method 2: binomial testing. A more conventional approach, which does not account for the uncertainty surrounding P_{t_0+1}, would involve conducting a standard binomial test at a significance level approximately equal to α (accounting for the discreteness of the binomial distribution). The decision rule would be as follows:

$$\text{“Reject } H_0 \text{ if } X_{t_0+1} > \tau_{2,t_0+1}\text{”},$$

where τ_{2,t_0+1} is the $(1-\alpha)$-quantile of the binomial distribution with parameters n_S and $\tilde{P}_{t_0+1}$.

Method 3: normal testing. When the sample size n_S is sufficiently large, the binomial distribution can be approximated by a normal distribution, allowing for a precise α significance level. As a result, a threshold similar, but not equal, to the previous one can be established following the decision rule:

$$\text{“Reject } H_0 \text{ if } X_{t_0+1} > \tau_{3,t_0+1} = n_S\tilde{P}_{t_0+1} + z_{1-\alpha}\sqrt{n_S\tilde{P}_{t_0+1}(1-\tilde{P}_{t_0+1})}\text{”.}$$

The underlying concept of these techniques is to leverage the primary benefit of control charts, which excel in identifying significant and abrupt deviations from the average [76], while simultaneously incorporating a forecasting technique capable of considering the temporal evolution of the characteristic of interest.

4.2.4 A naive fixed threshold

The three previously discussed methods contrast with a naive approach, where an alert is triggered if, at a given significance level α, the null hypothesis $H_0 : p_t = p_N$ is rejected at time $t \geq 0$ using binomial testing. In this approach, p_N represents a predetermined acceptable level of the characteristic's prevalence within the subpopulation. This procedure is formally equivalent to a binary attribute control chart [79] and does not consider the temporal evolution of the characteristic of interest over time.

Method 4: naive fixed threshold binomial testing. Binomial testing using a fixed threshold.

$$\text{“Reject } H_0 \text{ if } X_{t_0+1} > \tau_N\text{”,}$$

where τ_N , is the $(1-\alpha)$-quantile of the binomial distribution parameters n_S and p_N.

4.3 A case study: COVID-19 testing at Politecnico di Torino.

4.3.1 Current testing.

During the first semester of the academic year 2021-2022 at POLITO, an estimated 21,870 students were planning to attend in-person classes. To develop an outbreak prediction system, the university implemented a cluster detection mechanism in conjunction with regular screening and preventive measures mandated by the national system. Hereafter, we propose not to test the POLITO students sub-population stand-alone, but to compare it with to the whole Italian population in order to better detect possible clusters. The screening process involves conducting 250 oropharyngeal swabs every Monday, Wednesday, and Friday, totaling 750 swabs each week. Although the upcoming discussion will focus on the test using 250 swabs, in practice, the same test is repeated three times per week. The screening is carried out on a random sample of students who have scheduled to attend their lessons on the respective day. While students have the option to decline testing, the number of refusals has been minimal up to December 3, 2021. It is noteworthy that symptomatic students who test positive tend to avoid coming to campus, opting instead to attend online lessons. Therefore, a reliable screening procedure is crucial to identifying potential hidden clusters, such as those resulting from asymptomatic cases, within the university environment.

Following the methods described in Section 4.2, the probability of a positive test on a particular day $t \geq 0$ is denoted as p_t. Thus, in accordance with Method 4, an alert will be triggered if, at some significance level α for the type I error, the null hypothesis $H_0 : p_t = p_N$ is rejected at time $t \geq 0$. By utilizing $\alpha = 0.20, n_S = 250$, and $p_N = 0.015$, an approximate estimation of the average pandemic situation in the country in September 2021, we obtain $\tau_N = 5$ as the threshold (the reason for choosing such a large level of type I error will be explained below). This particular procedure is currently implemented in POLITO, with assistance from the authors who have contributed to its establishment. Up until December 3, 2021, the screening process has resulted in very few positive tests, which remain well below the threshold. This is mainly due to the stringent rules enforced for accessing the POLITO site.

The current procedure neglects the evolving nature of the pandemic over time. In Italy, the pandemic's progression is closely monitored by the Istituto Superiore di Sanità and Protezione Civile, which provide daily data on its evolution [83]. These data offer insights into how the pandemic is unfolding in the country. It is expected that if the pandemic worsens on a national scale, it will likely also worsen at POLITO, given that the assumption of homogeneity holds true. The use of national data instead of regional (Piemonte) data serves two purposes. Firstly, despite the current regulations categorizing Italian regions into different risk areas based on regional spread, restrictions on movements (such as traveling between regions) and many social constraints do not apply to individuals possessing the "green pass certification", which is mandatory to access physical university facilities. Secondly, just over a third of POLITO students residing in Italy originate from the Piemonte region. A significant number of students come from the south of Italy, and a considerable portion also hails from the center and north-east regions. Given the diverse regional composition of students planning to attend in-person classes and the freedom granted to green pass holders, exempting them from regional constraints, national data are considered more representative of the considered subpopulation.

In Figure 4.1, the weekly percentage of positive tests in Italy since the beginning of the pandemic is displayed on left, with a focus on the last weeks on the right. However, it is essential to acknowledge that these data do not accurately represent the proportion of infected individuals at time t due to the way they are collected in Italy.

The use of both molecular tests, which are more accurate, and antigenic tests in the data collection process creates certain complexities. Molecular tests are predominantly conducted to confirm cases reported via antigenic tests, which may result in multiple reports for the same case, thereby inflating the estimate of positive cases. On the other hand, the data from antigenic tests can lead to an underestimation of positive cases since they are often administered by unvaccinated individuals to meet work-related requirements in compliance with current regulations. Furthermore, the number of weekly antigenic tests is approximately twice that of molecular tests during this phase of the pandemic, but this ratio varies over time.

As a result, the percentage of positive tests likely overestimates the true proportion of infected individuals at time t. Quantifying the magnitude of this overestimation is challenging. However, we believe it is preferable to work with the original data and employ a large type I error. This choice follows the principle that prevention is better than cure, and a false alarm does not result in severe consequences. If the authors detect an alarming situation based on the screening process data, appropriate actions will be taken. In response to an alarm, a series of containment measures will be implemented, beginning with retesting positive cases using more precise molecular tests. Subsequent measures may include quarantining individuals and potentially transitioning to online teaching for entire classes. Accordingly, we will adopt a significant level of $\alpha = 0.20$ and consider the estimates $\hat{P}_t$ for all $t \geq 0$, as displayed in Figure 4.1. These estimates are obtained by dividing the total number of new cases reported in a specific week by the total number of tests conducted (including both antigen and molecular tests) during that week. Utilizing percentages rather than raw counts is justified because percentages are known to mitigate the impact of inconsistent case reporting, a well-known aspect in outbreak detection [75].

Utilizing the dataset from the Istituto Superiore di Sanità and Protezione Civile and the previously described methodology, we can generate a forecast for the percentage of positive tests in Italy for the upcoming week. The ARMA(p,q) model will offer an appropriate prediction if the pandemic is either worsening or improving nationwide. This prediction can then be utilized to enhance the accuracy of the alarm system compared to a fixed threshold. By using weekly time series data, we obtain a quantity that can be directly compared to the percentage of positive tests among students at POLITO. The data are aggregated on a weekly basis for two reasons. Firstly, the number of tests conducted during the week is not constant but depends on the day of the week. Secondly, the tests at Politecnico are not scheduled daily, unlike those conducted across the entire country.

Lastly, considering the highly non-stationary nature of the pandemic's evolution, influenced by various factors like restrictive measures, vaccination efforts, seasonality, and variations in screening intensity (e.g., different search rates for asymptomatic individuals), it is advisable to focus on the most recent portion of the time series. In this study, we utilize only the last 16 weeks of observations and dynamically discard previous data when analyzing different t_0 times.

Various more accurate methods for predicting the number of SARS-CoV-2 positives in a general population exist in the literature. These methods include the SIS model [84], the SIPRO model [85], the SIDARTHE model [86], and its extensions [87], the Covasim model [88], as well as several others (e.g., [89–91]). However, the challenges arise due to the fluctuating number of weekly tests in Italy and the continuous implementation of social and economic measures by the government to curb the pandemic. These factors make obtaining a precise estimation of P_t very challenging. In this context, it is important to note that the proposal in this work is not to present ARMA models as an elaborate and realistic model for epidemic prevalence. Instead, the suggestion is to use ARMA models as a practical and adaptive tool that performs acceptably well for making one-time step-ahead predictions, and not for predicting further into the future.

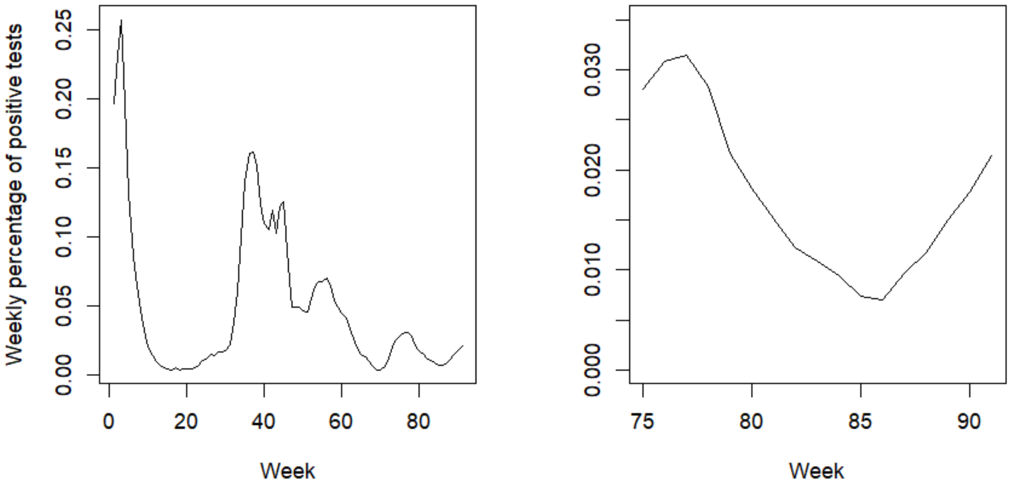


Fig. 4.1 Left: percentage of weekly positive SARS-CoV-2 tests in Italy since the beginning of the pandemic (Week 0 is 24-02-2020) up to the end of November 2021 (Week 91 is 28-11-2021). Right: focus on the weeks 75 to 91.

4.3.2 A proposal for adaptive testing

Applying the methodology described in Section 4.2, we use R version 4.1.2 and the forecast package [92] to analyze the dataset. The estimation process commences at week 74 after the beginning of the pandemic, which corresponds to July 25, 2021. At this date, the weekly percentage of positive tests in Italy is $\hat{P}_{74} = 0.025$, and the trend shows a slight increase. Using R, we can fit an ARMA(p,q) model to the dataset using the preceding 16 weeks' data.

Based on the BIC minimization, the best fitting model is ARMA(4,0). The estimated parameters are as follows: $\hat{a}_1 = 2.723$, $\hat{a}_2 = -3.334$, $\hat{a}_3 = 2.235$, $\hat{a}_4 = -0.705$, $\hat{K} = -3.876$, and $\hat{\sigma}^2 = 0.009$. Using this model, we can forecast $\tilde{P}_{75} = 0.026$. As anticipated, this value is higher than $\hat{P}_{74}$, indicating a slight worsening of the pandemic during this period. Additionally, we calculate the three thresholds discussed in subsection 4.2.3: $\tau_{1,75} = 8$, $\tau_{2,75} = 9$, and $\tau_{3,75} = 9$, with $\tau_{1,75}$ and $\tau_{3,75}$ rounded up to the next integer. Consequently, if the number of positive cases on Monday, Wednesday, or Friday of week 75 among the 250 swabs at Politecnico is greater than or equal to the specified threshold, an alert is triggered. This is because we reject the null hypothesis that the proportion of positive cases in the university p_{75} is equal to the proportion of positive cases in the whole country P_{75}, indicating evidence of an ongoing outbreak at POLITO.

After obtaining the national data for week 75, we can proceed with estimating the threshold for the following week and repeat this process over time. Figure 4.2 displays the three distinct thresholds calculated from week 75 to week 92 since the beginning of the pandemic, using all available data from the previous 16 weeks. Comparing this figure with Figure 4.1, it is evident that the model effectively incorporates the progress of the pandemic. The decrease in the weekly percentage of positive cases after week 77 is captured by the variable thresholds, as is the rising trend after week 86. Among the three different thresholds, τ_1 is the most conservative, resulting in the lowest number of positive cases needed to trigger an alarm. On the other hand, τ_2 and τ_3 yield very similar results, with τ_2 being slightly more conservative of the two. This similarity arises because the rationale behind the two thresholds is the same, but τ_3 is derived from a normal approximation of the binomial distribution, akin to the widely used asymptotic test for proportions, while τ_2 represents the exact quantile of a binomial distribution. In any case, all three methods outperform the fixed threshold τ_N, which fails to capture any fluctuations in the progress of the pandemic. Consequently, τ_N may result in either a lower or higher threshold compared to the other methods during different phases of the pandemic, leading to potential inefficiencies in detecting outbreaks.

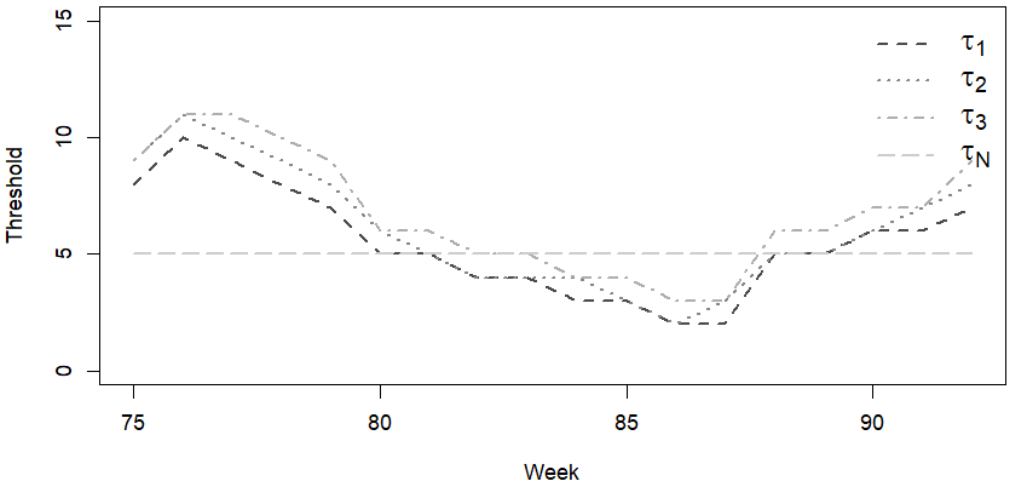


Fig. 4.2 The variable and fixed thresholds for the SARS-CoV-2 swabs at Politecnico di Torino.

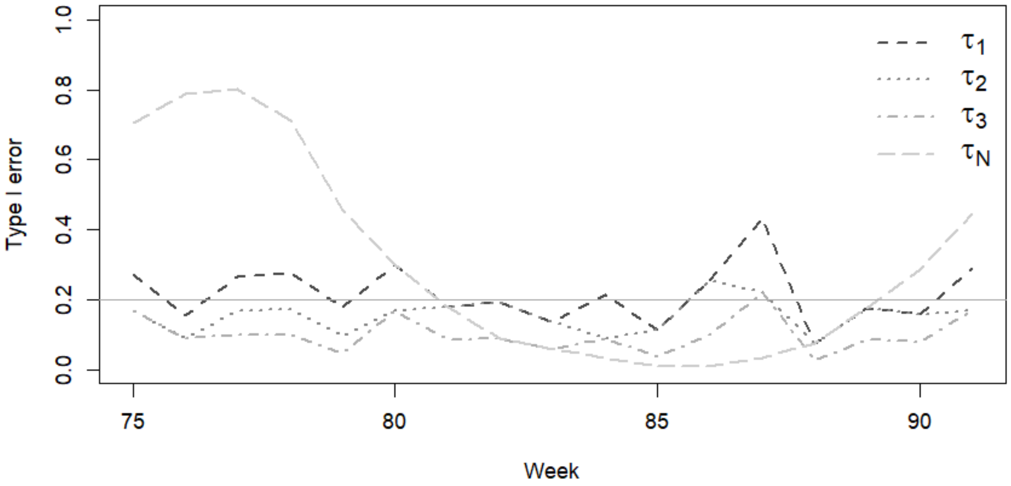


Fig. 4.3 Type I error for the different thresholds. The solid line is 0.2.

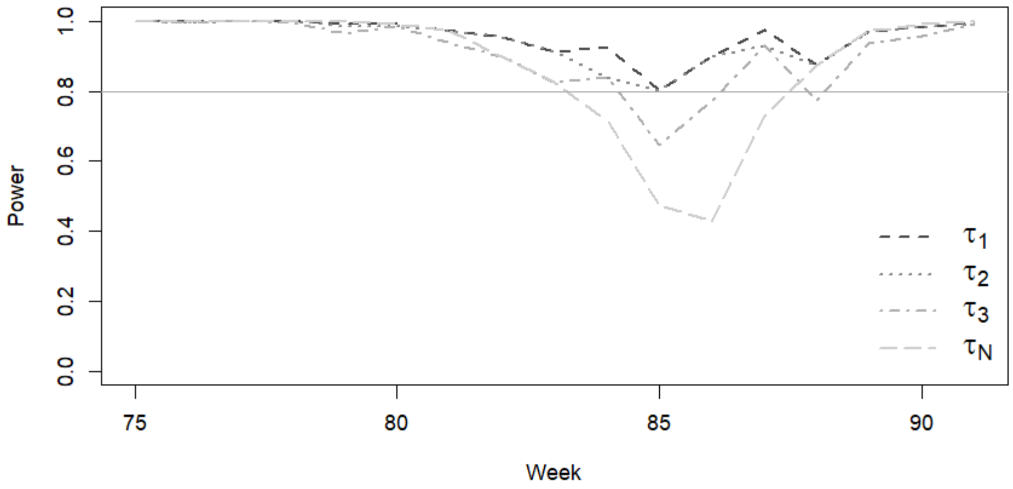


Fig. 4.4 Power for the different thresholds. The solid line is 0.8.

4.3.3 Operating characteristics of adaptive testing

With the available data on the pandemic's progress $\hat{P}_t$, we can calculate various operating characteristics for the different thresholds to assess their properties. As an example, let's start with the previously determined values $\tau_{1,75} = 8$, $\tau_{2,75} = 9$, $\tau_{3,75} = 9$, and $\tau_{N,75} = 5$, along with the actual percentage of positive tests in Italy for week 75, $\hat{P}_{75} = 0.028$. Using this information, we can determine the power and type I error of the thresholds for this week. The type I error represents the probability of surpassing the specified threshold when the true value of p_{75} is actually equal to $\hat{P}_{75}$, leading to a false alarm. On the other hand, the power is the probability of exceeding the given threshold when the true value of p_{75} is three times $\hat{P}_{75}$, resulting in a correct alarm. We assume that the positive tests at POLITO follow a binomial distribution with parameters n_S and p_{75}. The type I error can then be defined as follows:

$$\alpha_{i,75} = 1 - \sum_{x=1}^{\tau_{i,75}} \binom{n_S}{x} (\hat{P}_{75})^x (1 - \hat{P}_{75})^{n_S - x}$$

for $i = 1, 2, 3, N$; while the power can be defined as:

$$(1-\beta_{i,75}) = 1 - \sum_{x=1}^{\tau_{i,75}} \binom{ns}{x} (3 \cdot \hat{P}_{75})^x (1 - 3 \cdot \hat{P}_{75})^{ns-x}$$

for $i = 1, 2, 3, N$. The numeric example above leads to $\alpha_{1,75} = 0.271$, $\alpha_{2,75} = 0.168$, $\alpha_{3,75} = 0.168$, $\alpha_{N,75} = 0.705$, $(1-\beta_{1,75}) = 0.999$, $(1-\beta_{2,75}) = 0.998$, $(1-\beta_{3,75}) = 0.998$, $(1-\beta_{N,75}) = 1$.

In Figure 4.3 and Figure 4.4, we present the type I error and power for the different thresholds. These values are calculated using the available weekly data of $\hat{P}_t$ from week 75 to 91, as mentioned earlier. Here are some key observations from the analysis:

- The fixed threshold exhibits an extremely high type I error and a very high power when the pandemic is worsening. However, its type I error is controlled below 0.20, and its power decreases to 0.429 during regressive phases of the pandemic. This indicates that in expansive phases, there is a high risk of false alarms due to the more plausible high number of positives. On the other hand, during regressive phases, it is more realistic to not detect possible clusters within the university due to the high threshold.
- The τ_1 threshold maintains a type I error between 0.07 and 0.30, with a peak of 0.431 at week 87 (when the weekly percentage of positive tests changes its convexity and starts increasing again). However, its power remains above 0.8 throughout all weeks.
- Both τ_2 and τ_3 exhibit similar results. Their type I error is controlled below 0.20 for the most part, except around week 87, with τ_2 peaking at 0.256 at week 86 and at 0.221 at week 87, and τ_3 peaking at 0.221 at week 87. The power of τ_2 remains above 0.8 at all times, while the power of τ_3 is also above 0.8, except for weeks 86 and 88 where it decreases to about 0.640.

Overall, the variable thresholds τ_1, τ_2, and τ_3 demonstrate better performance compared to the fixed threshold, especially in capturing fluctuations in the progress of the pandemic and reducing the risk of false alarms.

A concise overview of the operating characteristics is presented in Table 4.1, revealing that the variable thresholds outperform the fixed threshold. The proposed methodology yields a reduced number of false alarms and better detection of potential outbreaks.

Table 4.1 Summary of the operating characteristics of the different thresholds.

Threshold	Fixed	Type I error	Power
τ_1	No	(0.07,0.44)	(0.80,1)
τ_2	No	(0.07,0.26)	(0.80,1)
τ_3	No	(0.02,0.23)	(0.64,1)
τ_N	Yes	(0.01,0.80)	(0.42,1)

4.4 Discussion

This study introduces a methodology aimed at assessing the similarity between a subpopulation and a general population concerning the distribution of a binary variable. The motivation for this approach stemmed from a case study on SARS-CoV-2 tests conducted at POLITO. The objective was to detect outbreaks within the university through a screening process involving oropharyngeal swabs on three days each week.

By utilizing a very general ARMA(p,q) model, three time-varying thresholds have been derived to test the equality of the proportion of individuals with a specific characteristic in both the subpopulation and the general population. Through the case study, it has been demonstrated that these variable thresholds effectively track the evolution of the underlying process, outperforming a fixed threshold in terms of operating characteristics. The three thresholds presented exhibit distinct properties: while threshold τ_1 demonstrates excellent power, it lacks control over type I error at the chosen significance level of $\alpha = 0.20$; on the other hand, thresholds τ_2 and τ_3 maintain controlled type I error around the $\alpha = 0.20$ level but possess slightly less power compared to τ_1.

This work acknowledges some limitations and potential areas for future improvement. One aspect pertains to the possibility of employing a more precise model for predicting the COVID-19 pandemic in Italy, which could yield more accurate thresholds for the case study. Nevertheless, the primary objective of this study is to maintain a broad and adaptable approach that can be applicable to diverse scenarios.

Chapter 5

A comparison of estimation methods in adaptive enrichment designs with time-to-event endpoints

Background

This chapter is published as:

Fulvio Di Stefano, Matthieu Pannaux, Anne Correges, Stephanie Galtier, Veronique Robert, Gaelle Saint-Hilary. *A comparison of estimation methods adjusting for selection bias in adaptive enrichment designs with time-to-event endpoints.* Statistics in Medicine, 2022. DOI: 10.1002/sim.9327.

5.1 Introduction

Over the past years, there has been significant progress in the development of adaptive designs (ADs) for clinical trials for improving drug development processes. These designs allow for pre-planned modifications to be made during the trial's interim analyses, which include actions such as adjusting the sample size, halting the entire trial or specific dosages in case of insufficient efficacy, discontinuing the entire trial in case of success, redistributing patients among different treatment groups, and selecting a population more likely to benefit from the treatment [5]. Enrichment

AD trials offer the possibility to select at interim analyses specific sub-populations that are expected to benefit the most from a treatment, thus optimizing resources by focusing on the most promising patient groups. In such trials, patients are categorized into sub-groups based on certain biomarkers or covariate values (e.g., tumor size, baseline heart rate). The treatment's efficacy and safety are evaluated both within each group and overall at interim analyses. If, based on pre-defined criteria, certain sub-groups show more significant benefits from the experimental drug compared to others, the trial's recruitment is then restricted to these particular patients for the remainder of the study.

While employing ADs appears highly promising, this increased flexibility does come with certain drawbacks. It is widely acknowledged that the selection rule used in ADs can lead to a biased estimation of the treatment effect [5, 93, 94]. Figure 5.1, obtained similarly to Pallmann et al [5], shows that when the lowest treatment effects are omitted while retaining the highest ones, the treatment effect tends to be overestimated. This occurs because an estimator that neglects the selection process yields positively biased results. Similarly, the estimation derived from the excluded data is biased as well, but in a negative direction. Indeed, in two-stage ADs, the stage 1 naive estimators (obtained using data before the interim analysis) are subject to bias due to the applied selection rule. On the other hand, the stage 2 naive estimators (derived from data after the interim analysis) provide unbiased estimations of the true treatment effect for the selected treatments [94, 95], as no selection rule is applied to them.

In their 2019 Guidance for Industry on Adaptive Designs for Clinical Trials of Drugs and Biologics [6], the FDA acknowledges the ethical advantages of using AD trials but emphasizes the importance of adhering to key principles for regulatory approval. Specifically, the sponsor must assess the potential bias in the estimates and, if available, pre-specify methods to adjust the estimates and minimize or eliminate this bias. This precautionary step is crucial to prevent an overly optimistic estimation of the treatment effect and to control the inflation of type I errors resulting from incorporating data used for selection in the final analysis. The regulatory requirement served as the driving force behind the research presented in this chapter.

In the subsequent sections, our emphasis lies on two-stage adaptive designs that involve sub-population selection, also known as enrichment designs, and time-to-event data. This particular design allows for the selection of the sub-population that

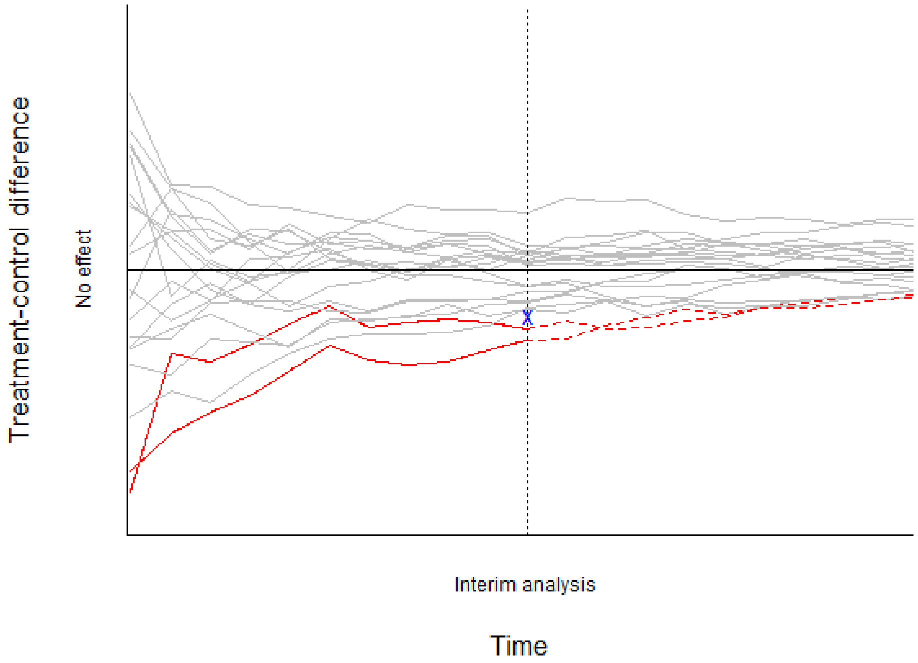


Fig. 5.1 Illustration of bias caused by early stopping for futility. Derived similarly to Pallmann et al [5]. In red there are two of 20 simulated two-arm trials with zero treatment effect that are excluded because of the threshold (blue cross), resulting in optimistic estimation of the effect.

gains the most benefit from the treatment during a single interim analysis conducted within the trial. Drawing from the research conducted by Kimani et al. [96] on adaptive threshold enrichment clinical trials with normally distributed endpoints, we explore various approaches extended to accommodate time-to-event data:

- Uniformly minimum variance conditional unbiased estimator (UMVCUE), developed to estimate the true treatment effect with no bias. A first unbiased estimator was developed for treatment selection and presented by Cohen and Sackrowitz in 1989 [97]. This work was continued by Bowden and Glimm [98]. Kimani et al [99] present a version adapted for sub-population selection with time-to-event data.
- Shrinkage estimators aim to mitigate bias without completely eliminating it. These estimators work by shrinking the stage 1 estimates towards the

stage 1 overall mean, thereby reducing the bias. In this study, we assess two approaches proposed by Carreras and Brannath [95] and Brückner et al [100].

- Bias-Adjusted Estimators are mainly developed by Whitehead [101] and Stallard and Todd [102]. The core concept behind these procedures is to iteratively calculate an estimation of the bias and subsequently subtract it from the original naive estimator.

Brückner et al [100] conducted a comparison on some of these estimators within the framework of multi-arm two-stage trials, involving treatment selection and time-to-event endpoints. Kunzmann et al [103] conducted a comparison of six other estimators within the context of adaptive enrichment designs, focusing on normally distributed endpoints. The study recommended a hybrid estimator, which combines the UMVCUE with a conditional moment estimator, as a general rule. In the context of enrichment adaptive design (AD) clinical trials with normally distributed endpoints, Kimani et al. [96] compared these estimators and suggested the unbiased estimator as a general rule. In a follow-up study [99], they derived expressions for an unbiased estimator in a two-stage adaptive design with time-to-event data and focused on the construction of confidence intervals. The objective of this study is to expand upon previous research and conduct a comprehensive comparison of six treatment effect estimators within the context of two-stage enrichment adaptive designs with time-to-event data. The paper aims to offer recommendations on the most suitable estimators that, in the authors' opinion, adhere well to regulatory requirements and effectively support internal decision-making. It is essential to highlight that the primary focus of this paper is on accurately estimating the treatment effect in the selected sub-populations. Consequently, the estimators provided are conditional on the selection made. While unconditional estimators may be relevant in other scenarios, it is crucial to recognize that reducing the unconditional bias does not guarantee a reduction in bias conditioned on a specific selection [104], which is the primary concern of this study.

The following of this chapter is organized as follows. The methodology to retrieve the different treatment effect estimators is presented in Section 5.2. In Section 5.3 the methods are applied to a case-study in cardiology. In Section 5.4 it is presented a comprehensive simulation study on the performances of the six estimators with regard to their bias, variance and mean squared error (MSE). A discussion concludes in Section 5.5.

5.2 Methods

We explore the context of adaptive clinical trials involving two treatment arms - an experimental drug and a control group - where sub-population selection occurs at a single interim analysis, and the data collected are of time-to-event nature. The approach involves dividing the patient population into different sub-populations, denoted by indices $i = 1, ..., K$, based on specific biomarker values, and analyzing these sub-populations separately. We establish sub-populations such that patients within each sub-population possess a biomarker value falling between predefined upper and lower thresholds, ensuring that the sub-populations are mutually exclusive. The data collected before the interim analysis is referred to as stage 1 data, while the data gathered after the interim analysis is referred to as stage 2 data. d_{ji} denotes the number of events in sub-population $i = 1, ..., K$ at stage $j = 1, 2$. In our setting, d_{2i} does not contain the events in d_{1i}.

In the context of time-to-event data, we focus on examining the log hazard ratio (HR) between the two treatment arms within each sub-population, defined as $\delta_i = \log\left(\frac{h_{ti}(t)}{h_{ci}(t)}\right)$ for $i = 1, ..., K$, where $h_{ti}(t)$ and $h_{ci}(t)$ are, respectively, the hazard functions of the treatment and the control in sub-population i. We employ a Cox proportional hazard model to calculate the estimates. In this context, a negative value of the log HR indicates a reduction in the risk of the event with the treatment, implying the treatment's efficacy compared to the control. Furthermore, if the log HR in one sub-population is lower than in another, it suggests that the treatment is more effective in that specific sub-population. The log HR is assumed to follow a normal distribution, and the stage 1 and stage 2 estimators are $\hat{\delta}_{1i} \sim N(\delta_i, \tau_{1i}^2)$ and $\hat{\delta}_{2i} \sim N(\delta_i, \tau_{2i}^2)$ for $i = 1, ..., K$, respectively. The variances $\hat{\tau}_{1i}^2$ and $\hat{\tau}_{2i}^2$ are also estimated from the Cox model.

We establish a selection rule at the interim analysis defined as follows: given a threshold value b for the log hazard ratio (HR) between treatment arms, each sub-population $i \in (1, ..., K)$ will not proceed to stage 2 if its stage 1 estimate is not lower than b ($\hat{\delta}_{1i} \geq b$). In case the stage 1 estimates for all sub-populations are greater than or equal to b, the trial is stopped for futility. We denote the set of indices corresponding to the selected sub-populations continuing to stage 2 as $\mathscr{S}$.

In this context, one notable aspect of adaptive clinical trials with time-to-event data is that some stage 1 patients might not have experienced the event of interest by the time of the interim analysis. If we include these patients in the stage 2 analysis, the test statistics from stage 1 and stage 2 will be correlated, leading to potential bias in the estimation. Indeed, the calculation of the forthcoming estimators relies on the assumption of an independent increment structure, where the test statistics from stage 1 and stage 2 are independent. However, if the same patients continue from stage 1 to stage 2, the independent increment structure holds only approximately, even when the timing of the interim and final analysis is independent of each other [99, 105]. To avoid the correlation, it would be necessary for patients from stage 1 to exit the study at the interim analysis. However, such a practice is clearly unethical and impractical, as patients cannot be asked to stop the study before completing a minimum treatment period. Instead, we employ an intermediate rule to address this issue, following Kimani et al [99] and Jenkins et al [105]. Let T_1 represent the time of the interim analysis, and T_2 denote the time of the final analysis, which are defined as the point when a specific number of patients have experienced the event of interest. We introduce $\tilde{T}_1$ (with $T_1 \leq \tilde{T}_1 \leq T_2$) as the time until which the stage 1 patients are followed up. By implementing this approach, we enhance the independent increment structure, leading to more precise estimations. Additionally, this method is realistic as in many therapeutic areas, it is possible to pre-specify a maximum follow-up time for all patients in the study protocol. It is worth to note that since T_2 is predetermined in terms of events, the number of events in stage 2 for each selected sub-population d_{2i} (where $i \in \mathscr{S}$) depends on both the hazard ratios and the number of selected partitions. However, the sum of events across all selected sub-populations, $\sum_{i \in \mathscr{S}} d_{2i}$, remains constant.

Following the approach outlined in Kimani et al. [99] and Bruckner et al. [100], the following methodology is employed to obtain the estimators. Initially, using data solely from stage 1, and utilizing survival times up to the time of the interim analysis (T_1), the estimators $\hat{\delta}_{1i}$ and $\hat{\tau}_{1i}^2$ are directly estimated through the score process of a Cox proportional hazard model. At the conclusion of the trial, considering all available evidence (i.e., using survival times up to the time of the final analysis T_2), the naive estimators $\hat{\delta}_{i,N}$ and $\hat{\tau}_{i,N}^2$ for the selected sub-populations are also directly estimated using a Cox proportional hazard model. Subsequently, the stage 2 estimators for the selected sub-populations are calculated as $\hat{\tau}_{2i}^2 = \left(\frac{1}{\hat{\tau}_{i,N}^2} - \frac{1}{\hat{\tau}_{1i}^2} \right)^{-1}$ and $\hat{\delta}_{2i} = \hat{\tau}_{2i}^2 \left(\frac{\hat{\delta}_{i,N}}{\hat{\tau}_{i,N}^2} - \frac{\hat{\delta}_{1i}}{\hat{\tau}_{1i}^2} \right) \forall i \in \mathscr{S}$.

5.2.1 Naive estimator

The naive estimator is calculated as:

$$\hat{\delta}_{i,N} = \frac{\hat{\tau}_{2i}^2 \, \hat{\delta}_{1i} + \hat{\tau}_{1i}^2 \, \hat{\delta}_{2i}}{\hat{\tau}_{1i}^2 + \hat{\tau}_{2i}^2} \qquad \forall i \in \mathscr{S}$$

and $\hat{\delta}_{i,N} = \hat{\delta}_{1i} \; \forall i \notin \mathscr{S}$.

This estimation is biased by the selection process [5, 93, 94], and our primary focus is to handle this bias. However, this estimator assumes the independent increment structure and pools all available data at the end of the study. Therefore, there is also a bias component stemming from the inclusion of stage 1 patients in stage 2, leading to a correlation between the stages. While this correlation bias is mostly mitigated by independently selecting T_1 and T_2 and fixing $\tilde{T}_1$ beforehand, it is not further adjusted in the subsequent calculations. As a result, this estimator is not entirely naive, as the correlation between stages has been reduced.

5.2.2 UMVCUE

Building upon the approach introduced by Kimani et al. [99], we compute the uniformly minimum variance conditional unbiased estimator (UMVCUE), designed to handle the selection bias. However, due to the correlation bias between stages 1 and 2, this estimator may not achieve perfect unbiasedness. For the selected sub-populations, it is:

$$\hat{\delta}_{i,U} = \hat{\delta}_{i,N} - \frac{\hat{\tau}_{2i}^2}{\sqrt{\hat{\tau}_{1i}^2 + \hat{\tau}_{2i}^2}} \frac{\phi(g(b))}{\Phi(g(b))}, \qquad \forall i \in \mathscr{S}$$

where ϕ and Φ denote the density and cumulative distribution functions of a standard normal distribution, respectively, and $g(x) = \frac{\sqrt{\hat{\tau}_{1i}^2 + \hat{\tau}_{2i}^2}}{\hat{\tau}_{1i}^2} \left(\hat{\delta}_{i,N} - x \right)$. This estimation, applied to the selected treatments and not available to the dropped ones, effectively eliminates the bias caused by the selection process. However, the bias eradication comes at the expense of an increased variance in these estimates.

5.2.3 Shrinkage estimators

The goal of these estimators is to reduce bias compared to a naive estimator, without increasing the variance. While stage 2 estimates offer an unbiased estimation of the treatment effects, stage 1 estimations are biased due to the selection process [94, 95]. The concept underlying these estimators involves shrinking the stage 1 estimates towards the overall average log HR of stage 1 to mitigate bias. We explore two shrinkage estimators. The first one, labeled as $S1$, was introduced by Carreras and Brannath [95]. We define $\hat{\delta}_{1\cdot} = \frac{1}{K}\sum_{i=1}^{K}\hat{\delta}_{1i}$ as the overall stage 1 average log HR and $t_i = \frac{d_{1i}}{d_{1i}+d_{2i}}$ as the information fraction at the time of the interim analysis. The shrinkage estimator $S1$ is calculated for the selected sub-populations as:

$$\hat{\delta}_{i,S1} = t_i\left[\hat{C}_i^+\hat{\delta}_{1i} + (1-\hat{C}_i^+)\hat{\delta}_{1\cdot}\right] + (1-t_i)\hat{\delta}_{2i} \qquad \forall i \in \mathscr{S}$$

while they are $\hat{\delta}_{i,S1} = [\hat{C}_i^+\hat{\delta}_{1i} + (1-\hat{C}_i^+)\hat{\delta}_{1\cdot}]\ \forall i \notin \mathscr{S}$, with $\hat{C}_i^+$ defined as follows. If $K \geq 4$:

$$\hat{C}_i^+ = max(0,\hat{C}_i), \qquad \hat{C}_i = 1 - \frac{(K-3)\hat{\tau}_{1i}^2}{\sum_{j=1}^{K}\left(\hat{\delta}_{1j} - \hat{\delta}_{1\cdot}\right)^2},$$

while if $K = 2,3$:

$$\hat{C}_i^+ = max(0,\hat{C}_i), \qquad \hat{C}_i = 1 - \frac{(K-1)\hat{\tau}_{1i}^2}{\sum_{j=1}^{K}\left(\hat{\delta}_{1j} - \hat{\delta}_{1\cdot}\right)^2}.$$

The second shrinkage estimator $S2$ was proposed by Brückner et al [100] and comes from a Bayesian framework. Suppose to have a prior distribution of the vector of true log HRs $\delta = (\delta_1, ..., \delta_K)$, which is a multivariate normal $MVN(\mu, \nu^2\mathbf{I}_K)$ ($\mathbf{I}_K$ is the KxK identity matrix). This is updated with the data $\hat{\delta}^{Stage1} \sim MVN(\delta, \Sigma)$ to get a posterior estimation for δ. The posterior log HR of δ is $C\hat{\delta}^{Stage1} + (\mathbf{I}_K - C)\mu$, where $C = \mathbf{I}_K - \Sigma(\nu^2\mathbf{I}_K + \Sigma)^{-1}$. Because sub-populations are disjoint, Σ is a diagonal matrix containing the τ_{1i}^2 on the diagonal. This matrix is unknown and we use an estimation $\hat{\Sigma}$ containing $\hat{\tau}_{1i}^2$ on the diagonal. We define the prior log HR μ as a vector of length K with the overall average stage 1 log HR $\hat{\delta}_{1\cdot}$. An estimate $\hat{\nu}^2$ of ν^2 is obtained iteratively :

- Step 1: Define an initial guess of $\hat{\nu}^2$.

- Step 2: Define weights $w_i = (\hat{\nu}^2 + \hat{\Sigma}^2_{ii})^{-1}$ for $i = 1, ..., K$.

- Step 3: Update the estimate calculating

$$\hat{\nu}^2 = \frac{\sum_{i=1}^{K} w_i \left[(\hat{\delta}_{1i} - \hat{\delta}_{1\cdot})^2 - \hat{\Sigma}^2_{ii} \right]}{\sum_{i=1}^{K} w_i}.$$

- Step 4: If $\hat{\nu}^2 < 0$, set $\hat{\nu}^2 = 0$.

- Step 5: Go back to step 2 using the updated $\hat{\nu}^2$, until convergence.

An estimate $\hat{\nu}^2$ is available when the approach converges and is used to calculate $\hat{C} = \mathbf{I}_K - \hat{\Sigma}(\hat{\nu}^2 \mathbf{I} + \hat{\Sigma})^{-1}$. The stage 1 estimator is:

$$\hat{\delta}^{Stage1}_{S2} = \hat{C}\hat{\delta}^{Stage1} + (\mathbf{I}_K - \hat{C})\mathbb{1}\hat{\delta}_{1\cdot}$$

where $\mathbb{1}$ is the vector with all entries equal to 1. The shrinkage estimator is calculated as:

$$\hat{\delta}_{i,S2} = t_i\,\hat{\delta}^{Stage1}_{i,S2} + (1 - t_i)\hat{\delta}_{2i} \qquad \forall i \in \mathscr{S},$$

and

$$\hat{\delta}_{i,S2} = \hat{\delta}^{Stage1}_{i,S2} \qquad \forall i \notin \mathscr{S}.$$

5.2.4 Bias-Adjusted Estimators

Finally, we explore bias-adjusted estimators based on the work of Whitehead [101] and Stallard and Todd [102]. The primary concept behind these estimators is to estimate the bias of the naive estimator and then subtract this bias from the naive estimator. In this study, we compare two approaches: the single-iteration estimator (SI) and the multi-iteration estimator (MI). The single-iteration estimator is computed as follows:

$$\hat{\delta}_{i,SI} = \hat{\delta}_{i,N} - \hat{b}_i(\hat{\delta}_{i,N})$$

where $\hat{b}_i(\hat{\delta}_{i,N})$ is an estimator of the bias for the naive estimator, where true log hazard ratios are replaced with the naive estimators themselves. In the multiple-iteration procedure, iterative values are used to replace the true log hazard ratios in the expression for bias. The iterative process begins with step 1, where naive estimates are used, making the single-iteration bias-adjusted estimator a special case of the multiple-iteration bias-adjusted estimator. In step 2, the single-iteration bias-adjusted estimator replaces the true log hazard ratios in the bias estimation. Subsequently, we calculate a new estimator by subtracting the newly estimated bias from the naive estimator and repeat the process until convergence is achieved. Given that the single-iteration approach is a special case of the multi-iteration, in the following, we demonstrate the calculation of the bias at a generic iteration with estimator $\widetilde{\delta}$:

$$\hat{b}_i(\widetilde{\delta}_i) = t_i(E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] - \hat{\delta}_{i,N}) \qquad \forall i \in \mathscr{S},$$

and

$$\hat{b}_i(\widetilde{\delta}_i) = (E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] - \hat{\delta}_{i,N}) \qquad \forall i \notin \mathscr{S}.$$

The $E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i]$ $i \in (1, ..., K)$ is calculated as follows:

$$E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] = \int_{-\infty}^{b} x\, \phi\left(\frac{x - \widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad \forall i \in \mathscr{S},$$

and

$$E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] = \int_{b}^{\infty} x\, \phi\left(\frac{x - \widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad \forall i \notin \mathscr{S}.$$

where ϕ is the probability density function of a normal distribution. This formula is derived with the consideration that a sub-population is selected if the treatment exhibits a log HR lower than b, whereas it is dropped if not.

5.3 Case-study

The analyses presented in this chapter were inspired by a real case-study in heart failure. The original study employed a group sequential design without population selection. However, after the analysis, certain subgroups with varying efficacy levels were identified, leading to the realization that the design could have been more effectively conducted as an adaptive enrichment design. Due to confidentiality reasons, the data used in this section are simulated data. The comparison is made between an experimental treatment and a placebo, with the initial patient population being divided into $K = 3$ sub-populations based on their baseline heart rate: low heart rate (below 75 bpm); medium heart rate (between 75 and 81 bpm); and high heart rate (above 81 bpm). The primary endpoint of the study is the time from randomization to either cardiovascular death or hospital admission for worsening of heart failure. The main analysis involves a Cox proportional hazard model that is adjusted for previous beta-blocker intake at randomization, and the treatment effect is estimated using the hazard ratio between the two treatment arms. The interim analysis is performed once 630 events have occurred, and the stage 1 patients are followed up for a maximum of 18 months after the analysis. The final analysis is conducted when a total of 1260 events have occurred, allowing for the detection of a hazard ratio of 0.85 with a power of 82% and a one-sided type 1 error of 2.5%. The futility threshold on the log-hazard ratio scale is set to $b = -0.1$, which corresponds to a hazard ratio of 0.9.

In Table 5.1, we present the stage 1 and stage 2 estimations obtained from the Cox proportional hazard model. During the interim analysis, the low heart rate sub-population is dropped as it falls below the futility threshold, while the other sub-populations continue to stage 2. In stage 1, the medium heart rate sub-population shows a substantial treatment effect, but this effect diminishes in stage 2 due to sampling variations. On the other hand, the high heart rate sub-population exhibits a significant treatment effect in both stage 1 and stage 2. Notably, since the stage 1 estimates are considerably different from b in both cases, we expect the various estimates to be quite close. At the conclusion of the study, we proceed to estimate the treatment effect in the selected sub-populations, and the results are displayed in Table 5.2:

- Among the sub-population with medium heart rate, the UMVCUE, single-iteration, multiple-iteration bias-adjusted, and second shrinkage estimators offer comparable and slightly more conservative estimations in comparison to the naive one. The differences among these estimators are minimal. However, the first shrinkage estimator yields even more conservative estimations.
- In the high heart rate sub-population, the naive estimator also shows the most optimistic result. However, the UMVCUE, single-iteration, and multiple-iteration bias-adjusted estimators remain comparable to each other and to

the naive one. The shrinkage estimators provide slightly more conservative estimations in this case.

Table 5.1 Case-study in heart failure: stage 1 and stage 2 MLE estimators of the log HR.

Log HR	**Low Heart Rate**	**Medium Heart Rate**	**High Heart Rate**
$\hat{\delta}_{1i}(\hat{\tau}_{1i})$	-0.075 (0.155)	-0.397 (0.150)	-0.358 (0.121)
$\hat{\delta}_{2i}(\hat{\tau}_{2i})$	-	-0.109 (0.109)	-0.313 (0.097)

Table 5.2 Case-study in heart failure: comparison of the estimators of the log HR.

Log HR Estimator	**Medium Heart Rate**	**High Heart Rate**
Naive estimator (N)	-0.209	-0.330
UMVCUE (U)	-0.188	-0.329
Shrinkage 1 (S1)	-0.180	-0.315
Shrinkage 2 (S2)	-0.189	-0.317
Single-iteration (SI)	-0.187	-0.327
Multiple-iteration (MI)	-0.191	-0.328

Table 5.3 Case-study in heart failure: Sidak[106], Bonferroni[107] and selection-adjusted[99] confidence intervals.

95% confidence intervals	**Medium Heart Rate**	**High Heart Rate**
Sidak	[-0.396;-0.021]	[-0.491;-0.170]
Bonferroni	[-0.421; 0.003]	[-0.511;-0.149]
Selection-adjusted	[-0.391; 0.045]	[-0.509;-0.140]

For comprehensive analysis, Table 5.3 displays 95% confidence intervals for the selected sub-populations. In addition to the Sidak [106] and Bonferroni [107] confidence intervals, we include the selection-adjusted confidence intervals from Kimani et al. [99]. It is observed that the Bonferroni confidence intervals are wider than the Sidak's intervals, as anticipated, and the one for the medium heart rate sub-population contains zero. However, it is crucial to note that these confidence intervals rely on the assumption that the overall estimate is normally distributed and do not account for the selection process. Consequently, the selection-adjusted confidence intervals appear more conservative. In the medium heart rate sub-population, the lower bound is higher compared to the other two intervals, and the upper bound is above zero, suggesting that there might be no significant difference between the treatment and the placebo. In the high heart rate sub-population, the lower bound is similar to Bonferroni's, while the upper bound is higher, leading to a wider confidence interval. It is important to highlight that the selection-adjusted confidence intervals are based on both the stage 1 and stage 2 estimates, and thus, the confidence interval for the medium heart rate sub-population incorporates zero due to its smaller effect in stage 2.

5.4 Simulation study

In this section, we conduct simulations to compare the performances of the six previously presented estimators in terms of bias, variance, and mean squared error (MSE). While bias is the most critical metric of interest, as these estimators aim to reduce or eliminate it, variance is also significant. An unbiased estimator with high variance may not be preferred over a slightly biased but more precise alternative, as the latter can better support decision-making. The MSE combines both bias and variance information, providing a comprehensive measure for evaluating the overall performance of the estimators.

5.4.1 Setting

Consider a scenario inspired by the case study with three sub-populations and two treatment arms: an experimental treatment and a control. The recruitment of patients takes place evenly from the three sub-populations over a maximum period of 3 years, and they are equally allocated to the treatment arms (randomization ratio 1:1). The maximum follow-up time is set at 9 months. Assuming that the hazard function remains constant for all treatments and is equal to $h_c = 0.0005$ for the control group, a total of 632 events is required to detect a hazard ratio of 0.8 with a power of 80% and a one-sided type 1 error of 2.5%. An interim analysis is scheduled when half of the total events are observed. Thus, we establish the following time points: the interim analysis T_1 occurs after 316 events; the stage 1 patients are followed up until stage 2 at $\tilde{T}_1$, which is set to 6 months after the interim analysis; and the trial concludes at

T_2 when a total of 632 events are observed, accounting for stage 1 patients followed up until $\tilde{T}_1$. Moreover, the number of events remains fixed regardless of the selected sub-populations, while the number of patients required to reach these events may vary depending on the scenarios and selected sub-populations. However, a maximum of 3000 patients is allocated to each sub-population.

We consider three cases of log HR: treatment being ineffective in all sub-populations $\delta = (0\ ,0\ ,0)$; treatment being effective in only one sub-population $\delta = (0\ ,0\ ,-0.3)$; and treatment exhibiting a linear effect on the sub-populations $\delta = (-0.1, -0.2, -0.3)$. For all cases, the threshold is fixed at $b = -0.1$. We conducted simulations for each scenario until we obtained 10000 simulated clinical trials with a stage 2, i.e., not stopped for futility at the interim analysis. Table 5.4 presents the empirical selection probabilities.

Table 5.4 Empirical probability of selection for the different sub-populations in the simulation study according to their log HR.

Log HR	$\delta_i = 0$	$\delta_i = -0.1$	$\delta_i = -0.2$	$\delta_i = -0.3$
Probability of selection	30%	49%	69%	84%

For completeness, the Supplementary Material includes additional simulation scenarios where: (1) the threshold is set to $b = 0$; (2) $\tilde{T}_1$ is set to 3 months after the interim analysis; and (3) we compare 4 sub-populations while keeping the other parameters consistent with this setting.

5.4.2 Results

This section presents the estimates for the bias, variance and MSE of the estimators in the simulation study. We define $\mathbb{S}_i$ as the set of simulations where sub-population i is selected, and $\hat{\delta}^s_{i,\cdot}$ as a generic estimator of δ_i in a simulation $s \in \mathbb{S}_i$. The bias in each sub-population is estimated via $\frac{1}{|\mathbb{S}_i|} \sum_{s \in \mathbb{S}_i} (\hat{\delta}^s_{i,\cdot} - \delta_i)$ and the MSE via $\frac{1}{|\mathbb{S}_i|} \sum_{s \in \mathbb{S}_i} (\hat{\delta}^s_{i,\cdot} - \delta_i)^2$; the variance is calculated as (MSE - bias^2). Figure 5.2 illustrates the outcomes of our analysis. Each row corresponds to one of the three settings examined: treatment being ineffective in all sub-populations (top row), treatment being effective only in one sub-population (center row), and linear effect on the sub-populations (bottom row). The columns present the three metrics of interest: bias (left column), variance (center column), and MSE (right column). In this Figure, if more than one sub-population is selected, the results for bias, variance, and MSE are averaged over all selected sub-populations.

We observe that the bias of the estimators is generally higher in the top row. In this scenario, the optimal decision would be to stop the entire trial for futility since none of the underlying log HRs is larger than the threshold. Consequently, a sub-population is selected only when the estimated effect significantly outperforms the true effect in that sub-population. Among the estimators, the naive estimator exhibits the greatest bias, followed by the shrinkage estimators, where S1 slightly outperforms S2. Next, the bias-adjusted estimators perform better, with the single-iteration outperforming the multiple-iteration. As expected, the UMVCUE provides an almost unbiased estimation in this case, but it tends to over-correct, leading to a positive bias. However, this bias is of lower magnitude compared to the other estimators' bias and may be attributable to the remaining correlation bias. Regarding variance, the naive estimator and the shrinkage estimators demonstrate the best performance, followed by the bias-adjusted estimators performing equally well, and lastly, the UMVCUE, which is the least precise. In terms of MSE, the S1 and S2 (in this order) are the best-performing estimators, followed by the single-iteration bias-adjusted estimator and the multi-iteration bias-adjusted estimator. The naive estimator ranks slightly worse in terms of MSE, while the UMVCUE shows the worst performance.

The second row corresponds to clinical trials where, at the interim analysis, the only sub-population on which the treatment is effective should be selected, and it notably differs from the other ones. Therefore, the extent of bias is lower than in the first row. In this case, the best-performing estimator in terms of absolute bias is the multiple-iteration bias-adjusted estimator, followed by S1, SI, S2, UMVCUE, and the naive estimator in that order. However, it is still evident that the UMVCUE over-corrects the bias, resulting in a positive bias of similar magnitude compared to the previous scenario. Regarding variance, the naive estimator performs the best, followed by the shrinkage estimators and the bias-adjusted estimators, with the UMVCUE ranking last. In terms of MSE, the shrinkage estimators and the naive estimator perform the best, followed by the bias-adjusted estimators performing equally well, and finally, the UMVCUE shows the least favorable performance.

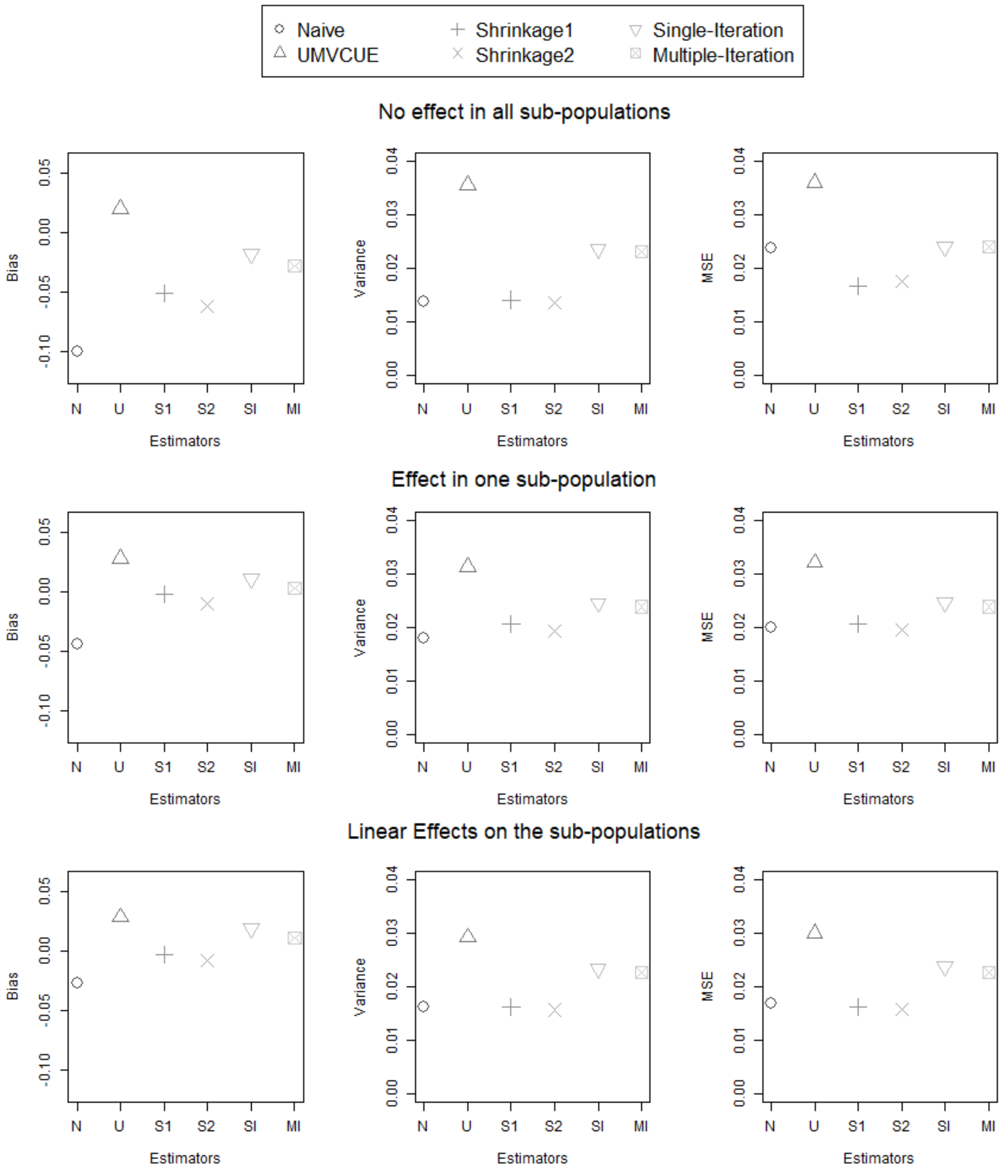


Fig. 5.2 Estimators' performances. Top row: treatment ineffective in all sub-populations $\delta = (0,0,0)$; Middle row: treatment effective only in one sub-population $\delta = (0\ ,0\ ,-0.3)$; Bottom row: linear effect on the sub-populations $\delta = (-0.1,-0.2,-0.3)$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

In the third row, the correct choice is to select the two sub-populations where the treatment is more effective. Again, the naive estimator exhibits the highest absolute bias, followed by UMVCUE, SI, and MI, with S1 and S2 performing equally well and showing the least bias. However, the single-iteration bias-adjusted estimator, multi-iteration bias-adjusted estimator, and the UMVCUE tend to over-correct the bias, resulting in a positive bias in their case. Regarding variance, the UMVCUE has the highest variance, followed by SI, MI, the naive estimator, and finally, the most precise estimators are the shrinkage estimators. Eventually, in terms of MSE, the unbiased estimator has the highest value, followed by SI and MI, the naive estimator, and the shrinkage estimators in that order.

At this point, we aim to gain a more precise understanding of the behavior of the estimators in each sub-population. Figure 5.3, shows the results in each sub-population in the case of $\delta = (-0.1, -0.2, -0.3)$. We first observe that the variance and MSE of the estimators show minimal variation from one sub-population to another, maintaining the same order as noticed in the bottom row of Figure 5.2, with only slight differences in their magnitudes across the rows. On the other hand, the bias varies substantially from one sub-population to another. The bias of the naive estimator is higher in the top row and gradually decreases with an increasing effect, as expected when moving from the threshold, approaching zero in the case of an effect equal to -0.3. In contrast, the bias of the UMVCUE remains constant across the three sub-populations and is positive, indicating an over-correction that may be attributable to the remaining correlation bias. The bias of the shrinkage estimators, which appears to be approximately zero when averaged over the sub-populations in the bottom row of Figure 5.2, is actually substantially variable. When the effect is equal to -0.1, the bias of the shrinkage estimators matches that of the naive estimator. For an effect equal to -0.2, the bias is almost zero, with S1 outperforming S2. However, when the effect is equal to -0.3, the bias becomes positive, with S1 showing a higher bias compared to the UMVCUE, and S2 having a similar performance to SI. Regarding the bias of the bias-adjusted estimators: in the top row, it is approximately zero for both, with the SI outperforming MI; in the middle row, it is positive but lower than the UMVCUE's for the SI, with the MI outperforming the SI in this case; in the bottom row, the results are similar to the middle row.

Supplementary Material presents the results of additional simulation scenarios. When setting $b = 0$, the bias is generally lower for all the estimators since they are further from the threshold. However, the overall order for the estimators' bias,

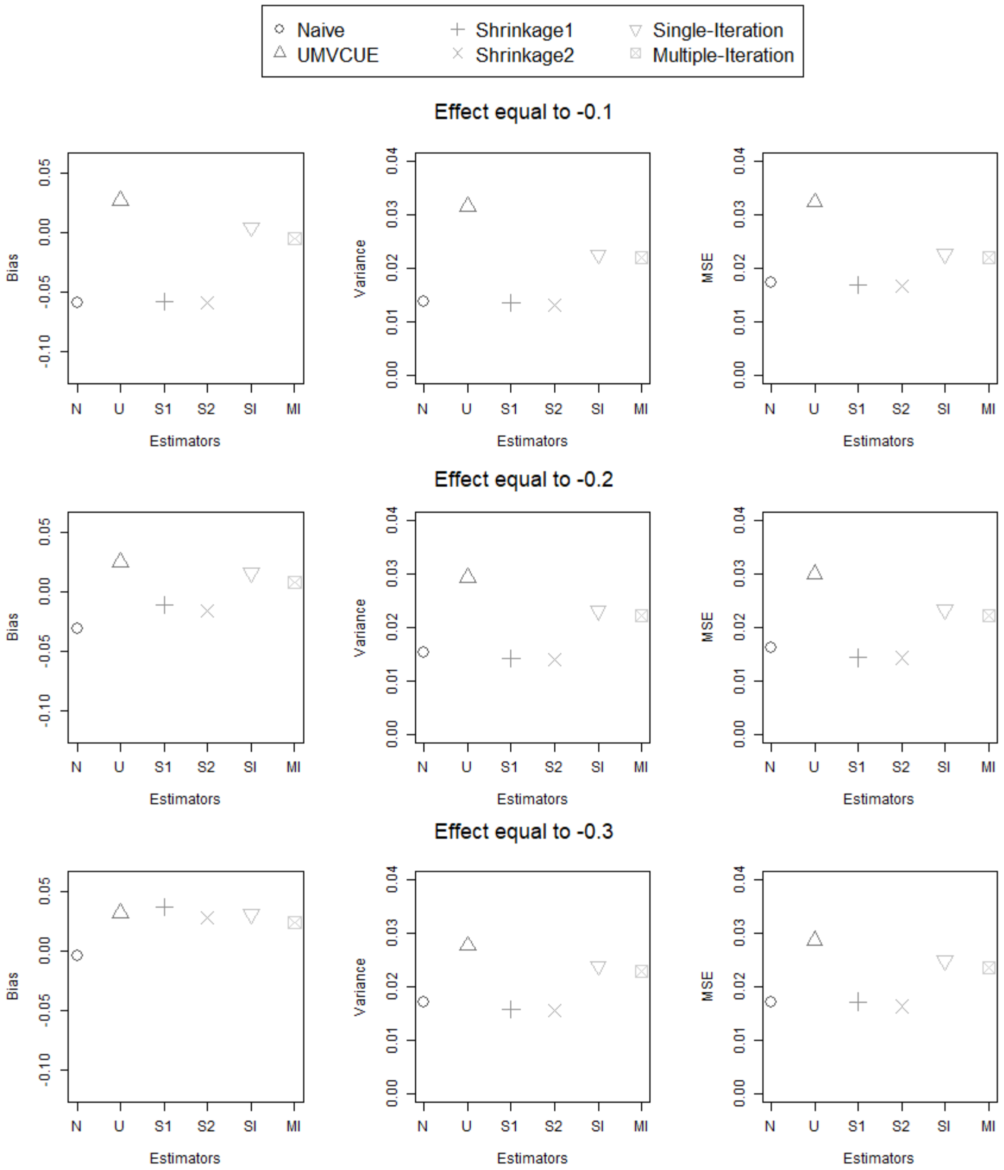


Fig. 5.3 Estimators' performances in each sub-population in case of linear effects on the sub-populations. Top row: effect equal to -0.1; Middle row: effect equal to -0.2; Bottom row: effect equal to -0.3. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

variance, and MSE remains the same. When $\tilde{T}_1$ is set to 3 months after the interim analysis, the performance metrics are similar to those presented in this study. In the scenario with 4 sub-populations, we observe a higher overall bias, and the S1 and UMVCUE perform better, with the latter being almost unbiased. This might be attributed to the fact that the decision is made with a smaller information fraction compared to the previous case where there were only three sub-populations. As a result, the stage 2 data have a greater impact on the overall estimates' derivation [95, 99]. In this case, we also analyze the results in each sub-population with linear effects, and the same pattern as seen in Figure 5.3 is obtained.

In summary, the performance of each estimator can be characterized as follows:

- The naive estimator (N) exhibits the highest bias but very low variance, resulting in a moderate MSE compared to the other estimators.
- The UMVCUE (U) shows a small constant positive bias (which is not zero due to correlation bias), and it has the highest variance, resulting in the highest MSE.
- The single-iteration (SI) bias-adjusted estimator has a small bias but higher variance compared to the naive estimator, resulting in a comparable MSE.
- The multiple-iteration (MI) bias-adjusted estimator tends to provide less conservative estimations than the single-iteration approach and has similar variance and MSE.
- The first shrinkage estimator (S1) demonstrates a noticeable bias that varies substantially across sub-populations, but it has very low variance, resulting in a very low MSE.
- The second shrinkage estimator (S2) performs similarly to S1.

These observations are consistent across all scenarios considered, and they align with results previously published in different settings [95, 96, 98, 100].

5.5 Discussion

This chapter focused on addressing the issue of selection bias [94] in the context of two-stage enrichment adaptive designs with time-to-event data. To achieve this, we explored various estimators available in the literature [95, 97–102]. This study was motivated by the FDA Guidance for Industry on Adaptive Design Clinical Trials for Drugs and Biologics [6], which requires sponsors to evaluate the reliability and extent of bias in treatment effect estimation using appropriate methods in adaptive design trials. We applied these different methods to a cardiology case-study and compared their performances in various scenarios with a selection rule based on a futility threshold. Comparative studies in the literature have primarily focused on adaptive designs with treatment selection [95, 100] or sub-population selection with normally distributed endpoints [96, 103].

We conducted a comprehensive simulation study to compare the performance of the estimators in terms of bias, variance, and mean squared error (MSE). While bias remains the primary focus, the variability can also influence the choice of one estimator over another. In addition to the maximum likelihood estimator (MLE) used in the primary analysis, we recommend presenting the unbiased estimator and the single-iteration bias-adjusted estimator in sensitivity analysis to address regulatory requirements. The unbiased estimator completely corrects the selection bias, but it may be highly variable compared to the naive estimator. On the other hand, the single-iteration bias-adjusted estimator is less biased than the naive estimator and only slightly more variable. Thus, including both these estimators allows us to assess the extent of bias in the naive estimator and provides a more precise and less biased estimation for decision-making purposes. To provide a comprehensive overview, a simulation study can be added as a supplement to identify the estimator that best suits the context and objectives of the trial, as suggested in the literature [104]. This additional analysis will further aid in making informed choices regarding estimator selection.

The current study has several limitations that should be acknowledged. Firstly, it focuses on clinical trials with a single interim analysis and only two arms, consisting of one experimental treatment and one control. Future extensions could involve incorporating multiple interim analyses or additional treatment arms to explore more complex scenarios. Another limitation is that we considered disjoint sub-populations, while some other studies have explored non-disjoint sub-populations [96, 99]. Including such scenarios could provide a more comprehensive understanding of the estimators' performance. In this study, we chose a selection rule based on comparing the hazard ratio (HR) to a pre-specified threshold. Alternative selection rules based on conditional power or predictive power could have been explored, but it is expected that similar conclusions would be reached [108]. Additionally, the literature offers various methods for calculating appropriate confidence intervals, such as simultaneous inference [109], bootstrap resampling [100], confidence regions based on orderings [102], or simultaneous inference based on the duality between hypothesis testing and confidence intervals [99]. Each of these methods has its advantages and limitations, and further exploration of these techniques could be beneficial. For more in-depth information, we refer readers to the relevant literature.

Despite its limitations, this study offers valuable insights into the performance of various estimators for mitigating bias in enrichment adaptive designs with time-to-event data. The findings provide essential information to enhance decision-making and meet regulatory requirements effectively.

Chapter 6

Conclusion and perspectives

This thesis contains the examination of some statistical methods for incorporating available evidence in order to support decision-making. The methods presented in this work have been thoroughly analyzed and discussed, with each chapter providing a detailed description of the methods, along with a case study demonstrating their application and a discussion of the topic and its strengths and weaknesses. The goal is to provide a thorough understanding of these methods and their potential applications, in order to aid in the decision-making process.

This report focuses on three main research axes: the incorporation of historical data in early phase trials, a novel methodology to test repeatedly and adaptively whether a certain sub-population is conformal with respect to the general population it is a subset of, and a comparison of estimation methods adjusting for selection bias in adaptive enrichment designs with time-to-event endpoints. The use of all available evidence to support decision-making is the aim of the proposed methodologies.

For the incorporation of historical data in early phase trials, it has been shown that the methods presented can assist in making more informed decisions in case where there is concordance between the historical data and the concurrent data. They also perform at least as good as the other methods which do not include any historical data in the study. Therefore, the use of historical data in early phase permits to improve decision making. Encouraging the use of quantitative data for constructing the prior can help mitigate potential controversies during result evaluation. Moreover, operating characteristics before the start of the trial are instrumental when combining historical data with concurrent data, in order to identify potential weaknesses and fallacies, and a sound planning and execution of the trial is also fundamental. Surely, the proposed methodologies have the potential to be applied to various other stages of early phase drug development and can be the foundation for further exploration and analysis by the scientific community. Alongside, the analyses presented here can serve as a starting point for those interested in investigating the topic.

For the application on the adaptive Covid-19 screening of a sub-population, it is clear that the use of all available evidence provides better results for the testing with respect to a classical fixed threshold. In fact, even using a model for prediction which is non-native for the problem presented, the results overcome the classical methodology in terms of operating characteristics. Certainly, the application shown is specific of the time in which it was designed but, in general, it can be applied to any problem in outbreak detection, finance or even statistical quality control. Of course, adaptations are needed in case of a different use of the methods, but valuable insides can be retrieved from the analyses here presented.

The study on the comparison of estimation methods in adaptive enrichment designs with time-to-event endpoints provides valuable insights on the topic. It identifies two estimators which have better operating characteristics with respect to a naive one, each with its own peculiarities. The first one completely eradicates the selection bias, at the price of an increase in variance. The second one is less biased with respect to the naive one, but only slightly more variable. By enhancing the understanding of these estimators and their trade-offs, this work not only advances our knowledge but also facilitates improved decision-making when choosing between them. Even acknowledging its limitations, such as focusing on trials with one interim analysis and two arms, and disjoint sub-populations, the study paves the way to future ones on the topic. They will enhance the possibility for the patients, both those enrolled in trials and those who struggle everyday, to receive the most appropriate treatment for their condition.

In conclusion, the statistical methods presented are shown to be effective in supporting decision-making by using available evidence. They provide performances at least as good as the ones where no other information were included, obtaining better results under certain appropriate conditions. The methods can potentially be applied to a variety of other problems, taking into account adaptations may be necessary. However, they provide a useful tool that can help in making more accurate and precise decision in the future.

References

[1] Food and Drug Administration. Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers - guidance for industry, 2005.

[2] European Medicines Agency. ICH topic E9 - Statistical principles for clinical trials, 1998.

[3] Food and Drug Administration. Guidance for the use of bayesian statistics in medical device clinical trials, 2010.

[4] Food and Drug Administration. Leveraging existing clinical data for extrapolation to pediatric uses of medical devices, 2016.

[5] Philip Pallmann, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1), February 2018. 10.1186/s12916-018-1017-7.

[6] Food and Drug Administration. Adaptive design clinical trials for drugs and biologics guidance for industry, 2019.

[7] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, et al. Safety and efficacy of the BNT162b2 mRNA covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615, December 2020.

[8] Lindsey R. Baden, Hana M. El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, Stephen A. Spector, Nadine Rouphael, C. Buddy Creech, John McGettigan, Shishir Khetan, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, 384(5):403–416, February 2021.

[9] Merryn Voysey, Sue Ann Costa Clemens, Shabir A Madhi, Lily Y Weckx, Pedro M Folegatti, Parvinder K Aley, Brian Angus, Vicky L Baillie, Shaun L Barnabas, Qasim E Bhorat, Sagida Bibi, Carmen Briner, , et al. Safety and

efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet*, 397(10269):99–111, January 2021.

[10] Jerald Sadoff, Mathieu Le Gars, Georgi Shukarev, Dirk Heerwegh, Carla Truyers, Anne M. de Groot, Jeroen Stoop, Sarah Tete, Wim Van Damme, Isabel Leroux-Roels, Pieter-Jan Berghmans, Murray Kimmel, et al. Interim results of a phase 1–2a trial of Ad26.COV2.S Covid-19 vaccine. *New England Journal of Medicine*, 384(19):1824–1835, May 2021.

[11] Edward Bradley. Incorporating biomarkers into clinical trial designs: points to consider. *Nature Biotechnology*, 30(7):596–599, July 2012. DOI: 10.1038/nbt.2296.

[12] R L Lalonde, K G Kowalski, M M Hutmacher, W Ewy, D J Nichols, P A Milligan, B W Corrigan, P A Lockwood, S A Marshall, L J Benincosa, T G Tensfeldt, K Parivar, M Amantea, P Glue, H Koide, and R Miller. Model-based drug development. *Clinical Pharmacology & Therapeutics*, 82(1):21–32, May 2007. DOI: 10.1038/sj.clpt.6100235.

[13] Igor Radanovic, Naomi Klarenbeek, Robert Rissmann, Geert Jan Groeneveld, Emilie M. J. van Brummelen, Matthijs Moerland, and Jacobus J. Bosch. Integration of healthy volunteers in early phase clinical trials with immuno-oncological compounds. *Frontiers in Oncology*, 12, August 2022. DOI: 10.3389/fonc.2022.954806.

[14] Phylinda L.S. Chan, Lynn McFadyen, Andrea Quaye, Heidi Leister-Tebbe, Victoria M. Hendrick, Jennifer Hammond, and Susan Raber. The use of extrapolation based on modeling and simulation to support high-dose regimens of ceftaroline fosamil in pediatric patients with complicated skin and soft-tissue infections. *CPT: Pharmacometrics & Systems Pharmacology*, 10(6):551–563, May 2021. DOI: 10.1002/psp4.12608.

[15] Reza Khosravan, Steven G. DuBois, Katherine Janeway, and Erjian Wang. Extrapolation of pharmacokinetics and pharmacodynamics of sunitinib in children with gastrointestinal stromal tumors. *Cancer Chemotherapy and Pharmacology*, 87(5):621–634, January 2021. DOI: 10.1007/s00280-020-04221-x.

[16] E. Niclas Jonsson, Fiona Macintyre, Ian James, Michael Krams, and Scott Marshall. Bridging the pharmacokinetics and pharmacodynamics of UK-279, 276 across healthy volunteers and stroke patients using a mechanistically based model for target-mediated disposition. *Pharmaceutical Research*, 22(8):1236–1246, August 2005. DOI: 10.1007/s11095-005-5264-x.

[17] Stefan Willmann, Eleonora Marostica, Nelleke Snelder, Alexander Solms, Markus Jensen, Maximilian Lobmeyer, Anthonie W. A. Lensing, Claudette Bethune, Erin Morgan, Rosie Z. Yu, Yanfeng Wang, Shiangtung W. Jung,

Richard Geary, and Sanjay Bhanot. PK/PD modeling of FXI antisense oligonucleotides to bridge the dose-FXI activity relation from healthy volunteers to end-stage renal disease patients. *CPT: Pharmacometrics & Systems Pharmacology*, 10(8):890–901, June 2021. DOI: 10.1002/psp4.12663.

[18] Fabiola La Gamba, Tom Jacobs, Helena Geys, Thomas Jaki, Jan Serroyen, Moreno Ursino, Alberto Russu, and Christel Faes. Bayesian sequential integration within a preclinical pharmacokinetic and pharmacodynamic modeling framework: Lessons learned. *Pharmaceutical Statistics*, April 2019. DOI: 10.1002/pst.1941.

[19] Fanni Natanegara, Beat Neuenschwander, John W. Seaman, Nelson Kinnersley, Cory R. Heilmann, David Ohlssen, and George Rochester. The current state of bayesian methods in medical product development: survey results and recommendations from the DIA bayesian scientific working group. *Pharmaceutical Statistics*, 13(1):3–12, September 2013. DOI: 10.1002/pst.1595.

[20] Mercedeh Ghadessi, Rui Tang, Joey Zhou, Rong Liu, Chenkun Wang, Kiichiro Toyoizumi, Chaoqun Mei, Lixia Zhang, C. Q. Deng, and Robert A. Beckman. A roadmap to using historical controls in clinical trials – by drug information association adaptive design scientific working group (DIA-ADSWG). *Orphanet Journal of Rare Diseases*, 15(1), March 2020. DOI: 10.1186/s13023-020-1332-x.

[21] Hans Ulrich Burger, Christoph Gerlinger, Chris Harbron, Armin Koch, Martin Posch, Justine Rochon, and Anja Schiel. The use of external controls: To what extent can it currently be recommended? *Pharmaceutical Statistics*, April 2021. DOI: 10.1002/pst.2120.

[22] 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(1):41–54, aug 2013. DOI: 10.1002/pst.1589.

[23] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, January 2010. DOI: 10.1177/1740774509356002.

[24] 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(4):1023–1032, October 2014. DOI: 10.1111/biom.12242.

[25] Jessica Lim, Rosalind Walley, Jiacheng Yuan, Jeen Liu, Abhishek Dabral, Nicky Best, Andrew Grieve, Lisa Hampson, Josephine Wolfram, Phil Woodward, Florence Yong, Xiang Zhang, and Ed Bowen. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Therapeutic*

Innovation & Regulatory Science, 52(5):546–559, September 2018. DOI: 10.1177/2168479018778282.

[26] 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(10):3167–3182, February 2017. DOI: 10.1177/0962280217694506.

[27] Claire L. Smith, Zachary Thomas, Nathan Enas, Katharine Thorn, Michael Lahn, Karim Benhadji, and Ann Cleverly. Leveraging historical data into oncology development programs: Two case studies of phase 2 bayesian augmented control trial designs. *Pharmaceutical Statistics*, 19(3):276–290, January 2020. DOI: 10.1002/pst.1990.

[28] Elias Laurin Meyer, Peter Mesenbrink, Cornelia Dunger-Baldauf, Ekkehard Glimm, Yuhan Li, and Franz König and. Decision rules for identifying combination therapies in open-entry, randomized controlled platform trials. *Pharmaceutical Statistics*, 21(3):671–690, January 2022. DOI:10.1002/pst.2194.

[29] Arthur Christopoulos and Esam E. El-Fakahany. Qualitative and quantitative assessment of relative agonist efficacy. *Biochemical Pharmacology*, 58(5):735–748, September 1999. DOI: 10.1016/s0006-2952(99)00087-8.

[30] Meina Liang, Martin Schwickart, Amy K. Schneider, Inna Vainshtein, Christopher Del Nagro, Nathan Standifer, and Lorin K. Roskos. Receptor occupancy assessment by flow cytometry as a pharmacodynamic biomarker in biopharmaceutical development. *Cytometry Part B*, 90(B):117–127, July 2016. DOI: 10.1002/cyto.b.21259.

[31] F Junker, P Gulati, U Wessels, S Seeber, KG Stubenrauch, L Codarri-Deak, C Markert, C Klein, P Camillo Teixeira, and H Kao. A human receptor occupancy assay to measure anti-pd-1 binding in patients with prior anti-pd-1. *Cytometry A*, 99(8):832–843, 2021. DOI: 10.1002/cyto.a.24334.

[32] H Jones and K Rowland-Yeo. Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development. *CPT Pharmacometrics Syst Pharmacol*, 2(8):e63, 2013. DOI: 10.1038/psp.2013.41.

[33] N Srinivas, K Maffuid, and ADM Kashuba. Clinical pharmacokinetics and pharmacodynamics of drugs in the central nervous system. *Clin Pharmacokinet*, 57(9):1059–1074, 2018. DOI: 10.1007/s40262-018-0632-y.

[34] Yanguang Cao and William J. Jusko. Incorporating target-mediated drug disposition in a minimal physiologically-based pharmacokinetic model for monoclonal antibodies. *Journal of Pharmacokinetics and Pharmacodynamics*, 41(4):375–387, July 2014. DOI: 10.1007/s10928-014-9372-2.

[35] N Best, RG Price, IJ Pouliquen, and ON Keene. Assessing efficacy in important subgroups in confirmatory trials: An example using bayesian dynamic borrowing. *Pharmaceutical Statistics*, 20(3):551–562, 2021. DOI: 10.1002/pst.2093.

[36] Gaelle Saint-Hilary, Valentine Barboux, Matthieu Pannaux, Mauro Gasparini, Veronique Robert, and Gianluca Mastrantonio. Predictive probability of success using surrogate endpoints. *Statistics in Medicine*, 38(10):1753–1774, December 2018. DOI: 10.1002/sim.8060.

[37] B Neuenschwander, N Rouyrre, N Hollaender, E Zuber, and Branson M. A proof of concept phase ii non-inferiority criterion. *Statistics in Medicine*, 30(13):1618–27, 2011. DOI: 10.1002/sim.3997.

[38] C Chuang-Stein and S Kirby. *Quantitative Decisions in Drug Development*. Cham, Switzerland: Springer International Publishing AG, 2017. DOI: 10.1007/978-3-319-46076-5.

[39] Beat Neuenschwander, Sebastian Weber, Heinz Schmidli, and Anthony O'Hagan. Predictively consistent prior effective sample sizes. *Biometrics*, 76(2):578–587, April 2020. DOI: 10.1111/biom.13252.

[40] Paul Frewer, Pat Mitchell, Claire Watkins, and James Matcham. Decision-making in early clinical drug development. *Pharmaceutical Statistics*, 15(3):255–263, March 2016. DOI: 10.1002/pst.1746.

[41] S Roychoudhury, N Scheuer, and B Neuenschwander. Beyond p-values: A phase ii dual-criterion design with statistical significance and clinical relevance. *Clinical Trials*, 15(5):452–461, 2018. DOI: 10.1177/1740774518770661.

[42] G Saint-Hilary, V Robert, and M Gasparini. Decision-making in drug development using a composite definition of success. *Pharm Stat*, 17(5):555–569, 2018. DOI: 10.1002/pst.1870.

[43] Edmund S. Kostewicz, Bertil Abrahamsson, Marcus Brewster, Joachim Brouwers, James Butler, Sara Carlert, Paul A. Dickinson, Jennifer Dressman, René Holm, Sandra Klein, James Mann, Mark McAllister, et al. In vitro models for the prediction of in vivo performance of oral dosage forms. *European Journal of Pharmaceutical Sciences*, 57:342–366, June 2014. DOI: 10.1016/j.ejps.2013.08.024.

[44] Food and Drug Administration. Physiologically based pharmacokinetic analyses — format and content guidance for industry, 2018.

[45] Food and Drug Administration. The use of physiologically based pharmacokinetic analyses — biopharmaceutics applications for oral drug product development, manufacturing changes, and controls - draft guidance for industry, 2020.

[46] Jennifer E. Sager, Jingjing Yu, Isabelle Ragueneau-Majlessi, and Nina Isoherranen. Physiologically based pharmacokinetic (PBPK) modeling and simulation approaches: A systematic review of published models, applications, and model verification. *Drug Metabolism and Disposition*, 43(11):1823–1837, August 2015. DOI: 10.1124/dmd.115.065920.

[47] Paul Rolan. The contribution of clinical pharmacology surrogates and models to drug development-a critical appraisal. *British Journal of Clinical Pharmacology*, 44(3):219–225, September 1997. DOI: 10.1046/j.1365-2125.1997.t01-1-00583.x.

[48] Luca Richeldi, Arata Azuma, Vincent Cottin, Christian Hesslinger, Susanne Stowasser, Claudia Valenzuela, Marlies S. Wijsenbeek, Donald F. Zoz, Florian Voss, and Toby M. Maher. Trial of a preferential phosphodiesterase 4b inhibitor for idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 386(23):2178–2187, June 2022. DOI: 10.1056/nejmoa2201737.

[49] European Medicines Agency. Ich guideline s9 on nonclinical evaluation for anticancer pharmaceuticals, 2008.

[50] Vijay Sharma and John H McNeill. To scale or not to scale: the principles of dose extrapolation. *British Journal of Pharmacology*, 157(6):907–921, July 2009. DOI: 10.1111/j.1476-5381.2009.00267.x.

[51] Nina Magnolo, Külli Kingo, Vivian Laquer, John Browning, Adam Reich, Jacek C. Szepietowski, Deborah Keefe, Philemon Papanastasiou, Weibin Bao, Pascal Forrer, and Manmath Patekar. Efficacy of secukinumab across subgroups and overall safety in pediatric patients with moderate to severe plaque psoriasis: Week 52 results from a phase III randomized study. *Pediatric Drugs*, 24(4):377–387, June 2022. DOI: 10.1007/s40272-022-00507-0.

[52] Heinz Schmidli, Beat Neuenschwander, and Tim Friede. Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Computational Statistics & Data Analysis*, 113:100–110, sep 2017. DOI: 10.1016/j.csda.2016.08.007.

[53] Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1), feb 2000. DOI: 10.1214/ss/1009212673.

[54] Ming-Hui Chen and Joseph G. Ibrahim. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3), September 2006. DOI: 10.1214/06-ba118.

[55] Joseph G. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, sep 2015. DOI:10.1002/sim.6728.

[56] Timothy Mutsvari, Dominique Tytgat, and Rosalind Walley. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 15(1):28–36, November 2015. DOI: 10.1002/pst.1722.

[57] Beat Neuenschwander, Michael Branson, and David J. Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, sep 2009. DOI: 10.1002/sim.3722.

[58] Byron J. Gajewski. Comments on 'a note on the power prior'. *Statistics in Medicine*, 29(6):708–709, feb 2010. DOI: 10.1002/sim.3824.

[59] Beat Neuenschwander, Michael Branson, and Thomas Gsponer. Critical aspects of the bayesian approach to phase i cancer trials. *Statistics in Medicine*, 27(13):2420–2439, 2008. DOI:10.1002/sim.3230.

[60] Thomas Jaki, Sally Clive, and Christopher J. Weir. Principles of dose finding studies in cancer: a comparison of trial designs. *Cancer Chemotherapy and Pharmacology*, 71(5):1107–1114, January 2013. DOI: 10.1007/s00280-012-2059-8.

[61] Lei Nie, Eric H. Rubin, Nitin Mehrotra, José Pinheiro, Laura L. Fernandes, Amit Roy, Stuart Bailey, and Dinesh P. de Alwis. Rendering the 3 + 3 design to rest: More efficient approaches to oncology dose-finding trials in the era of targeted therapy. *Clinical Cancer Research*, 22(11):2623–2629, May 2016. DOI: 10.1158/1078-0432.ccr-15-2644.

[62] Haiyan Zheng, Lisa V Hampson, and Simon Wandel. A robust bayesian meta-analytic approach to incorporate animal data into phase i oncology trials. *Statistical Methods in Medical Research*, 29(1):94–110, January 2019. DOI: 10.1177/0962280218820040.

[63] Haiyan Zheng and Lisa V. Hampson. A bayesian decision-theoretic approach to incorporate preclinical information into phase i oncology trials. *Biometrical Journal*, 62(6):1408–1427, April 2020. DOI: 10.1002/bimj.201900161.

[64] James Babb, André Rogatko, and Shelemyahu Zacks. Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10):1103–1120, may 1998. DOI: 10.1002/(sici)1097-0258(19980530)17:10<1103::aid-sim793>3.0.co;2-9.

[65] Hongtao Zhang, Alan Y. Chiang, and Jixian Wang. Improving the performance of bayesian logistic regression model with overdose control in oncology dose-finding studies. *Statistics in Medicine*, apr 2022. DOI: 10.1002/sim.9402.

[66] Isaac Gravestock and Leonhard Held and. Adaptive power priors with empirical bayes for clinical trials. *Pharmaceutical Statistics*, 16(5):349–360, jun 2017. DOI: 10.1002/pst.1814.

[67] Adrien Ollier, Satoshi Morita, Moreno Ursino, and Sarah Zohar. An adaptive power prior for sequential clinical trials – application to bridging studies. *Statistical Methods in Medical Research*, 29(8):2282–2294, nov 2019. DOI: 10.1177/0962280219886609.

[68] Bradley Carlin and Thomas Louis. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman and Hall/CRC, June 2000. DOI: 10.1177/1740774510382799.

[69] Harry Olson, Graham Betton, Denise Robinson, Karluss Thomas, Alastair Monro, Gerald Kolaja, Patrick Lilly, James Sanders, Glenn Sipes, William Bracken, Michael Dorato, Koen Van Deun, Peter Smith, Bruce Berger, and Allen Heller. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory Toxicology and Pharmacology*, 32(1):56–67, aug 2000. DOI: 10.1006/rtph.2000.1399.

[70] Heng Zhou, Ying Yuan, and Lei Nie. Accuracy, safety, and reliability of novel phase i trial designs. *Clinical Cancer Research*, 24(18):4357–4364, September 2018. DOI: 10.1158/1078-0432.ccr-18-0168.

[71] Yuan Ji, Ping Liu, Yisheng Li, and B. Nebiyou Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):653–663, October 2010. DOI: 10.1177/1740774510382799.

[72] 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(3):1047–1056, March 2011. DOI: 10.1111/j.1541-0420.2011.01564.x.

[73] Anthony O'Hagan. Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81, March 2019.

[74] European Center for Disease prevention and Control. Covid-19 clusters and outbreaks in occupational settings in the eu/eea and the uk, 2020.

[75] David L. Buckeridge, Howard Burkom, Murray Campbell, William R. Hogan, and Andrew W. Moore. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38(2):99–113, April 2005.

[76] Brice Leclère, David L. Buckeridge, Pierre-Yves Boëlle, Pascal Astagneau, and Didier Lepelletier. Automated detection of hospital outbreaks: A systematic review of methods. *PLOS ONE*, 12(4):e0176438, April 2017.

[77] John W Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co., 1977.

[78] Douglas M Hawkins. *Identification of Outliers*. Springer, 1980.

[79] Douglas C Montgomery. *Introduction to Statistical Quality Control*. John Wiley & Sons, 2019.

[80] Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, 2009.

[81] Petre Stoica and Yngve Selen. Model-order selection. *IEEE Signal Processing Magazine*, 21(4):36–47, July 2004.

[82] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), March 1978.

[83] https://github.com/pcm-dpc/COVID-19/. Accessed 20 Dec 2021.

[84] William O. Kermack and McKendrick Anderson G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, August 1927.

[85] Martina Amongero, Enrico Bibbona, and Gianluca Mastrantonio. *Analysing the Covid-19 pandemic in Italy with the SIPRO model.* In *Book of short papers - SIS 2021*. Pearson, 2021.

[86] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, 26(6):855–860, April 2020.

[87] Giulia Giordano, Marta Colaneri, Alessandro Di Filippo, Franco Blanchini, Paolo Bolzern, Giuseppe De Nicolao, Paolo Sacchi, Patrizio Colaneri, and Raffaele Bruno. Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in italy. *Nature Medicine*, 27(6):993–998, April 2021.

[88] Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, et al. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):e1009149, July 2021.

[89] Alessio Farcomeni, Antonello Maruotti, Fabio Divino, Giovanna Jona-Lasinio, and Gianfranco Lovison. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in italian regions. *Biometrical Journal*, 63(3):503–513, November 2020.

[90] A. S. Fokas, N. Dikaios, and G. A. Kastis. Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. *Journal of The Royal Society Interface*, 17(169):20200494, August 2020.

[91] Stephen M. Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H. Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 368(6493):860–868, May 2020.

[92] https://cran.r-project.org/web/packages/forecast/index.html. Accessed 20 Dec 2021.

[93] Dirk Bassler. Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *JAMA*, 303(12):1180, March 2010. DOI: 10.1001/jama.2010.310.

[94] Peter Bauer, Franz Koenig, Werner Brannath, and Martin Posch. Selection and bias-two hostile brothers. *Statistics in Medicine*, 29:1–13, 2009. DOI: 10.1002/sim.3716.

[95] Maximo Carreras and Werner Brannath. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine*, 32(10):1677–1690, June 2012. DOI: 10.1002/sim.5463.

[96] Peter K. Kimani, Susan Todd, Lindsay A. Renfro, and Nigel Stallard. Point estimation following two-stage adaptive threshold enrichment clinical trials. *Statistics in Medicine*, 37(22):3179–3196, May 2018. DOI: 10.1002/sim.7831.

[97] Arthur Cohen and Harold B. Sackrowitz. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8(3):273–278, August 1989. DOI: 10.1016/0167-7152(89)90133-8.

[98] Jack Bowden and Ekkehard Glimm. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal*, 50(4):515–527, August 2008. DOI: 10.1002/bimj.200810442.

[99] Peter K. Kimani, Susan Todd, Lindsay A. Renfro, Ekkehard Glimm, Josephine N. Khan, John A. Kairalla, and Nigel Stallard. Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. *Statistics in Medicine*, 39(19):2568–2586, May 2020. DOI: 10.1002/sim.8557.

[100] Matthias Brückner, Andrew Titman, and Thomas Jaki. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. *Statistics in Medicine*, 36(20):3137–3153, June 2017. DOI: 10.1002/sim.7367.

[101] John Whitehead. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581, 1986. DOI: 10.1093/biomet/73.3.573.

[102] Nigel Stallard and Susan Todd. Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference*, 135(2):402–419, December 2005. DOI: 10.1016/j.jspi.2004.05.006.

[103] Kevin Kunzmann, Laura Benner, and Meinhard Kieser. Point estimation in adaptive enrichment designs. *Statistics in Medicine*, 36(25):3935–3947, August 2017. DOI: 10.1002/sim.7412.

[104] David S. Robertson, Babak Choodari-Oskooei, Munya Dimairo, Laura Flight, Philip Pallmann, and Thomas Jaki. Point estimation for adaptive trial designs i: A methodological review. *Statistics in Medicine*, 42(2):122–145, November 2022. DOI: 10.1002/sim.9605.

[105] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356, December 2010. DOI: 10.1002/pst.472.

[106] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, June 1967. DOI: 10.1080/01621459.1967.10482935.

[107] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961. DOI: 10.1080/01621459.1961.10482090.

[108] Paul Gallo, Lu Mao, and Vivian H. Shih. Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics*, 24(5):976–993, August 2014. DOI: 10.1080/10543406.2014.932285.

[109] Ekkehard Glimm. Adjusting for selection bias in assessing treatment effect estimates from multiple subgroups. *Biometrical Journal*, 61(1):216–229, November 2018. DOI: 10.1002/bimj.201800097.

Appendix A

Incorporation of healthy volunteers data into a phase II proof-of-concept trial: supplementary material

A.1 mPBPK model

The extrapolation of the healthy volunteer model to the patients was based on several hypothesis:

- Affinity association and dissociation constants were considered similar in healthy volunteers and patients, as noted from internal data;
- Same quantity of membran-bound target in plasma for patients and healthy volunteers;
- Presence of a median of 95 interleukin receptors in healthy volunteers and 347 in patients, which is the rationale behind the λ parameter.

$$\frac{\partial C_{total}}{\partial t} = \frac{Input}{V_p} + \frac{1}{V_p}\left[C_{lymph} \cdot L - C_P \cdot L_1 \cdot (1-\sigma_1) - C_P \cdot L_2 \cdot (1-\sigma_2) - CL_p \cdot C_P - \frac{V_{max} \cdot C_P}{K_m + C_P}\right]$$

$$\frac{\partial C_{tight}}{\partial t} = \frac{1}{V_{tight}}\left[C_P \cdot L_1 \cdot (1-\sigma_1) - C_{tight} \cdot L_1 \cdot (1-\sigma_L)\right]$$

$$\frac{\partial C_{leaky}}{\partial t} = \frac{1}{V_{leaky}}\left[C_P \cdot L_2 \cdot (1-\sigma_2) - C_{leaky} \cdot L_2 \cdot (1-\sigma_L) - \frac{V_{max} \cdot C_{leaky}}{K_m + C_{leaky}}\right]$$

$$\frac{\partial C_{lymph}}{\partial t} = \frac{1}{V_{lymph}}\left[C_{tight} \cdot L_1 \cdot (1-\sigma_L) + C_{leaky} \cdot L_2 \cdot (1-\sigma_L) - C_{lymph} \cdot L\right]$$

$$\frac{\partial R_s}{\partial t} = k_{syn} - k_{deg} \cdot R_s + \frac{k_{deg} \cdot R_s \cdot C_P}{K_{SS} + C_P}$$

$$C_P = 0.5 \cdot \left[(C_{total} - R_s - K_{SS}) + \sqrt{(C_{total} - R_s - K_{SS})^2 + 4 \cdot K_{SS} \cdot C_{total}}\right]$$

$$RO = \frac{C_{leaky}}{C_{leaky} + K_m}$$

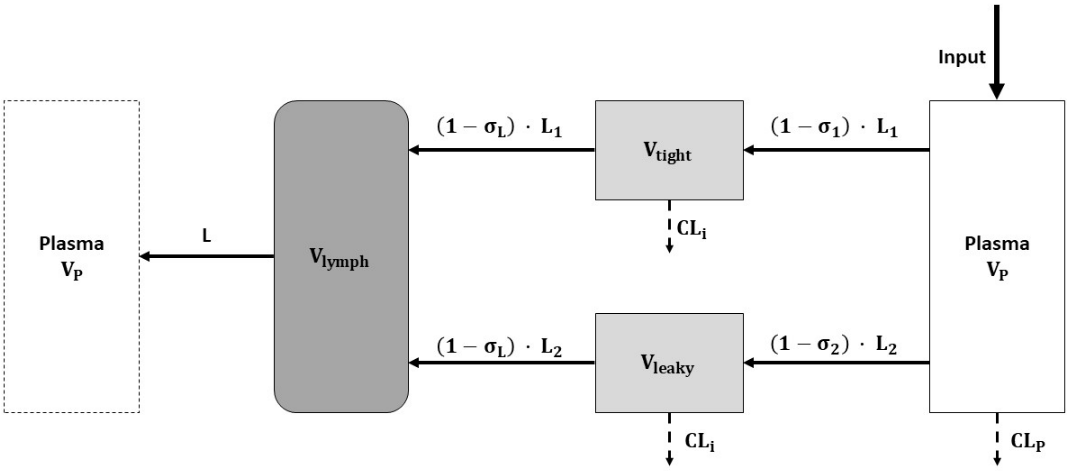


Fig. A.1 Representation of the mPBPK model.

Table A.1 Key parameters of the mPBPK model.

Parameter	Description
$Input$	Input concentration of the drug
C_{total}	Total concentration of the drug
C_{lymph}	Concentration of the drug in the lymph volume
C_P	Concentration of the drug in the plasma
C_{tight}	Concentration of the drug in the tight compartment
C_{leaky}	Concentration of the drug in the leaky compartment
V_P	Plasmatic volume
V_{max}	Maximum binding capacity in the binding site
V_{tight}	Volume of the distribution of the drug in the tight compartment
V_{leaky}	Volume of the distribution of the drug in the leaky compartment
V_{lymph}	Total lymph volume
L	Lymph flow
L_1	Flow in the tight compartment
L_2	Flow in the leaky compartment
σ_1	Vascular reflection coefficient for the volume of tight compartment
σ_2	Vascular reflection coefficient for the volume of leaky compartment
σ_L	Lymphatic reflection coefficient
CL_P	Systemic clearance
K_m	Concentration of the free (not bound) drug
K_{syn}	Constant for soluble receptor degradation
K_{deg}	Free soluble receptor degradation rate
K_{SS}	Quasi-stationarity constant
R_s	Amount of soluble receptor

A.2 Histogram of ROs

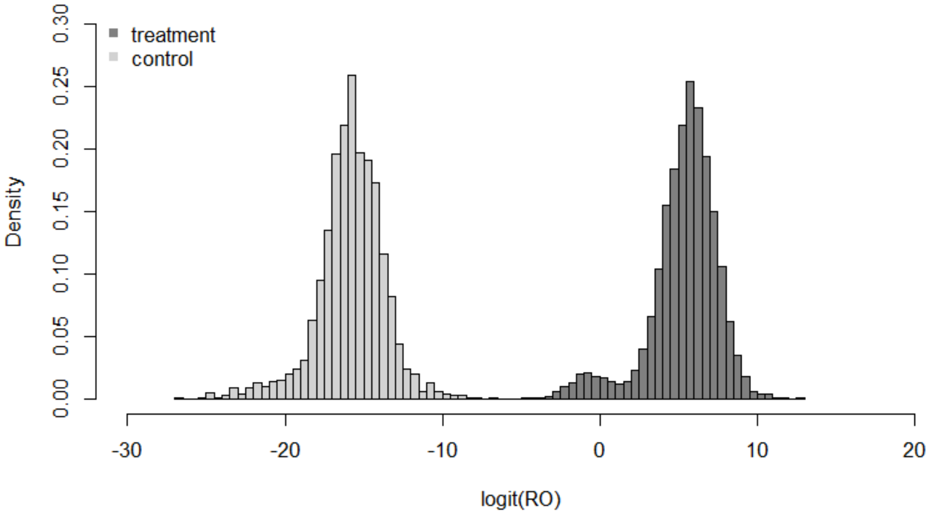


Fig. A.2 Histogram of the distribution of $\gamma = \text{logit(RO)}$ in the treatment and control arm.

A.3 Heatmaps

In this supplementary material we present additional simulation scenarios cited in the main article. We start from the heatmaps on type I error, maximum type I error and power. We considered the first efficacy criteria only, the second efficacy criteria only and both efficacy criteria.

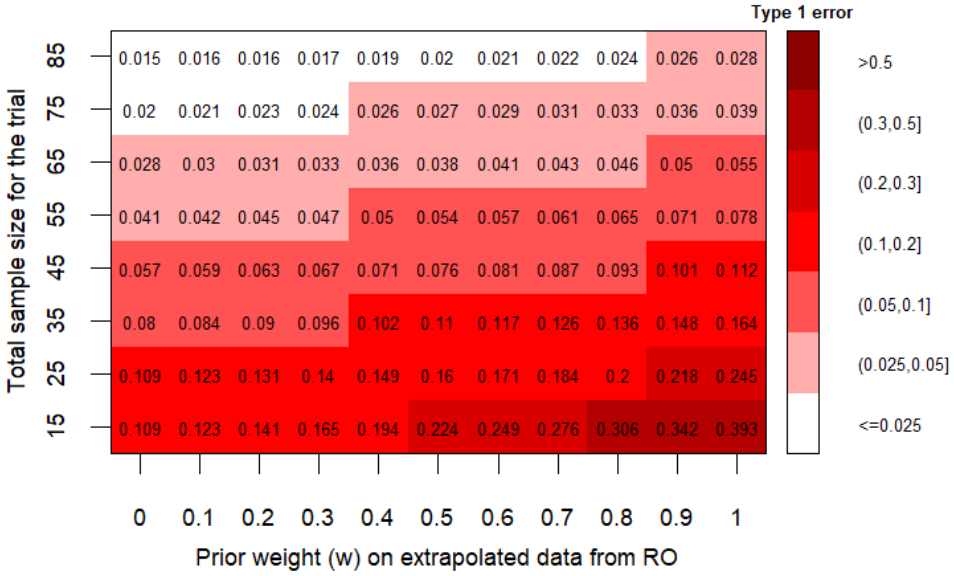
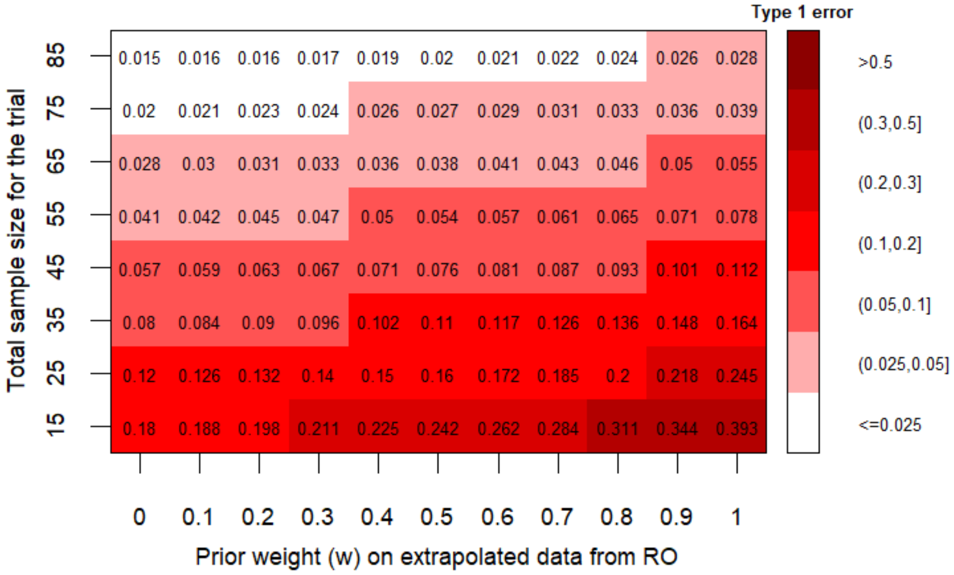
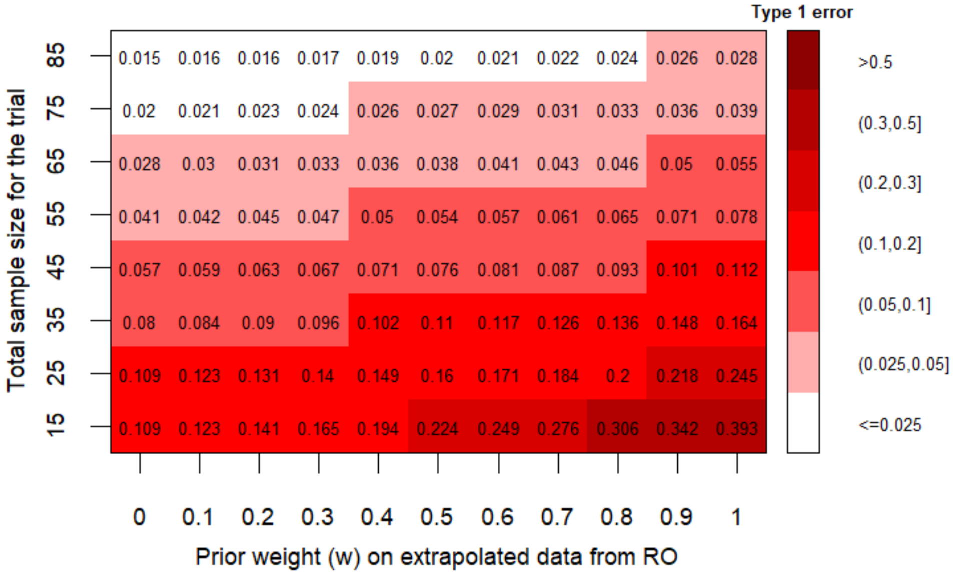


Fig. A.3 Heatmap of Type I Error with the first efficacy criteria only for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect (-1).



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
85	0.015	0.016	0.016	0.017	0.019	0.02	0.021	0.022	0.024	0.026	0.028
75	0.02	0.021	0.023	0.024	0.026	0.027	0.029	0.031	0.033	0.036	0.039
65	0.028	0.03	0.031	0.033	0.036	0.038	0.041	0.043	0.046	0.05	0.055
55	0.041	0.042	0.045	0.047	0.05	0.054	0.057	0.061	0.065	0.071	0.078
45	0.057	0.059	0.063	0.067	0.071	0.076	0.081	0.087	0.093	0.101	0.112
35	0.08	0.084	0.09	0.096	0.102	0.11	0.117	0.126	0.136	0.148	0.164
25	0.12	0.126	0.132	0.14	0.15	0.16	0.172	0.185	0.2	0.218	0.245
15	0.18	0.188	0.198	0.211	0.225	0.242	0.262	0.284	0.311	0.344	0.393

Fig. A.4 Heatmap of Type I Error with the second efficacy criteria only for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect (-1).



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
85	0.015	0.016	0.016	0.017	0.019	0.02	0.021	0.022	0.024	0.026	0.028
75	0.02	0.021	0.023	0.024	0.026	0.027	0.029	0.031	0.033	0.036	0.039
65	0.028	0.03	0.031	0.033	0.036	0.038	0.041	0.043	0.046	0.05	0.055
55	0.041	0.042	0.045	0.047	0.05	0.054	0.057	0.061	0.065	0.071	0.078
45	0.057	0.059	0.063	0.067	0.071	0.076	0.081	0.087	0.093	0.101	0.112
35	0.08	0.084	0.09	0.096	0.102	0.11	0.117	0.126	0.136	0.148	0.164
25	0.109	0.123	0.131	0.14	0.149	0.16	0.171	0.184	0.2	0.218	0.245
15	0.109	0.123	0.141	0.165	0.194	0.224	0.249	0.276	0.306	0.342	0.393

Fig. A.5 Heatmap of Type I Error considering both efficacy criteria for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect (-1).

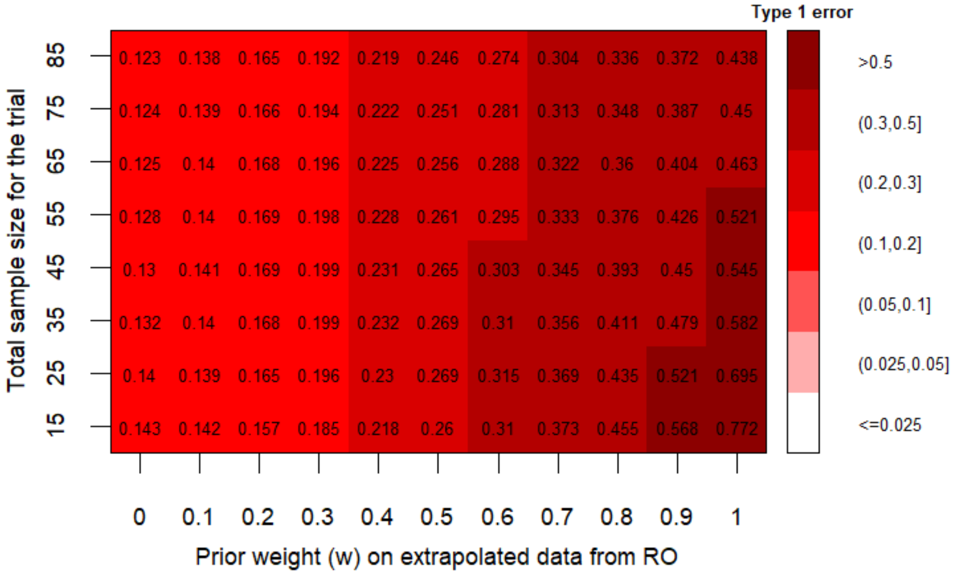


Fig. A.6 Heatmap of the maximum Type I Error with the first efficacy criteria only for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect and spans in $[-7.3;4.9]$.

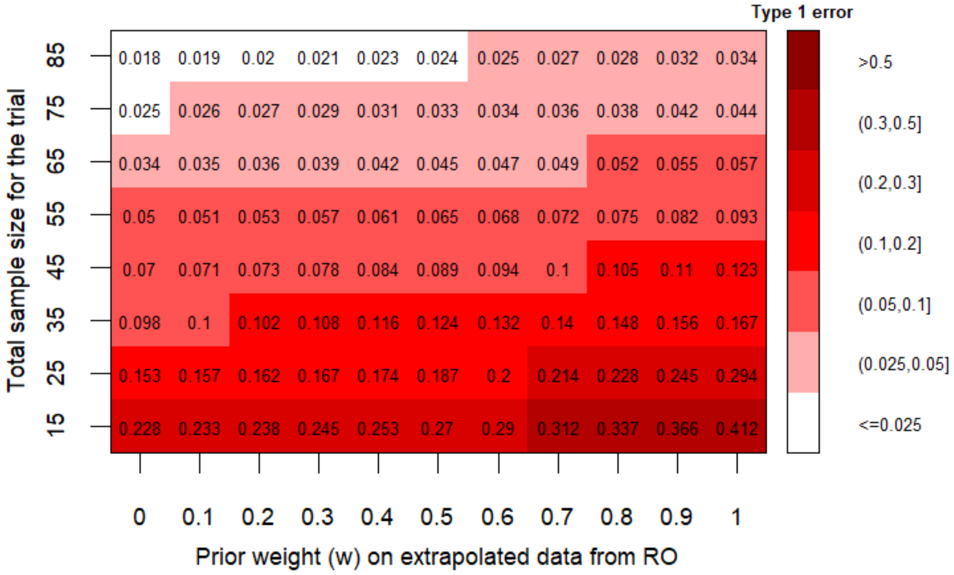


Fig. A.7 Heatmap of the maximum Type I Error with the second efficacy criteria only for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect and spans in $[-7.3;4.9]$.

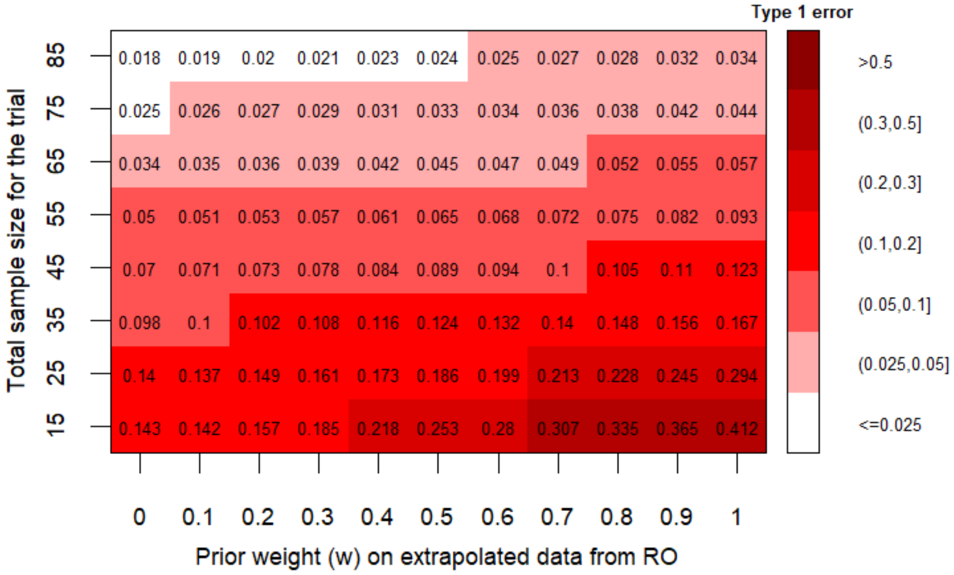


Fig. A.8 Heatmap of the maximum Type I Error with both efficacy criteria for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment effect is equal to the expected control effect and spans in $[-7.3;4.9]$.

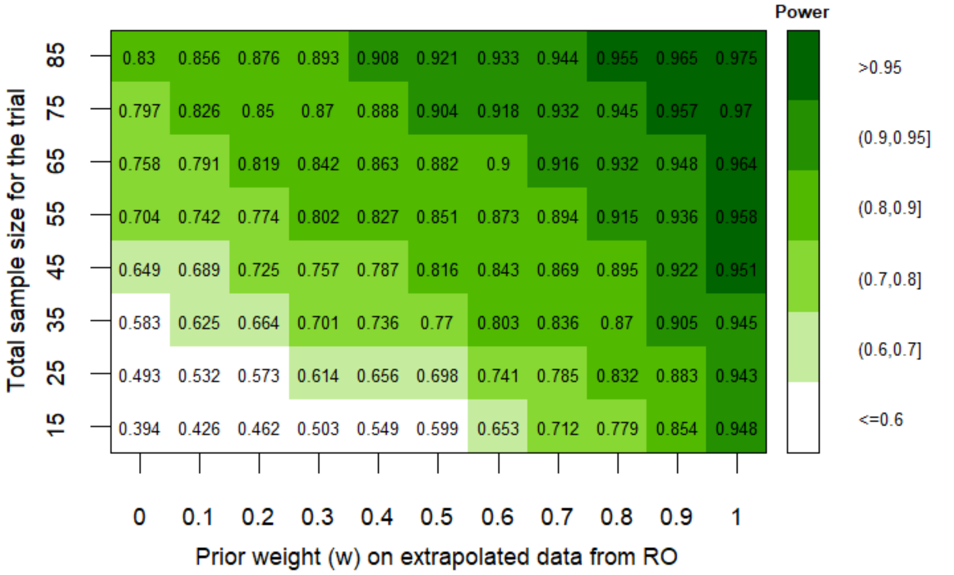


Fig. A.9 Heatmap of power reached with the first efficacy criteria for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment (-4) is more effective than the control (-1).

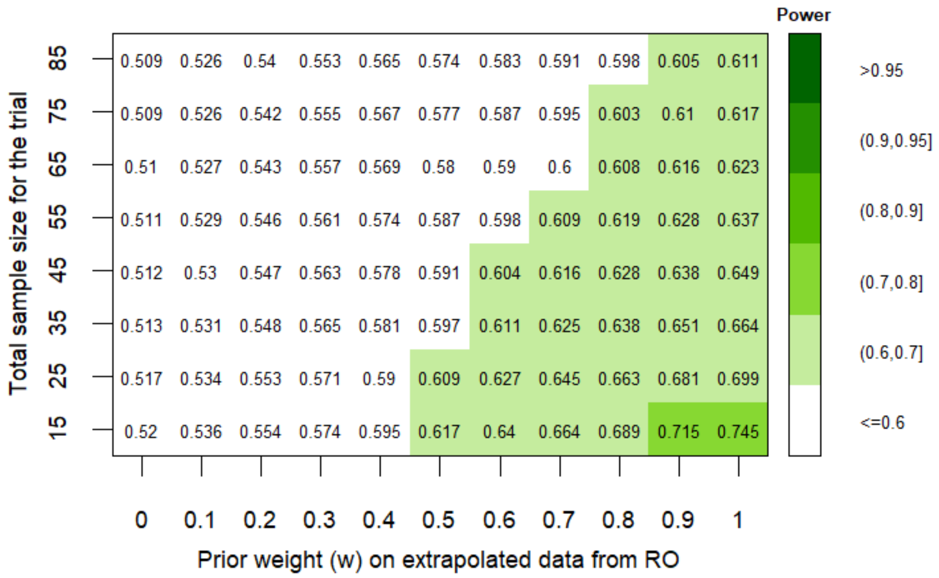


Fig. A.10 Heatmap of power reached with the second efficacy criteria for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment (-4) is more effective than the control (-1).

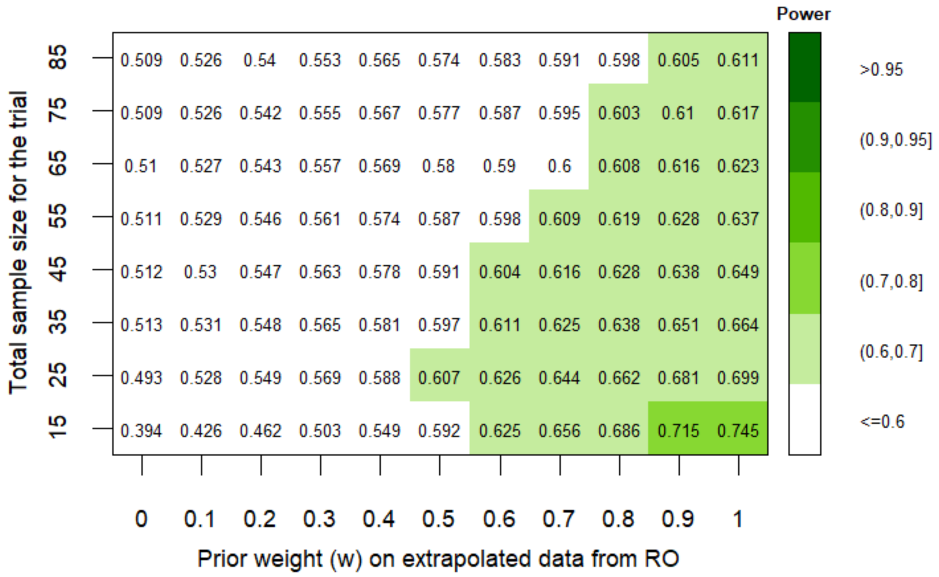


Fig. A.11 Heatmap of power reached with both efficacy criteria for robust priors with different weights varying the total sample size with (2:1) randomization. The treatment (-4) is more effective than the control (-1).

A.4 Long-run operating characteristics

Here we present the long-run operating characteristics for the BDB design with $w = 0.8$ and different true responses for the treatment and the control. We provide the type I error, the power, the average posterior weight, the average gain in precision and the average bias in posterior estimation. These are calculated via 100000 simulations. The average gain in precision is defined as 1 - (width of the 80% credibility interval / width of the 80% confidence interval). The bias in posterior estimation is the difference between the posterior mean and the true response.

Table A.2 Long-run operating characteristics of BDB design with w=0.8, using different assumptions for the true treatment effect. We have considered a control response of -1 and a study with 45 (2:1) patients and a patient standard deviation of 6.

Assumed true treatment response	**Drift relative to the prior treatment mean**	**Pointwise type I error** (assuming treatment equal to control)	**Power** (assuming control response equal to -1)	**Average posterior weight on prior distribution**	**Average gain in precision** (with respect to a design without borrowing)	**Average bias in posterior estimation**
-8	-4.21	9.7%	99.1%	0.39	-12%	0.7
-6	-2.21	9.6%	91.2%	0.80	26%	0.8
-4	-0.21	10.2%	62.8%	0.92	26%	0.1
-2	1.79	10.3%	22.7%	0.85	23%	-0.7
0	3.79	7.3%	2.7%	0.49	-14%	-0.8

Table A.3 Long-run operating characteristics of BDB design with w=0.8, using different assumptions for the true control effect. We have considered a treatment response of -4 and a study with 45 (2:1) patients and a patient standard deviation of 6.

Assumed true control response	**Drift relative to the prior control mean**	**Pointwise type I error** (assuming treatment equal to control)	**Power** (assuming control response equal to -1)	**Average posterior weight on prior distribution**	**Average gain in precision** (with respect to a design without borrowing)	**Average bias in posterior estimation**
-3	-2.98	10.5%	25%	0.78	25%	1.1
-2	-1.98	10.3%	41.9%	0.85	26%	0.8
-1	-0.98	9.3%	62.8%	0.89	26%	0.4
0	0.02	7.3%	80.3%	0.90	26%	0
1	1.02	4.7%	91.6%	0.88	21%	-0.4

A.5 Posterior probability distribution

Here we present the posterior distributions of the BDB design with $w = 0.8$ and different observed responses for the treatment and the control. We provide: the point estimate and the 80% credibility interval for the treatment or control in the design without borrowing; the point estimate and the 80% credibility interval for the treatment difference in the design without borrowing; the posterior weight on the prior distribution; the posterior median and 80% credibility interval for the treatment or control; the posterior median and 80% credibility interval for the treatment difference.

Table A.4 Summary of the estimated posterior distributions of a BDB design with w=0.8 and the design without borrowing properties of the treatment response. We have considered a control response of -1, a study with 45 (2:1) patients and a patient standard deviation of 6.

Assumed observed treatment response	**Design without borrowing point estimate of the treatment response [80% CI]**	**Design without borrowing point estimate of the difference in response [80% CI]**	**Posterior weight on prior distribution in BDB design**	**Point estimate (posterior median) of the treatment response [80% CrI]**	**Point estimate (posterior median) of the difference in response [80% CrI]**
-8	-8 [-9.4;-6.6]	-7 [-9.4;-4.6]	0.36	-7.2 [-8.9;-5.5]	-6.6 [-8.9;-4.4]
-6	-6 [-7.4;-4.6]	-5 [-7.4;-2.6]	0.86	-5.1 [-6.2;-4]	-4.5 [-6.4;-2.7]
-4	-4 [-5.4;-2.6]	-3 [-5.4;-0.6]	0.94	-3.9 [-4.9;-2.9]	-3.3 [-5.1;-1.5]
-2	-2 [-3.4;-0.6]	-1 [-3.4;1.4]	0.89	-2.8 [-3.8;-1.7]	-2.2 [-4;-0.4]
0	0 [-1.4;1.4]	1 [-1.4;3.4]	0.52	-1.0 [-2.5;0.8]	-0.5 [-2.6;1.8]

Table A.5 Summary of the estimated posterior distributions of a BDB design with w=0.8 and the design without borrowing properties of the control response. We have considered a treatment response of -4, a study with 45 (2:1) patients and a patient standard deviation of 6.

Assumed observed control response	**Design without borrowing point estimate of the control response [80% CI]**	**Design without borrowing point estimate of the difference in response [80% CI]**	**Posterior weight on prior distribution in BDB design**	**Point estimate (posterior median) of the control response [80% CrI]**	**Point estimate (posterior median) of the difference in response [80% CrI]**
-3	-3 [-5;-1]	-1 [-3.4;1.4]	0.84	-1.8 [-3.4;-0.2]	-2.2 [-4;-0.2]
-2	-2 [-4;0]	-2 [-4.4;0.4]	0.89	-1.1 [-2.6;0.4]	-2.8 [-4.6;-0.9]
-1	-1 [-3;1]	-3 [-5.4;-0.6]	0.91	-0.6 [-2;0.9]	-3.3 [-5.1;-1.5]
0	0 [-2;2]	-4 [-6.4;-1.6]	0.92	0 [-1.5;1.5]	-3.9 [-5.7;-2.1]
1	1 [-1;3]	-5 [-7.4;-2.6]	0.91	0.5 [-0.9;2]	-4.4 [-6.3;-2.6]

A.6 Approximation by mixtures of two normal distribution

The histogram displaying the empirical distribution of $\gamma = \text{logit(RO)}$ reveals a bimodal distribution in the treatment arm. This bimodality arises from the variation in the rate of decay of RO following the last treatment dose among individual patients. Some patients experience a rapid decline, while others exhibit a slower decay, depending on their unique patient-specific parameters.

An intriguing possibility to address this observed bimodal distribution is to extend the mixture prior introduced in the main paper. This extension involves adding an extra informative component to better capture this bimodal shape. Utilizing the data presented in the main paper and modeling the treatment arm data as a mixture of two Normal distributions, we obtain the following results:

$$\pi_T \sim 0.612 \times N(-3.467, 1.108^2) + 0.388 \times N(-4.290, 1.022^2)$$

$$\pi_C \sim N(-0.018, 1.595^2)$$

Figure A.12, A.13, and Table A.6 replicate the analysis results presented in the main paper for the scenario described above. These results closely resemble those previously discussed, with the only noticeable difference occurring in the type I error rates at the extreme levels, where, beyond the defined plausible range, the two distributions exhibit significant differences in their tails.

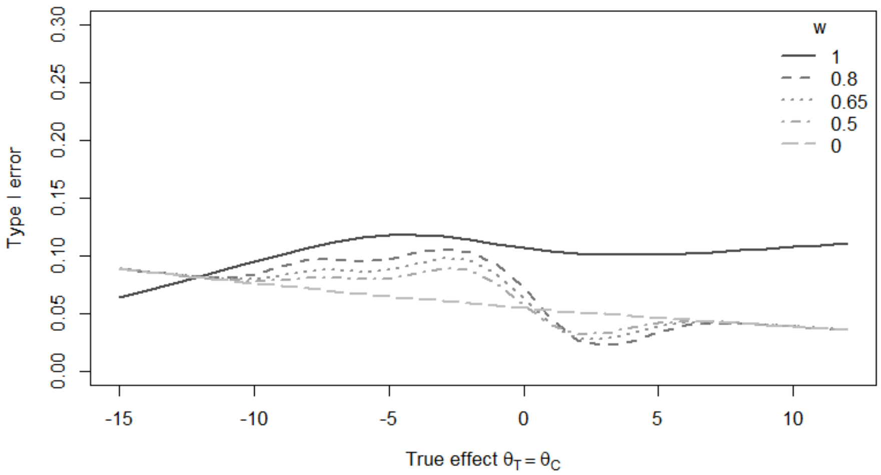


Fig. A.12 The plot depicts the relationship between the Type I error and the effect size in both arms, assuming equal effect sizes, while considering 30 patients in the treatment arm, 15 patients in the control arm, and varying prior weights (w) assigned to the informative component.

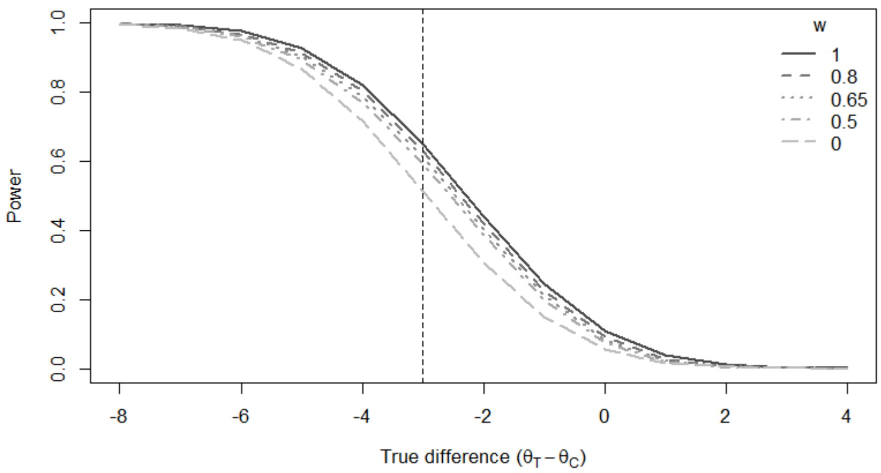


Fig. A.13 The plot displays the power within the BDB design as a function of the discrepancy between the effects in each arm. This analysis assumes a control effect of $\theta_C = -1$, involves 30 patients in the treatment arm, 15 patients in the control arm, and various prior weights (w) assigned to the informative component. It's worth noting that a smaller difference between the treatment and control effects indicates a more favorable treatment outcome. The vertical dotted line signifies the minimum clinically significant difference of -3.

Table A.6 Summary of the operating characteristics for BDB designs with 30 patients in the treatment arm, 15 patients in the control arm, varying prior weights (w) on the informative component, and design without borrowing. The plausible for the treatment and control effects is specified as [-7.3, 4.9].

Design	**Type I error when** $\theta_T = \theta_C = -1$	**Maximum type I error over the plausible range** (value at which occurs)	**Range of values where type I error is greater than 10%** (probability under π_T and π_C)	**Power when** $\theta_T = -4$ and $\theta_C = -1$
BDB with w=1	11.0%	11.8% (-4.4)	[-7.3, 4.9] (99.8%)	65.1%
BDB with w=0.8	9.3%	10.6% (-2.9)	[-4.4,-1.6] (10.7%)	63.0%
BDB with w=0.65	8.4%	9.8% (-2.8)	-	61.2%
BDB with w=0.5	7.6%	9.0% (-2.6)	-	59.3%
BDB with w=0	5.7%	7% (-7.3)	-	51.1%
Frequentist	5.2%	5.2% (all)	-	50%

A.7 Fictive and sensitivity analysis in a false positive scenario

Similar to the fictive analysis presented in the main paper, a case-study, if combined with an inappropriate application of the methods, could lead to a false positive decision. Such a scenario is illustrated in Table A.7. In this case, the final results are heavily influenced by the extrapolated data, which do not provide additional evidence to the phase II study. This observation underscores the significance of the tipping point sensitivity analysis shown in Figure A.14. It becomes evident that both success criteria are far from being met in the phase II study alone, and a weight as high as 0.8 (the weight chosen for incorporation) on the informative component is required to satisfy both criteria, unlike the favorable case presented in the main paper. In such a trial, a Go decision should not be made.

Table A.7 Summary of the primary analysis on the treatment difference, treatment and control response, utilizing hypothetical but realistic data. The lower row showcases the simulated observed data from a design without borrowing, clearly illustrating a failure to meet the success criteria. In contrast, the upper row presents the outcomes achieved by combining fictive observed data and informative components using a BDB design with a weight of $w = 0.8$ demonstrating the fulfillment of the success criteria.

Evidence Source	**Treatment difference [80%CrI]**	**Treatment effect [80%CrI]**	**Control effect [80%CrI]**
Phase I + phase II	-3 [-5.6;-0.1]	-3.4 [-4.9;-1.9]	-0.4 [-2.6;1.7]
Phase II only (frequentist)	0 [-6.1;6.1]	-2 [-5.5;-1.5]	-2 [-7;3]

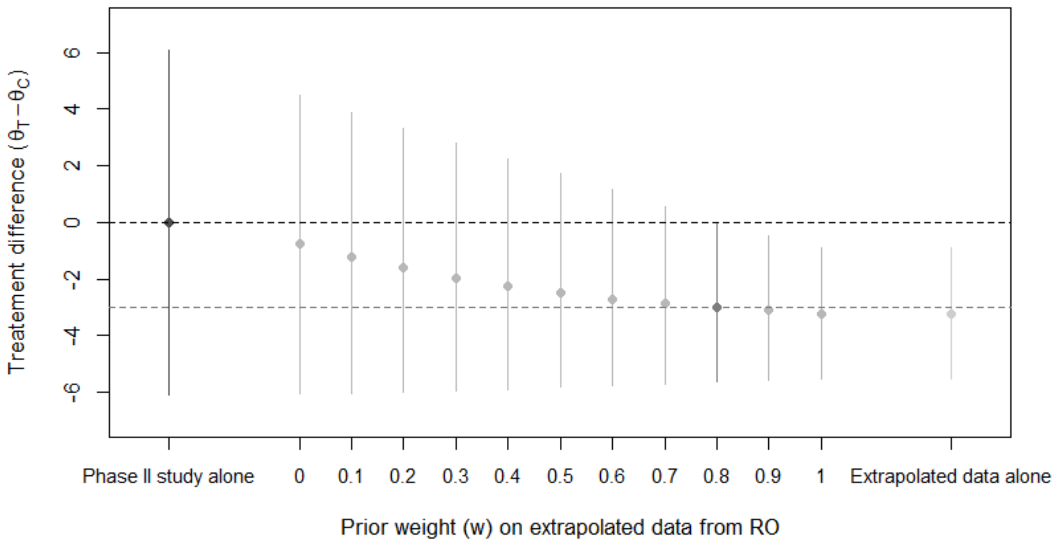


Fig. A.14 Sensitivity analysis conducted using hypothetical but realistic data, displaying the posterior mean and 80% credible interval (CrI) for the estimated treatment difference in relation to the prior weight. The two dashed lines on the graph represent the two success criteria thresholds: $(P[(\theta_T - \theta_C) < 0] > 0.9)$ and $(P[(\theta_T - \theta_C) < -3] > 0.5)$.

Appendix B

A comparison of estimation methods in adaptive enrichment designs with time-to-event endpoints: supplementary material

B.1 Additional simulation scenarios

In this supplementary material we present additional simulation scenarios cited in the main article. In Figure B.1 we set the threshold to $b = 0$. In Figure B.2 we set $\tilde{T}_1$ to 3 months after the interim analysis. We also compare 4 sub-populations while keeping the other parameters as in the main setting and in Figure B.3 we present the results averaged in all sub-populations, while in Figure B.4 we show the performance of the estimators in each sub-population in the case of linear effects on the sub-populations. In Tables B.1, B.2 and B.3 we present also the empirical probabilities of selection for the different sub-populations in the different scenarios.

Table B.1 Empirical probability of selection for the different sub-populations in the simulation study when $b = 0$, according to their log HR.

Log HR	$\delta_i = 0$	$\delta_i = -0.1$	$\delta_i = -0.2$	$\delta_i = -0.3$
Probability of selection	50%	70%	84%	93%

Table B.2 Empirical probability of selection for the different sub-populations in the simulation study when $\tilde{T}_1 = T_1 + 90$ *days*, according to their log HR.

Log HR	$\delta_i = 0$	$\delta_i = -0.1$	$\delta_i = -0.2$	$\delta_i = -0.3$
Probability of selection	30%	50%	69%	83%

Table B.3 Empirical probability of selection for the different sub-populations in the simulation study when 4 sub-populations are included, according to their log HR.

Log HR	$\delta_i = 0$	$\delta_i = -0.1$	$\delta_i = -0.2$	$\delta_i = -0.3$
Probability of selection	33%	50%	65%	80%

B.2 Threshold equal to zero

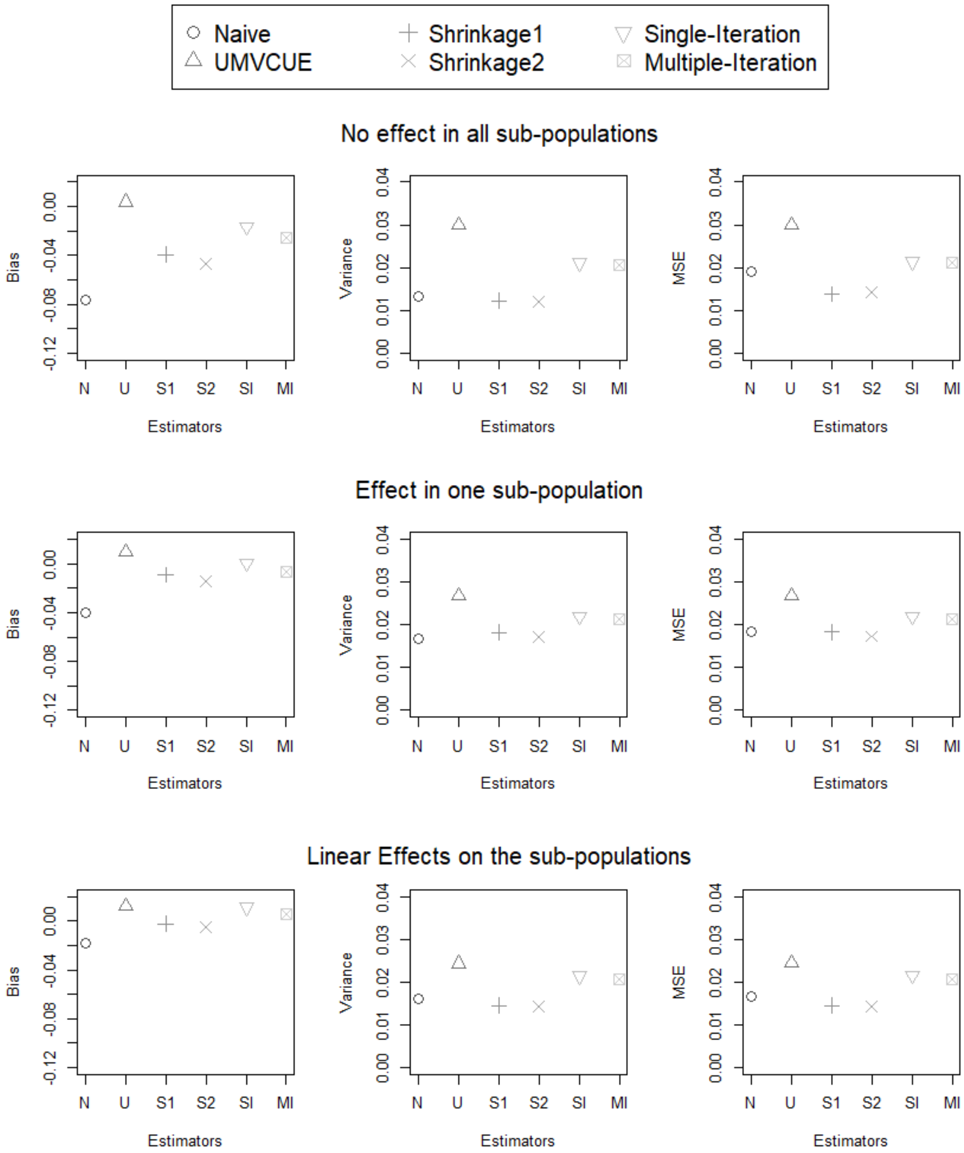


Fig. B.1 Estimators' performances in case of three sub-populations and $b = 0$. Top row: treatment ineffective in all sub-populations $\delta = (0,0,0)$; Middle row: treatment effective only in one sub-population $\delta = (0\ ,0\ ,-0.3)$; Bottom row: linear effect on the sub-populations $\delta = (-0.1,-0.2,-0.3)$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

B.3 Stage 1 patients followed for 90 days after interim analysis

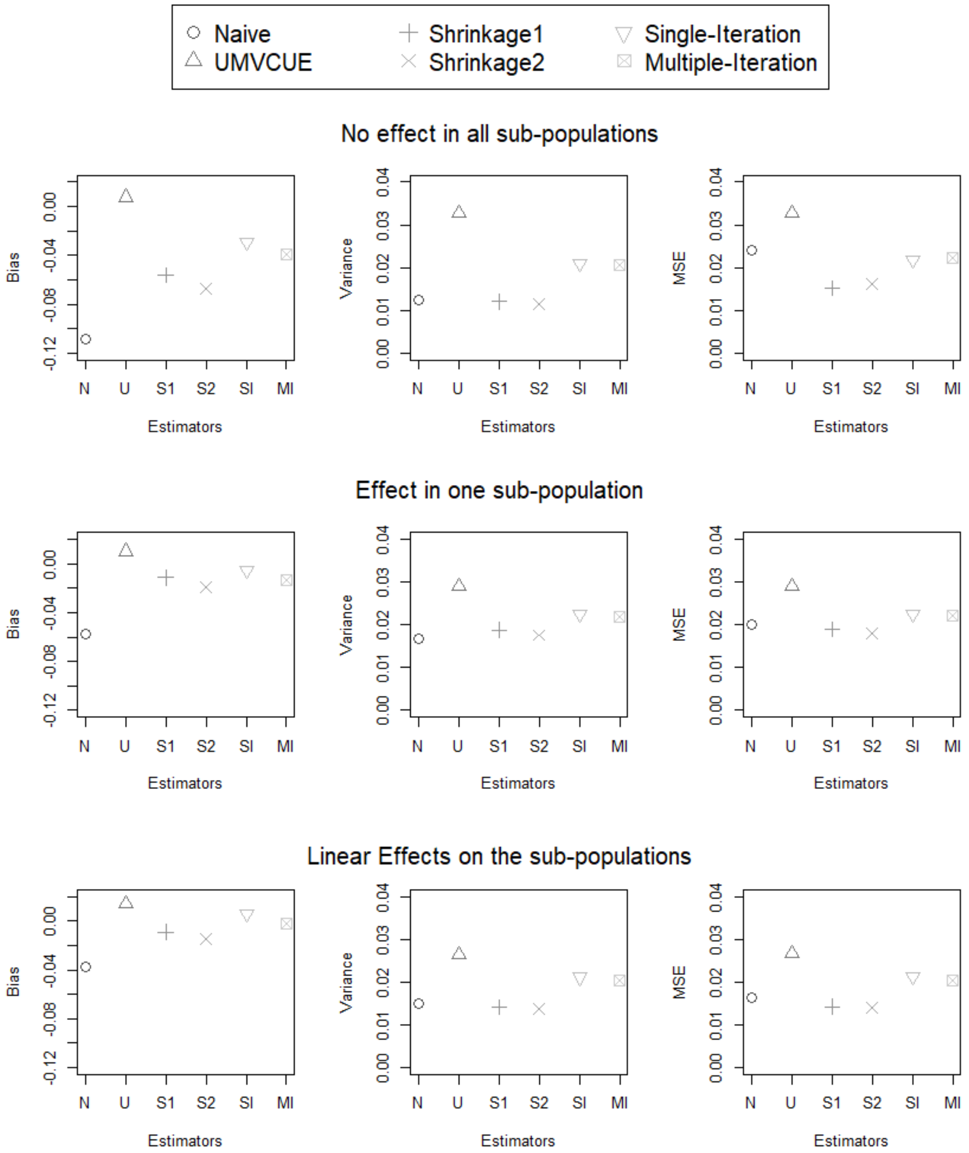


Fig. B.2 Estimators' performances in case of three sub-populations and $\tilde{T}_1 = T_1 + 90\ days$. Top row: treatment ineffective in all sub-populations $\delta = (0,0,0)$; Middle row: treatment effective only in one sub-population $\delta = (0\ ,0\ ,-0.3)$; Bottom row: linear effect on the sub-populations $\delta = (-0.1,-0.2,-0.3)$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

B.4 4 sub-populations

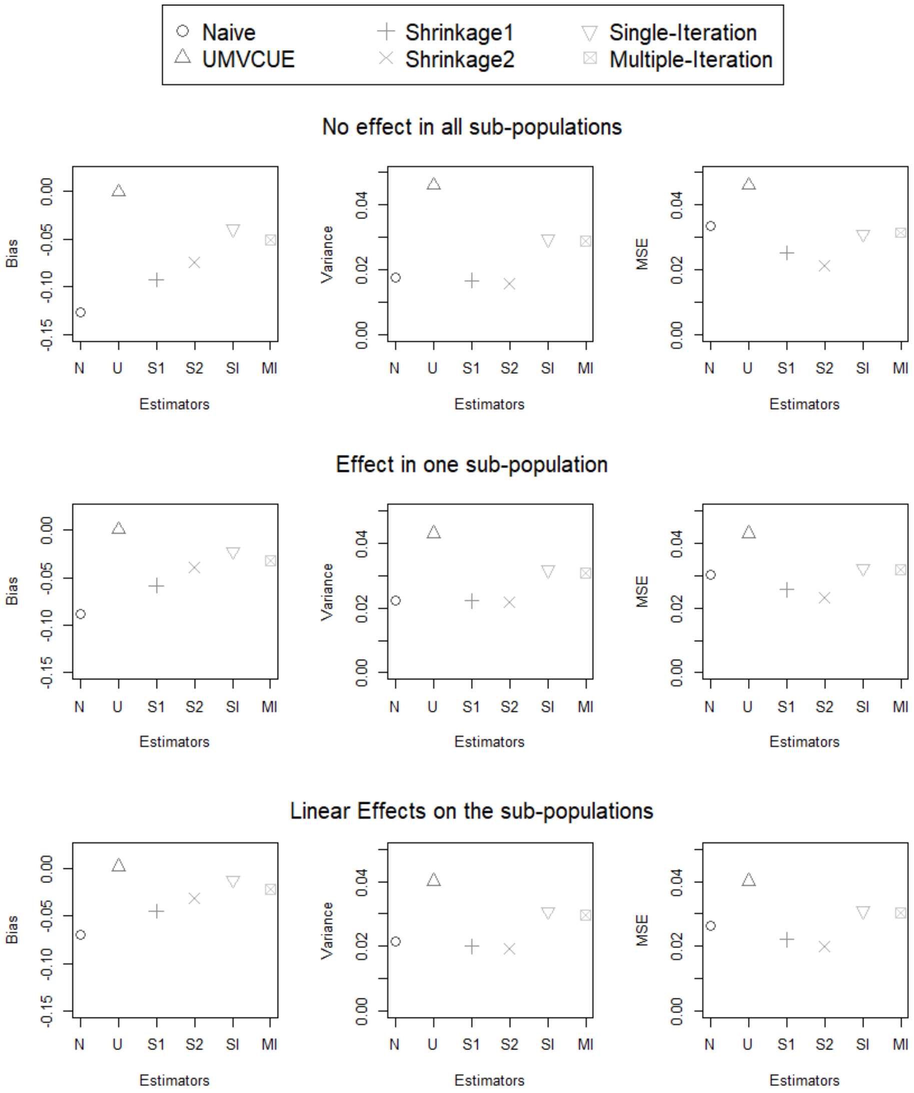


Fig. B.3 Estimators' performances in case of four sub-populations. Top row: treatment ineffective in all sub-populations $\delta = (0,0,0,0)$; Middle row: treatment effective only in one sub-population $\delta = (0\ ,0\ ,0\ ,-0.3)$; Bottom row: linear effect on the sub-populations $\delta = (0\ ,-0.1,-0.2,-0.3)$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

B.5 Sub-population specific bias, variance and MSE with 4 sub-populations

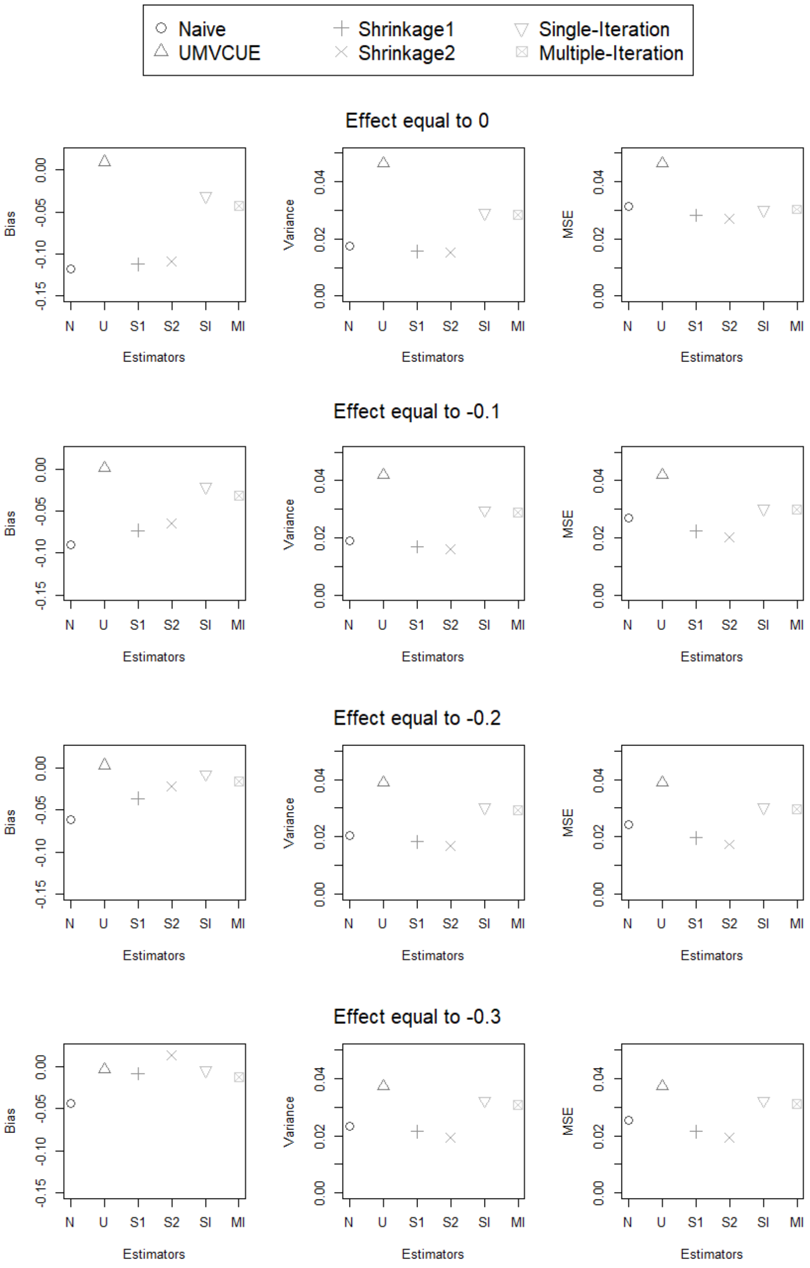


Fig. B.4 Estimators' performances in each sub-population in case of four sub-populations and linear effects on the sub-populations. From top row to bottom row effect equal to: $0, -0.1, -0.2, -0.3$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

B.6 Boxplots of the estimators

For completeness, we present also boxplots for the estimators.

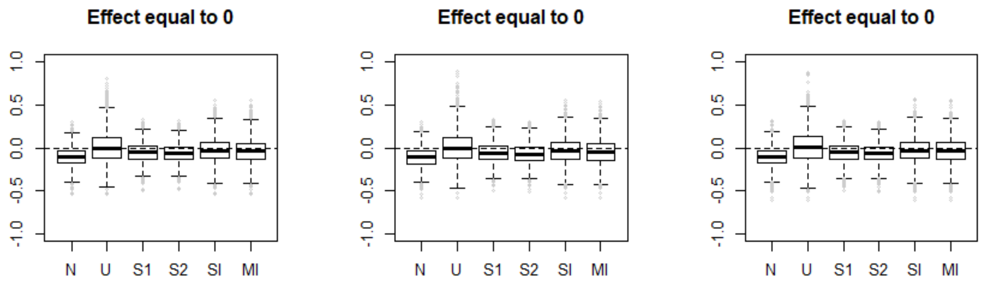


Fig. B.5 Estimators' boxplots for the different sub-populations in case of three sub-populations and effect equal to: $\delta = (0, 0, 0)$.

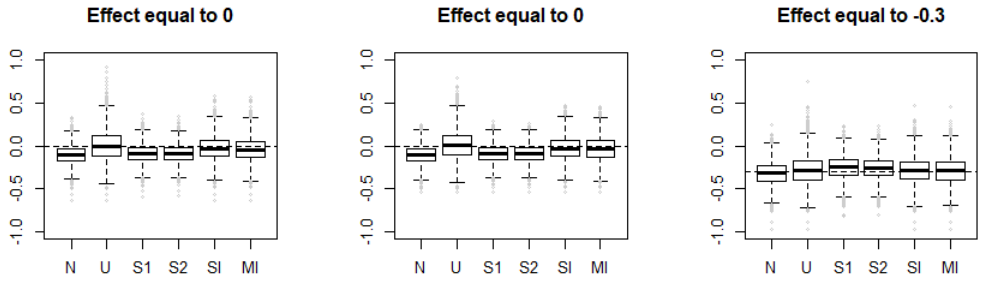


Fig. B.6 Estimators' boxplots for the different sub-populations in case of three sub-populations and effect equal to: $\delta = (0, 0, -0.3)$.

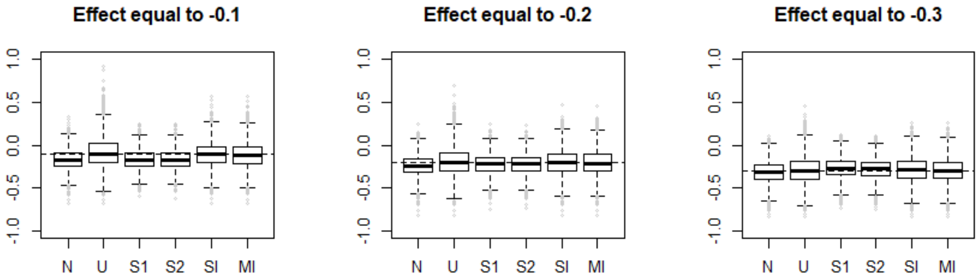


Fig. B.7 Estimators' boxplots for the different sub-populations in case of three sub-populations and effect equal to: $\delta = (-0.1, -0.2, -0.3)$.

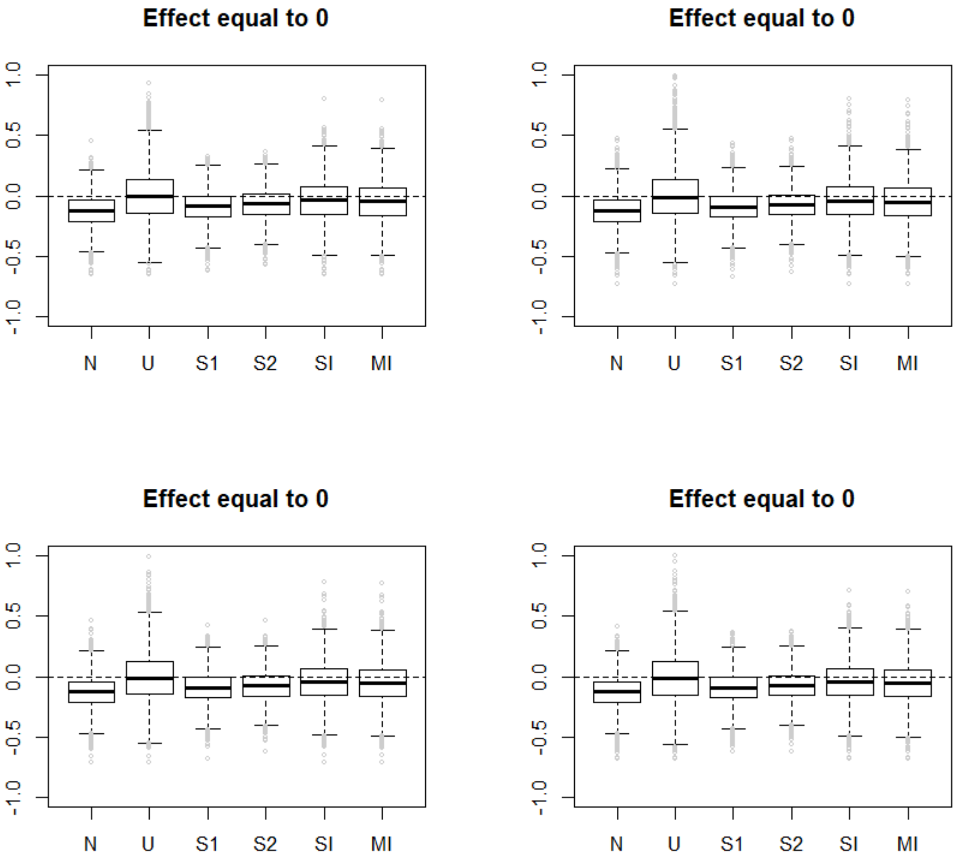


Fig. B.8 Estimators' boxplots for the different sub-populations in case of four sub-populations and effect equal to: $\delta = (0, 0, 0, 0)$.

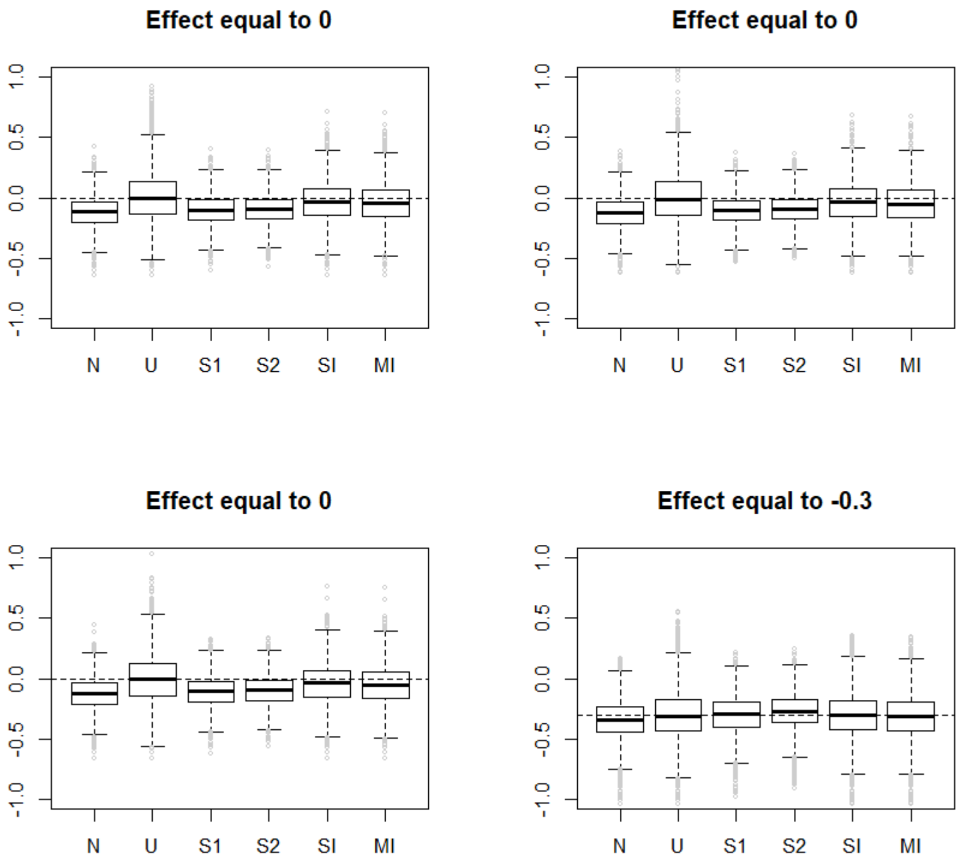


Fig. B.9 Estimators' boxplots for the different sub-populations in case of four sub-populations and effect equal to: $\delta = (0, 0, 0, -0.3)$.

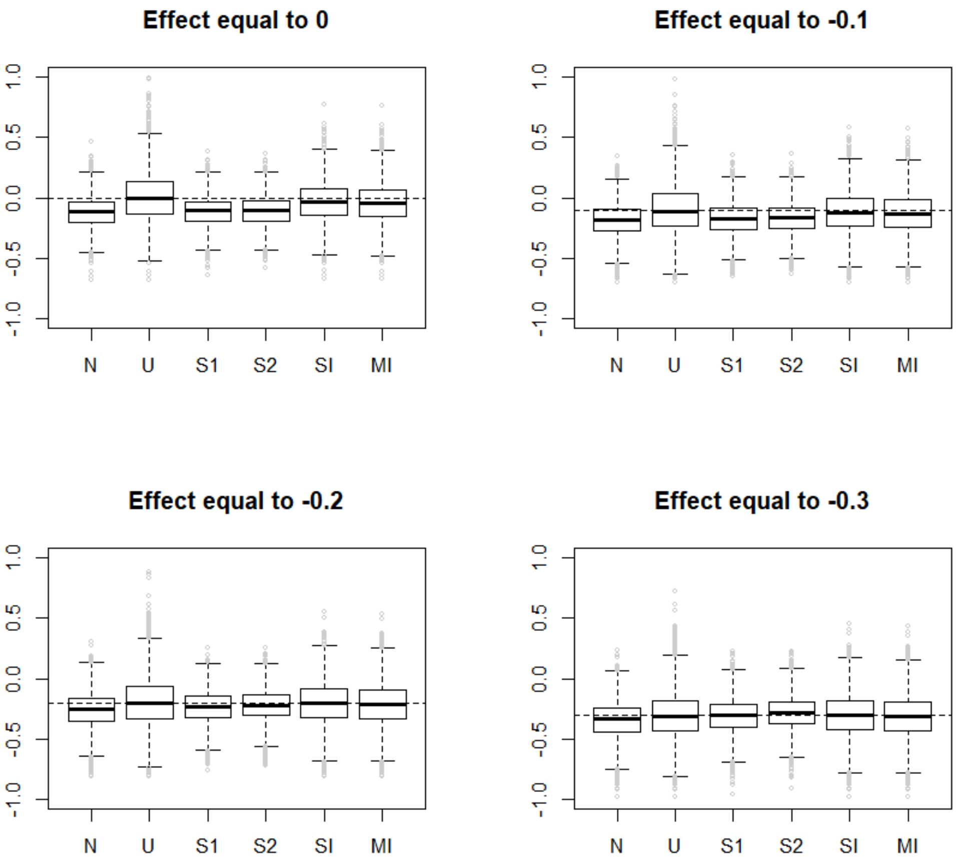


Fig. B.10 Estimators' boxplots for the different sub-populations in case of four sub-populations and effect equal to: $\delta = (0, -0.1, -0.2, -0.3)$.