SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Medicine and Surgery

PhD program in Public Health

Cycle XXXVII

Curriculum in Biostatistics and Epidemiology

RARE DISEASES: A COMPARISON OF BAYESIAN METHODS FOR BASKET TRIALS AND AN INNOVATIVE SEQUENTIAL DESIGN

Candidate: RISCA GIULIA
Registration number: 891803

Tutor: Prof.ssa Stefania Galimberti

Coordinator: Prof. Luigi Badano

ACADEMIC YEAR 2023-2024

Contents

Abstract

Rare diseases present unique challenges in the design of clinical trials due to limited patient populations, making it difficult to recruit enough participants to have potential clinical benefits also statistically significant. This thesis explores innovative trial designs, with a particular focus on the use of Bayesian methods and Master Protocols to address these challenges and enhance the research in this field. Bayesian methods offer several advantages for trials involving small sample sizes. By incorporating prior knowledge, enabling continuous updates, and supporting adaptive trial designs, the Bayesian framework provides flexibility and precision that is particularly valuable in the context of rare diseases. Additionally, Master Protocols, such as Basket and Platform trials, allow for the simultaneous evaluation of multiple treatments across different patient subgroups, thereby improving patient enrolment efficiency, optimizing resource allocation, and accelerating the process of evaluating novel therapies.

This thesis has two main objectives. First, we aim to assess the robustness of three standard Bayesian methods, specifically adapted to continuous outcomes, in the analysis of basket trials within the context of rare diseases. The second objective, motivated by a context of rare diseases, is to develop a novel, sequential, single-arm Platform-Basket trial design.

The results of the simulation study to assess the value of Bayesian models for the analysis of basket trials in rare diseases are very promising. In particular, when clear treatment effects are present within subpopulations, the three methods can be confidently applied even when sample sizes are very small (i.e. 4-7). However, when the treatment effects are mild or when there is a relevant heterogeneity in treatment responses, careful consideration of sample size and final cut-off probability for final decision is crucial during the planning phase of the trial.

The novel two stage design we have proposed combines innovative elements to plan a complex clinical study involving a platform-basket trial back-bone, reinforced by the use of external information (e.g. from a completed trial on a similar disease) and an interim evaluation for futility. The use of interim analyses is impactful, enabling early decisions to either halt unpromising trials or expand the study to include additional subgroups based on predictive probabilities. We assessed the

performance of this design through simulations that resulted sensitive to the choice of some parameters (e.g. prior weights, cut-off probability for the interim analysis and the final decision), even if it was in general robust when there are strong believes of highly effective treatments.
It is important to emphasise that in the set-up of the protocol it is of paramount importance to study the properties of the proposed design through simulations. Indeed, the operating characteristics should be carefully evaluated by extensive clinical trial simulations whose results should be always part of the study protocol.
In conclusion, the work presented here serves as a foundational starting point for enhancing our understanding of how to design a trial in the context of rare diseases and additional work is needed for a more comprehensive evaluation.

Keywords

Rare diseases; Basket trial; Sequential design; Adaptive design; Interim analysis.

Riassunto

Le malattie rare presentano particolari sfide per quanto riguarda la progettazione di trial clinici a causa della limitata disponibilità di pazienti. Ciò rende difficile il reclutamento di un numero sufficiente di partecipanti allo studio che abbiano un potenziale beneficio clinico che sia anche statisticamente significativo. Questa tesi esplora dei disegni di studio innovativi, con focus particolare sull'uso dei metodi bayesiani e dei Master Protocols, per affrontare le difficoltà e contribuire alla ricerca in questo campo.

I metodi bayesiani offrono diversi vantaggi per gli studi che coinvolgono campioni di piccole dimensioni. Incorporando conoscenze pregresse, permettendo aggiornamenti continui e supportando disegni adattivi, l'approccio bayesiano offre flessibilità e precisione, risultando particolarmente utile nel contesto delle malattie rare. Inoltre, i Master Protocols, come i Basket Trial e i Platform Trial, consentono di valutare contemporaneamente più trattamenti in diversi sottogruppi di pazienti, migliorando così l'efficienza nel reclutamento, ottimizzando l'allocazione delle risorse e accelerando il processo di valutazione di terapie nuove.

Questa tesi ha due obiettivi. Il primo è quello di valutare la robustezza di tre metodi bayesiani standard, specificamente adattati alla presenza di endpoint continui, nell'analisi di un basket trial nel contesto di malattie rare. Il secondo obiettivo, motivato da un contesto di malattie rare, è quello di sviluppare un nuovo disegno di studio per un Platform-Basket trial, sequenziale, con il solo braccio di trattamento.

I risultati dello studio di simulazione per valutare le performance dei modelli bayesiani nell'analisi dei basket trials in malattie rare sono molto promettenti. In particolare, in presenza di rilevanti effetti del trattamento nei vari sottogruppi di malattia, i tre metodi risultano già ottimali con campioni molto piccoli (ovvero 4-7 pazienti). Tuttavia, quando l'effetto del trattamento è lieve o c'è eterogeneità nelle risposte ai trattamenti, è fondamentale scegliere attentamente la numerosità campionaria e i cut-off decisionali durante la fase di pianificazione del trial.

Il nuovo disegno a due fasi che abbiamo proposto combina elementi innovativi per progettare uno studio clinico complesso, utilizzando come base un trial Platform-Basket, rafforzato dall'uso di informazioni esterne (per esempio da un trial su una malattia simile già completato) e da una valutazione ad interim per futilità. L'uso

delle analisi intermedie è particolarmente importante, in quanto consente di interrompere precocemente trial che non si dimostrano promettenti o di espandere lo studio includendo sottogruppi aggiuntivi, basandosi su delle probabilità predittive. Abbiamo valutato le performance di questo disegno attraverso simulazioni, che sono risultate sensibili alla scelta di alcuni parametri (come ad esempio i pesi scelti a priori, i cut-off per le decisioni ad interim e finale), anche se, in generale, il disegno è risultato robusto in presenza di forti evidenze sull'efficacia dei trattamenti.
È importante sottolineare che nella definizione del protocollo è fondamentale studiare le proprietà del disegno che abbiamo proposto attraverso simulazioni. Infatti, le caratteristiche operative devono essere valutate attentamente tramite estese simulazioni di trial clinici i cui risultati devono sempre essere riportati nel protocollo di studio.
In conclusione, questa tesi è solo il punto di partenza per migliorare la metodologia relativa alla pianificazione di studi clinici nel contesto delle malattie rare e ulteriori sviluppi saranno necessari per una valutazione più approfondita e completa.

Parole chiave

Malattie rare; Basket trial; Disegni sequenziali; Disegni Adattivi; Analisi ad interim.

Introduction

Rare diseases are defined as conditions affecting less than 5 in 10000 individuals in the EU and less than 7.5 in the US, with estimates of thousands currently identified. Technological advances, especially in genomics, will result in an increase in this number, raising interest in personalised medicine and targeted therapies. However, the limited patient populations pose significant challenges in the design of clinical trials, often leading to a preference for single arm studies due to difficulties in recruitment and ethical issues. To overcome these issues, researchers are considering the adoption of innovative trial designs that explore the use of external data as control group or supplementary information derived from trials in the same class of disease and/or in the same treatment. Indeed, regulatory bodies are actively involved in this innovation process and they recommend a continuous dialogue to optimise study designs in the planning stages, emphasising the need for adaptations and Bayesian statistical methods to enhance research efforts.

In particular, the Bayesian framework offers several advantages in clinical trials, particularly when dealing with very small sample sizes. First, Bayesian methods allow for the incorporation of prior information, which can be especially valuable when historical data on the same or similar treatment and/or disease, or expert opinions, are available, enhancing the robustness of the analysis. Second, they provide a flexible framework for dynamic updating of findings as new data are collected, allowing for adjustments in ongoing trials. In fact, Bayesian methods also facilitate adaptive trial designs, enabling researchers to make informed decisions about continuing, modifying, or stopping trials based on interim analyses. Additionally, Bayesian approaches can yield more precise estimates of treatment effects by borrowing information between multiple subpopulations. This is crucial in rare disease research, where sample sizes may be insufficient to reach the scientific validity required in traditional study designs. In addition, they allow for a more intuitive interpretation of results, presenting findings as probabilities rather than p-values, which can be more meaningful to clinicians and stakeholders.

An supplementary approach that has been recently introduced with the aim to improve clinical trial research and particularly suited to the context of rare diseases is related to the development of Master Protocols. They are innovative trial designs

that allow a simultaneous evaluation of multiple treatments across defined patient subgroups based on specific biomarkers or genetic mutations. They include three types of design (i.e. basket trials, umbrella trials and platform trials) that are able to enhance patient enrollment, reduce costs, and accelerate development through adaptive designs and interim analyses. However, due to their complexity they require robust statistical methodologies both at the planning and at the analysis stage and a constant regulatory assistance and compliance.
We have integrated all these elements (i.e. Bayesian methods, adaptive interim analyses, internal and external information borrowing) in facing the design of the clinical trial that has motivated this thesis. Our goal was to design a Phase I-II single arm trial for three rare subtypes of a lysosomal pediatric disease. A specific gene-mutation identifies the three patient subgroup of interest and they have in common the type/nature of treatment and thus the operational infrastructure and the primary endpoint of activity, represented by a continuous variable. Acknowledging the potential for differential effectiveness among the enrolled subpopulations by a basket design, we assume that heterogeneity in the outcome exists as an intrinsic help in evaluating treatment effects. Additionally, since we have access to results from a completed clinical trial on the same treatment in a similar disease, we aim to leverage that information in the current trial.
This thesis has two main objectives:

1. assess the robustness of bayesian methods in the analysis of a basket trial in the setting of rare diseases;

2. develop a new design for a sequential single arm platform-basket trial.

The evaluation of the standard bayesian methodologies for the design and analysis of basket trials is the starting point to show their validity also with respect to small sample sizes. All the main stakeholders (i.e. researchers involved in the trial, regulatory agencies and patients) should be aware whether "small" is "too small" and when traditional methods eventually fail. This is why we used three standard methods (i.e. Bayesian hierarchical model [1], the Exchangeability-Noexchangeability method [2] and the Treatment response Borrowing [3] approach) and assessed their statistical performances in a simulation study that considered

a single arm basket trial with few subgroups and sample sizes. Since these models were originally proposed to deal with a binary outcome, they were adapted in this thesis for the analysis of a continuous endpoint.

Then, we proposed a new design that accounts for all the complexities of the motivating clinical context. In particular, we developed a sequential single arm platform-basket trial that combines the leveraging of external information and a sequential design to enhance the traditional basket trial design (in rare diseases). Two stages are outlined: the first one involves an interim analysis on the data from the two initial subgroups, and then, based on the results, the second stage may also include a third subgroup to obtain the final evaluation. The interim analysis considers a stopping rule for futility on predictive prior probability that the new subgroup exceeds a specific threshold with a certain probability, given the data of the two initial subgroups. This innovative approach aims to optimise resource allocation and enhances the patient outcomes power in rare diseases. Ultimately, it includes ethical considerations allowing for multiple subgroups to enter into the trial and improve considerations about the decision-making criteria to strengthen the evidence for treatment effectiveness.

The thesis is organised as follows. In Chapter 1 we reviewed the design of clinical trials in rare diseases, provided an overview on the advantages of the Bayesian framework in context with small sample sizes and we introduced the Master Protocols. In Chapter 2 we introduced the three standard Bayesian methods for the analysis of basket trials, their adaptation to continuous endpoints, the simulation study to assess their performances in the presence of limited sample size and their results. Chapter 3 presents the motivating clinical context and the novel study design we propose together with the simulation study to assess its robustness and the results. Conclusions are in Chapter 4.

1 The design of clinical trials in rare diseases

A rare disease for the European Union (EU) is a condition with a prevalence of less than 5 subjects in 10000, whereas for the United States (US) is less than 7.5 subjects in 10000 [4]. Currently, the number of rare diseases is estimated between 6000 and 8000 in EU and over 7000 in US. However, technological advances, especially in the field of genomics, will lead to an increasing number of these conditions. In the meantime, the interest in the personalized medicine [5] had sharply risen, promoting the development of more specific drugs for individual genetic mutations. Indeed, one important aim of personalized medicine is to provide drugs and cures for rare or ultra rare diseases. The regulatory agencies like European Medicines Agency (EMA) and Food and Drug Administration (FDA) together to many associations (e.g. International Rare Diseases Research Consortium) have developed over the years guidance and plans to assist this field of research, also from an economic point of view. Besides, FDA is now more open to non traditional designs when the investigation regards very rare diseases [6] or orphan drugs [7]. Furthermore, many projects were funded to institute multidisciplinary research groups with the goal to work on trial design for small populations, see the INSPIRE and the IDeAIInSPiRe EU funded projects. Nevertheless, the methodological standards required for a rare disease trial are equivalent to those for a non-rare disease [8], creating many issues that complicate the development, the registration and the diffusion of new treatments. The most impacting complication is certainly the limited number of subjects needed to demonstrate the efficacy of the novel treatment. As a consequence, the feasibility of a study is jeopardised and both the time to enroll patients and the interval of follow-up greatly increase. Consequently, the costs of supporting a clinical study on a rare disease also increase. However, one option to overcome the problem of small sample size is to conduct longitudinal studies, which require fewer subjects and allow to collect a more representative response [9]. Additionally, they are widely used in rare diseases to analyse pre-clinical studies [10], useful for guiding the conduction of early phase trials. Alternatively, longitudinal studies are also widely used in rare diseases for the analysis of secondary endpoints [11]. Another aspect to keep in mind working in a very rare disease setting regard the unfeasibility to conduct a

standard randomized clinical trial (RCT) due to the recruitment difficulties, the absence of a standard of care and the ethical aspects. Many times, there are so few people with a rare or ultra-rare condition (and who fit the eligibility criteria) that it is not possible to make a control group as well. Finally, a rare disease linked to a genetic mutation can often be very disabling, so from an ethical point of view the aim is to allow everyone to receive a promising treatment within the trial. For these reasons, a single arm trial without a concurrent control group is often conducted and accepted by the regulatory agencies. Bell and Tudur Smith in 2014 [12] had reviewed clinical trials registered in ClinicalTrials.gov counting a 63% of single arm studies in rare diseases versus a 29% in no-rare diseases. Even so, the implicit assumption behind a non comparative trial is that any changes measured in the outcomes are a consequence of the treatment and are not intrinsic of the disease [13]. For this reason, the conduction of a study on the natural history of the disease of interest is a good start to understand its course and to help the development of new therapies. There are some examples of studies with this aim in the literature [14], however the resources required to achieve sufficient information from a long-term natural history study are considerable. In the recent years, the possibility of replacing the internal control group with external data has been extensively studied. However, the selection of external data, like historical data or data from registries, to use in the current trial need appropriate considerations to verify that they match each other [15]. Nevertheless, the integration of external data is spreading a lot because it allows to reinforce the power of the study and to avoid the risk of bias, in addition to reduce the trial size [16]. Even so, the inclusion of external controls into clinical trials requires careful analysis and expert considerations. Some other options are already considered to avoid the limited number of patients when rare diseases are involved. For example, some methodological development like sample size calculation based on a decision-theoretic approach [17] or sample size re-estimation [18] or seamless trial [7] or crossover design [19] or again extrapolation from adults to pediatric clinical trials (FDA Guidance) are recent aspects aiming to reduce the sample size. However, when it comes to non-standard clinical trials for rare diseases, the regulatory agencies' guidelines underline that their decisions will be taken on a case-by-case. In addition, they recommend to consult regulators in advance to optimize the drug development process (see FDA

and EMA Guidelines). Finally, a very recent option is to rely on in silico clinical trials, which are based on complex and interdisciplinary models that mimic the responses of the parties involved into the studies in order to minimise the in vivo experiments [20].
Given the complexity of clinical trials on rare diseases, an efficient network of scientific societies and patients' associations at global level is a fundamental resource to share information on patients, clinical and methodological strategies. Biobanks, clinical database and registries should be encouraged to increment the knowledge on the natural history or the prevalence/incidence of rare diseases at international level. Also, they allow to reach all the potential stakeholders in the development of a new drug, especially the orphan ones, both from the point of view of researchers and patients [21]. Last but not least, the statistical methodology plays an important role to adequate the classic methods to a more extreme area of research. In particular, the Bayesian framework seems to be a viable solution to deal with many issues previously mentioned [22, 23].
In section 1.1 we introduce the Bayesian framework in clinical trials in a setting of rare diseases. Then, in section 1.2 we present the different type of Master Protocols (i.e. Basket, Umbrella and Platform trial).

1.1 The Bayesian framework in clinical trials on rare diseases

In the Bayesian framework the measure of interest (e.g. the measure of treatment effect) is treated as a random variable arising from a prior distribution, defined on previous and subjective knowledge. Then, after data collection the prior distribution is updated to a posterior distribution through the likelihood function. Therefore, the statistical inference is based on the posterior distribution. For example, in a clinical trial the decision rules about the success of a new intervention are based on the posterior probability, replacing the well known p-values. More details regarding the Bayesian framework are reported in Supplementary material 5.1. Due to the challenges that can be potentially encountered in the practice, for example the prior elicitation, the Bayesian framework is not widely used in standard clinical trials. Nonetheless, there is a lot of literature in favor of its use

because Bayesian methods in some contexts may offer more attractive designs, lower sample size [24] and more direct answers to clinical questions [25]. In particular, it has recently become an useful instrument in rare diseases due to its flexibility, especially in early phase trials. Bayesian inference remains robust to changes in study design because it is fundamentally based on the data collected rather than the specific design itself. As a consequence, the Bayesian process perfectly reflects the concept of adaptive designs. Such designs represent a great opportunity to avoid the limited knowledge available in the planning phase of trial, especially on rare disease. They consist in a decision-making process defined a priori, according to which is possible to adjust sample size [26], target population [27], treatments, endpoints, and randomization ratios [28]. All the decisions are based on the measured responses at specified recruitment steps or time-points by interim analysis. So, the interim results guide the progress of the trial and direct the structure of the study design. Sometimes, the decision rules may include a stop for futility of the treatment under investigation or a stop for efficacy. So, the trial terminates earlier than planned due to ineffectiveness or because it has reached a successful level of evidence on the treatment effect. The advantage to consider futility/efficacy stopping rules is evident from an ethical point of view and sample size reduction. These are features also important in the setting of rare diseases.
Similarly, the sequential designs that are an alternative to adaptive design,offer the possibility to have a non-predefined number of patients to enroll at the beginning of the study and to evaluate the need to recruit further or not, applying interim analysis at specific steps [19]. This kind of design is advantageous in case of rare diseases because thank to the futility or efficacy interim evaluations it allows to reduce the number of enrolled patients [29]. Moreover, stopping the inferior treatment group they preserve future patients and eventually offer them the opportunity to join the most promising treatment. Furthermore, sequential design sometimes facilitate and accelerate the process between different phases, usually phase II and III, selecting the superior treatment in phase II and confirming it in phase III [30].
In addition, the idea of borrowing information from external sources is well-established in Bayesian framework, given that it naturally incorporates different level of hierarchy in the construction of the priors to account for external evidences.

For example, data from other studies on the same treatment and/or target population may be included to inform the prior of current trial. Beyond that, there are many ways of borrowing data according to different scopes, for example to transfer knowledge from larger target population to smaller one [31], depending on the similarity between populations/periods and treatment. Furthermore, many approaches exist to borrow historical data to use as external control group in randomized trials [32] by different kind of prior (e.g. power [33], non informative, commensurate [16] or robust meta-analysis predictive prior [34]). Also, multiple external sources may be simultaneously considered as control group taking into account the heterogeneity between the studies to preserve the type I error and power [35]. Additionally, an adjustment by baseline covariates can be introduced to align the current and the historical trials [36]. In any case, sensitivity analysis are needed to investigate the impact of borrowing on the results, especially on type I error and power.
Then, besides the borrowing from external data, it is possible to borrow information within the trial reducing the total sample size needed. Data of the same trial belonging to different target populations can share the information to provide the evidence of efficacy of the same treatment in different population subgroups. Otherwise, it is possible to borrow between different treatment arms that involve the same target population. Both those cases make it possible to exploit the information already present in the current trial to increase the strength of the response on all the studied subgroups. The latter two designs fall within the so-called Master Protocol and will be the subject of the next section.

1.2 Master Protocols

Master Protocols are innovative designs, proposed in the context of personalized medicine, to study simultaneously the effect of single/multiple drugs on defined multiple/single subgroups. They aim to answer the question: "What are the (and are there any) target populations in which the treatment is effective?". Master Protocols can respond to multiple questions together, reducing time and the number of recruited patients. Recent advances, particularly in omic research, have shifted the focus of the study design from the drug to the patient, evolving to-

wards a patient-centered perspective designs. The definition of a biomarker or the detection of specific DNA's mutation have allowed to develop patient-specific drugs, improving the sensitivity of the treatment. The treatments and the patients are matched according to specific genetic, biological or clinical conditions. Hence, subjects candidate to enter in a clinical trial undergo medical screening to detect the biomarker or the genetic mutation of interest and based on their tests they are assigned to the corresponding subgroup/treatment arm. As a result, Master Protocols give patient a better chance of being enrolled into one of the subgroups, having multiple marker/treatment ongoing at the same time [37]. Sometimes, a subgroup is also considered for those patients who tested negative for all the interested biomarkers, still giving them a chance to receive potential beneficial treatment. Thus, Master Protocols allow to allocate patients more effectively to treatment subgroups for which they are eligible, trying to enhance their outcome. Additionally, one single protocol is required to evaluate all at once different subgroups, reducing costs, regulatory bureaucracy and accelerating the clinical development. Usually, they consider short-term endpoints in order to implement frequent futility/efficacy interim analyses, reducing the number of patients assigned to treatment arms that turn out to be ineffective/toxic. Furthermore, the use of a common infrastructure to collect and analyse data can improve the efficiency and the quality of the results [38]. On the other hand, the operational and methodological aspects that are faced in this setting require skilful adjustment. Moreover, there may be certain limitations in this approach, such as the scientific validity, the balance between benefits and risks and the misunderstanding of the informed consent [39]. In fact, while the possibility of including into the trials as many subjects as possible gives hope to many patients, it can also disappoint them easily.

Master Protocols make it possible to include subgroups with a rare target mutation into the study. However, if the included sample is too small due to the low prevalence in the population, the validity of the results may be questioned. Also, the time to recruit a sufficient sample size may be long when the target population for one treatment arm is limited, delaying the progress of the whole trial. Nevertheless, the flexibility and the affordability of this new designs encourage the implementation of these approaches in the clinical research and also encourage regulatory

agencies to define accurate guidelines to protect the ethicality of the studies. For example, FDA has drawn up a guidance to regulate clinical trials using master Protocols within industrial research (Guidance for Industry: FDA-2018-D-3292), describing all operational features to follow. The guidance also includes the use of not randomised designs, suggesting that they be combined with interim analyses to avoid carrying out ineffective studies.
In the past ten years, the number of Master Protocols has increased and this resulted in many reviews [40]-[41] where the advantages and limitations of these innovative approaches are presented. A comprehensive paper, which summarised the clinical trials planned and/or conducted based on master protocol was drafted by Fountzilas *et al.*[42]. The main area of application of master protocols is in oncology, for the design of phase II trials, especially single arm studies using binary outcomes and frequentist decision rules [41]. However, they are also used in phase III trials with various endpoints (i.e. time-to-event, continuous and longitudinal). Three kind of designs are typically included into Master Protocols: basket trial, umbrella trial and platform trial. However many combinations of them are possible, mixing their features (e.g. NCI-MATCH trial). Figure 1.2.1 shows the structure of basket, umbrella and platform trials, while their characteristics are briefly summarised below:

- Basket Trials: test a single treatment across various diseases or sub-types, often without a control group, making them suitable for rare conditions [43].
- Umbrella Trials: focus on a single disease with multiple treatments based on biomarker profiles, usually incorporating randomization and a control group [44].
- Platform Trials: assess multiple treatments against a common control over time, allowing for the exclusion of ineffective arms or the addition of new treatments as the trial progresses [45].

1.2.1 Basket, Umbrella and Platform trials

Basket trials

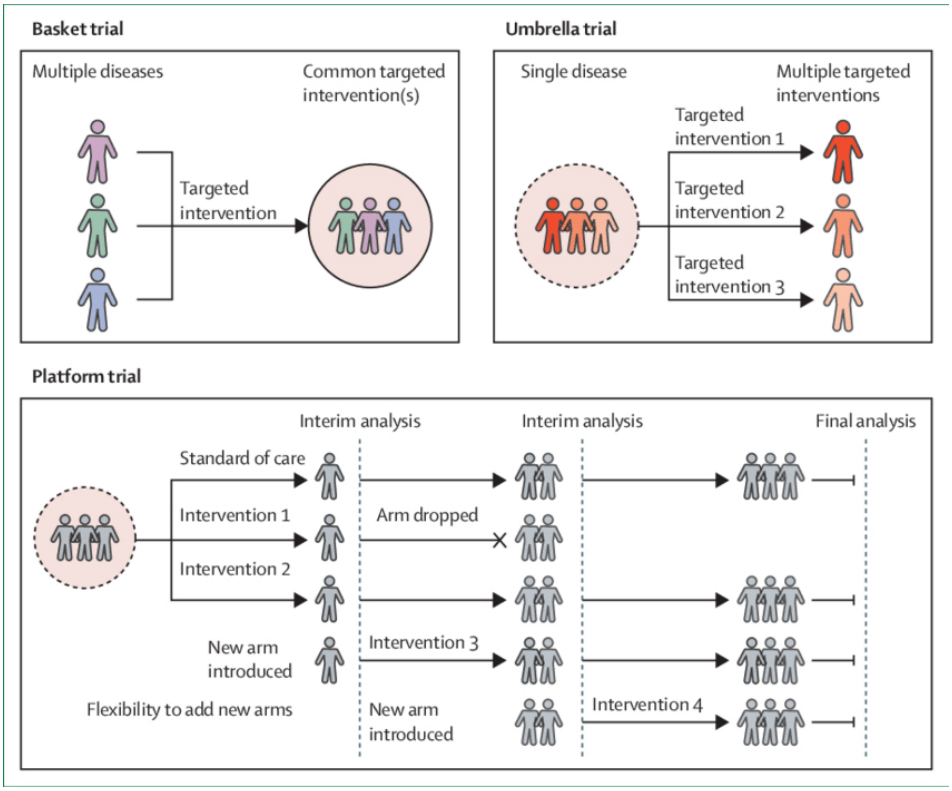


Figure 1.2.1: Master protocols: basket trials, umbrella trials, and platform trials from *Randomised trials at the level of the individual*, by J. Park, 2021, The Lancet Global Health, 9(5), p. e695. [46].

Basket trials represent a cutting-edge approach in clinical research, focusing on testing a single experimental treatment across various patient subgroups defined by genetic mutations, cancer types, or biomarker responses. The primary aim of these trials is to discern the effectiveness of the treatment for specific subgroups, rather than treating all patients as a homogeneous group.

Traditionally, researchers have employed two main strategies to analyze targeted treatments for multiple disease sub-types. The first strategy involves pooling patients together, which overlooks the nuances of individual subgroups. This can lead to misleading conclusions, as the distinct responses of different populations may be masked. The second strategy analyzes each subgroup independently, but this can result in missed opportunities to glean insights from the interactions between groups, potentially limiting the understanding of treatment efficacy. Hence, bas-

ket trials introduced the concept of borrowing information between the treatment arms to take into account all the available information about the treatment under investigation. Consequently, they are particularly advantageous in the study of rare diseases because they allow greater exploitation of the information gathered from individual subgroups. Usually, basket trials are used in the early phase of clinical research and the control group is not considered. In fact, there are different diseases at stake, which often have different standards of care (whether existing). However, the randomization to assign the intervention is also possible.

Umbrella trials

In umbrella trials, the patients with a single disease are assigned to different subgroups on the basis of biomarker screening. Then, they receive a target therapy according to the subgroup for which they are eligible. In this case, it is more common that the interventions and the control group are randomised, given that the disease is the same for all the subgroups as well as the standard of care. Indeed, depending on the disease, an individual may be eligible for various treatments. In any case, an accurate biomarker test is needed to differentiate between the subgroups, and an adequate sample size is required. If the biomarker's prevalence is low, either the basket trial or one treatment arm of umbrella trial may be compromised. Thus, it is very important to design in advance the timing and the strategies of recruitment to reach the planned sample size, especially in case of rare sub-types or mutations. In general, it may be easier to enrol the required number of subjects in a basket than in umbrella trial. In basket trial recruitment is done over several diseases and the sample size calculation takes the entire trial into account, since only one treatment needs validation. Whereas, in the umbrella trial only one diseases is involved and the sample size is treatment arm specific [47]. Master protocols are several time combined with adaptive designs to update the proportion of patients into subgroups [48] according to the interim results, favouring the most promising or limiting and eventually stopping the ineffective subgroups [49]. Besides the ethical benefit, the inclusion of interim analyses to adjust the sample size/proportion or add/drop arms saves money and time for the dissemination of the (interim) results. In the end, despite the fact that the first umbrella trial was completed in 2011 [50], there is no extensive literature from the methodological point of view, as there is for basket trials. In fact, the idea of basket

trials to borrow information within the subgroups required more methodological developments to define the characteristics of the borrowing.

Platform trials

In platform trials the objective is to search the best treatment for a disease analysing multiple experimental groups with respect to a control arm. A platform trial may be active for a long period and the standard of care can be tested several times by interim analysis. Some decision rules must be defined by design to determine whether the treatment arms still continue or drop out. Additionally, a new treatment arm may be added to the trial changing the number of concurrent subgroups each time.. An advantage of platform trials that improves comparability between subgroups is the shared control group for all treatment arms [51]. This allows for a smaller sample size and is particularly favourable for rare diseases. In addition, the use of a common infrastructure to analyse multiple experimental agents under a single protocol saves costs and favours cooperation between multidisciplinary research groups and companies. Also, more options are available for the patients given that the aim is to find the best treatment (or combination of treatments) for each sub-population affected by the same disease. The treatment arms may be heterogeneous among themselves and follow dissimilar eligibility criteria. Consequently, to design a platform trial there are many methodological considerations to keep in mind. For example, the multiplicity of tests to assess the treatment effect of multiple arms at the same time or the timing definition of the interim analysis or the mechanism to add a new treatment arm. These issues and others are discussed in depth in the review of Park *et al.* [52]. However, in recent years, several platform trials have been carried out developing various approaches regarding different methodological aspects, such as eligibility criteria [53], adaptivity [54, 55], the borrowing from historical data [56], non-concurrent control [57, 58] and the general implication of adding arms on ongoing trials [59, 60], because sometimes it might be not advantageous [61]. Some tools were also implemented to facilitate the use of platform trials [62].

In conclusion, in spite of the numerous challenges, master protocols represent a valid solution to conduct trial more efficiently, from different points of view, i.e. cost, quality, time, benefit for the patients and required sample size. In the framework of personalised medicine, with the collaboration of multidisciplinary

investigators, they allow to best-fit the target therapies to patients. In particular, they are advantageous in rare diseases because the single comprehensive infrastructure helps to collect different stakeholders, improving the recruitment, and consequently the time and the feasibility of the study.

2 Basket trials in rare diseases

Basket trials are novel designs that allow to borrow information among multiple arms to evaluate the effectiveness of a new agent in different sub-populations (see Chapter 1, section 1.2.1). The advantage of borrowing information allows for a promising result in one arm to influence the result of all the other arms, and the same holds true for an unpromising result. Thus, statistical power to detect that a treatment is beneficial for a specific subgroup increased in basket trials, and this is particularly important in trials involving rare subgroups.

The growing use of basket trials is well documented in oncology, in particular in phase I/II trials with the purpose of detecting indications that a treatment works on binary outcomes. In this chapter, we will focus on basket trials that evaluate a single treatment for multiple diseases that share the same mechanism of drug activity/efficacy. Our research is motivated by the need to plan a trial which evaluates the efficacy of an innovative therapy in a limited number of similar rare diseases. The challenges here are the rarity of the disease and the use of a continuous endpoint to gather as much evidence as possible to support promising efficacy. Several statistical methods that account have been proposed for potential heterogeneity in basket trials, especially with binary outcomes, but those based on the Bayesian approach are more powerful and flexible.

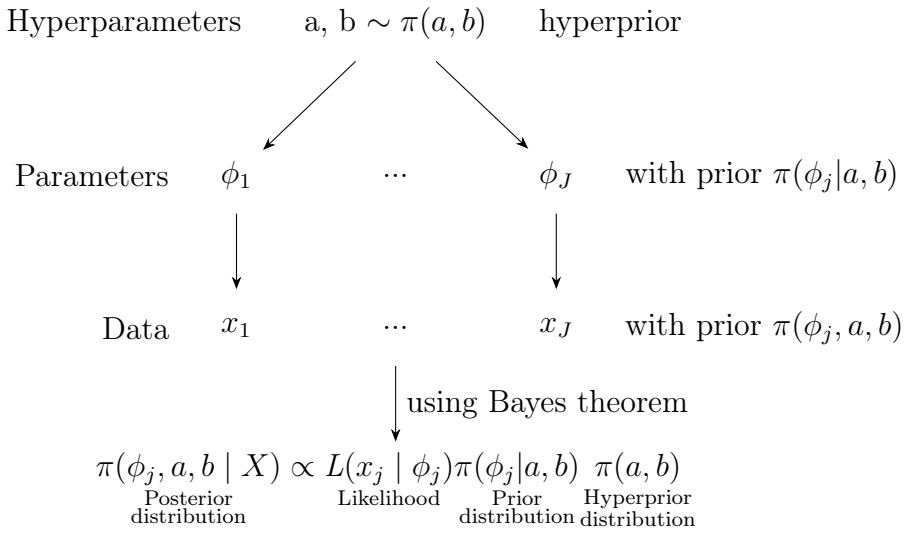
Our aim was to verify the robustness of three standard Bayesian methods for basket trials in a setting of rare diseases, including an extension of the models to handle continuous outcomes. The scope was to give an idea of how accurate the results are when the sample size is very limited. This evaluation can be useful to save costs, time and from an ethical point of view, given the problems of recruitment in rare diseases.

In Section 2.1 we introduce three standard Bayesian methods for the analysis of basket trials (i.e. Bayesian Hierarchical Model, Exchangeability-Noexchangeability model, and Treatment Response Borrowing), while a more technical presentation of these methodologies and their adaptation to the case of a continuous outcome are in Section 2.2. The protocol of the simulation study conducted to assess the modifications we made to the three methods is described in Section 2.3. The simulation results are reported in Section 2.4 and a discussion is provided in Section

2.5.

2.1 Bayesian methods

In 2003, Thall *et al.* [63] developed a method for the analysis of designs similar to what we now refer to as basket trials, even before the term was officially introduced. They implemented a Bayesian hierarchical model (BHM) [64, 65, 66] for a phase II single arm design in the context of multiple subtypes of disease and the effect of treatment among subtypes was assumed to be exchangeable and correlated. The hierarchical model allowed the information to be borrowed between disease subtypes influencing simultaneously the posterior distribution of all the arm specific treatment effect. We briefly summarised the structure of the Bayesian hierarchical model in a general context below. In practice, all outcomes $(\phi_j,\ j = 1 \dots J)$ arise from the same distribution $\pi(\phi_j, a, b)$ that is characterised by the hyperparameters a and b [67]. In brief, there is a group level of sharing given by the common distribution of the random variable of interest (outcomes) and a second level, where the prior parameters of the shared prior distribution arise from other shared hyperprior distributions.



Thall *et al.* considered two examples, one with a binary and one with a time to event outcome and they also explored a more complex outcome given by a combination of clinical response and toxicity. They also introduced a stopping rule

on each subtype disease to determine an early termination when a treatment arm was failing. However, the novelty was about the strength of the information borrowed among the treatment arms. Subsequently, the first basket trial based on a Bayesian hierarchical model was introduced by Berry *et al.* in 2013 [1]. They considered a phase II trial with a binary outcome, representing the tumour's response to the experimental treatment. Five subgroups of different cancers were enrolled and they modelled the log-odds of the outcome as a normal random variable. They also extended the hierarchical model to allow adaptive decisions with an addition of five patients after each interim analysis in all the promising arms. These promising arms were identified through the overcoming of a futility rule concerning the estimated posterior probability of success. Finally through the use of the hierarchical model, the posterior probability at each interim analysis was based on the information shared among the subgroups. Indeed, they emphasised the advantages of their approach over Simon's Optimal Two-Stage [68] design and over a Bayesian adaptive design, outlining the potential of the information borrowing. In particular, they observed a reduction of type I error, an increase of the power, and the reduction of the total sample size due to the inclusion of the interim analysis as compared to the other methods.
However, the main assumption behind the Bayesian hierarchical model is the exchangeability between the subgroups. The definition states that:

$$\begin{gathered}\text{K random variables} \quad \theta_1, \theta_2 \ldots, \theta_K \quad \text{are exchangeable if} \\ f_{\theta_1,\theta_2\ldots,\theta_K}(t_1, t_2 \ldots, t_K) \overset{distr.}{=} f_{\theta_{\pi_1},\theta_{\pi_2}\ldots,\theta_{\pi_K}}(t_1, t_2 \ldots, t_K) \\ \text{for any permutation} \quad (\pi_1, \pi_2, \ldots, \pi_K) \quad \text{of the indices} \quad 1, 2, \ldots, K.\end{gathered}$$

If this assumption is not met, the borrowing can yield bias and inflate the type I error.
In later years, many researchers developed different approaches that relaxed the exchangeability hypothesis (e.g. the multi-source exchangeability model (MEM) [69]). Neuenschwander *et al.* [2] introduced the exchangeability-nonexchangeability method (EXNEX) which was the most famous and easy extension of the Bayesian hierarchical model comprising non-exchangeable subgroups. Each subgroup is assigned a probability reflecting its similarity to the other subgroups, based on a

priori clinical knowledge. The weights assigned to different prior distributions help to regulate the borrowing of information among subgroups that are not exchangeable. This model protects against the inflation of the type I error also allowing for standard borrowing in hierarchical models when the subgroups are fully exchangeable. It enables a more nuanced approach to borrowing information, ensuring that the assumptions of exchangeability are carefully considered. In fact, a team of experts must thoroughly evaluate the prior probability of exchangeability among the subgroups.

Other considerations and extensions to the hierarchical Bayesian model have been made over time, many of which are well summarised in the 2021's methodological review on basket trials by Pohl *et al.* [70]. The authors revised the structures used to share the information between subgroups, dividing in Bayesian and Frequentist technique, the presence of futility and/or efficacy interim analysis and the final decision rule of the overall trial. Particular attention was also paid to the error rate and power estimation, given their importance in the regulatory context, especially in late phases. In fact, even in clinical trials with Bayesian design the final decision takes into account the results on the frequentist concept of type I error and power. However, the majority of applications of basket trials with a Bayesian design are in early phases, given that they are not yet commonly used and accepted in advanced phases by regulatory agencies [71]. The issues related to this limitation are several, starting with the nature of the typical endpoints of phase III studies, which usually are a time to event or a continuous outcomes. Instead, many statistical methods for basket trials refer to binary endpoints for ease of choice of the prior distributions and construction of interim decision rules. Moreover, in phase III clinical trials, the randomised control group is strongly demanded to approve new treatments, but in the basket trials the control group generally is subgroup-specific. Nevertheless, Zheng *et al.* in 2019 [72] and Ouma *et al.* in 2022 [3] proposed two Bayesian statistical models to analyse randomised basket trials. The first one works on the commensurability of the treatment effect in one couple experimental-control subgroup with respect to the others couples. It borrows the information at treatment effect level and it is called Treatment Effect Borrowing (TEB). Whereas, the second one works firstly on the similarity across experimental subgroups and secondly across control subgroups and then it

measures the treatment effect as difference in responses into each experimental-control couple. It borrows at responses per treatment arms level across subgroups and is called Treatment Response borrowing (TRB). In their paper, Ouma *et al.* compared the methods above with a no borrowing approach and considered two settings of sample size (about 105 and 15 per subgroup, respectively) in the simulation study. They showed the advantages of borrowing over a standard approach, especially if there is high homogeneity among the responses, and observed better results for the TRB method if the scenarios involve a small sample size.
As a consequence, to test the performances of basket trials in the context of rare diseases we focused on the BHM (by Berry *et al.*), EXNEX and TRB methods. The primary motivation for conducting a clinical trial in the context of a rare disease is the high expectation about the benefits of the new therapy. So, a typical scenario for rare diseases involves a single arm trial with a continuous outcomes aimed at maximising the potential positive effect of a new treatment. Consequently, we adapted the methods considered to fit these scenarios and we conducted simulation studies to assess their robustness as the sample size decreases.

2.2 Methodological adaptation

We considered a single arm basket trial with few K subgroups $k = 1 \dots K$ and a small number of patients for each subgroup (n_k). A continuous outcome is usually preferred because it is more informative and might be even more informative when the innovative treatment is highly effective, as expected in the motivating clinical context. We adapted the two methods originally proposed for binary outcomes (i.e. BHM and EXNEX) to a continuous outcome and we extended the TRB method, developed for a continuous variable but in a randomised two arms context to a single arm trial.
In general, to assess the results of a single arm basket trial the following system of hypothesis on the benefit of the treatment under investigation is defined:

$$H_0 : \theta_k \leq \delta \quad \text{and} \quad H_1 : \theta_k > \delta \quad k = 1 \dots K \tag{2.1}$$

where θ_k represents the treatment effect in the k-subgroup and δ the threshold to declare the effectiveness of the new therapy. In general, it is possible to assign

different thresholds to the different subgroups in case it is deemed necessary to construct different criteria for the final evaluation. However, in our development, we opted for a common threshold. The final decision rule is guided by the estimated posterior probability that θ_k exceeds δ with high probability, i.e. $\mathbb{P}(\theta_k > \delta) > 95\%$. The 95% cut-off is a standard for the final decision criteria, but levels of 90% or 97.5% are also possible. The decision about this cut-off depends on how conservative one wants to be in assessing that subgroup k satisfies the efficacy criteria, obviously taking into account the opinion of clinical experts in the subject matter.

BHM

The adaptive Bayesian hierarchical method by Berry *et al.* [1] considered a binomial outcome modelled in terms of the log-odds of the probability responses, θ_k (adding an adjustment for a target rate p_1). In details, for each subgroup k the n_k observations y_{ik}, where i identified the subject, are observations from a random variable with a binomial distribution with subgroup-specific probability of success p_k. The log-odds were modelled by a normal distribution sharing the mean μ and the standard deviation σ at subgroups level. Then, on these parameters, μ and σ, is defined the hierarchical structure by the assumptions on their hyperpriors as follows:

$$\begin{aligned} Y_{ik} \sim B(n_k, p_k) \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K \\ \theta_k = \log\left(\frac{p_k}{1-p_k}\right) - \log\left(\frac{p_1}{1-p_1}\right) \\ \theta_k | \mu, \sigma \sim N(\mu, \sigma^2) \quad k = 1, \ldots, K \\ \mu \sim N(-1.34, 10^2), \quad \sigma^2 \sim InvGamma(0.0005, 0.000005) \end{aligned} \tag{2.2}$$

In the Bayesian hierarchical model, the information collected in the subgroups is borrowed according to their degree of similarity in order to construct a common posterior distribution. We underline the importance of σ^2 as it is the parameter indicating how much the responses of the subgroups vary from each other. Whereas, the prior distribution of μ is a non-informative prior due to the large standard deviation. The choice of the hyper-priors' parameters is discussed in [1], in combination with a sensitivity analysis that underlines the robustness of the

hyperpriors to parameters' changes.
To adapt the above model to a single arm basket trial with a continuous outcome, we made the following assumptions:

$$\begin{aligned} Y_{ik} \sim N(\theta_k, s^2), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K \\ \theta_k | \mu, \sigma \sim N(\mu, \sigma^2), \quad s \sim HN(1) \\ \mu \sim N(0, 10^2), \quad \sigma \sim HN(0.5^2) \end{aligned} \tag{2.3}$$

We replaced the inverse gamma distribution of σ with a Half-Normal (HN) prior distribution to be more general and avoid possible issues related to σ closed to zero, i.e. perfect exchangeability [73]. Similarly, we adopted a Half-Normal distribution for the inter patients variance s^2, so we assumed to have the same prior distribution regardless of the group the patients belong to. The Half-Normal distribution with scale parameter equals to 1 or 0.5 covers very large 95% intervals (i.e. 0.03-2.24 and 0.02-1.12, respectively), allowing for a large heterogeneity among the mean subgroups θ_k. This is a weakly informative (hyper-)prior distribution recommended where there is no reliable a priori information for between-subgroups heterogeneity and exchangeability.

EXNEX

Neuenschwander *et al.* in 2016 extended the BHM adding an a priori weight p_k on the exchangeability assumption between the subgroups. The log-odds of the responses follow a hierarchical model as defined in the model 2.2 with a prior probability p_k. Then, with a probability $1 - p_k$ they follow a subgroup-specific normal distributed prior where each mean and variance (m_k and v_k^2, respectively)

are pre-specified according to prior knowledge.

$$
\begin{gathered}
Y_{ik} \sim B(n_k, p_k) \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K \\
\theta_k = \log\left(\frac{p_k}{1-p_k}\right) - \log\left(\frac{p_1}{1-p_1}\right) \\
\text{EX with } p_k : \theta_k | \mu, \sigma \sim N(\mu, \sigma^2) \\
\mu \sim N(-1.73, 2.616^2) \text{ and } \sigma \sim HN(1) \\
\text{NEX with } 1 - p_k : \theta_k \sim N(m_k, v_k^2) \\
m_k \text{ and } v_k \text{ fixed a priori}
\end{gathered}
$$

Also in this case, for the choices concerning the parameters we refer to the original paper [2], however no extreme weights' values are suggested by the authors when poor a priori information is available.
Now, to adapt the EXNEX model to a continuous outcome we followed the idea of Neuenschwander *et al.* and we added the prior weights and the no-exchangeable part to the BHM model 2.3. For what concern the choice of the prior distributions and their parameters we kept the choices made on the BHM model so that they were comparable. In addition, we set a very low informative prior for the noexchangeability prior distribution of all the subgroups. However, different choices are possible according to the information available for each subgroup. Moreover, for ease of interpretation and following what the authors did in their *nugget* scenario in [2], we studied the model by setting the prior weights equal for each subgroup.

$$
\begin{gathered}
Y_{ik} \sim N(\theta_k, s^2), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K \\
\text{EX with } p_k : \quad \theta_k | \mu, \sigma \sim N(\mu, \sigma^2), \quad s \sim HN(1) \\
\mu \sim N(0, 10^2), \quad \sigma \sim HN(0.5^2) \\
\text{NEX with } 1 - p_k : \quad \theta_k \sim N(m_k, v_k^2), \quad m_k = 0, \quad v_k = 10 \quad \forall k
\end{gathered}
$$

Despite the choices made on the parameters of the subgroup-specific distributions and the use of prior weights, we would like to emphasise that such a model, with equal parameters and prior weights across the subgroups, implicitly means a strong exchangeability of θ_ks, since *"the joint distribution of strata parameters is permutation invariant"* [2]. Therefore, when using the model in real case studies,

we recommend to pay attention to the choice of weights and parameters of the robust distribution in order to make the most of the concept of exchangeability-noexchangeability.

TRB

Finally, we extended the TRB method to the case of single arm trial, and we concentrate only on the experimental subgroups. The main idea of the authors was to consider the mean response of the control (C) and the experimental (E) groups separately and to borrow the information intra treatment groups. So, they defined θ_{jk} as the mean response of the subgroup k for treatment j, (j= C, E) and then by a commensurate normal predictive prior (CPP) [16, 35] they linked the mean response of current interest θ_{jk^*} with the other mean responses θ_{jk} with $k \neq k^*$:

$$\theta_{jk^*}|\theta_{jk}, \nu^{(j)}_{kk^*} \sim N\left(\theta_{jk}, \frac{1}{(\nu^{(\mathrm{j})}_{kk^*})^2}\right), \quad j = E, C \quad k = 1, \ldots, K \tag{2.4}$$

where $\nu^{(j)}_{kk^*}$ parameterises the consistency between the responses θ_{jk} and θ_{jk^*} in either the experimental and control groups under investigation. Specifically, their commensurability is modelled by a spike and slab prior [74] and the a priori weight that allocates probability to the slab or the spike is defined via the Hellinger distance [75] between the mean responses posterior operational distributions (the posterior distributions of two subgroups under investigation). Hence, as the distance is small, the commensurability is high and there is strong borrowing, whereas if the distance is big there is not borrowing. After each pairwise evaluation, all the discrepancy measurements were collected to reach a single prior for each θ_{jk^*}. A decreasing function based on the discrepancies within subgroups was applied to balance the commensurability and the smallest weight was assigned to a subgroup more distant [76, 72]. For more details, see [3].

Accordingly, we implemented the model fixing $j = E$ and borrowing information on the continuous endpoint on the basis of the similarity between the experimental responses. At the end, we compared the results with the null hypothesis (2.1). In addition, to be comparable with the results on BHM and EXNEX methods we

modelled the inter subjects variance by a Half-Normal distribution as described in model (2.3) to replace the inverse gamma distribution chosen by the authors. Thus, similarly to the EXNEX method the information is not borrowed a priori (under strong biological and clinical evidence of course), but according to the commensurability between the subgroups. This allows to limit the sharing of information when the subgroups responses are different from each other given that the same treatment might act in different ways on different diseases.

2.3 Simulation study on small samples

We implemented a simulation study to assess the performance of the three methods considering three subgroups (K=3) and the following sample size n_k in each subgroup: 4, 7, 15, 50 (benchmark). We assumed a hypothetical reference value for the treatment effect equal to 1 ($\delta = 1$) and defined six different combinations of true responses (θ_k), as reported in Table 2.3.1:

Table 2.3.1: Simulations' scenarios on treatment effects

	Scenario							
True effect	1	2	3	4	5	6	7	8
θ_1	3	1.2	0.5	0	3	3	1.2	1
θ_2	3	1.2	0.5	0	0.5	1.2	0.5	1
θ_3	3	1.2	0.5	0	3	3	3	1

The aim was to investigate what happens in scenarios with different amount of heterogeneity/homogeneity and for different intensity of treatment effects. Specifically, we assessed scenarios where the treatment effect is very high ($\theta_k = 3\ \forall k$, Scenario 1), low ($\theta_k = 1.2\ \forall k$, Scenario 2) or there is no effect ($\theta_k = 0.5$ or $0\ \forall k$, Scenarios 3 and 4, respectively) and a mixture of the previous ones (Scenarios 5, 6 and 7) to have heterogeneous responses within subgroups. Lastly, we implemented the Scenario 8, where the true responses are equal to the reference value to investigate how the probability of success works in a very tricky, but not effective, scenario.

For simplicity, we set the prior weights of the EXNEX model equal to 0.5 for the

EX and NEX part, respectively. Subsequently, to investigate the impact of the prior weights, especially in the heterogeneous scenarios, we assumed $p_k = 0.8$ or 0.2 and, consequently, $1 - p_k = 0.2$ and 0.8 for all the involved subgroups.
We simulated M=10000 replicates ($m = 1 \dots M$) of each scenario using two parallel chains with 13000 MCMC iterations and 3000 burn-in (i.e. number of iterations to discard at the beginning). The convergency of the MCMC method were assessed by the Gelman-Rubin statistics [77] and the simulations' results were evaluated in terms of bias and MSE:

$$\text{Bias}(\theta_k) \approx \frac{1}{M} \sum_{m=1}^{M} (\bar{\theta}_k^m - \theta_k)$$

$$\text{MSE}(\theta_k) \approx \frac{1}{M} \sum_{m=1}^{M} (\bar{\theta}_k^m - \theta_k)^2$$

where $\bar{\theta}_k^m$ is the posterior mean of θ_k on the iterations for the mth simulated trial. Moreover, we reported the median of the posterior probability (PP) that θ_k exceeds δ across all the replicates and in a frequentist perspective, the marginal power and type I error. The power and the type I error were defined as the proportions of simulated trials with PP that guides to a correct/incorrect final decision in the subgroups under the alternative/null hypothesis (i.e. $\theta_k > 1$ or $\theta_k \leq 1$, respectively).
The simulation study was performed using the open-source R software v.4.3.2 (R Foundation for Statistical Computing, Vienna, Austria) and the R2jags package. All the replicates were parallelised on 89 cores to reduce the computational time.
In recent years a lot of software tools have been developed to analyse basket trials of various types. In the R environment the *bhmbasket* package was developed to apply the BHM and the EXNEX methods, while the packages *bmabasket* and *basket* implement the Psioda *et al.* [78], the MEM [69] and Kaizer *et al.* [79] methods, respectively. There is also a web-based integrated platform Trialdesign that collects some very useful tools for the design and analysis of bayesian clinical trials. Finally, we mention the FACTS tool, developed by Berry consultants to implement adaptive designs (and not only) including basket trials and the East Bayes software by Cytel, Inc. (Cambridge, MA, USA). However, we want to

emphasise that the simulation study was implemented with R scripts developed ad-hoc, without making use of existing R packages (see Appendix 6).

2.4 Results

The results obtained from the three designs that we adapted to the case of a single arm basket trial with a continuous endpoint are reported for each sample size and scenario. We started considering the situation of an homogeneous high treatment effect ($\theta_k = 3\ \forall k$, Figure 2.4.1), and we found that, with a sample size of 50, the bias in each subgroup is nearly zero across all approaches, and the mean squared error (MSE) is minimal, particularly for TRB. When sample sizes are limited, the bias remains negligible and comparable, whether there are 4, 7 or 15 patients per subgroup. However, the MSE varies with sample size, albeit remaining at low level, with TRB exhibiting smaller values, while BHM and EXNEX show similar MSE levels.

Similar results were obtained for scenarios characterised by an homogeneous treatment effect within subgroups, but less distant to the reference value $\delta = 1$ ($\theta = 1.2,\ 0.5,\ 0$ and 1) (Figures 2.4.2-2.4.5 for scenarios 2-4 and 8, respectively).

In Scenario 5 and 6 we added increasing levels of heterogeneity considering two subgroups highly effective and one subgroup less effective (Scenario 6) or with an effect with different direction (Scenario 5). In these situations, the BHM and EXNEX approaches showed optimal behaviours both in terms of bias and MSE, while TRB sistematically overestimated the treatment effect in subgroup 2, that affected also the level of MSE. Very little differences were reported for different sample sizes (see Figures 2.4.6 and 2.4.7).

The highest degree of heterogeneity was considered in Scenario 7, characterised by three levels of efficacy, reflecting a weak difference with respect to the reference value (subgroup 1) and a lower/higher effect (subgroups 2 and 3, respectively). In this scenario the subgroups with the most diverging effects result biased, regardless of the sample size, only when analysed by the TRB method. The bias/MSE here are lower than those obtained in Scenarios 5 and 6, but still high.

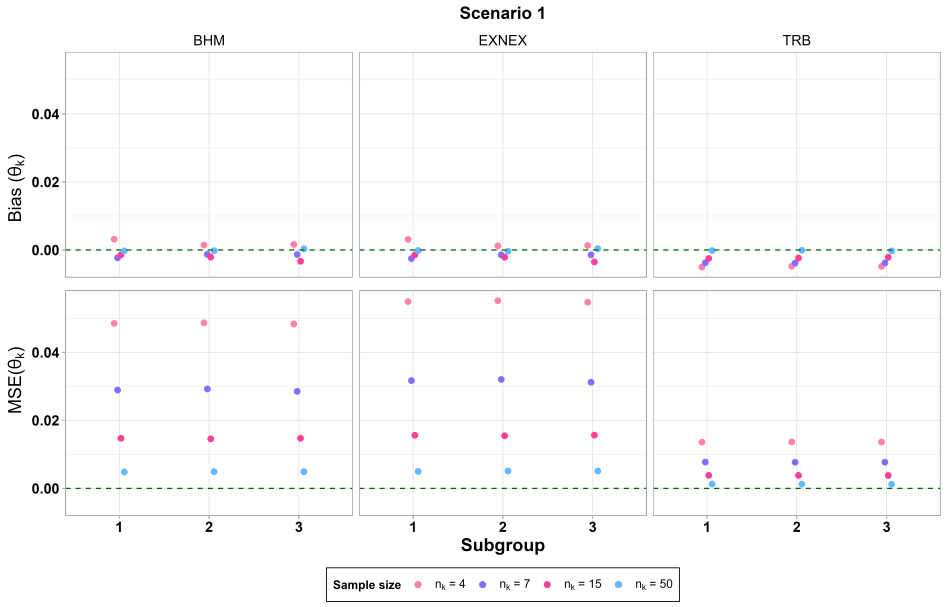


Figure 2.4.1: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 1 ($\theta_k = 3\ \forall k$).

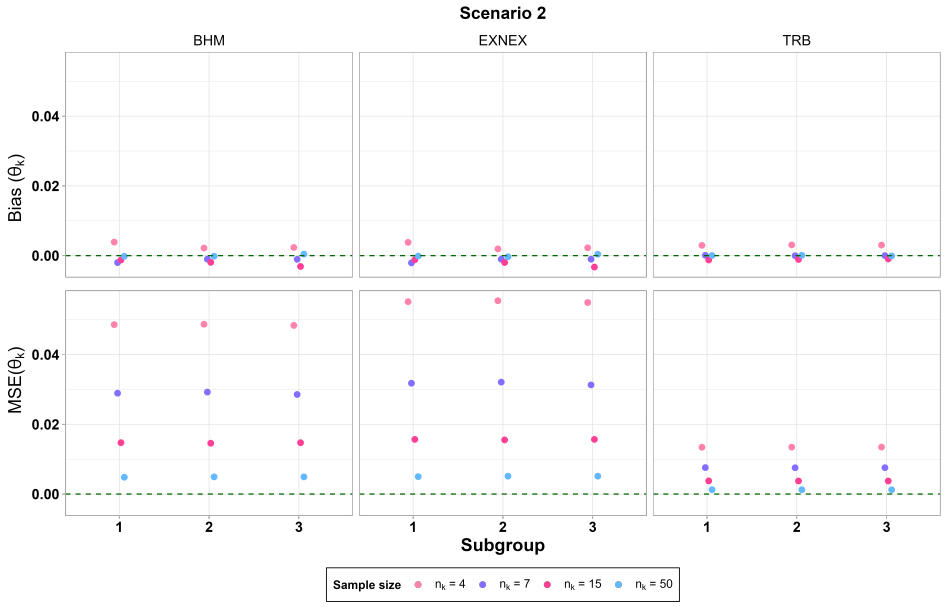


Figure 2.4.2: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 2 ($\theta_k = 1.2\ \forall k$).

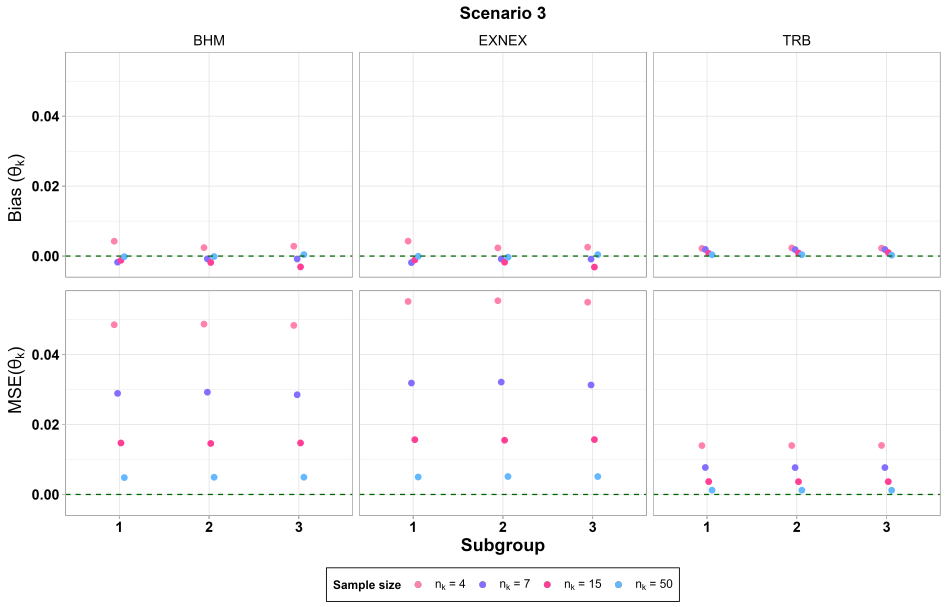


Figure 2.4.3: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 3 ($\theta_k = 0.5\ \forall k$).

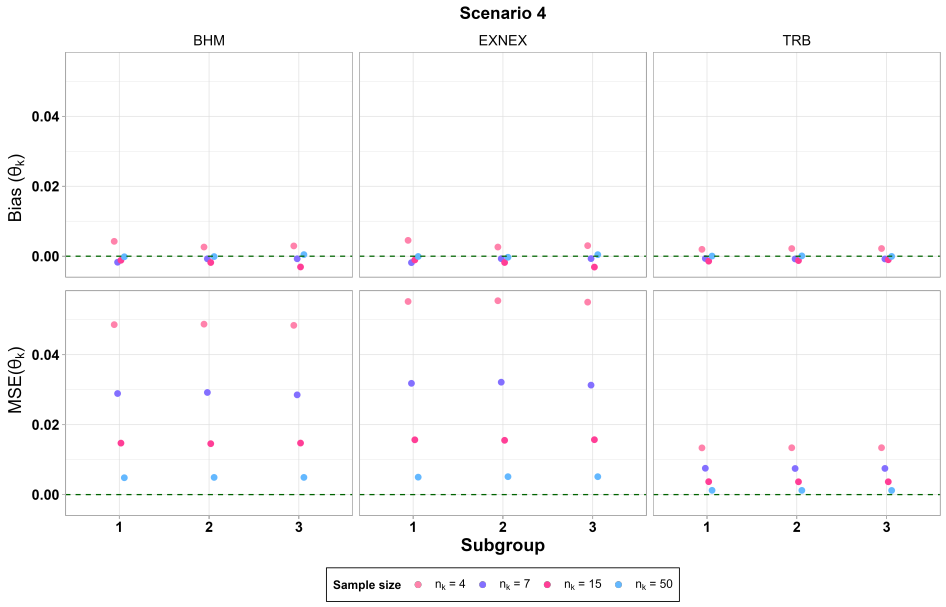


Figure 2.4.4: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 4 ($\theta_k = 0\ \forall k$).

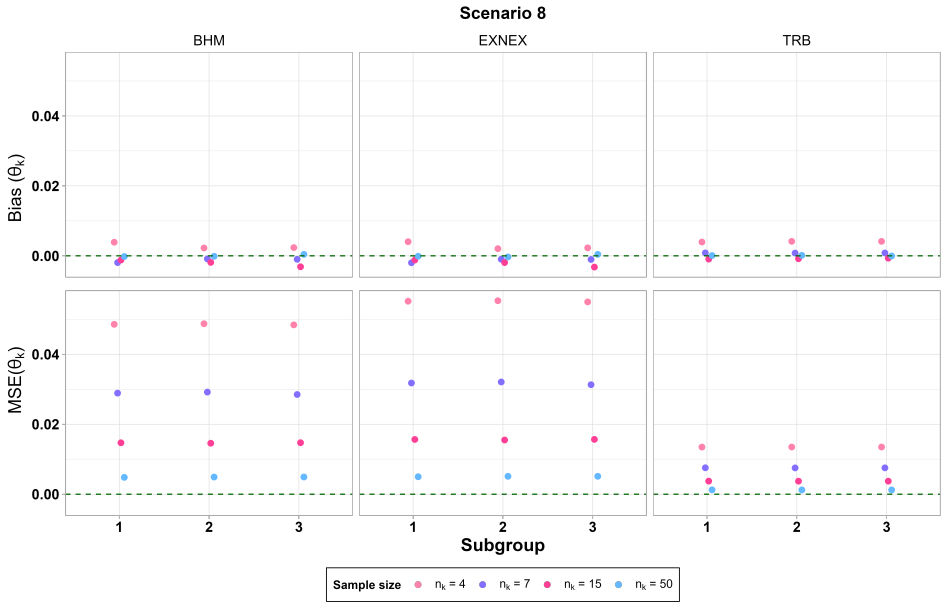


Figure 2.4.5: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 8 ($\theta_k = 1\ \forall k$).

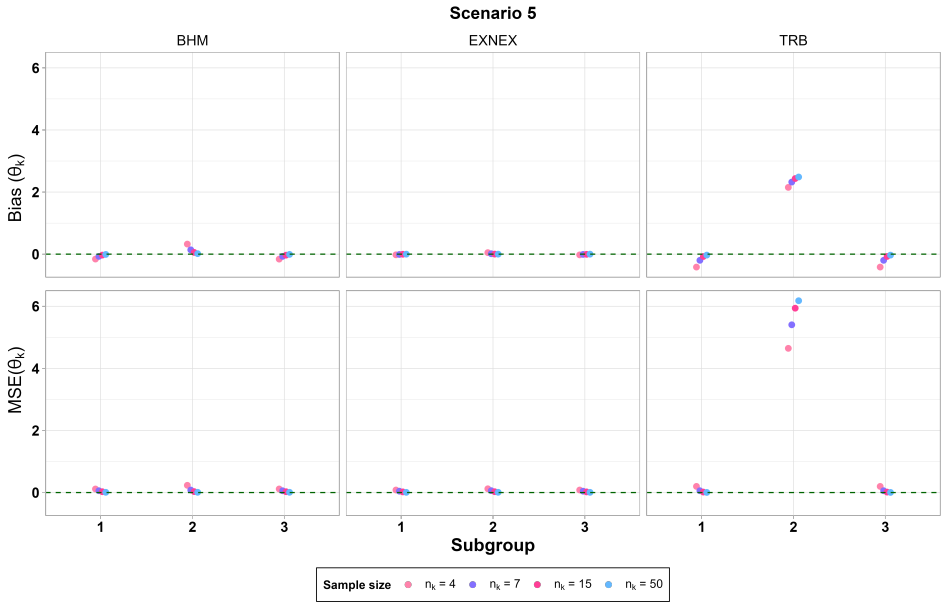


Figure 2.4.6: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 5 ($\theta_{1,3} = 3$ and $\theta_2 = 0.5$).

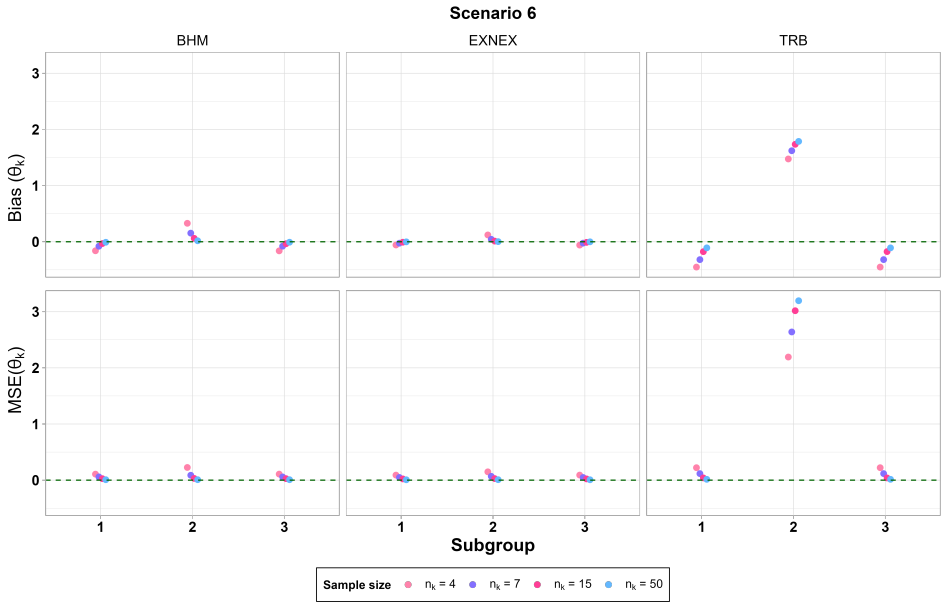


Figure 2.4.7: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$).

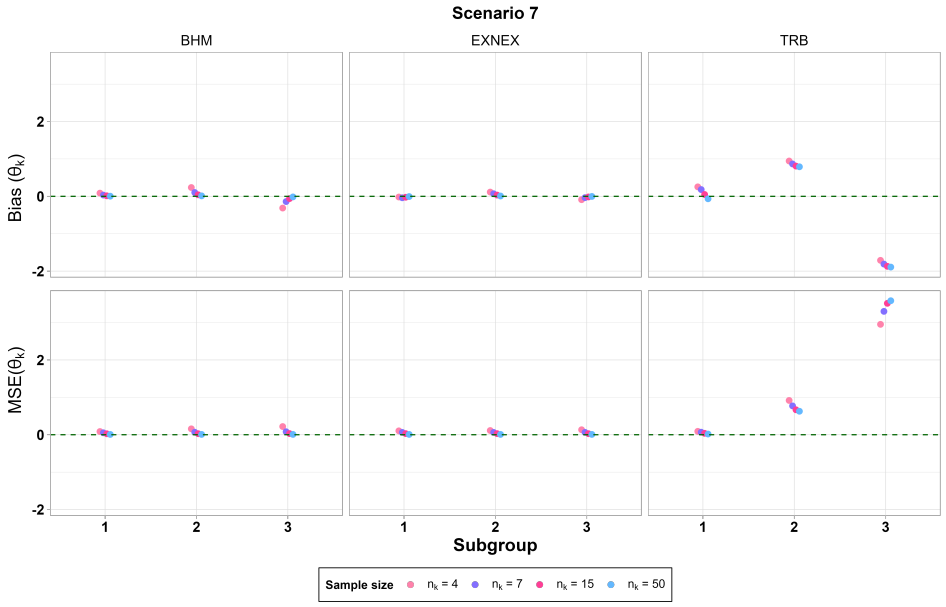


Figure 2.4.8: Bias and MSE of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 7 ($\theta_1 = 1.2$, $\theta_2 = 0.5$ and $\theta_3 = 3$).

In Table 2.4.1 we show the median posterior probability (PP) that the treatment effect exceeds the δ threshold in Scenarios 2, 5, 6 and 7. The remaining results are shown in the Supplementary material 5.2, as they do not require any particular comment.

Table 2.4.1: Median posterior probability of the BHM, EXNEX and TRB methods according to different sample sizes (7, 15 and 50) in Scenarios 2 ($\theta_k = 1.2\ \forall k$), 5 ($\theta_{1,3} = 3$ and $\theta_2 = 0.5$), 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$) and 7 ($\theta_1 = 1.2$, $\theta_2 = 0.5$ and $\theta_3 = 3$). The median PPs that exceed the 95% cut-off are reported in green.

Sample	Method	Subgroup	PP_2 (%)	PP_5 (%)	PP_6 (%)	PP_7 (%)
4	BHM	1	79	100	100	81
4	BHM	2	79	28	92	23
4	BHM	3	79	100	100	100
4	EXNEX	1	79	100	100	68
4	EXNEX	2	78	9	80	13
4	EXNEX	3	78	100	100	100
4	TRB	1	78	100	100	81
4	TRB	2	78	97	99	83
4	TRB	3	78	100	100	77
7	BHM	1	85	100	100	84
7	BHM	2	85	9	92	6
7	BHM	3	85	100	100	100
7	EXNEX	1	84	100	100	73
7	EXNEX	2	84	3	83	5
7	EXNEX	3	85	100	100	100
7	TRB	1	87	100	100	77
7	TRB	2	87	95	99	84
7	TRB	3	87	100	100	74
15	BHM	1	93	100	100	90
15	BHM	2	93	0	94	0
15	BHM	3	93	100	100	100
15	EXNEX	1	92	100	100	84
15	EXNEX	2	92	0	90	0

Continued on next page

Table 2.4.1 – continued from previous page

Sample	Method	Subgroup	PP_2 (%)	PP_5 (%)	PP_6 (%)	PP_7 (%)
15	EXNEX	3	92	100	100	100
15	TRB	1	97	100	100	71
15	TRB	2	97	93	98	91
15	TRB	3	97	100	100	71
50	BHM	1	99	100	100	99
50	BHM	2	99	0	99	0
50	BHM	3	99	100	100	100
50	EXNEX	1	99	100	100	98
50	EXNEX	2	99	0	99	0
50	EXNEX	3	99	100	100	100
50	TRB	1	100	100	100	75
50	TRB	2	100	93	98	98
50	TRB	3	100	100	100	76

In presence of a limited homogeneous effect (Scenario 2), the values of PP increase as the sample size increases and those corresponding to $n = 7$ are high, even if still below the 95% cut-off of success.

In Scenario 5, the presence of heterogeneity of the effect between subgroups (two subgroups highly effective and one ineffective) does not influence the median posterior probability of success of the promising subgroups, whereas it impacts on the ineffective one. Indeed, the BHM and EXNEX methods coherently show poor PP values when the sample size is very low or low (n=4 and 7, respectively). A different behaviour is shown for the TRB approach which has PP levels around 95%, regardless of the sample size for the subgroup with $\theta = 0.5 < \delta$. The 97% of PP in subgroup 2 with 4 subjects is a clear example of PP's overestimation given the very small number of patients per subgroup.

When there is heterogeneity, but the three subgroups are pointing in the direction of an effective treatment (i.e. Scenario 6), the presence of two very effective subgroups impacts on the subgroup with mild effect. The increase of the probability of success is well evident even when the sample size is very small. Indeed, there is

a consistent gain as compared to the performance of Scenario 2, especially for the BHM and the TRB approach. For example when $n = 7$ there was an improvement from 85% to 92% for BHM and from 87% to 99% for TRB. The TRB in subgroup 2 has always PP values greater than the 95% cut-off so we can declare the success of the subgroup, while both BHM and EXNEX require larger sample size.
In the complex Scenario 7, which is characterised by very different treatment effects, both in intensity and direction, the BHM and EXNEX approaches always resulted in a PP of 100% in the very effective subgroup, whereas the subgroup with a mild treatment effect exceeds the decision cut-off only when 50 patients are involved. Finally, the non effective subgroup has always very low PP, ranging from 0% to 6% when the sample size is 7 or 15 and between 13% and 23% when the sample size is 4. In contrast, the TRB method concludes for the benefit of the ineffective subgroup reaching the cut-off of efficacy with the largest sample size. In the remaining subgroups, the PPs are very similar, but lower than obtained with TRB, and they seem to be not affected by the sample size.
The three methods were also compared in terms of power, in a frequentist perspective. We measured the power in Scenarios 1, 2, 6 and in subgroups 1 and 3 of Scenarios 5 and 7, which were under the alternative hypothesis that $\theta_k > \delta$. In all the subgroups where we assumed a very effective benefit given by the new therapy the power was 100% (see Figure 2.4.9, 2.4.11, 2.4.12 and 2.4.13, except Scenario 7 subgroup 3). This is true with all the methods and regardless of the number of subjects in the subgroup. The behaviour of BHM and EXNEX is in general very similar, with the power always influenced by the sample size. Moreover, in the scenarios presenting a positive, but heterogeneous effect, there is a borrowing effect also in terms of power that increases for the BHM approach more than for EXNEX (see Figure 2.4.12). The presence of an ineffective subgroup beside two highly effective subgroups has no influence at all in Scenario 5 (see Figure 2.4.11).

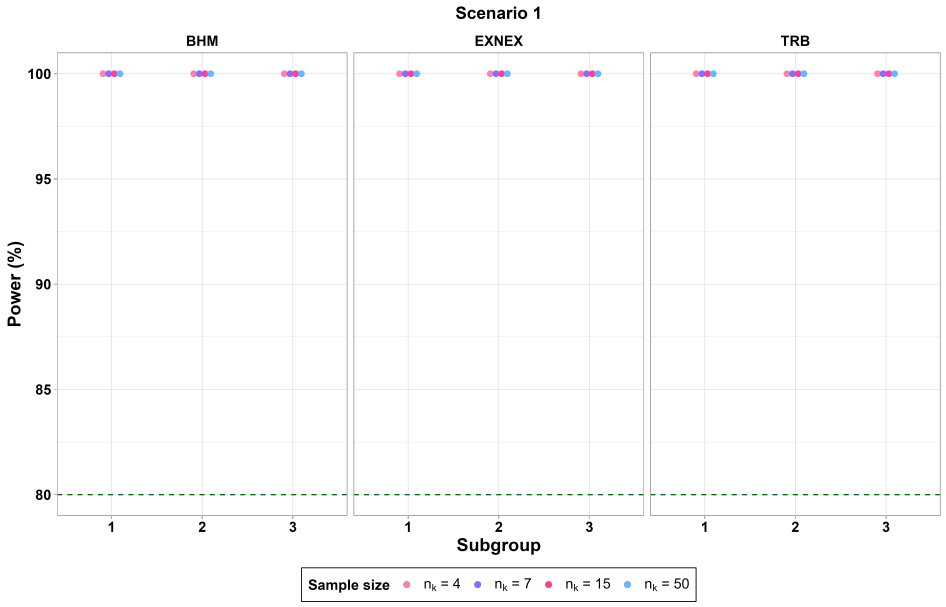


Figure 2.4.9: Power in the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 1 ($\theta_k = 3 \quad \forall k$). The green dashed line indicates 80% power.

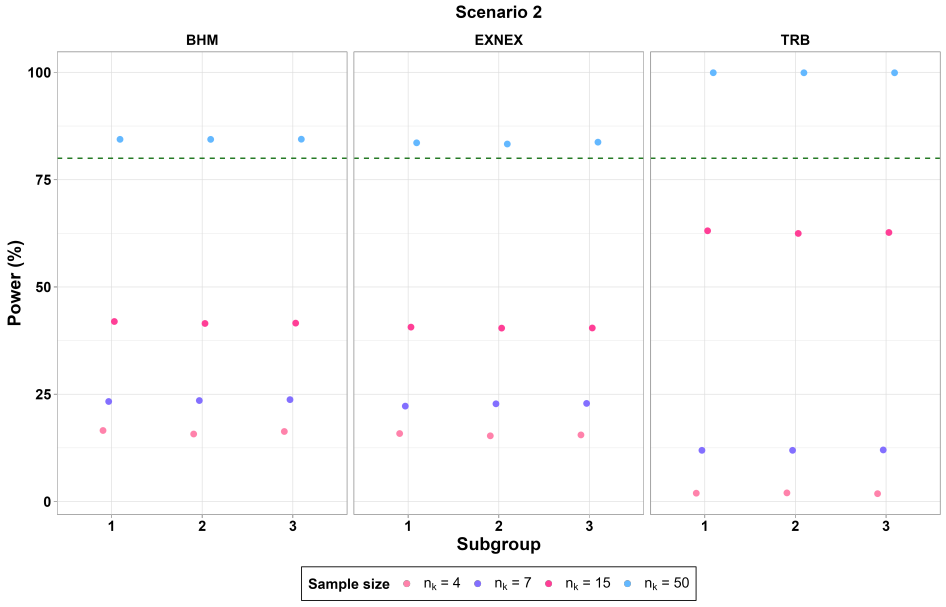


Figure 2.4.10: Power in the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 2 ($\theta_k = 1.2 \quad \forall k$). The green dashed line indicates 80% power.

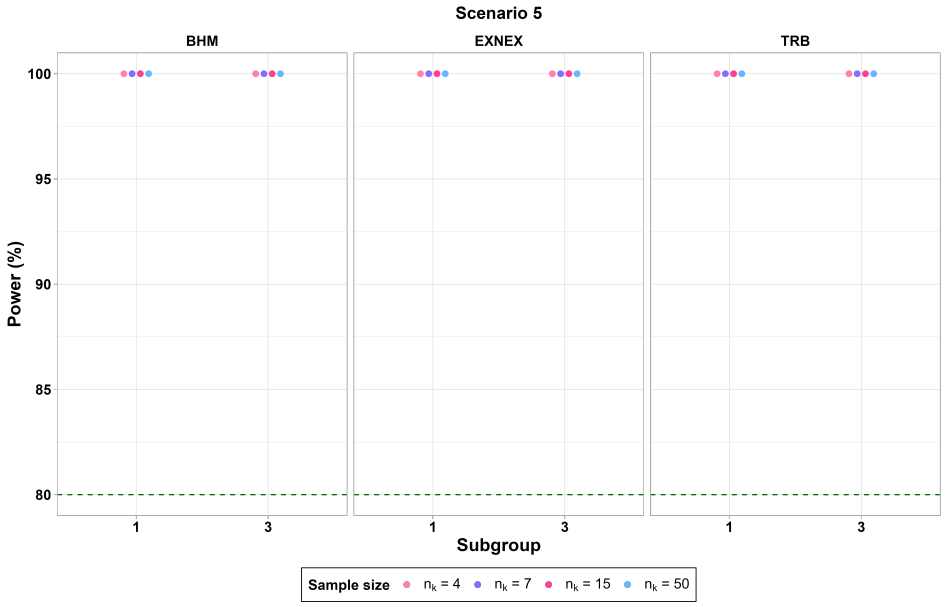


Figure 2.4.11: Power in the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 5 ($\theta_{1,3} = 3$). The green dashed line indicates 80% power.

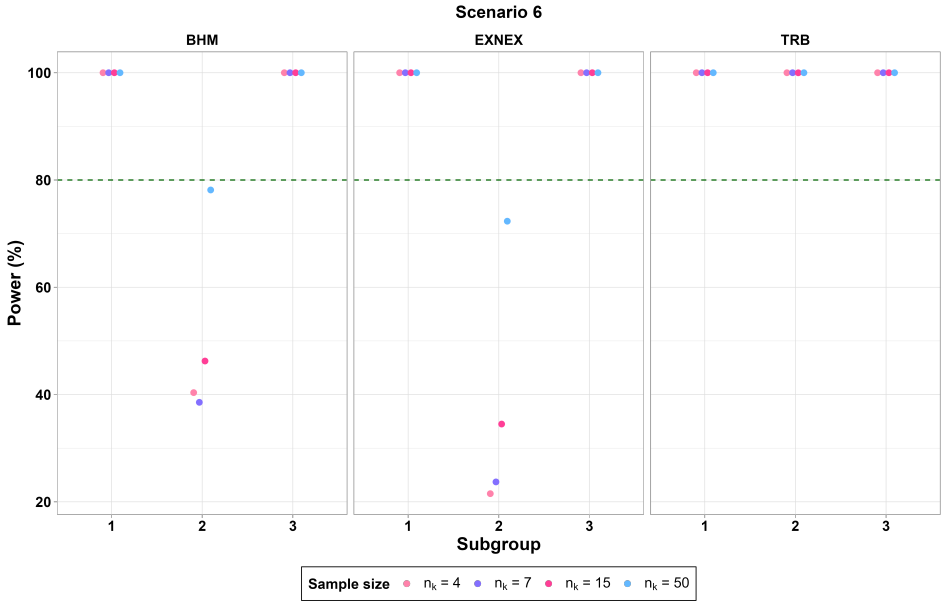


Figure 2.4.12: Power in the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$). The green dashed line indicates 80% power.

Particularly interesting are the results in Figure 2.4.13 from the seventh scenario. The comparison between BHM and EXNEX across sample sizes shows no relevant differences in terms of power, even if it is slightly lower with the EXNEX method. In contrast, the TRB method concludes with a power very close to zero in subgroups 1 and 3 ($\theta_k = 1.2$ and 3, respectively) in spite of the sample size considered. Hence, the high heterogeneity in treatment effect values does not impact on the power of the very effective subgroup in BHM and EXNEX methods, but impacts a lot on the TRB approach.

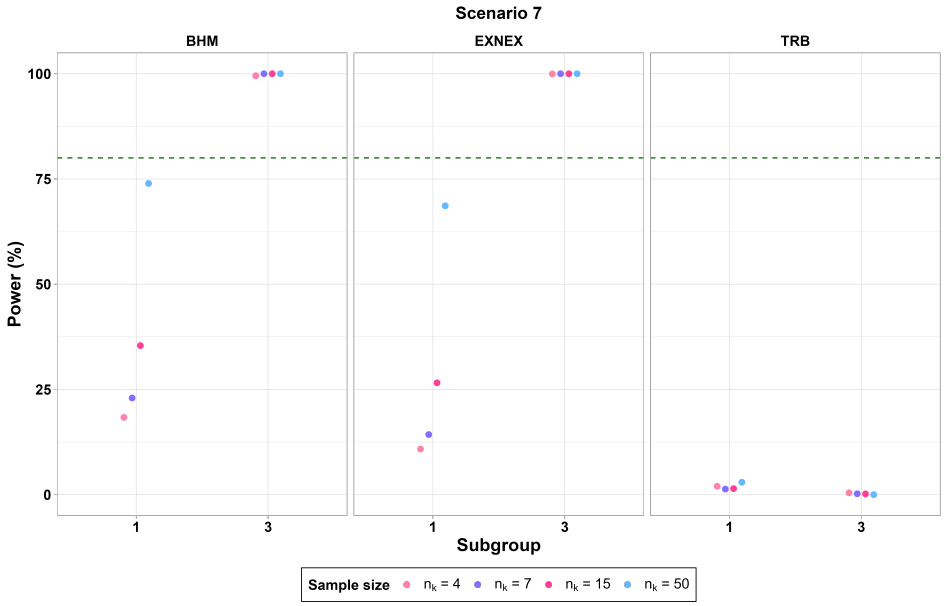


Figure 2.4.13: Power in the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenario 7 ($\theta_1 = 1.2$ and $\theta_3 = 3$). The green dashed line indicates 80% power.

We also measured the probability of making a type I error (i.e. the proportion of simulations erroneously rejected under $\theta_k = 1$) in Scenario 8, which reflects a situation with no effect ($\theta_k = 1\ \forall k$) in all three subgroups. The type I error rate remains consistently below 5% across all methods and sample sizes, as shown in Table 2.4.2. Notably, the TRB method demonstrates the best control over the probability of rejecting the null hypothesis when it is true.

Table 2.4.2: Type I error in the BHM, EXNEX and TRB methods according to different sample sizes (7, 15 and 50) in Scenario 8 ($\theta_k = 1 \forall k$).

Sample	Method	Subgroup	Type I error (%)
4	BHM	1	2.96
4	BHM	2	3.20
4	BHM	3	3.30
4	EXNEX	1	3.10
4	EXNEX	2	3.22
4	EXNEX	3	3.33
4	TRB	1	0.04
4	TRB	2	0.04
4	TRB	3	0.04
7	BHM	1	2.71
7	BHM	2	2.91
7	BHM	3	2.70
7	EXNEX	1	2.84
7	EXNEX	2	3.07
7	EXNEX	3	2.84
7	TRB	1	0.00
7	TRB	2	0.02
7	TRB	3	0.02
15	BHM	1	2.93
15	BHM	2	2.92
15	BHM	3	2.82
15	EXNEX	1	3.05
15	EXNEX	2	3.20
15	EXNEX	3	2.91
15	TRB	1	0.24
15	TRB	2	0.09
15	TRB	3	0.18
50	BHM	1	3.24

Continued on next page

Table 2.4.2 – continued from previous page

Sample	Method	Subgroup	Type I error (%)
50	BHM	2	3.04
50	BHM	3	3.79
50	EXNEX	1	3.27
50	EXNEX	2	3.27
50	EXNEX	3	3.73
50	TRB	1	1.00
50	TRB	2	0.89
50	TRB	3	1.00

Finally, to explore the impact of the prior weights in the EXNEX approach, we implemented a simulation study considering the same scenarios, but with different levels of p_k ($p_k = 0.8$ and 0.2). The idea was to better guide the final decision in the heterogeneous scenarios investigating the results when we are confident of more exchangeable or not subgroups. The results are reported in Supplementary materials 5.2.1 and are similar to those ones we obtained with $p_k = 0.5$, except in Scenarios 5-7, where a lower prior weights on the exchangeability assumption ($p_k = 0.2$) decreases the bias and the MSE. On the other hand, a strong assumption of no-exchangeable subgroups in the homogeneous ones does not substantially affect the measurement error.

2.5 Discussion

It is well known that the basket trials, which borrow information among subgroups, provide more advantages with respect to standard designs, especially when the treatment effects within subgroups are similar [1, 2, 3, 70]. Consequently, they might also represent a valid solution in the case of rare diseases where the need to borrow information is dictated by the limited sample available.
With this in mind we investigated the robustness of small sample sizes of three well-known Bayesian methods for the analysis of basket trials which we adapted to face a continuous outcome, using 50 subjects per subgroup as benchmark. The

results of our simulation study, which considered various scenarios of treatment effects, suggest that basket trials are also feasible in the context of rare diseases, when the evidence on the treatment effect is promising. However, when the effect of the treatment is mild or there is heterogeneity in the subgroup treatment effects the choice of the sample size and the final cut-off of trial success should be carefully considered in the planning phase of the trial (see Table 2.4.1). In particular, in the presence of homogeneous mild effect the results are promising and severely affected by the sample sizes. Moreover, in scenarios where the mild treatment effect is combined with very effective treatment the posterior probabilities increased as a result of the borrowing of information, regardless of the methods considered. In contrast, when there is high heterogeneity and the treatment effects are not uniform in the direction (e.g. Scenario 7), the TRB method's borrowing of information affects the accuracy of the results, leading to wrong conclusions. On the other hand, the BHM and the EXNEX are able to borrow information more accurately, even in the presence of heterogeneous treatment effects. These methods accurately identify the success of a highly effective subgroup with 100% power using only a few subjects (see Figure 2.4.13). Moreover, with a sample size of 50, the subgroup exhibiting a mild effect achieves a posterior probability exceeding 95%, even if its power remains below 80%.

As final remark, the three methods are also able to handle situations in which there is no evidence of treatment effect, with a strict control of type I error. Concerning the three Bayesian approaches, we observed that the TRB method is accurate only in the presence of homogeneity across the subgroups. Our results are consistent with those reported by Ouma *et al.* [3], highlighting the importance of a strong biological or clinical motivation to use a borrowing approach. The use of the Hellinger's distance for the assessment of how much borrowing can be done is a very successful method when the treatment effects go in the same direction, as it allows the evidence of treatment effect to be strengthened in the subgroups with milder effects, even if the sample size is small. In fact, with this distance approach, the degree of borrowing is driven by the data and not assumed a priori, as happens with other approaches.

However, it is not always easy to know in advance how great the effect of the treatment will be. Therefore, it is essential to choose methods carefully and ad-

here closely to regulatory guidelines. Conducting simulation studies to explore potential scenarios of interest and clearly articulating the operational characteristics is crucial. While preclinical and natural history studies can provide valuable insights [15], careful attention to the study design and compliance with regulatory guidelines remain of paramount importance.

We are aware that these results are far to be exhaustive and additional work should be planned. Indeed, the simulation study should be extended to explore whether these results are sensitive to unbalanced subgroup sizes, different number of subgroups and different prior weights. We have chosen prior distributions widely used in the literature for similar situations and in particular, uninformative distributions to leave to the data the determination of the degree of borrowing among subgroups. However, working with small sample sizes, we are also going to investigate how the results are sensitive to changes and prior distributions in parameters. For the moment, the set-up of the current simulation study is in line with the clinical context that motivated this work, where an innovative highly effective treatment (based on pre-clinical data) will be investigated in three rare subgroups of the same class of disease.

In conclusion, we can safely use a basket trial design based on a Bayesian model on a continuous outcome even when the subgroup sample size is small if a clear treatment effect is present. Indeed, the design seems to preserve the determination of trial/subgroups success, the power and the type I error.

3 Innovative designs in rare diseases

The progressive advancement of precision medicine presents increasing opportunities for developing clinical trials focused on rare diseases. Consequently, study designs are evolving toward greater flexibility and adaptability to the diverse scenarios that can arise. In particular, as previously mentioned, adaptive designs and information borrowing are helping to address some limitations of standard study designs and are particularly useful in the context of rare diseases. We have integrated these elements to create a new study design tailored to the context of rare diseases, incorporating innovative aspects that address the needs of an upcoming clinical trial. This trial will assess an innovative treatment strategy that should offer strong effect to all putative clinical indications, with expected heterogeneity only in the intensity when evaluated in different diseases of the same class. Acknowledging the potential for differential effectiveness among the enrolled subpopulations by design, we assume that heterogeneity exists as an intrinsic help in evaluating treatment effects. Given these premises and the results of Chapter 2, demonstrating the utility of a basket design in streamlining drug development in rare diseases, we considered the framework of basket trial for the development of a new design.

This Chapter is organised as follows. In Section 3.1 we introduce the two key ingredients of the new design: the fundamental of the interim analysis (for futility) and the potentiality in the use of external information in a Bayesian context. The motivating clinical context and the novel strategy to develop an innovative treatment in rare diseases is described in Section 3.2, while the design of a sequential platform-basket trial is elaborated in its technicality in Section 3.3. The simulation study to evaluate the operating characteristics of the proposed design is described in Section 3.4 and the results are shown and discussed in Section 3.5 and 3.6.

3.1 Interim analysis (for futility) and external information

Adaptive designs offer a promising alternative to classical approaches because they can expedite the development process, while maintaining validity and efficacy. Specifically, we focused on sequential designs [80] to monitor the course of the trial

in terms of treatment effect and adapt the design accordingly. Interim analysis for futility is a common tool for eliminating treatment arms with weak outcomes, when there is a strong evidence to reject the null hypothesis before the end of the trial. The early identification of ineffectiveness of the new therapy in some specific subgroup of patients helps to reduce resource allocation in unproductive studies, allowing those resources to be redirected toward more promising therapies. The investigators usually define a decision rule based on the evidence that the probability of rejecting the null hypothesis at the end of study is high, given the data available at the interim stage (D_I).

$$\text{if} \quad \mathbb{P}(\text{rejecting } H_0|D_I) \leq q \quad \text{stop the trial}$$

where q is the pre-specified boundary level.
Hence, the design plans to adapt the sample size, the treatment arms or stop the trial according to the amount of this predictive probability with respect to q [81]. However, since variations in sample size and multiple testing can inflate type I error or reduce power, Demets and Lan [82] and Pampallona *et al.* [83] introduced the alpha and beta spending functions [84, 85, 86] to control the type I and II errors of sequential designs. The strategies consisted into the definition of the predictive probability boundary values by a function of the information fraction available at the interim stage t. The information fraction (N_t/N) is commonly defined by the number of patients enrolled at interim step (N_t) on the overall sample size (N). The general idea is that with a small sample size it is preferable to implement a more tolerant stopping rule that relies on lower thresholds to prevent prematurely halting the trial due to negative results obtained by chance. However, as the trial continues and more data are collected, a strict criterion with higher cut-off values is necessary to accurately detect ineffective treatments. Indeed, the spending functions satisfy the following properties:

$$f(0) = 0 \quad \text{and} \quad f(1) = \alpha \quad \text{or} \quad \beta$$

where α and β are the defined level of type I or II error, respectively. Generally, all monotonically decreasing functions fit well [87] and a lot of methods were

implemented to define the boundary levels of the decision rules, both in frequentist and in Bayesian statistics [88].
First of all, the suggestion is to pre-define, in the planning phase of a trial, a value sufficiently large to stop the trial based on strong evidence of failure, but sufficiently small not to stop excessively [49, 27]. A widely used method to determine the boundary level is the calibration approach by simulation study, according to the nature of the test and the outcome considered [89, 90]. Then, different metrics can be used, for example conditional or predictive power in the Bayesian framework [91]. However, the rationale of the futility rule must take into account prior belief about the expectations of success of the treatment and ethical guidelines on the trade-off between additional enrolment and potential benefits. So, especially in situations with very small sample size the expectations of success of the treatment are very high and the adaptive design helps avoid wasting already limited resources.
When information on safety and efficacy are available on similar treatment and/or disease, there is the opportunity to exploit these sources of information for optimizing the study designs in other diseases. However, the process of incorporating historical data must be well-defined to avoid introducing bias or producing misleading results into the design of a clinical trial [92]. Statistical solutions include covariates adjustment, meta-analytic predictive priors [93], or weighting models based on the relevance of external information. We have referred to the latter approach in developing our study design. In practice, the idea of borrowing external information can be an appealing strategy to save costs, accelerate timelines, and reduce sample sizes. However, it is crucial to consider how closely related the external information are to the current trial, and the Bayesian framework can help in this evaluation. Additionally, the Bayesian process of accumulating and updating knowledge is particularly crucial in the context of rare diseases. Given the inherent limitations of small sample sizes, it is essential to leverage existing information from other trials as resources for the current trial.
We introduced an approach to trial design that integrates the adaptive flexibility of sequential design with the borrowing of external information to address challenges of small sample sizes clinical trials.

3.2 The motivating clinical context

The novel study design that we proposed responds to a real case study taking into consideration interim futility analysis, external and internal borrowing of information. Our goal was to design a Phase I-II single arm trial for three ultra-rare subtypes of a lysosomal pediatric disease. A specific gene-mutation identifies the three patients subgroups of interest, which also share the therapeutic rationale, the product manufacturing process and product control strategy. Indeed, the same experimental treatment under investigation is used in all subgroups and the primary endpoint is an outcome of activity, represented by a continuous variable.

Due to various factors, including the rarity of the disease, the absence of a standard of care, and ethical issues, a control group was not feasible. Given the need to test a single new treatment across multiple disease subgroups with very small sample sizes, a basket trial approach seemed reasonable. Additionally, since we have access to results from a completed clinical trial on the same treatment in a similar disease, we aim to leverage that information in our current trial. To enhance the scientific evidence of the treatment effect, we considered the possibility to borrow information from both an external source, called "co-data", and intra-subgroups under investigation.

Furthermore, we have considered to implement a sequential approach based on two stages to allow for the initiation of the trial in the last subgroup only when the evidence on the treatment effect available on the first two subgroups is convincing. Specifically, motivated by the need to begin the trial with only two of the three subgroups, we have planned an interim analysis at the end of the first stage, upon completion of the trial for the two available subgroups. The low incidence, extended enrolment timelines for one subgroup and the high cost of the new therapy suggested to resort to a platform trial which has the advantage to use a single infrastructure to investigate consecutive trials. Hence, we included in the first stage the two available subgroups until their complete follow-up to subsequently perform an interim analysis for futility. Then, based on the result, the design allows for either enrolling patients from the additional subgroup or stopping the study early. Thus, the interim results guide the decision on whether to enrol the third subgroup according to a statistical decision rule. If the study continues, the

second stage involves the analysis of the complete basket trial to achieve a final overall evaluation. Conversely, if it is stopped for futility, the preliminary findings from the two initial subgroups become the final results.
A stopping rule for efficacy was not included, because, even in the presence of an innovative treatment potentially highly effective, we aim to recover the maximum amount of information in each disease subgroup, thus opening the possibility also to the third subgroup of patients to be treated.
The advantages of a platform-basket trial design include the ability to approve a single protocol for the multiple diseases under investigation. Additionally, this design reduces the time required to have preliminary results on subgroups that are more readily available, providing stronger evidence to determine whether it is ethical and reasonable to commit time and resources to enrol patients in additional, potentially more complex subgroups. However, if the trial does not stop for futility, the time to obtain the final results for the subgroups enrolled in the first step are extended. Consequently, it is up to a team of experts to assess the appropriate trade-offs for applying such a design.
The novel design can be easily generalised to include schemes that consider more than two subgroups in the first step, allowing for decisions about the inclusion of one or more additional subgroups.
To design the model and perform the simulation study we did some strict assumptions based on the knowledge about the motivating clinical study, however, for reasons of confidentiality, we will continue to talk about it in a general way.

3.3 A sequential single arm Platform-Basket trial design

We propose here a sequential single arm platform-basket trial leveraging the information from co-data available from a previous trial in a similar disease. We considered three subgroups ($K = 3$) and the same number of patients n_k in each subgroup ($n_1 = n_2 = n_3$) for a total sample of N subjects. The primary endpoint Y_{ik} ($i = 1 \dots n_k$ and $k = 1 \dots 3$) is assumed continuous and normally distributed, with θ_k the subgroup-specific treatment effect. Considering the EXNEX model as reference approach [2], our aim was to incorporate the external information available from the co-data into the new model. Thus, an additional level of borrowing

was considered, introducing a prior weight w_{co_k} that explains how the subgroups under investigation and the completed trial are similar.
First of all, we considered the co-data endpoint Y_{0l} ($l = 1, \ldots, N_0$, N_0 number of patients), normally distributed and a prior normal distribution for the treatment effect θ_0. Then, given the results of the completed trial ($\bar{y}_0$ and s_0^2, the sample mean and variance, respectively), we determined the posterior distribution of θ_0 thank to the conjugate prior property of the normal distribution. Secondly, we used the posterior parameters of θ_0 to define the prior parameters of the normal distribution of θ_k under w_{co_k} in the co-data component. We called the posterior parameters of θ_0 μ_0 and σ_0^2, respectively, and we assumed an overlap in the prior variance of θ_k and the posterior variance of θ_0 (i.e. $\lambda = 0$).

$$
\begin{gathered}
Y_{0l} \sim N(\theta_0, s_0^2) \quad l = 1 \ldots N_0 \\
\text{and} \quad \theta_0 \sim N(\hat{\mu_0}, \hat{\sigma_0}^2) \\
\theta_0 | \mathbf{y_0} \sim N\Big(\underbrace{\frac{s_0^2 \hat{\mu_0} + N_0 \hat{\sigma_0}^2 \bar{y}_0}{s_0^2 + N_0 \hat{\sigma_0}^2}}_{\mu_0}, \underbrace{\frac{s_0^2 \hat{\sigma_0}^2}{s_0^2 + N_0 \hat{\sigma_0}^2}}_{+\lambda = \sigma_0^2} \Big)
\end{gathered}
$$

where $\mathbf{y_0}$ are the co-data observed.
Then, we used a Bayesian hierarchical model to estimate θ_k under the assumption of exchangeability (EX) with prior weight w_{ex_k}. In particular, we assumed that θ_k arises from a normal distribution with prior parameters μ and σ, which arise from a hyperprior normal distribution and a Half-Normal distribution, respectively, where $\bar{\mu}$, $\bar{\sigma}^2$ and $\bar{\tau}^2$ are the hyperparameters. Finally, we assumed a normal prior distribution with prior weight $w_{r_k} = 1 - w_{co_k} - w_{ex_k}$ to define the robust component. The prior parameters m_0 and v_0^2 were fixed across subgroups.

$$
\begin{gathered}
Y_{ik} \sim N(\theta_k, s_k^2), \quad s_k \sim HN(\tau^2) \quad i = 1 \dots n_k \quad k = 1, 2, 3 \\
\text{CO-DATA} \quad w_{co_k} : \quad \theta_k \sim N(\mu_0, \sigma_0^2) \\
\text{EX} \quad w_{ex_k} : \theta_k | \mu, \sigma^2 \sim N(\mu, \sigma^2) \\
\mu \sim N(\bar{\mu}, \bar{\sigma}^2), \quad \text{and} \quad \sigma \sim HN(\bar{\tau}^2) \\
\text{ROBUST} \quad w_{r_k} = 1 - w_{co_k} - w_{ex_k} : \quad \theta_i \sim N(m_0, v_0^2)
\end{gathered}
\tag{3.1}
$$

The design provides two different stages. In the first one, the two initial subgroups undergo the interim analysis using the model defined above and according to the following stopping rule:

$$\mathbb{P}(\theta_3 > \delta | D_I) > q$$

the second stage will begin or not. So, we are checking whether the prior predictive probability (PPP) that the third subgroup exceeds a threshold (δ) given the data available from the two initial subgroups (D_I) is greater than a certain cut-off q. If this is true, the third subgroup is included into the trial and the final analysis will be based on the model (3.1). Otherwise, the trial is stopped for futility and no patient is enrolled in the third subgroup.

In the end, since the aim was to test the null hypothesis that the treatment effect is greater than a pre-specified threshold ($H_0 : \theta_k \leq \delta$) for each subgroup, we defined the following final decision rule to apply to each subgroup k to declare the benefit of that treatment: $\mathbb{P}(\theta_k > \delta) > 95\%$.

We summarise the design of the 2+1 sequential platform-basket trial in Figure 3.3.1.

The most critical issue of this innovative design was the definition of the cut-off q. We considered various strategies to define this value and we evaluated the different choices in the simulation study in the next section.

3.4 Simulation study

A simulation study was performed to assess the performance of the proposed design, assuming different true treatment effects and prior weights in the model (3.1).

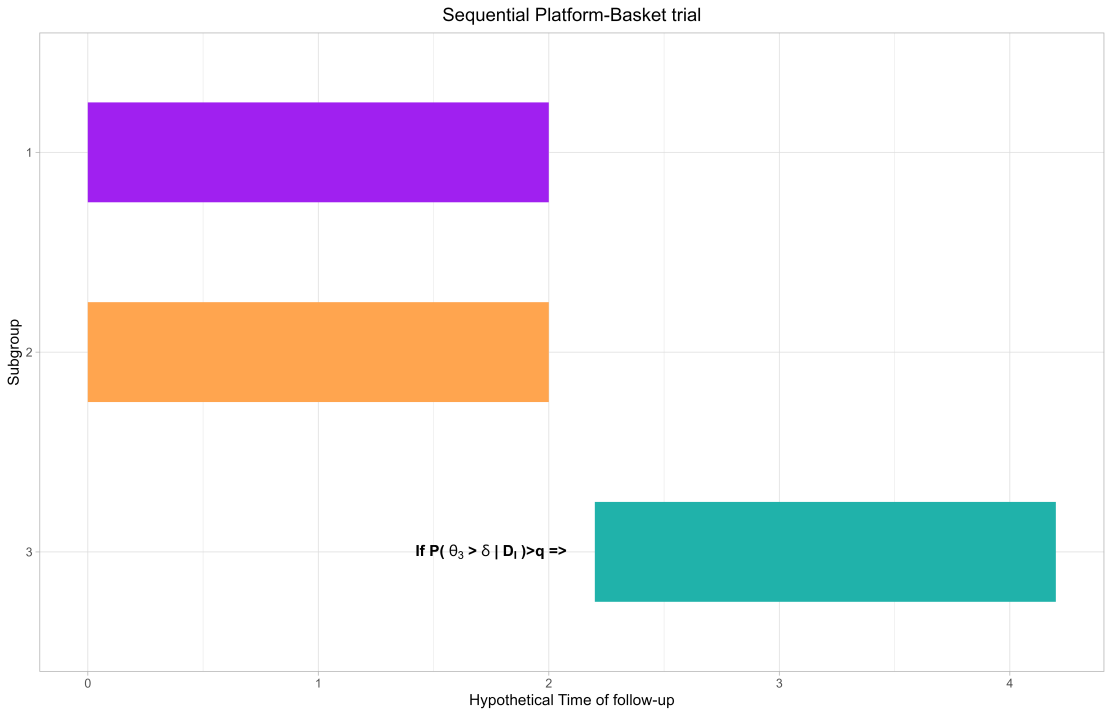


Figure 3.3.1: A sequential Platform-Basket design: given the data of subgroup 1 and 2 (D_I) if the predictive prior probability that the treatment effect of the third subgroup exceeds a threshold (δ) is greater than a cut-off (q), the trial is not stopped for futility and the third subgroup is included into the trial.

We considered eight scenarios for the true values of θ_1 and θ_2 (Table 3.4.1), the two subgroups in the first stage, and 20 patients for each subgroup. We simulated very high, weak and ineffective treatment with respect to the threshold $\delta = 1$ in Scenarios 1, 2, 3, 4 and 8 and also some of their combinations in Scenarios 5, 6 and 7.

Table 3.4.1: Simulations' scenarios

		Scenario							
Stage	True effect	1	2	3	4	5	6	7	8
I	θ_1	3	1.2	0.5	0	3	3	1.2	1
I	θ_2	3	1.2	0.5	0	0.5	1.2	0.5	1
II	θ_3	3	3	3	3	3	3	3	3

Then, we applied six systems of weights on the prior probabilities w_{co_k}, w_{ex_k} and w_{r_k} (Table 3.4.2), considering different nuance of the exchangeability assumption. The Model A favours equally the exchangeable and the co-data component, Model B and F favours the similarity with co-data, whereas Model C and E the similarity within subgroups and Model D the robust component. For simplicity, we considered the system of weight pre-defined at the planning phase without of an update at stage I. Moreover, we assumed equal prior weights for each subgroup, even if different choices can be explored.

Table 3.4.2: Prior weights

Model	w_{ex_k}	w_{co_k}	w_{r_k}
A	0.5	0.5	0
B	0.2	0.5	0.3
C	0.5	0.2	0.3
D	0.2	0.2	0.6
E	0.7	0.2	0.1
F	0.2	0.7	0.1

We simulated M=10000 ($m = 1 \dots M$) replicates of each scenario using two parallel chains with 13000 MCMC iterations and 3000 burn-in (i.e. number of iterations to discard at the beginning, B=10000, $b = 1 \dots B$). The convergency of the MCMC method were assessed by the Gelman-Rubin statistics [77].
We assumed that the inter-patients standard deviation is distributed as a Half-Normal with parameter 1, $s_k \sim HN(1)$ (median = 0.67, 95% CI=0.03-2.24), to be not very stringent. Moreover, to calculate μ_0 and σ_0^2, we considered a hypothetical very promising co-data trial. We assumed the following prior for the co-data treatment effect, $\theta_0 \sim N(4, 5)$, the sample mean $\bar{y}_0$ equal to 3.5 and low inter-patients variance $s_0^2 = 0.5$.
The configuration of the co-data mimics the findings of the clinical trial that was conducted with the same treatment in a similar disease of the motivating clinical context. In the real example, the experimental treatment reached supra-physiologic levels in all the treated patients, with a mean level of the presenting a

10-fold increase with respect to the healthy controls. However, in the simulation study we were more conservative considering a smaller effect and we set the same number of patients of the completed trial ($N_0 = 8$).
The hyperpriors in the exchangeable part were not so informative to allow the data to determine the posterior distribution: $\mu \sim N(0, 10^2)$ and $\sigma \sim HN(0.5^2)$ (95% CI= $0.02 - 1.12$). Finally, the robust part was assumed to be centered on the threshold value ($m_0 = 1$) with large variance ($v_0 = 10$) to be uninformative about the tails of the prior distribution.
The cut-off q that regulates the access to the second stage was defined by different approaches. The first option was to pre-specify it depending on how much the investigators want to be conservative/liberal, and we decided for a value of 0.5. Secondly, using the power spending function $0.95 * t^2$ ([94]) we obtained q=0.42, considering as the information fraction $t = \frac{40}{60} = 0.67$, where $40 = n_1 + n_2$ and 60 is the final sample size ($n_3 = 20$). In this case, we assumed $u = 95\%$ as the cut-off for the final decision rule. In the third approach, the idea was to use the simulations' results to measure the maximum and the minimum values expected of the PPPs and to define the q cut-off as the mean value. We expected the maximum value of the PPPs under the alternative hypothesis ($\theta_k > \delta$) by the Scenario 1 and the minimum value under the null hypothesis ($\theta_k \leq \delta$) by the Scenario 4, assuming strong exchangeability within subgroups (i.e. Model E). Thus, we simulated Scenario 1 and 4 under Model E at stage I and then calculated the mean among their PPPs (i.e. $q = 0.58$). Alternatively, we used the Youden Index (J-Index, [95]) to calculate the cut-off ($q = 0.60$) that optimised the count of correct rejections of H_0 in Scenario 1E (sensitivity) and the correct no-rejection in Scenario 4E (specificity), across all simulation replicates. Finally, according to the interim decision rule on different q values, we implemented or not the second stage assuming a high benefit of treatment in subgroup 3 ($\theta_3 = 3$).
We evaluated the performance of the innovative design through the bias and the mean square error (MSE) of the estimated treatment effect. Then, given the posterior mean on B iterations of θ_k ($\bar{\theta}_k$), we also reported their median across all the replicates ($\hat{\theta}_k$) and the 95% credible interval defined by the median of the

2.5th and 97.5th percentiles ($Q_{\theta_k}(.)$) of the posterior θ_k.

$$\text{Bias}(\theta_k) \approx \frac{1}{M}\sum_{m=1}^{M}(\bar{\theta}_k^m - \theta_k)$$

$$\text{MSE}(\theta_k) \approx \frac{1}{M}\sum_{m=1}^{M}(\bar{\theta}_k^m - \theta_k)^2$$

$$\hat{\theta}_k \approx \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}_k^m$$

$$95\% \quad \text{Credible interval lwr} \approx \frac{1}{M}\sum_{m=1}^{M} Q_{\theta_k}^m(0.025)$$

$$95\% \quad \text{Credible interval upr} \approx \frac{1}{M}\sum_{m=1}^{M} Q_{\theta_k}^m(0.975)$$

where $\bar{\theta}_k^m$ is the posterior mean of θ_k for the mth simulated trial. Moreover, we also estimated for each subgroup the median of the posterior probability (PP) that θ_k exceeds δ across the replicates, the marginal power and type I error (in a frequentist perspective). The power and type I error were defined as the proportions of simulated trials with PP that guides to a correct/incorrect final decision in the subgroups under the alternative/null hypothesis (i.e. $\theta_k > 1$ or $\theta_k \leq 1$).
The simulations were performed using the open-source R software v.4.3.2 (R Foundation for Statistical Computing, Vienna, Austria) and the R2jags package. All replicates were parallelised on 89 cores to reduce the computational time.

3.5 Results

The prior predictive probability that θ_3 exceeds δ given the simulated data of subgroups 1 and 2 is reported in Table 3.5.1. We have highlighted in red the scenarios/models that stop after stage I due to futility, and in green those that continue to the second stage, regardless of the chosen cut-off q. Consequently, in what follows, when we refer to the scenarios that have stopped, we consider the results of the first stage as final results, while when we refer to those that continued, we considered the results at the end of the second stage. The remaining PPPs were between the different cut-offs, so according to different choices of q the scenarios

are stopped or not in the first stage.
The PPPs values highly depend on the different allocation weights. Very unpromising scenarios (e.g. Scenarios 3 and 4) have a lot of variability in PPPs depending on whether one favours the assumption of exchangeability within subgroups (Model C and E) or with co-data (Model B and F). In detail, in Scenario 3 very low PPPs were estimated under Model C and E, consistent with the true treatment effects. On the other hand, under Model B and F, which favour the leveraging of the co-data, the PPPs are very encouraging in the direction to continue the study. Similarly, but slightly lower results are observed in the null Scenario 4.

Table 3.5.1: Prior predictive probability (%) on θ_3 at stage I

	Model					
Scenario	A	B	C	D	E	F
1	85	76	77	62	89	86
2	89	80	73	63	79	90
3	54	68	40	54	31	77
4	51	66	37	53	27	76
5	59	72	53	59	45	81
6	86	76	68	61	76	88
7	69	73	55	58	52	83
8	75	75	60	60	60	85

The median PPPs that do not exceed the interim analysis are reported in red and those that exceed it in green, regardless of the q choices.

Now, we will concentrate on the results at the second stage for the more liberal value of q (i.e. 0.42), while those considering the remaining cut-off values are reported in the Supplementary material 5.3.1.
In Figure 3.5.1 and 3.5.2 we compare the bias and the MSE of the different models in the eight scenarios. All models seem to accurately estimate the true treatment effect in each subgroup, even for $\theta_k = 3$ where the amount of relative bias reaches approximately 4.2% under certain exchangeability conditions. The highest degree of relative bias considering low ($\theta_k = 1.2$) or ineffective ($\theta_k = 0.5$) treatment effect is observed in Scenario 7 (4% and approximately 10%, respectively)..

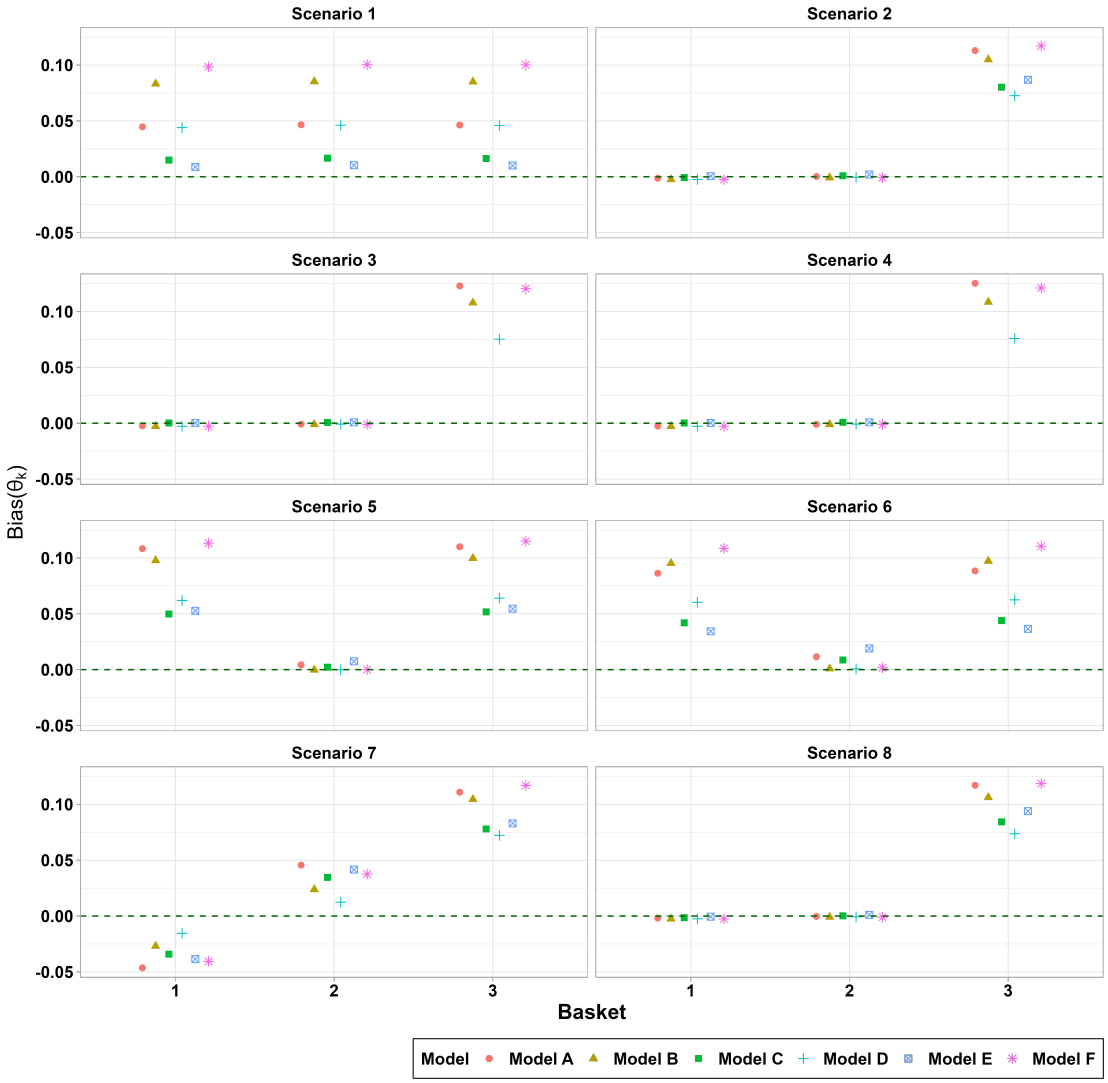


Figure 3.5.1: Comparison of different weights' model across scenarios in terms of bias.

Models A, B and F generally have a higher bias and MSE, instead C and E the lower ones except in those scenarios where the assumption of exchangeability was incorrectly stated a priori, e.g. in Scenario 7. In this case, Model D, which focuses on the robust part, is the most accurate. Nevertheless, Model C, D and E better balance the bias and MSE between subgroups across all the scenarios.
In Supplementary materials 5.3 we show the posterior median estimates of the

treatment effects and their 95% credible intervals in Scenarios 1-8 across models A-F. It can be seen that there is no marked difference in the width of the credible intervals between the different scenarios and models. However, it is slightly lower in Models A and F in most scenarios and subgroups, whereas in Scenario 1 it is narrower in Models C and E, as expected, since they are the ones that favour exchangeability between subgroups.

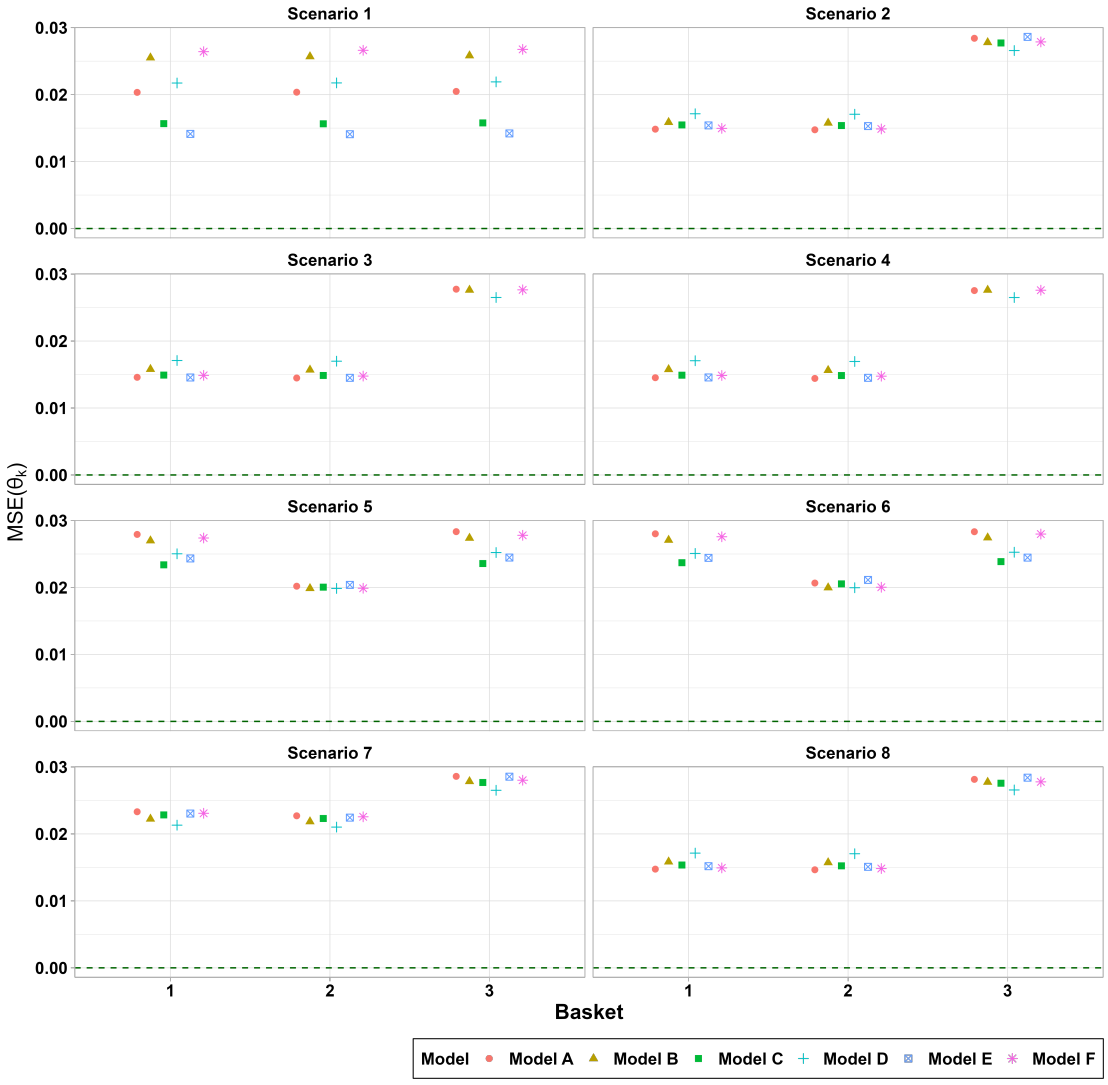


Figure 3.5.2: Comparison of different weights' model across scenarios in terms of MSE.

We summarised in Table 3.5.2 the median posterior probability of success estimated in each scenario/model combination for each subgroup in the trial. According to the final decision rule (2.2) we highlighted in green those subgroups that reached the success. The posterior probability of the subgroup with high true treatment effect is the only that exceeds the 95% cut-off needed to declare the success of the new therapy. Posterior probabilities very close to the cut-off were estimated in Scenarios 2 and 6, but they were not big enough to reach a successful result. Furthermore, the variability across models was very low. Thus, despite the influence of a very positive outcome in one or two subgroups (Scenario 2 and 6, respectively) and a high borrowing (Model F) with the effective co-data, the model fails to have more than 95% certainty that a moderate treatment effect ($\theta_k = 1.2$) is actually effective.

Table 3.5.2: Median posterior probability of success (%) in subgroups (1 to 3) across the scenarios.

Model	Scenario	1	2	3	Scenario	1	2	3
Model A	1	100	100	100	5	100	0	100
Model B	1	100	100	100	5	100	0	100
Model C	1	100	100	100	5	100	0	100
Model D	1	100	100	100	5	100	0	100
Model E	1	100	100	100	5	100	0	100
Model F	1	100	100	100	5	100	0	100
Model A	2	94	94	100	6	100	93	100
Model B	2	93	93	100	6	100	92	100
Model C	2	93	94	100	6	100	92	100
Model D	2	92	93	100	6	100	92	100
Model E	2	94	94	100	6	100	93	100
Model F	2	93	94	100	6	100	92	100
Model A	3	0	0	100	7	85	0	100
Model B	3	0	0	100	7	88	0	100
Model C	3	0	0		7	87	0	100
Model D	3	0	0	100	7	90	0	100

Continued on next page

Table 3.5.2 – continued from previous page

Model	Scenario	1	2	3	Scenario	1	2	3
Model E	3	0	0		7	86	0	100
Model F	3	0	0	100	7	86	0	100
Model A	4	0	0	100	8	49	50	100
Model B	4	0	0	100	8	49	50	100
Model C	4	0	0		8	49	50	100
Model D	4	0	0	100	8	49	50	100
Model E	4	0	0		8	50	50	100
Model F	4	0	0	100	8	49	50	100

The marginal power was estimated in those subgroups under the alternative hypothesis ($\theta_k > 1$) and reported in Figure 3.5.3. A 100% power was reached for highly treatment effects, regardless the models considered, while when θ_k was equal to 1.2 the power was very low (around $40 - 45\%$), without substantial differences between the models.

We measured the Type I error in Scenario 8 and found that the probability to reject the null hypothesis was closed to 4%. The highest type I error was measured in Model D, where the robust prior is favoured and the lower one in Model A. So, we can state that the probability to do the uncorrected decision under the null hypothesis is low, see Table 3.5.3.

Considering different cut-off q to stop for futility the sequential platform-basket trial, various scenarios are stopped early. The results about the comparison of stopping at the first stage instead that at the second one are reported in Supplementary 5.3.1. Briefly, we observed that very little increases in bias and MSE and decreases in median posterior probability when $q = 0.58$ to stop at Stage I are notable. More obvious changes can be seen in favour of Stage II in the power calculation, but not relevant for decision-making purposes. The marginal type I error is always higher, but well controlled.

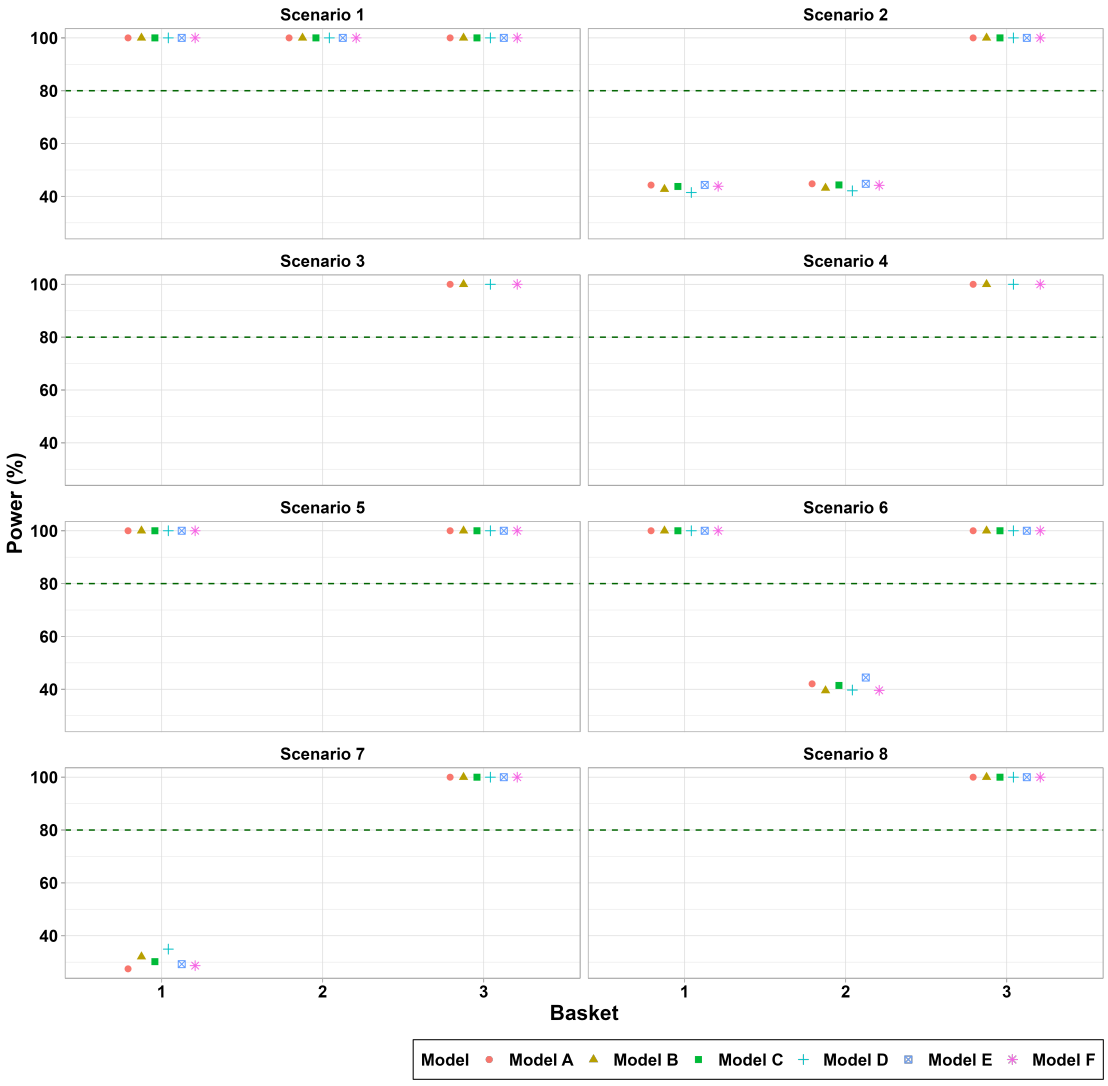


Figure 3.5.3: Comparison of different weights' model across scenarios in terms of frequentist power. The green dashed line represents the 80% of power.

Table 3.5.3: Type I error (%) in Scenario 8 ($\theta_k = 1 \quad \forall k$).

Model	1	2
A	3.83	3.73
B	3.84	3.90
C	4.02	3.89
D	4.12	4.16
E	4.03	3.91
F	3.79	3.71

1 and 2 are referred to the subgroup number.

3.6 Discussion

We proposed a novel design that combines innovative elements to plan a complex clinical study involving a platform-basket design with the use of external information and an interim evaluation (for futility). This design is intended for rare diseases, where it is essential to incorporate all available information, given the limitations on sample size. In particular, being aware that information available from treatment arms alone may be insufficient to achieve scientific validity due to the constrained sample size, we integrated previously validated information from a completed trial to strengthen the results of the current study. While the borrowing of information from various sources is already common in clinical trials, our approach specifically aims to guide the posterior distribution of the current responses, particularly when it is known that the treatment arms are similar or dissimilar to the co-data. Additionally, we considered a sequential design in which, based on the initial results in few subgroups, we can decide whether to include other subgroups in the trial, following the principles of platform trials. This allows a single protocol to analyse multiple diseases by making better use of available information. The maintenance and enrichment of the platform is regulated and strongly depend on the data that will be produced during the development. Thus, the Bayesian framework and the sequential approach, which are key elements of the design we proposed, are essential in this context.

In particular, we considered a two-stage design, where in the first stage we apply

an interim analysis for futility on the initially available subgroups, or more generally, on those expected to be the most promising or easiest to recruit (in our motivating study, they were two). Then, depending on these preliminary results, the second stage might (or might no) include subgroups that are more difficult or more vulnerable to recruit (one in our motivating study). However, such a design can also be used in standard situations where there are no issues with patient enrolment, and it can also be generalised to different schemes involving multiple interim analyses, either to add new subgroups or to include more patients in the existing subgroups.

We want to point out that our innovative design includes an interim analysis that determines whether to add an additional subgroup, but it does not specify the appropriate sample size for inclusion. Estimating sample size in the context of Bayesian designs remains an area of ongoing research and existing work aimed at addressing this issue [96], and the use of external information in rare diseases is very limited. In the motivating case study, the sample size is pre-fixed and primarily dictated by the feasibility of recruitment due to the rare nature of the diseases involved.

We specify that at the end of the first stage, the sub-trial on the first subgroups are terminated and the number of patients can not be increased. Thus, if we choose a q cut-off that is not too stringent, even the less promising subgroups (e.g., Scenarios 3 and 4) can access the second stage. However, we do not compromise the ethical integrity of the study by retaining the data of subgroups 1 and 2 in the second stage, as we are not enriching their sample size at this stage. We only used their data, combined with that of the third subgroup, to strengthen the evidence of the trial overall.

The interim analysis was based on the prior predictive probability of the third subgroup, given the data from the first and second subgroups. This was a specific feature of our design, as we did not use the predictive probability in the usual manner [97, 98]. Instead, we used the collected data to decide whether or not including an additional subgroup, rather than use the predictive prior probability on the same subgroups from which it is derived.

We conducted a simulation study with six system of weights to investigate the impact on the prior predictive probability of the third subgroup of the prior believes

on the similarity with co-data and exchangeability between subgroups, . Relevant differences were observed in the more heterogeneous scenarios (Scenarios 5-7, see Table 3.5.1), and very contrasting results were obtained in scenarios of very ineffective subgroups in stage I (Scenarios 3 and 4). We therefore emphasise how important is the choice of weights because they have a relevant impact on whether or not the trial proceeds. In the event of excellent results from external data and a strong likelihood that the current trial is promising, they make the study continue, even in the case of weak results at stage I, unless a very high cut-off for the interim analysis is fixed.

As a consequence, different choices of stopping rule cut-off may lead to different conclusions. In particular, choosing a q too small may lead to the inclusion of a new subgroup in an unpromising trial, also resulting in a loss of time and resources. Conversely, a q that is too large may lead to an early stop of the trial, preventing the investigation of the new therapy in a further group of patients who could benefit from it. Nevertheless, we noted that weak results in some subgroups did not impact on very effective results, e.g. subgroup 3 in Scenario 3, 4 and 8, in terms of power. So, if you have very positive expectations from the subgroup that you add at the second stage, it may be worth extending the trial as you are not allocating more subjects to the unpromising subgroups. Indeed, since one is working in the field of rare diseases, the reasonable expectation is to have a very effective treatment. On the other hand, choosing a small cut-off, the study will proceed if it is very likely that the experimental treatment has evident efficacy, given the current data.

We observed not relevant differences in results at Stage II across models and between Stage I and II, given different cut-offs, in terms of error rates.

Moreover, if we generalised the design to the inclusion of more than one subgroup in the second stage, their PPPs would arise from the same distribution. For this reason, the results obtained in the interim analysis can be useful in providing guidance for the choice of weights and cut-offs even in situations of study designs involving more than one disease subgroup at the second stage.

Regarding the choice of weights, we a priori set them equal for the three subgroups, but it is also possible to consider subgroups-specific weights to account for varying similarities with the co-data. Furthermore, we can adopt a dynamic approach that

evolves as trial data accumulate. In particular, the weights can change between the first and second stage based on the exchangeability between the initial subgroups and the added ones.

The quantities considered in the simulation study reflect the expert opinions on the threshold δ and the treatment effects, which reflect their optimistic expectations regarding treatment benefits. However, to evaluate the design more comprehensively, it is necessary to conduct a simulation study with more moderate treatment effects and perform additional sensitivity analyses. While we considered a balanced design with the same sample size future work will involve testing the robustness of an unbalanced design and the robustness to different sample sizes. Specifically, since we plan to apply this study design in the context of ultra-rare diseases, we will investigate how extremely small sample sizes can impact the results. Finally, further evaluations will be done considering varying numbers of subgroups in stage I.

4 Conclusions

Clinical trials in rare diseases require innovative and efficient trial designs to maximise the information gathered from a limited number of patients, since traditional designs may not be feasible or effective. To address this challenge, researchers are increasingly adopting adaptive designs, Master protocols and bayesian methods for the analyses. These approaches allow for high flexibility, enabling researchers to eventually adjust the trial parameters based on already collected data. Moreover, the borrowing of information within patient subgroups or the use of external information can provide insights from a methodological point of view and prioritise the use of comprehensive data. The ultimate aim of these efforts is to ensure that patients with rare and fatal diseases receive equitable access to potentially life-saving treatments. It is crucial that the rarity of a disease does not result in a reduced focus or research funding, which could otherwise lead to a lack of effective therapies. By developing and implementing innovative and more efficient clinical trial designs, the scientific community can enhance the understanding of rare diseases.

Motivated by the need to plan a trial that evaluates the efficacy of an innovative therapy in three similar rare diseases, we evaluated the robustness of three standard Bayesian methods for the analysis of basket trials. The results of the simulation study we performed are very promising. In particular, in the presence of clear treatment effects in the subpopulations, the three methods can be safely used with samples of 4-7 patients. However, when the effect of the treatment in the subpopulations is mild or there is heterogeneity in the treatment effect, the choice of the sample size and the final cut-off for decision should be carefully considered in the planning phase of the trial.

The novel two-stage design we proposed combines innovative elements to plan a complex clinical study involving a platform-basket trial back-bone, reinforced by the use of external information and an interim evaluation for futility. We assessed the performance of this design through simulations, which resulted sensitive to the choice of certain parameters (e.g., prior weights, q cut-offs for the interim analysis and the final cut-off), although it was generally robust when there was strong belief in a highly effective treatment. While, our design was motivated

by a research in rare diseases, its principles can be generalised to other contexts. However, it is important to emphasise that in the set up of the protocol, it is always of paramount importance the evaluation of the properties of the proposed design through simulations. Indeed, the operating characteristics should be carefully evaluated by extensive clinical trial simulations that should always be part of the study protocol.

In conclusion, the work presented here serves as a foundational starting point for enhancing our understanding of how to design a trial in the context of rare diseases and additional work is needed for a more comprehensive evaluation. In this context, fostering a collaborative dialogue between clinicians and statisticians is essential, alongside a strict connection with regulatory authorities. This integrated approach is crucial for advancing research and improving outcomes in this challenging field.

5 Supplementary

In the following chapters, we briefly introduce the Bayesian framework and the MCMC method with appropriate references to refer to. Then, we provide additional figures and tables related to Chapters 2 and 3.

5.1 Bayesian framework and MCMC method

In the Bayesian framework a probability function is assumed to be the prior distribution of a parameter of interest (θ, regarded as random variable) given previous knowledge [99]. Then, through data observation (y), the prior distribution $\pi(\theta)$ is updated to the posterior distribution $\pi(\theta|y)$ thank to Bayes' Theorem [100]. When the unknown parameter is continuous the the Bayes' Theorem is formulated as follows:

$$\pi(\theta|y) = \frac{\pi(\theta)L(y|\theta)}{\int \pi(\theta)L(y|\theta)d\theta}$$

and $L(y|\theta)$ is likelihood, i.e. the conditional probability of the data given θ. So, it might be summarised in $\pi(\theta|y) = \pi(\theta)L(y|\theta)$.
There is a class of prior distributions for the posterior distributions belongs to the same probability family, such as the beta distribution and normal with known variance. In these cases, the estimation of the a posteriori distribution is straightforward, however, one generally has to deal with distributions whose a posteriori distribution is unknown. In these situations, estimation by means of approximations or simulations is used. The method most commonly used in the literature is the Markov Chain Monte Carlo (MCMC) method [101]. A Markov Chain is a sequence of random variables $\theta_i, \ldots,$ where the value of θ_i depends only on the value of θ_{i-1} and not on $\theta_1, \ldots, \theta_{i-2}$ and the stationary distribution of the Markov Chain is the distribution that remains unchanged for all θ_i, where i is larger than some value M. Once Markov chain convergence has been achieved, sampling from the stationary distribution of the Markov Chain is equivalent to sampling from the a posteriori distribution sought. In our simulation studies, we implemented the MCMC method using the r2jags package [102, 103, 104] of R software. The Markov chain was performed by means of the mixture of two approaches: Gibbs

sampling [105] and Metropolis-Hasting sampling [106]. Generally, the chains start with an initial value decided by the investigators, that might be not closed to the true value however, after time and a number of iterations the chains hopefully reach the convergence. Then, a discarding of some initial iteration is applied, named *burn-in*. At the end, in addition to the posterior distribution from which the parameter of interest is sampled, its credible interval is also obtained. It is defined as the interval that contains the parameter with a certain probability.

5.2 Supplementary material chapter 2

In this subsection we report the tables regarding the median posterior probabilities in Scenario 1, 3, 4 and 8 and type I error of Scenarios 8, comparing the different methods used to analyse them.

Table 5.2.1: Median posterior probability of the BHM, EXNEX and TRB methods according to different sample sizes (4, 7, 15 and 50) in Scenarios 1, 3, 4 and 8.

Sample	Method	Subgroup	PP_1 (%)	PP_3 (%)	PP_4 (%)	PP_8 (%)
4	BHM	1	100	3	0	50
4	BHM	2	100	3	0	50
4	BHM	3	100	3	0	50
4	EXNEX	1	100	4	0	51
4	EXNEX	2	100	4	0	50
4	EXNEX	3	100	4	0	50
4	TRB	1	100	4	0	50
4	TRB	2	100	4	0	50
4	TRB	3	100	4	0	50
7	BHM	1	100	1	0	50
7	BHM	2	100	1	0	50
7	BHM	3	100	1	0	50
7	EXNEX	1	100	1	0	50
7	EXNEX	2	100	1	0	50
7	EXNEX	3	100	1	0	50
7	TRB	1	100	1	0	50
7	TRB	2	100	1	0	50
7	TRB	3	100	1	0	50
15	BHM	1	100	0	0	49
15	BHM	2	100	0	0	49
15	BHM	3	100	0	0	49
15	EXNEX	1	100	0	0	49
15	EXNEX	2	100	0	0	49

Continued on next page

Table 5.2.1 – continued from previous page

Sample	Method	Subgroup	PP_1 (%)	PP_3 (%)	PP_4 (%)	PP_8 (%)
15	EXNEX	3	100	0	0	49
15	TRB	1	100	0	0	50
15	TRB	2	100	0	0	50
15	TRB	3	100	0	0	50
50	BHM	1	100	0	0	50
50	BHM	2	100	0	0	50
50	BHM	3	100	0	0	50
50	EXNEX	1	100	0	0	50
50	EXNEX	2	100	0	0	50
50	EXNEX	3	100	0	0	50
50	TRB	1	100	0	0	50
50	TRB	2	100	0	0	50
50	TRB	3	100	0	0	50

5.2.1 EXNEX results considering different prior weights

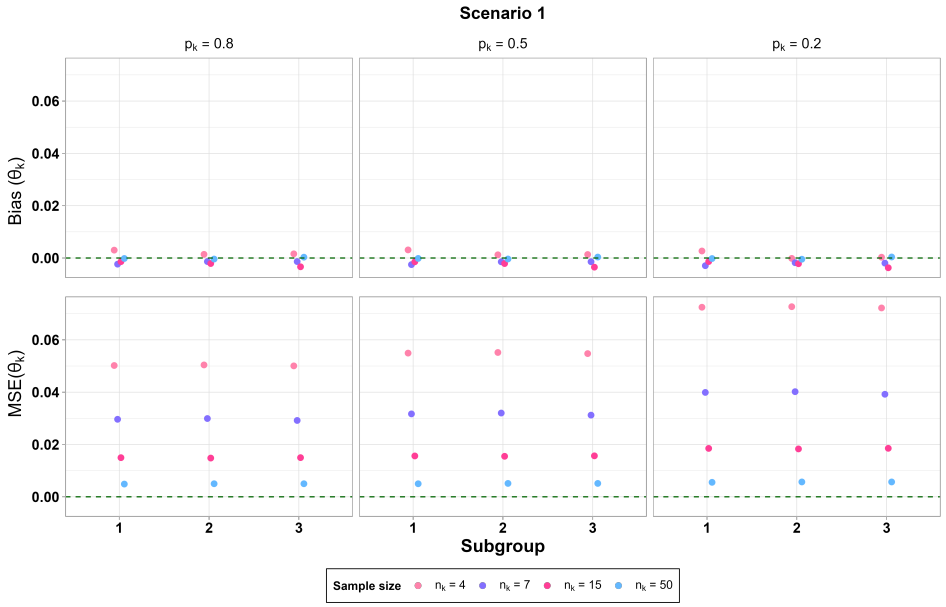


Figure 5.2.1: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 1 ($\theta_k = 3\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

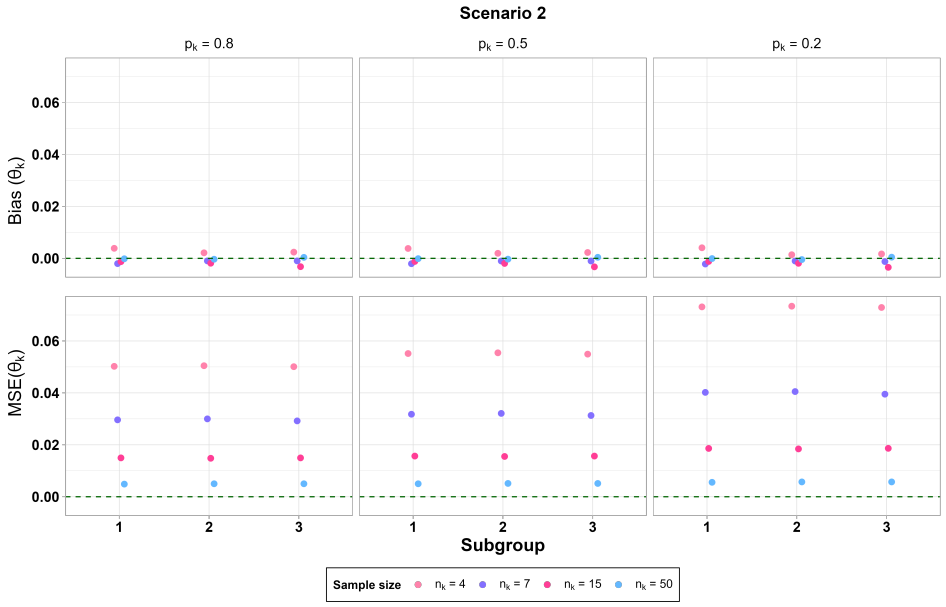


Figure 5.2.2: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 2 ($\theta_k = 1.2\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

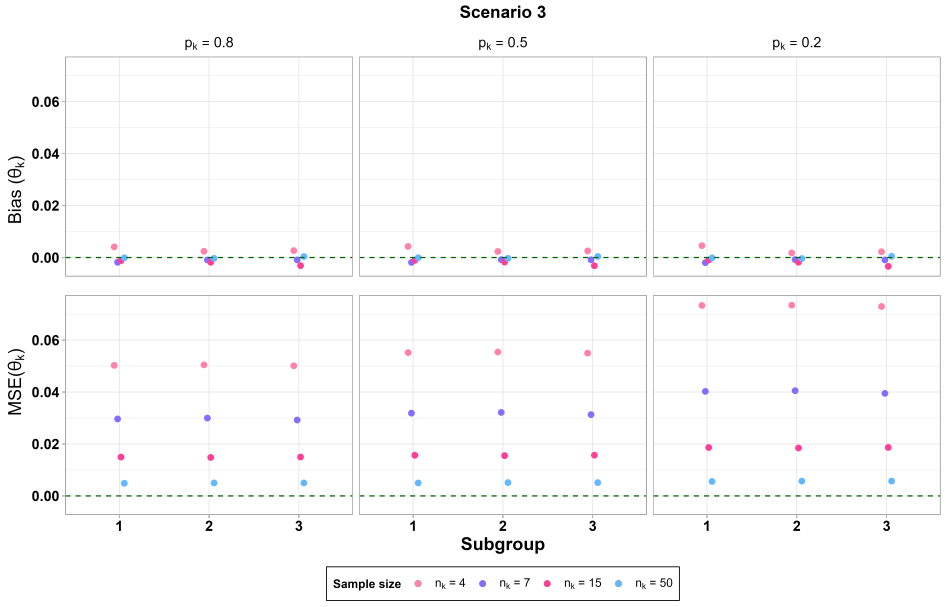


Figure 5.2.3: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 3 ($\theta_k = 0.5\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

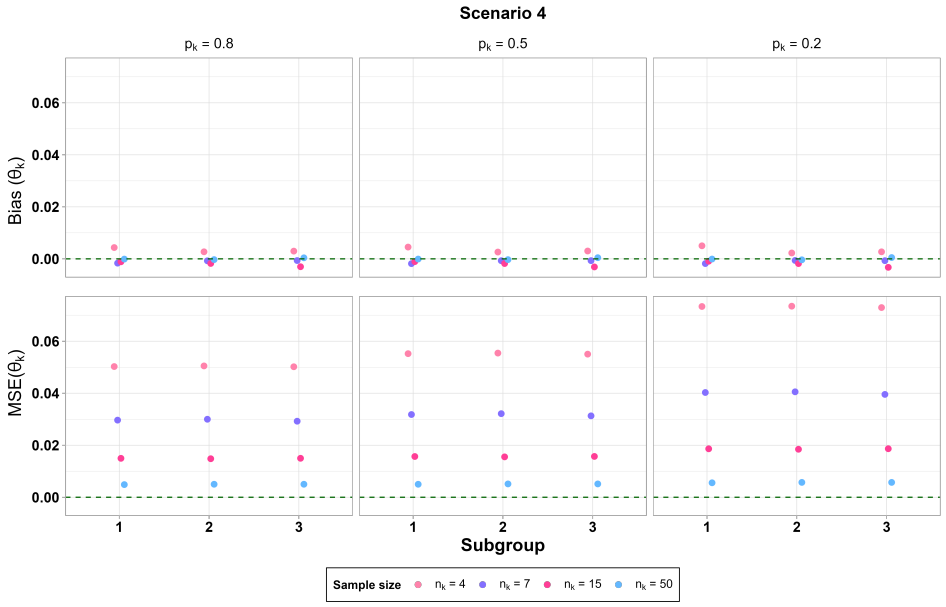


Figure 5.2.4: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 4 ($\theta_k = 0\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

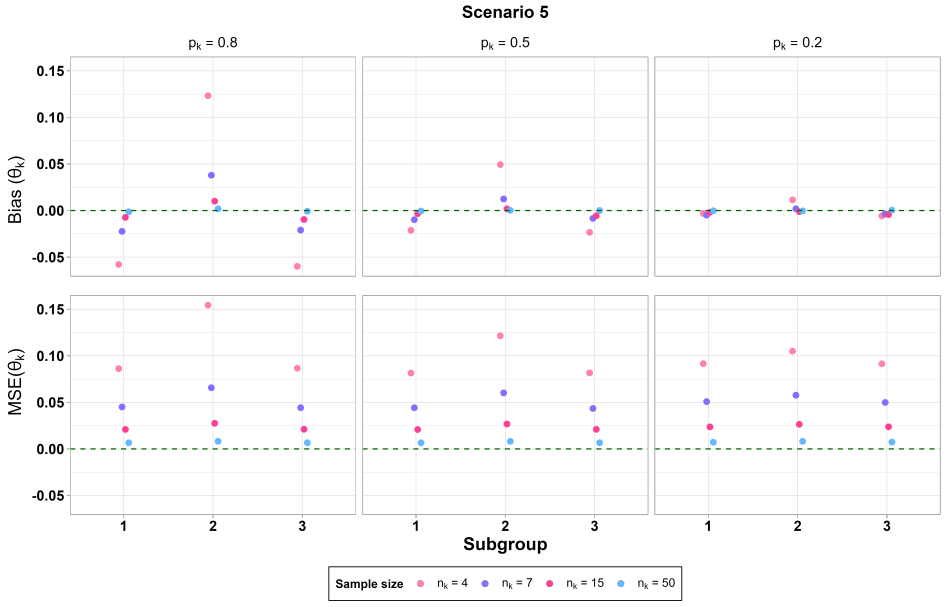


Figure 5.2.5: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 5 ($\theta_{1,3} = 3$ and $\theta_2 = 0.5$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

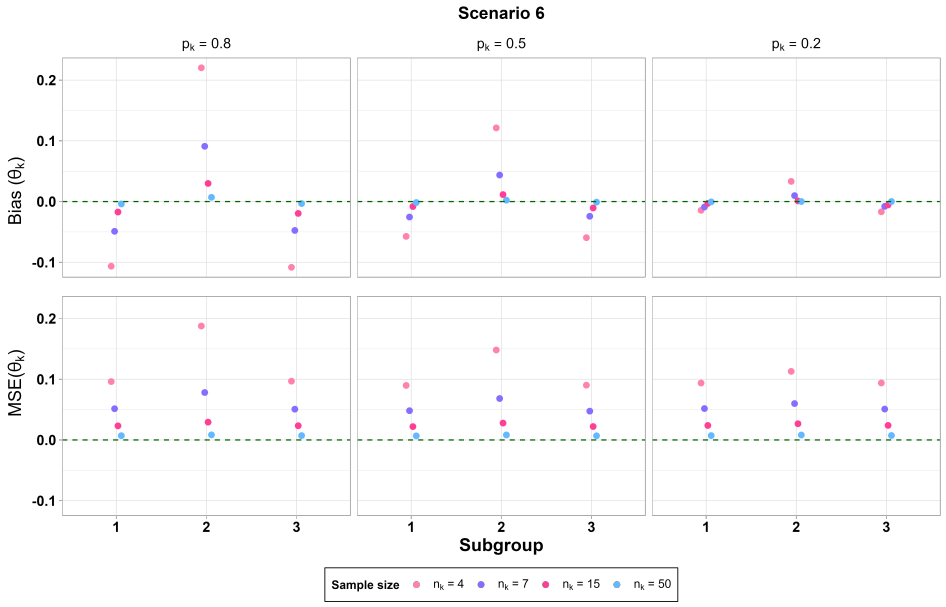


Figure 5.2.6: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

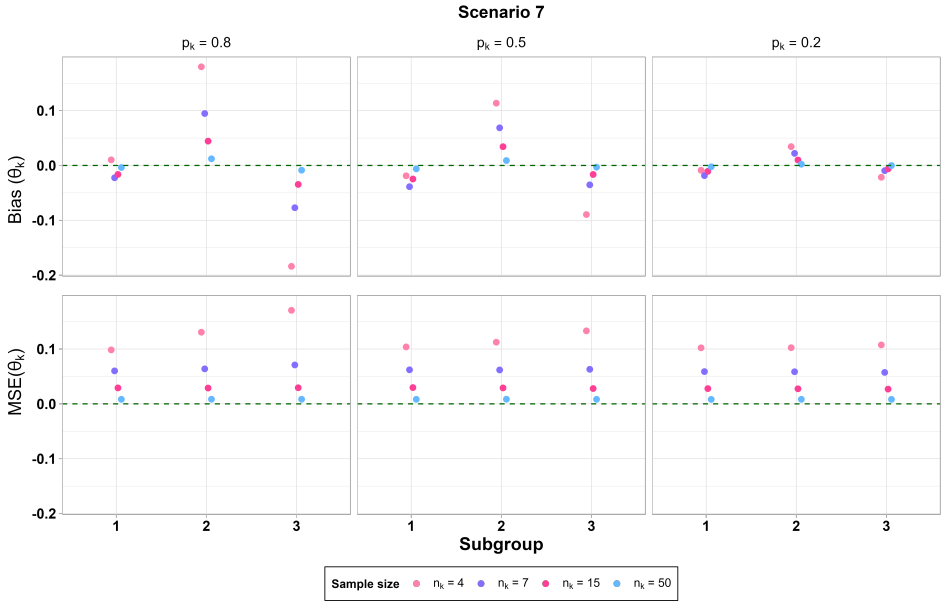


Figure 5.2.7: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 7($\theta_1 = 1.2$, $\theta_2 = 0.5$ and $\theta_3 = 3$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

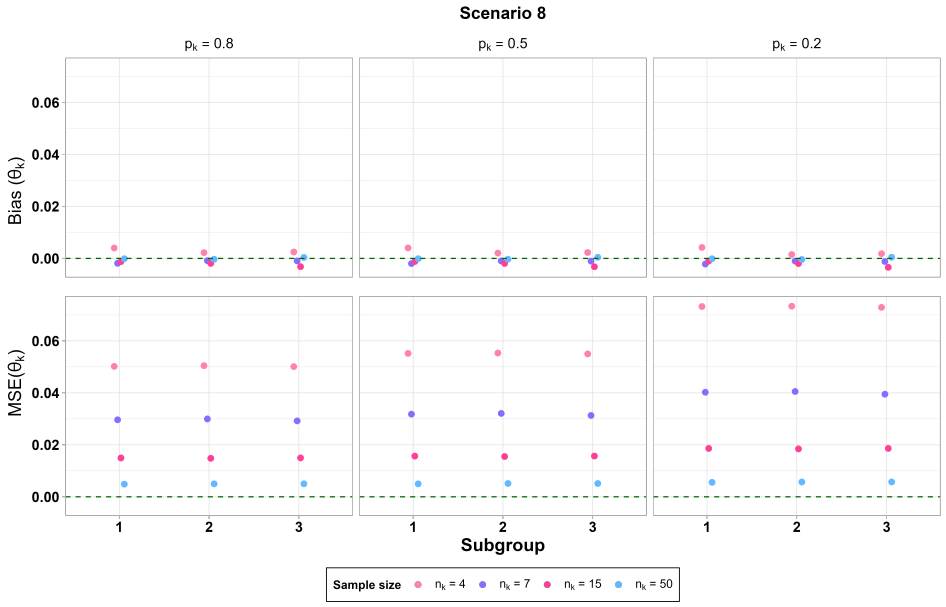


Figure 5.2.8: Bias and MSE according to different sample sizes (4, 7, 15 and 50) in Scenario 8 ($\theta_k = 1\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively).

The median posterior probabilities with p_k weights equal to 0.8 and 0.2 do not change with respect to the EXNEX results in Scenario 1, 3, 4 and 8, see Table 5.2.1. The remaining results are reported in Table 5.2.2, with EX identifying the scenarios with prior weight $p_k = 0.8$ and NEX identifying those $p_k = 0.2$.

Table 5.2.2: A comparison in terms of median posterior probability according to different sample sizes (4, 7, 15 and 50) in Scenarios 2, 5, 6 and 7 considering different p_k weights (0.8, 0.5 and 0.2, respectively). EX are the scenarios with prior weight $p_k = 0.8$ and NEX those with $p_k = 0.2$.

Sample	Method	Subgroup	PP_2 (%)	PP_5 (%)	PP_5 (%)	PP_7 (%)
4	EX	1	79	100	100	72
4	EX	2	79	14	86	17
4	EX	3	79	100	100	100
4	EXNEX	1	79	100	100	68
4	EXNEX	2	78	9	80	13
4	EXNEX	3	78	100	100	100

Continued on next page

Table 5.2.2 – continued from previous page

Sample	Method	Subgroup	PP_2 (%)	PP_5 (%)	PP_6 (%)	PP_7 (%)
4	NEX	1	76	100	100	71
4	NEX	2	76	8	75	9
4	NEX	3	76	100	100	100
7	EX	1	85	100	100	76
7	EX	2	85	4	87	6
7	EX	3	85	100	100	100
7	EXNEX	1	84	100	100	73
7	EXNEX	2	84	3	83	5
7	EXNEX	3	85	100	100	100
7	NEX	1	83	100	100	77
7	NEX	2	82	3	80	3
7	NEX	3	83	100	100	100
15	EX	1	93	100	100	86
15	EX	2	93	0	91	0
15	EX	3	93	100	100	100
15	EXNEX	1	92	100	100	84
15	EXNEX	2	92	0	90	0
15	EXNEX	3	92	100	100	100
15	NEX	1	91	100	100	87
15	NEX	2	91	0	89	0
15	NEX	3	91	100	100	100
50	EX	1	99	100	100	98
50	EX	2	99	0	99	0
50	EX	3	99	100	100	100
50	EXNEX	1	99	100	100	98
50	EXNEX	2	99	0	99	0
50	EXNEX	3	99	100	100	100
50	NEX	1	99	100	100	99
50	NEX	2	99	0	99	0
50	NEX	3	99	100	100	100

Continued on next page

Table 5.2.2 – continued from previous page

Sample	Method	Subgroup	PP_2 (%)	PP_5 (%)	PP_6 (%)	PP_7 (%)

We show in Figure 5.2.9-5.2.13 the power in Scenario 1, 2 and 5-7 considering different p_k weights.

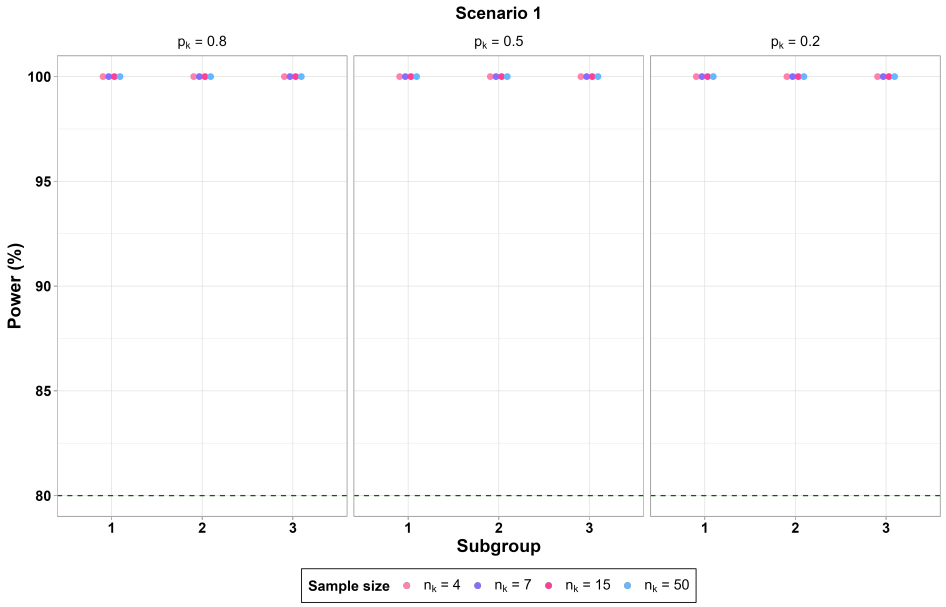


Figure 5.2.9: Power according to different sample sizes (4, 7, 15 and 50) in Scenario 1 ($\theta_k = 3\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively). The green dashed line indicates 80% power.

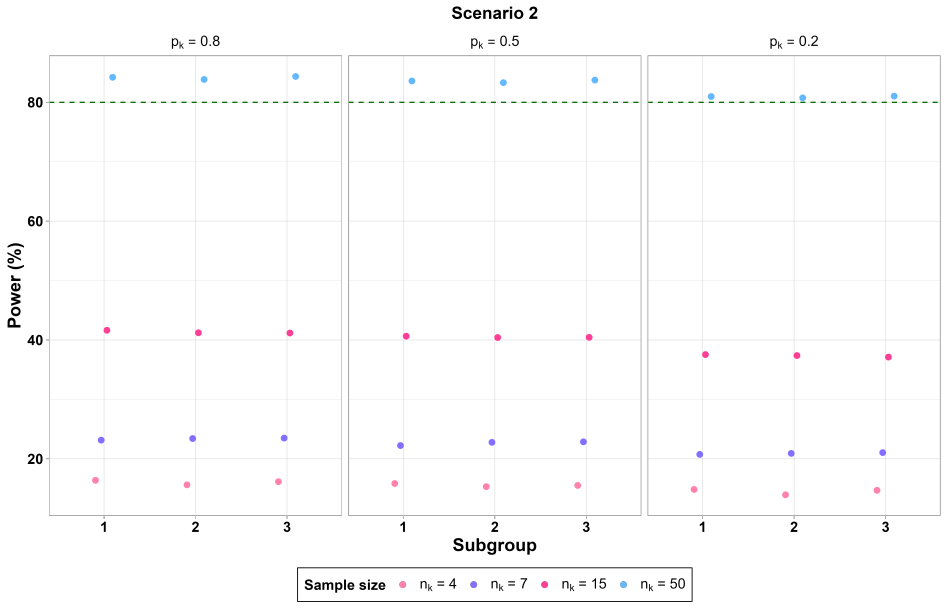


Figure 5.2.10: Power according to different sample sizes (4, 7, 15 and 50) in Scenario 2 ($\theta_k = 1.2\ \forall k$), considering different p_k weights (0.8, 0.5 and 0.2, respectively). The green dashed line indicates 80% power.

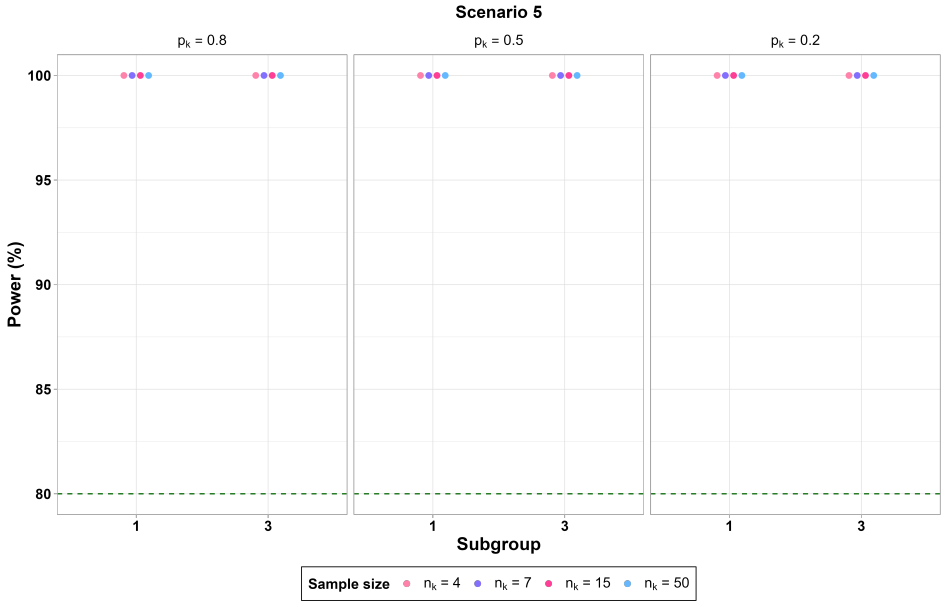


Figure 5.2.11: Power according to different sample sizes (4, 7, 15 and 50) in Scenario 5 ($\theta_{1,3} = 3$), considering different p_k weights (0.8, 0.5 and 0.2, respectively). The green dashed line indicates 80% power.

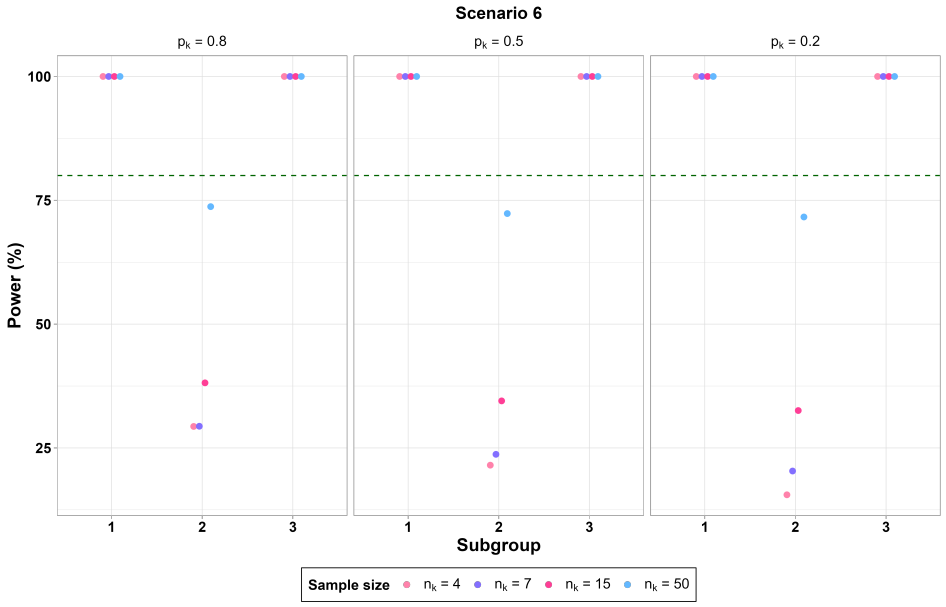


Figure 5.2.12: Power according to different sample sizes (4, 7, 15 and 50) in Scenario 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$), considering different p_k weights (0.8, 0.5 and 0.2, respectively). The green dashed line indicates 80% power.

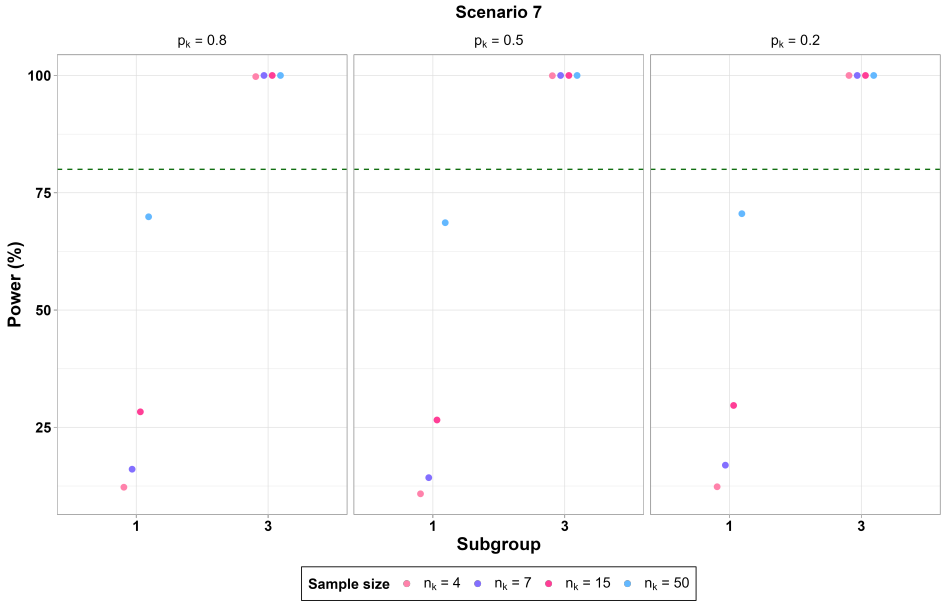


Figure 5.2.13: Power according to different sample sizes (4, 7, 15 and 50) in Scenario 7 ($\theta_1 = 1.2$ and $\theta_3 = 3$), considering different p_k weights (0.8, 0.5 and 0.2, respectively). The green dashed line indicates 80% power.

The type I errors with p_k weights equal to 0.8 and 0.2 in Scenario 8 are reported in Table 5.2.3, they are still well controlled but a little bit more high under the assumption of noexchangeability.

Table 5.2.3: A comparison in terms of type I error according to different sample sizes (4, 7, 15 and 50) in Scenario 8, considering different p_k weights (0.8, 0.5 and 0.2, respectively). EX are the scenarios with prior weight $p_k = 0.8$ and NEX those with $p_k = 0.2$

Sample	Method	Subgroup	Type I error (%)
4	EX	1	3.08
4	EX	2	3.11
4	EX	3	3.27
4	EXNEX	1	3.10
4	EXNEX	2	3.22
4	EXNEX	3	3.33
4	NEX	1	3.46
4	NEX	2	3.71
4	NEX	3	3.68
7	EX	1	2.70
7	EX	2	3.01
7	EX	3	2.78
7	EXNEX	1	2.84
7	EXNEX	2	3.07
7	EXNEX	3	2.84
7	NEX	1	3.28
7	NEX	2	3.57
7	NEX	3	3.33
15	EX	1	3.07
15	EX	2	2.94
15	EX	3	2.95
15	EXNEX	1	3.05
15	EXNEX	2	3.20

Continued on next page

Table 5.2.3 – continued from previous page

Sample	Method	Subgroup	Type I error (%)
15	EXNEX	3	2.91
15	NEX	1	3.60
15	NEX	2	3.32
15	NEX	3	3.42
50	EX	1	3.17
50	EX	2	3.20
50	EX	3	3.77
50	EXNEX	1	3.27
50	EXNEX	2	3.27
50	EXNEX	3	3.73
50	NEX	1	3.56
50	NEX	2	3.46
50	NEX	3	4.01

5.3 Supplementary material chapter 3

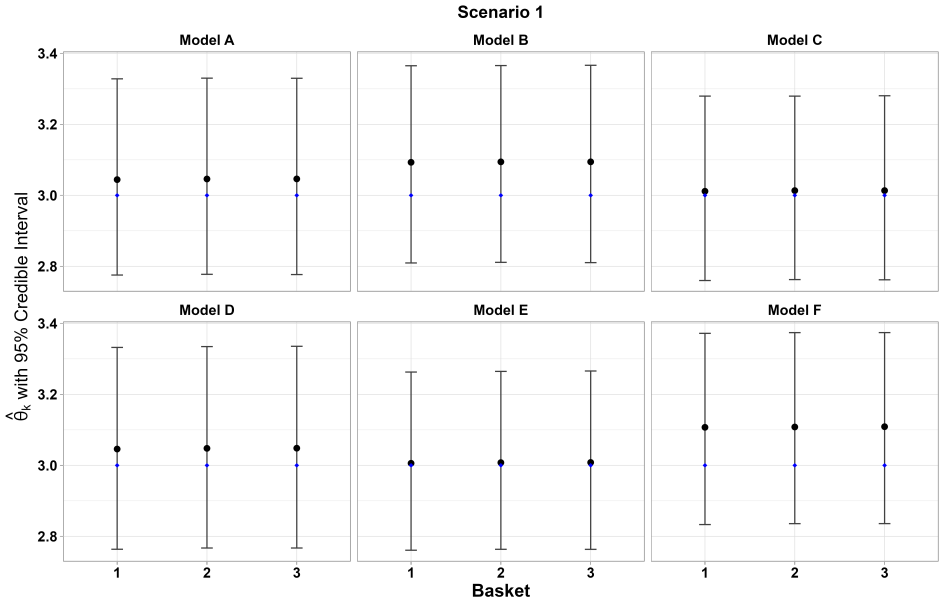


Figure 5.3.1: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 1 ($\theta_k = 3\ \forall k$) across the weights models.

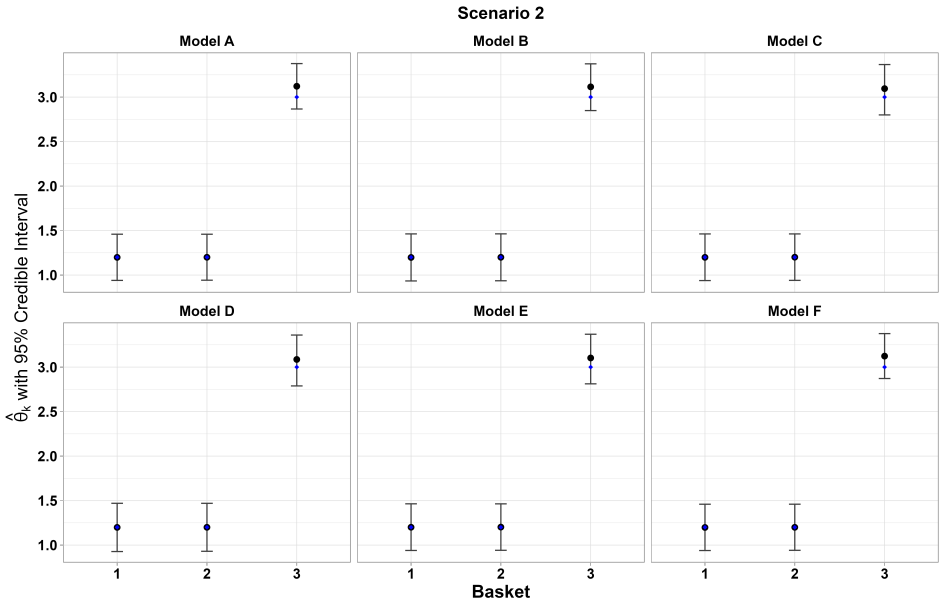


Figure 5.3.2: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 2 ($\theta_k = 1.2\ \forall k$) across the weights models.

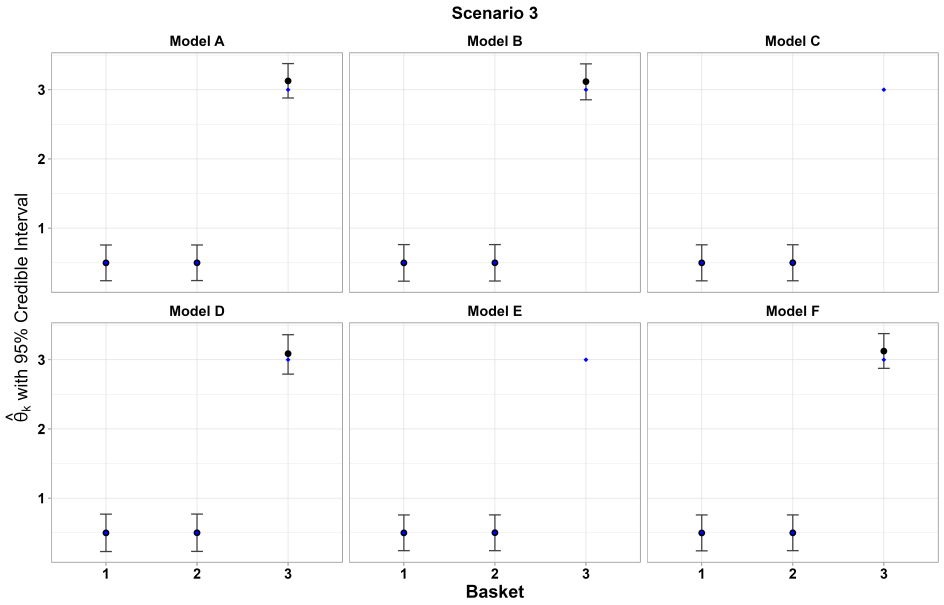


Figure 5.3.3: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 3 ($\theta_k = 0.5\ \forall k$) across the weights models.

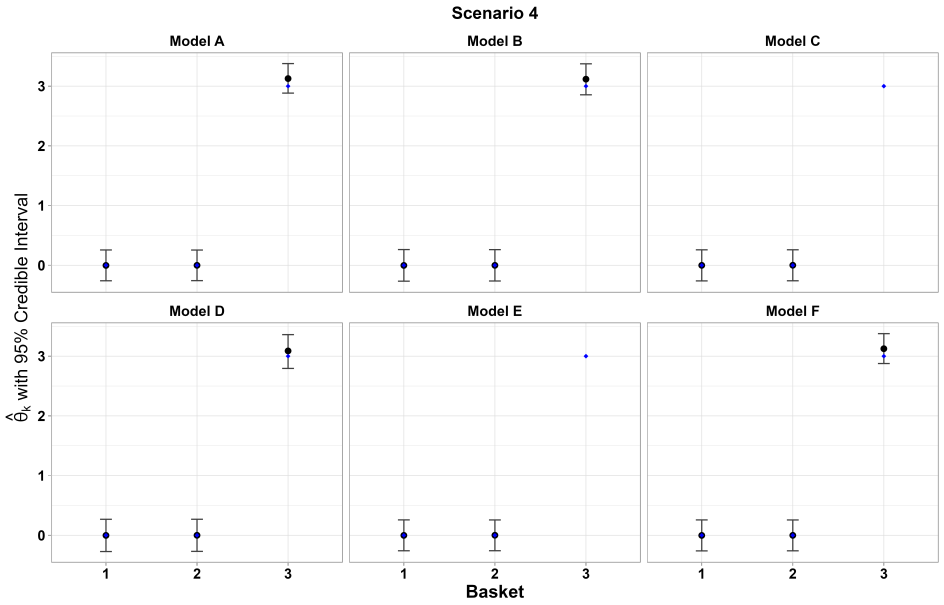


Figure 5.3.4: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 4 ($\theta_k = 0\ \forall k$) across the weights models.

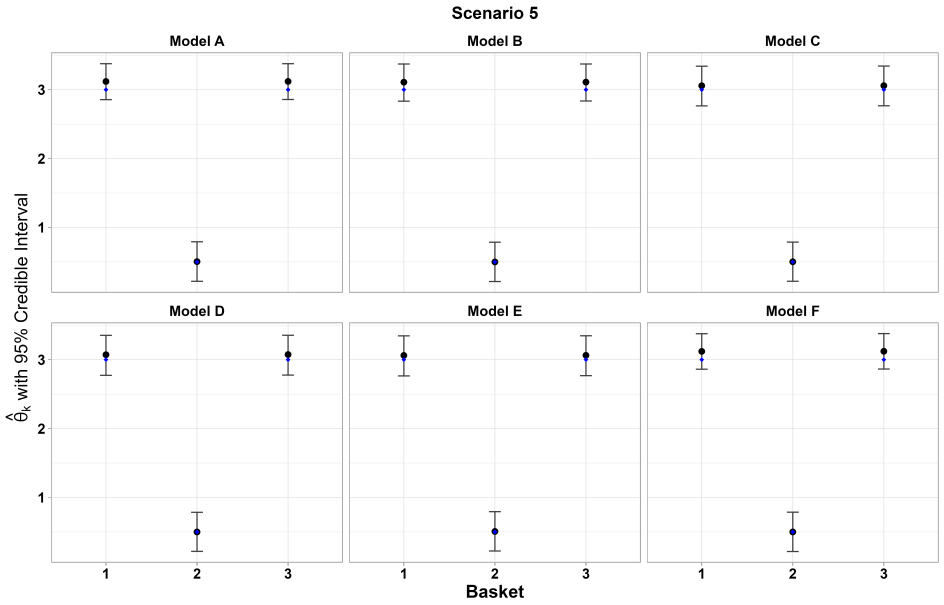


Figure 5.3.5: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 5 ($\theta_{1,3} = 3$ and $\theta_2 = 0.5$) across the weights models.

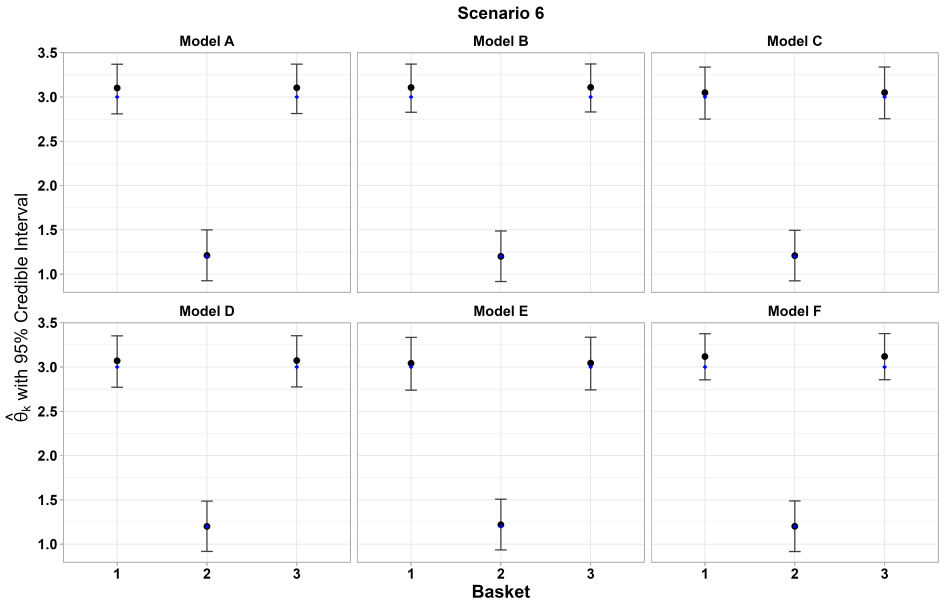


Figure 5.3.6: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 6 ($\theta_{1,3} = 3$ and $\theta_2 = 1.2$) across the weights models.

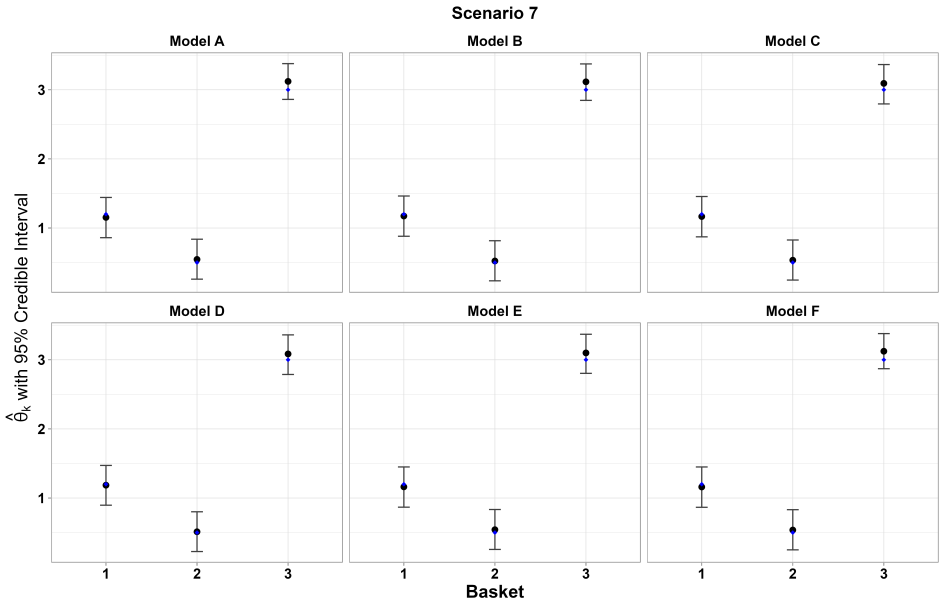


Figure 5.3.7: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 7 ($\theta_1 = 1.2$, $\theta_2 = 0.5$ and $\theta_3 = 3$) across the weights models.

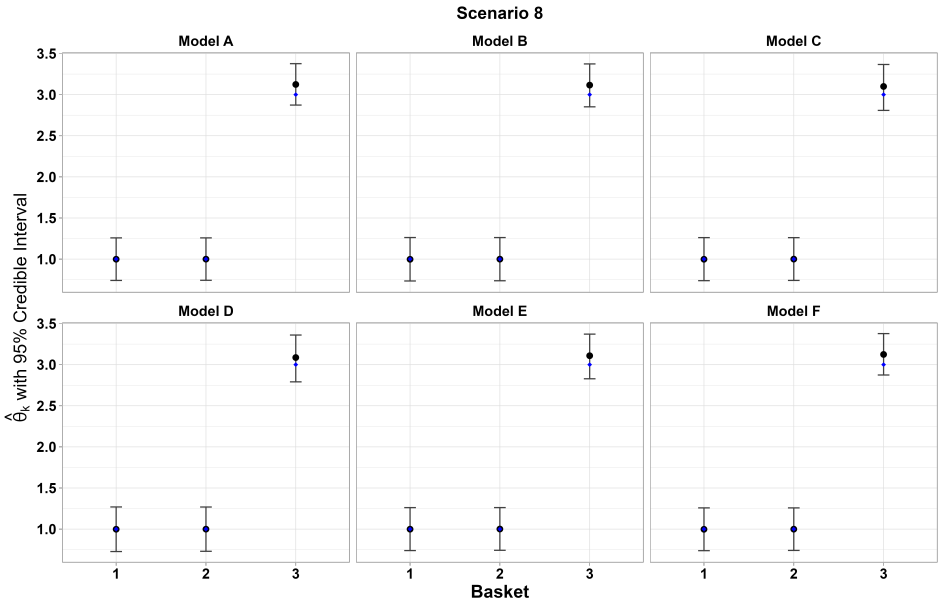


Figure 5.3.8: Median of posterior estimate of the treatment effects and 95% of credible interval in Scenario 8 ($\theta_k = 1\ \forall k$) across the weights models.

5.3.1 Comparison results in terms of different q values.

Table 5.3.1: Comparison of bias and MSE in Scenario 5 Model E, according to different q values, 0.42 (Stage II) and 0.50 (Stage I), respectively.

Subgroup	Stage	bias	mse
1	Stage II	0.05	0.02
1	Stage I	0.09	0.03
2	Stage II	0.01	0.02
2	Stage I	0.00	0.02

Table 5.3.2: Comparison of Bias and MSE, according to different q values, 0.42 (Stage II) and 0.58 (Stage I), respectively.

Scenario	Model	Subgroup	Stage	Bias	MSE
3	Model A	1	Stage II	-0.00	0.01
3	Model A	1	Stage I	0.00	0.01

Continued on next page

Table 5.3.2 – continued from previous page

Scenario	Model	Subgroup	Stage	Bias	MSE
3	Model A	2	Stage II	-0.00	0.01
3	Model A	2	Stage I	0.00	0.01
3	Model A	3	Stage II	0.12	0.03
3	Model A	3	Stage I		
3	Model D	1	Stage II	-0.00	0.02
3	Model D	1	Stage I	0.00	0.02
3	Model D	2	Stage II	-0.00	0.02
3	Model D	2	Stage I	0.00	0.02
3	Model D	3	Stage II	0.08	0.03
3	Model D	3	Stage I		
4	Model A	1	Stage II	-0.00	0.01
4	Model A	1	Stage I	0.00	0.01
4	Model A	2	Stage II	-0.00	0.01
4	Model A	2	Stage I	0.00	0.01
4	Model A	3	Stage II	0.13	0.03
4	Model A	3	Stage I		
4	Model D	1	Stage II	-0.00	0.02
4	Model D	1	Stage I	0.00	0.02
4	Model D	2	Stage II	-0.00	0.02
4	Model D	2	Stage I	0.00	0.02
4	Model D	3	Stage II	0.08	0.03
4	Model D	3	Stage I		
5	Model C	1	Stage II	0.05	0.02
5	Model C	1	Stage I	0.08	0.03
5	Model C	2	Stage II	0.00	0.02
5	Model C	2	Stage I	0.00	0.02
5	Model C	3	Stage II	0.05	0.02
5	Model C	3	Stage I		
7	Model C	1	Stage II	-0.03	0.02
7	Model C	1	Stage I	-0.04	0.02

Continued on next page

Table 5.3.2 – continued from previous page

Scenario	Model	Subgroup	Stage	Bias	MSE
7	Model C	2	Stage II	0.03	0.02
7	Model C	2	Stage I	0.04	0.02
7	Model C	3	Stage II	0.08	0.03
7	Model C	3	Stage I		
7	Model D	1	Stage II	-0.02	0.02
7	Model D	1	Stage I	-0.01	0.02
7	Model D	2	Stage II	0.01	0.02
7	Model D	2	Stage I	0.01	0.02
7	Model D	3	Stage II	0.07	0.03
7	Model D	3	Stage I		
7	Model E	1	Stage II	-0.04	0.02
7	Model E	1	Stage I	-0.05	0.02
7	Model E	2	Stage II	0.04	0.02
7	Model E	2	Stage I	0.05	0.02
7	Model E	3	Stage II	0.08	0.03
7	Model E	3	Stage I		

Table 5.3.3: Comparison of bias and MSE, according to different q values, 0.42 (Stage II) and 0.60 (Stage I), respectively.

Scenario	Subgroup	Model	Stage	Bias	MSE
5	1	Model A	Stage II	0.11	0.03
5	1	Model A	Stage I	0.12	0.03
5	1	Model D	Stage II	0.06	0.03
5	1	Model D	Stage I	0.07	0.03
5	2	Model A	Stage II	0.00	0.02
5	2	Model A	Stage I	0.00	0.02
5	2	Model D	Stage II	-0.00	0.02
5	2	Model D	Stage I	0.00	0.02
5	3	Model A	Stage II	0.11	0.03

Continued on next page

Table 5.3.3 – continued from previous page

Scenario	Subgroup	Model	Stage	Bias	MSE
5	3	Model A	Stage I		
5	3	Model D	Stage II	0.06	0.03
5	3	Model D	Stage I		
7	1	Model D	Stage II	-0.02	0.02
7	1	Model D	Stage I	-0.01	0.02
7	2	Model D	Stage II	0.01	0.02
7	2	Model D	Stage I	0.01	0.02
7	3	Model D	Stage II	0.07	0.03
7	3	Model D	Stage I		
8	1	Model C	Stage II	-0.00	0.02
8	1	Model C	Stage I	0.00	0.01
8	2	Model C	Stage II	0.00	0.02
8	2	Model C	Stage I	0.00	0.01
8	3	Model C	Stage II	0.08	0.03
8	3	Model C	Stage I		
8	1	Model D	Stage II	-0.00	0.02
8	1	Model D	Stage I	0.00	0.02
8	2	Model D	Stage II	-0.00	0.02
8	2	Model D	Stage I	0.00	0.02
8	3	Model D	Stage II	0.07	0.03
8	3	Model D	Stage I		
8	1	Model E	Stage II	-0.00	0.02
8	1	Model E	Stage I	0.00	0.01
8	2	Model E	Stage II	0.00	0.02
8	2	Model E	Stage I	0.00	0.01
8	3	Model E	Stage II	0.09	0.03
8	3	Model E	Stage I		

Table 5.3.4: Comparison of the median of the posterior probabilities in Scenario 5 Model E, according to different q values, 0.42 (Stage II) and 0.50 (Stage I), respectively.

Stage	1	2
Stage II	1.00	0.00
Stage I	1.00	0.00

The third subgroup was not showed given that it was not present in the trial at Stage I. 1 and 2 are referred to the subgroup number.

Table 5.3.5: Comparison of the median of the posterior probabilities, according to different q values, 0.42 (Stage II) and 0.58 (Stage I), respectively.

Scenario	Model	Stage	1	2
3	Model A	Stage II	0.00	0.00
3	Model A	Stage I	0.00	0.00
3	Model D	Stage II	0.00	0.00
3	Model D	Stage I	0.00	0.00
4	Model A	Stage II	0.00	0.00
4	Model A	Stage I	0.00	0.00
4	Model D	Stage II	0.00	0.00
4	Model D	Stage I	0.00	0.00
5	Model C	Stage II	1.00	0.00
5	Model C	Stage I	1.00	0.00
7	Model C	Stage II	0.87	0.00
7	Model C	Stage I	0.86	0.00
7	Model D	Stage II	0.90	0.00
7	Model D	Stage I	0.90	0.00
7	Model E	Stage II	0.86	0.00
7	Model E	Stage I	0.85	0.00

The third subgroup was not shown given that it was not present in the trial at Stage I. 1 and 2 refer to the subgroup number.

Table 5.3.6: Comparison of the median of the posterior probabilities, according to different q values, 0.42 (Stage II) and 0.60 (Stage I), respectively.

Scenario	Model	Stage	1	2
5	Model A	Stage II	1.00	0.00
5	Model A	Stage I	1.00	0.00
5	Model D	Stage II	1.00	0.00
5	Model D	Stage I	1.00	0.00
7	Model D	Stage II	0.90	0.00
7	Model D	Stage I	0.90	0.00
8	Model D	Stage II	0.49	0.50
8	Model D	Stage I	0.49	0.50
8	Model C	Stage II	0.49	0.50
8	Model C	Stage I	0.49	0.50
8	Model E	Stage II	0.50	0.50
8	Model E	Stage I	0.50	0.50

The third subgroup was not shown given that it was not present in the trial at Stage I. 1 and 2 refer to the subgroup number.

Table 5.3.7: Power (%) in Scenario 5 Model E, according to different q values, 0.42 (Stage II) and 0.50 (Stage I), respectively.

Subgroup	Stage II	Stage I
1	100.00	100.00

The third subgroup was not showed given that it was not present in the trial at Stage I.

Table 5.3.8: Power (%), according to different q values, 0.42 (Stage II) and 0.58 (Stage I), respectively.

Scenario	Model	Subgroup	Stage II	Stage I
5	Model C	1	100.00	100.00
7	Model C	1	30.18	28.92
7	Model D	1	34.92	34.94
7	Model E	1	29.23	26.90

The third subgroup was not showed given that it was not present in the trial at Stage I.

Table 5.3.9: Power (%), according to different q values, 0.42 (Stage II) and 0.60 (Stage I), respectively.

Scenario	Model	Subgroup	Stage II	Stage I
5	Model A	1	100.00	100.00
5	Model D	1	100.00	100.00
7	Model D	1	34.92	34.94

The third subgroup was not showed given that it was not present in the trial at Stage I.

Table 5.3.10: Comparison of Type I error (%) in Scenario 8, according to different q values, 0.42 (Stage II) and 0.58 (Stage I), respectively.

Scenario	Model	Subgroup	Stage II	Stage I
8	Model D	1	4.12	4.38
8	Model D	2	4.16	4.11
8	Model C	1	4.02	4.04
8	Model C	2	3.89	3.66
8	Model E	1	4.03	3.94
8	Model E	2	3.91	3.62

6 Appendix

In the following sections we report the BUGS code about the Bayesian models and the R functions used to compile the simulation studies described in Chapters 2 and 3. The *R2jags* package is required and we set the following seed: 12345.

6.1 R code for simulation study of Chapther 2

The operational characteristics regarding the scenarios are reported in Section 2.3.

6.1.1 BUGS code and R function of the BHM

```
model {
        ## sampling
        for (i in 1:N){
                    y[i] ~ dnorm(mu[BasketIndex[i]], invs2)

        }

        ## priors
        for (j in 1:n_basket){
                    mu[j] ~ dnorm(mu0, prec.sigma)
                 success[j]<- 1-step(delta[j]-mu[j])
        }

        ## Prior of s (sd of yi)
        invs2 <- pow(s, -2)
        s ~ dnorm(Prior.s[1],prec.s2)I(0.001,)
        prec.s2<- pow(Prior.s[2], -2)

        ## hyperpriors
        mu0 ~ dnorm (HPrior.mu[1], prec.mu)
        prec.mu <- pow(HPrior.mu[2], -2)
```

```
        prec.sigma <- pow(sigma, -2)
        sigma ~ dnorm(0, prec.HPrior.sigma)I(0.001,)
        prec.HPrior.sigma<- pow(HPrior.sigma, -2)
}


##Function for Basket trial with Continuous outcome in BHM
BasketHM_c<-function(n_basket,n,delta,mean_y,sd_y){

  result<-list()

  y<-NULL
  #Assembling of true responses
  for (i in 1:n_basket) {
    y<-c(y,rnorm(n[i], mean = mean_y[i], sd = sd_y[i]))
  }

  N = sum(n)
  BasketIndex=rep(seq(1,n_basket),n)

  #Hyperprior of mu and sigma
  HPrior.mu = c(0, 10)
  HPrior.sigma = 0.5

  #prior of s_k
  Prior.s = c(0,1)

  inits <- function(){
    list(
      s=1,
      mu0 = 0,
      sigma = 1
    )
```

```
  }

  data<-list("n_basket","N","y","delta","BasketIndex",
             "HPrior.mu","HPrior.sigma","Prior.s")

  resultBHM_c<-jags(data = data,
  parameters.to.save = c("mu","success","s"),
                     inits = inits, model.file = "BHM_c.txt",
                     n.chains = 2, n.burnin = 3000,
                     n.iter = 13000)

  #cat("Results:")
  result<-resultBHM_c$BUGSoutput$summary

  return(result)

}
```

6.1.2 BUGS code and R function of the EXNEX method

```
model{
        ## sampling
        for (i in 1:N){
                  y[i] ~ dnorm(mu[BasketIndex[i]], invs2)

        }

        ## likelihood/sampling model for the basket trials:
        for(i in 1:n_basket){

        # pick up mu
        mu[i] <- mix.mu[exch.ind[i], i]
        success[i]<- 1-step(delta[i]-mu[i])
```

```
        # Assume one EX distribution
        mix.mu[1, i] <- mu0 + re.mu[i]
        re.mu[i] ~ dnorm(0, prec.re)

        # NEX distributions
        mix.mu[2, i] ~ dnorm(nex.mu, nex.prec.sig)

        # Prior mixture weight
        exch.ind[i] ~ dcat(pMix[1:2])

                for(j in 1:2){
                        each[i, j] <- equals(exch.ind[i], j)
                        }
        }

        ## Prior of s (sd of yi)
        invs2 <- pow(s, -2)
        s ~ dnorm(Prior.s[1],prec.s2)I(0.001,)
        prec.s2<- pow(Prior.s[2], -2)


        mu0 ~ dnorm (HPrior.mu[1], prec.mu.theta)
        prec.mu.theta <- pow(HPrior.mu[2], -2)

        prec.re <- pow(sigma, -2)
        sigma ~ dnorm(0, prec.sig)I(0.001,)
        prec.sig <- pow(HPrior.sigma, -2)

        nex.prec.sig <- pow(nex.sig, -2)
}
```

##Function for exnex with Continuous outcome in EXNEX

```
BasketEXNEX_c<-function(n_basket,n,delta,
mean_y,sd_y,pMix){

  result<-list()

  y<-NULL
  #Assembling of true responses
  for (i in 1:n_basket) {
    y<-c(y,rnorm(n[i], mean = mean_y[i], sd = sd_y[i]))
  }

  N = sum(n)
  BasketIndex=rep(seq(1,n_basket),n)

  #Hyperprior of mu and sigma
  HPrior.mu = c(0, 10)
  HPrior.sigma = 0.5

  #prior of s_k
  Prior.s = c(0,1)

  nex.mu = 0
  nex.sig = 10

  inits <- function(){
    list(
      s=1,
      mu0 = 0,
      sigma = 1
    )
  }

  data<-list("n_basket","N","y","BasketIndex","pMix","delta",
```

```
                "HPrior.mu","HPrior.sigma","Prior.s",
                "nex.mu","nex.sig")

  resultEXNEX_c<-jags(data = data,
  parameters.to.save = c("mu","success","s"),
                      inits = inits, model.file = "EXNEX_c.txt",
                      n.chains = 2, n.burnin = 3000,
                      n.iter = 13000)

  #cat("Results:")
  result<-resultEXNEX_c$BUGSoutput$summary

  return(result)

}
```

6.1.3 BUGS code and R function of the TRB method

```
model{
  for (i in 1:N) {

    #Likelihood
        y[i] ~ dnorm(mu[i], tau[BasketIndex[i]])
           mu[i] <- beta0[BasketIndex[i]]
  }

  # Priors
  for (k in 1:n_basket) {
    beta0[k] <- mu.beta0 + re.beta0[k]
    re.beta0[k] ~ dnorm(0, prec.re.beta0)

    tau[k] ~ dnorm(Prior.s[1],prec.s2)I(0.001,)
```

```
  }
  prec.s2 <- pow(Prior.s[2], -2)

  mu.beta0 ~ dnorm(prior.mu.beta0[1], prec.mu.beta0)
  prec.mu.beta0 <- pow(prior.mu.beta0[2], -2)

  prec.re.beta0 <- pow(sigma, -2)
  sigma ~ dnorm(0, prec.sig) T(0.01, )
  prec.sig <- pow(prior.sig.HN, -2)

for (j in 1:n_basket) {

    for (q in 1:(n_basket)) { #5
      phiH[q, j] <- sqrt(1/2*pow(sqrt(abs(beta0[ j])) -
      sqrt(abs(beta0[Gmodule[q]])), 2))

      norm.phiH[q, j] <- exp(-phiH[q, j]/s0)

      wss[q, j] <- phiH[q, j] - equals(Gmodule[q], j)

      D01[q, j] ~ dunif(0.1, 1)
      nu0[q, j] <- lSlab + (uSlab - lSlab)*1/(0.0001+wss[q, j])*D01[q, j]
      nu[q, j] <- nu0[q, j]*step(wss[q, j] - nu0[q, j]) +
      spike*step(nu0[q, j] - wss[q, j])

      pred.mu.theta[q, j] <- beta0[j] + cms.mu[q, j]*1/nu[q, j]*
      (1 - equals(Gmodule[q], j))
      cms.mu[q, j] ~ dnorm(0, 1)
    }

  }
```

```
  # We restart another j loop here
  for (m in 1:(n_basket)) {
    sum.norm.phiH[m] <- sum(norm.phiH[m,]) - 1
  }

  for (j in 1:n_basket) {

    for(q in 1:(n_basket)){
      p0[q, j] <- norm.phiH[q, j]/(0.00001+sum.norm.phiH[q])
      V[q, j] <- 1 - equals(Gmodule[q], j)
      pmix[q, j] <- inprod(p0[q, j], V[q, j])
      r.theta.star[q, j] <- inprod(pmix[q, j], pred.mu.theta[q, j])
    }
  }

  for (i in 1:(n_basket)) {
    theta[i] <- sum(r.theta.star[i, ])
  }


  #Posterior prediction
      for (i in 1:n_basket) {
      success[i] <- 1 - step(delta[i] - theta[i])
  }

}
```

```
TRB_onearm <- function(true.beta0, Module, Trt,
cutoff.theta,n_basket,n,delta){

  N = sum(n)
```

```
BasketIndex=rep(seq(1,n_basket),n)


y <- numeric(length = N)
#simulate of true response y:
for (i in 1:N) {
  y[i] <- true.beta0[BasketIndex[i]]  +
  rnorm(1, mean = 0, sd = 0.4)
}


inits <- function(){
  list(
    tau = rep(10,n_basket),
    sigma = 1
  )
}

#prior of s_k
Prior.s = c(0,1)

s0 = 0.15; lSlab = 0.01;  uSlab = 1;  spike = 100
prior.mu.beta0 = c(0, 5);  prior.sig.HN = 0.5#1
Gmodule=c(1,2,3,4,5)

data <- list("n_basket", "N", "y", "Gmodule",
"BasketIndex",
             "s0", "lSlab", "uSlab", "spike",
             "prior.mu.beta0",
             "prior.sig.HN", "delta","Prior.s")

parameters <- c("theta", "success", "phiH")
```

```
  TRB_onearm <- jags(data = data, inits = inits,
  parameters.to.save = parameters,
  model.file = "TRB_singlearm.txt",
                    n.chains = 2, n.burnin = 3000,
                    n.iter = 13000)

  return(summarystats = TRB_onearm$BUGSoutput$summary)
}
```

6.2 R code for simulation study of Chapter 3

The operational characteristics regarding the scenarios are reported in Section 3.4. We now report the Bugs code and the R function to simulate the proposed innovative sequential platform-basket design.

```
model {
        ## responses
        for (i in 1:N){
                  y[i] ~ dnorm(theta[BasketIndex[i]], invs2)

        }

        ## priors of theta
        for (k in 1:n_basket){
                   mix.Nor[k,1] ~ dnorm(mu, prec.sigma)
                mix.Nor[k,2] ~ dnorm(mu0, prec.sigma0)
                mix.Nor[k,3] ~ dnorm(m0, prec.s0)

                theta[k] <- mix.Nor[k,which[k]]
                which[k] ~ dcat(wMix[k,1:3])

                success[k]<- 1-step(delta[k]-theta[k])
        }
```

```
        ## Prior of s (sd of yi)
        invs2 <- pow(s, -2)
        s ~ dnorm(Prior.s[1],prec.s2)I(0.001,)
        prec.s2<- pow(Prior.s[2], -2)

        ## Construction of theta by mixed normal distributions:

        ## hyperpriors of mu and sigma:
        mu ~ dnorm (HPrior.mu[1], prec.mu)
        prec.mu <- pow(HPrior.mu[2], -2)

        prec.sigma <- pow(sigma, -2)
        sigma ~ dnorm(0, prec.HPrior.sigma)I(0.001,)
        prec.HPrior.sigma<- pow(HPrior.sigma, -2)

        ##precision of sig0 from co-data
        prec.sigma0<- pow(sigma0, -2)

        ##precision of s0 from robust prior
        prec.s0<- pow(s0, -2)

}



##Function for Sequential Platform-Basket trial with Co-data
BasketHM_HD<-function(n_basket,n,wMix,delta,mean_y,sd_y,
mu0,sigma0,m0,s0){

  result<-list()

  y<-NULL
  #Assembling of true responses
```

```
for (i in 1:n_basket) {
  y<-c(y,rnorm(n[i], mean = mean_y[i], sd = sd_y[i]))
}

N = sum(n)
BasketIndex=rep(seq(1,n_basket),n)

#Hyperprior of mu and sigma
HPrior.mu = c(0, 10)
HPrior.sigma = 0.5

#prior of s_k
Prior.s = c(0,1)

inits <- function(){
  list(
    s=1,
    mu = 0,
    sigma = 1
  )
}


data<-list("n_basket","N","y","wMix","delta",
"BasketIndex", "HPrior.mu","HPrior.sigma","mu0",
"sigma0","m0","s0","Prior.s")

resultBHM_HD<-jags(data = data,
parameters.to.save = c("theta", "success","s"),
                   inits = inits, model.file = "MyMod.txt",
                   n.chains = 2, n.burnin = 3000,
                   n.iter = 13000)
```

```
  #cat("Results:")
  result$interim<-resultBHM_HD$BUGSoutput$summary

  return(result)

}
```

References

[1] S. M. Berry, K. R. Broglio, S. Groshen, and D. A. Berry, "Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase ii oncology clinical trials," in *Clinical Trials*, vol. 10, pp. 720–734, 10 2013.

[2] B. Neuenschwander, S. Wandel, S. Roychoudhury, and S. Bailey, "Robust exchangeability designs for early phase clinical trials with multiple strata," *Pharmaceutical Statistics*, vol. 15, pp. 123–134, 3 2016.

[3] L. O. Ouma, M. J. Grayling, J. M. Wason, and H. Zheng, "Bayesian modelling strategies for borrowing of information in randomised basket trials," *Journal of the Royal Statistical Society. Series C: Applied Statistics*, vol. 71, pp. 2014–2037, 11 2022.

[4] J. W. Dear, P. Lilitkarntakul, and D. J. Webb, "Are rare diseases still orphans or happily adopted? the challenges of developing and using orphan medicinal products," 9 2006.

[5] I. HA, "Human cancer classification: a systems biology- based model integrating morphology, cancer stem cells, proteomics, and genomics.," *J Cancer.*, vol. 2, p. 107–15, 2011.

[6] J. Gobburu and D. Pastoor, "Drugs against rare diseases: Are the regulatory standards higher?," 10 2016.

[7] J. Maca, S. Bhattacharya, V. Dragalin, P. Gallo, and M. Krams, "Adaptive seamless phase ii/iii designs—background, operational aspects, and examples," *Drug Information Journal*, vol. 40, no. 4, pp. 463–473, 2006.

[8] W. Zhang, F. Yan, F. Chen, and S.-C. Chow, "Advanced statistics in regulatory critical clinical initiatives," 2016.

[9] S. Day, A. H. Jonker, L. P. L. Lau, R. D. Hilgers, I. Irony, K. Larsson, K. C. Roes, and N. Stallard, "Recommendations for the design of small population clinical trials," *Orphanet Journal of Rare Diseases*, vol. 13, 11 2018.

[10] R. Ferla, F. Dell'Aquila, M. Doria, M. Ferraiuolo, A. Noto, F. Grazioli, V. Ammendola, F. Testa, P. Melillo, C. Iodice, G. Risca, N. Tedesco, P. le Brun, E. Surace, F. Simonelli, S. Galimberti, M. Valsecchi, J.-B. Marteau, P. Veron, S. Colloca, and A. Auricchio, "Efficacy, pharmacokinetics, and safety in the mouse and primate retina of dual aav vectors for usher syndrome type 1b," *Molecular Therapy - Methods and Clinical Development*, vol. 28, 2023.

[11] G. Consiglieri, F. Tucci, M. D. Pellegrin, B. Guerrini, A. Cattoni, G. Risca, S. Scarparo, M. Sarzana, S. Pontesilli, R. Mellone, S. Gasperini, S. Galimberti, P. Silvani, C. Filisetti, S. Darin, G. Forni, S. Miglietta, L. Santi, M. Facchini, A. Corti, F. Fumagalli, M. P. Cicalese, V. Calbi, M. Migliavacca, F. Barzaghi, F. Ferrua, V. Gallo, S. Recupero, D. Canarutto, M. Doglio, L. Tedesco, N. Volpi, A. Rovelli, G. L. Marca, M. G. Valsecchi, S. Zancan, F. Ciceri, L. Naldini, C. Baldoli, R. Parini, B. Gentner, A. Aiuti, and M. E. Bernardo, "Early skeletal outcomes after hematopoietic stem and progenitor cell gene therapy for hurler syndrome," *Sci. Transl. Med*, vol. 16, p. 8214, 5 2024.

[12] S. A. Bell and C. T. Smith, "A comparison of interventional clinical trials in rare versus non-rare diseases: an analysis of clinicaltrials.gov," *Orphanet journal of rare diseases*, vol. 9, p. 170, 11 2014.

[13] F. Lasch, K. Weber, M. M. Chao, and A. Koch, "A plea to provide best evidence in trials under sample-size restrictions: the example of pioglitazone to resolve leukoplakia and erythroplakia in fanconi anemia patients," *Orphanet Journal of Rare Diseases*, vol. 12, 5 2017.

[14] F. Testa, E. Carreño, L. I. van den Born, P. Melillo, I. Perea-Romero, V. D. Iorio, G. Risca, C. M. Iodice, R. J. Pennings, M. Karali, S. Banfi, A. Auricchio, S. Galimberti, C. Ayuso, and F. Simonelli, "Multicentric longitudinal prospective study in a european cohort of myo7a patients: Disease course and implications for gene therapy," *Investigative Ophthalmology and Visual Science*, vol. 65, 6 2024.

[15] Y. Li and R. Izem, "Novel clinical trial design and analytic methods to tackle challenges in therapeutic development in rare diseases," *Annals of Translational Medicine*, vol. 10, pp. 1034–1034, 9 2022.

[16] B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, and D. J. Sargent, "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials," *Biometrics*, vol. 67, pp. 1047–1056, 2011.

[17] F. Miller, S. Zohar, N. Stallard, J. Madan, M. Posch, S. W. Hee, M. Pearce, M. Vågerö, and S. Day, "Approaches to sample size calculation for clinical trials in rare diseases," *Pharmaceutical Statistics*, vol. 17, pp. 214–230, 5 2018.

[18] M. Ursino and N. Stallard, "Bayesian approaches for confirmatory trials in rare diseases: Opportunities and challenges," 2 2021.

[19] S. Gupta, M. E. Faughnan, G. A. Tomlinson, and A. M. Bayoumi, "A framework for applying unfamiliar trial designs in studies of rare diseases," 10 2011.

[20] A. Carlier, A. Vasilevich, M. Marechal, J. D. Boer, and L. Geris, "In silico clinical trials for pediatric orphan diseases," *Scientific Reports*, vol. 8, 12 2018.

[21] P. G. Casali, P. Bruzzi, J. Bogaerts, J. Y. Blay, M. Aapro, A. Adamous, A. Berruti, J. Bressington, B. Bruzzi, R. Capocaccia, F. Cardoso, J. E. Celis, A. Cervantes, F. Ciardiello, C. Claussen, M. Coleman, S. Comis, S. Craine, D. D. Boltz, F. D. Lorenzo, A. P. D. Tos, G. Gatta, J. Geissler, R. Giuliani, E. Grande, A. Gronchi, S. Jezdic, B. Jonsson, L. Jost, H. Keulen, D. Lacombe, G. Lamory, Y. L. Cam, S. L. di Priolo, L. Licitra, F. Macchia, A. Margulies, S. Marreaud, G. McVie, S. Narbutas, K. Oliver, N. Pavlidis, J. Pelouchova, G. Pentheroudakis, M. Piccart, M. A. Pierotti, G. Pravettoni, K. Redmond, P. Riegman, M. P. Ruffilli, D. Ryner, S. Sandrucci, M. Seymour, V. Torri, A. Trama, S. V. Belle, G. Vassal, M. Wartenberg, C. Watts,

A. Wilson, and W. Yared, "Rare cancers europe (rce) methodological recommendations for clinical studies in rare cancers: A european consensus position paper," 2 2015.

[22] U. Garczarek, N. Muehlemann, F. Richard, P. Yajnik, and E. Russek-Cohen, "Bayesian strategies in rare diseases," 5 2023.

[23] K. M. Kidwell, S. Roychoudhury, B. Wendelberger, J. Scott, T. Moroz, S. Yin, M. Majumder, J. Zhong, R. A. Huml, and V. Miller, "Application of bayesian methods to accelerate rare disease drug development: scopes and hurdles," 12 2022.

[24] L. V. Hampson, J. Whitehead, D. Eleftheriou, and P. Brogan, "Bayesian methods for the design and interpretation of clinical trials in very rare diseases," *Statistics in Medicine*, vol. 33, pp. 4186–4201, 10 2014.

[25] J. J. Lee and C. T. Chu, "Bayesian clinical trials in action," *Statistics in Medicine*, vol. 31, pp. 2955–2972, 11 2012.

[26] S. J. Wang, H. M. J. Hung, and R. T. O'Neill, "Adaptive patient enrichment designs in therapeutic trials," *Biometrical Journal*, vol. 51, pp. 358–374, 2009.

[27] W. Brannath, E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon, "Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology," *Statistics in Medicine*, vol. 28, pp. 1445–1463, 5 2009.

[28] B. DA, "Bayesian clinical trials.," *Nat Rev Drug Discov*, vol. 5, no. 1, pp. 27–36, 2006.

[29] J. H. van der Lee, J. Wesseling, M. W. Tanck, and M. Offringa, "Efficient ways exist to obtain the optimal sample size in clinical trials in rare diseases," 4 2008.

[30] N. Stallard and S. Todd, "Sequential designs for phase iii clinical trials incorporating treatment selection," 3 2003.

[31] R.-D. Hilgers, F. König, G. Molenberghs, and S. Senn, "Design and analysis of clinical trials for small rare disease populations," 2016.

[32] W. MN, H. SL, S. N, M. SJ, M.-H. N, and M. AS, "Statistical approaches for the integration of external controls in a cystic fibrosis clinical trial: a simulation and an application.," *Am J Epidemiol*, 2024.

[33] J. G. Ibrahim and M.-H. Chen, "Power prior distributions for regression models," 2000.

[34] H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander, "Robust meta-analytic-predictive priors in clinical trials with historical control information," *Biometrics*, vol. 70, pp. 1023–1032, 12 2014.

[35] B. P. Hobbs, D. J. Sargent, and B. P. Carlin, "Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models," *Bayesian Analysis*, vol. 7, pp. 639–674, 2012.

[36] B. Han, J. Zhan, Z. J. Zhong, D. Liu, and S. Lindborg, "Covariate-adjusted borrowing of historical control data in randomized clinical trials," *Pharmaceutical Statistics*, vol. 16, pp. 296–308, 7 2017.

[37] L. A. Renfro and D. J. Sargent, "Statistical controversies in clinical research: Basket trials, umbrella trials, and other master protocols: A review and examples," *Annals of Oncology*, vol. 28, pp. 34–43, 1 2017.

[38] J. Woodcock and L. M. LaVange, "Master protocols to study multiple therapies, multiple diseases, or both," *New England Journal of Medicine*, vol. 377, pp. 62–70, 7 2017.

[39] K. Strzebonska and M. Waligora, "Umbrella and basket trials in oncology: Ethical challenges," 8 2019.

[40] J. J. Park, E. Siden, M. J. Zoratti, L. Dron, O. Harari, J. Singer, R. T. Lester, K. Thorlund, and E. J. Mills, "Systematic review of basket trials, umbrella trials, and platform trials: A landscape analysis of master protocols," 9 2019.

[41] E. L. Meyer, P. Mesenbrink, C. Dunger-Baldauf, H. J. Fülle, E. Glimm, Y. Li, M. Posch, and F. König, *The Evolution of Master Protocol Clinical Trial Designs: A Systematic Literature Review*, vol. 42, pp. 1330–1360. Excerpta Medica Inc., 7 2020.

[42] E. Fountzilas, A. M. Tsimberidou, H. H. Vo, and R. Kurzrock, "Clinical trial design in the era of precision medicine," *Genome Medicine*, vol. 14, 12 2022.

[43] A. Kasim, N. Bean, S. J. Hendriksen, T. T. Chen, H. Zhou, and M. A. Psioda, "Basket trials in oncology: a systematic review of practices and methods, comparative analysis of innovative methods, and an appraisal of a missed opportunity," 2023.

[44] L. O. Ouma, J. Wason, and H. Zheng, "Design and analysis of umbrella trials: Where do we stand?," 2022.

[45] S. M. Berry, J. T. Connor, and R. J. Lewis, "The platform trial: An efficient strategy for evaluating multiple treatments," 4 2015.

[46] J. J. Park, N. Ford, D. Xavier, P. Ashorn, R. F. Grais, Z. A. Bhutta, H. Goossens, K. Thorlund, M. E. Socias, and E. J. Mills, "Randomised trials at the level of the individual," *The Lancet Global Health*, vol. 9, pp. e691–e700, 5 2021.

[47] J. J. H. Park, G. Hsu, E. G. Siden, K. Thorlund, and E. J. Mills, "An overview of precision oncology basket and umbrella trials for clinicians," *CA: A Cancer Journal for Clinicians*, vol. 70, pp. 125–137, 3 2020.

[48] J. M. Wason, J. E. Abraham, R. D. Baird, I. Gournaris, A. L. Vallier, J. D. Brenton, H. M. Earl, and A. P. Mander, "A bayesian adaptive design for biomarker trials with linked treatments," *British Journal of Cancer*, vol. 113, pp. 699–705, 9 2015.

[49] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. J. Lee, "Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine," *Clinical Trials*, vol. 5, pp. 181–193, 2008.

[50] E. S. Kim, R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, C. M. Alden, S. Liu, X. Tang, F. R. Khuri, H. T. Tran, B. E. Johnson, J. V. Heymach, L. Mao, F. Fossella, M. S. Kies, V. Papadimitrakopoulou, S. E. Davis, S. M. Lippman, and W. K. Hong, "The battle trial: Personalizing therapy for lung cancer," *Cancer Discovery*, vol. 1, pp. 44–53, 6 2011.

[51] S. M. Gold, M. B. Roig, J. J. Miranda, C. Pariante, M. Posch, and C. Otte, "Platform trials and the future of evaluating therapeutic behavioural interventions," 1 2022.

[52] J. J. Park, O. Harari, L. Dron, R. T. Lester, K. Thorlund, and E. J. Mills, "An overview of platform trials with a checklist for clinical readers," 9 2020.

[53] S. Ventz, B. M. Alexander, G. Parmigiani, R. D. Gelber, and L. Trippa, "Designing clinical trials that accept new arms: An example in metastatic breast cancer," *JOURNAL OF CLINICAL ONCOLOGY J Clin Oncol*, vol. 35, pp. 3160–3168, 2017.

[54] D. C. Angus, B. M. Alexander, S. Berry, M. Buxton, R. Lewis, M. Paoloni, S. A. Webb, S. Arnold, A. Barker, D. A. Berry, M. J. Bonten, M. Brophy, C. Butler, T. F. Cloughesy, L. P. Derde, L. J. Esserman, R. Ferguson, L. Fiore, S. C. Gaffey, J. M. Gaziano, K. Giusti, H. Goossens, S. Heritier, B. Hyman, M. Krams, K. Larholt, L. M. LaVange, P. Lavori, A. W. Lo, A. J. London, V. Manax, C. McArthur, G. O'Neill, G. Parmigiani, J. Perlmutter, E. A. Petzold, C. Ritchie, K. M. Rowan, C. W. Seymour, N. I. Shapiro, D. M. Simeone, B. Smith, B. Spellberg, A. D. Stern, L. Trippa, M. Trusheim, K. Viele, P. Y. Wen, and J. Woodcock, "Adaptive platform trials: definition, design, conduct and reporting considerations," *Nature Reviews Drug Discovery*, vol. 18, pp. 797–807, 10 2019.

[55] G. Gao, B. J. Gajewski, J. Wick, J. Beall, J. L. Saver, C. Meinzer, C. Derdeyn, D. Fiorella, T. Jovin, P. Khatri, E. Mistry, J. Mocco, R. Nogueira, and A. Siddiqui, "Optimizing a bayesian hierarchical adaptive platform trial design for stroke patients," *Trials*, vol. 23, 12 2022.

[56] J. Normington, J. Zhu, F. Mattiello, S. Sarkar, and B. Carlin, "An efficient bayesian platform trial design for borrowing adaptively from historical control data in lymphoma," *Contemporary Clinical Trials*, vol. 89, 2 2020.

[57] K. M. Lee and J. Wason, "Including non-concurrent control patients in the analysis of platform trials: Is it worth it?," *BMC Medical Research Methodology*, vol. 20, 6 2020. It is necessary to have a control group.

[58] M. Bofill Roig, C. Burgwinkel, U. Garczarek, F. Koenig, M. Posch, Q. Nguyen, and K. Hees, "On the use of non-concurrent controls in platform trials: a scoping review," 12 2023.

[59] K. M. Lee, L. C. Brown, T. Jaki, N. Stallard, and J. Wason, "Statistical consideration when adding new arms to ongoing clinical trials: the potentials and the caveats," 12 2021.

[60] B. R. Saville, D. A. Berry, N. S. Berry, K. Viele, and S. M. Berry, "The bayesian time machine: Accounting for temporal drift in multi-arm platform trials," *Clinical Trials*, vol. 19, pp. 490–501, 10 2022.

[61] K. M. Lee, J. Wason, and N. Stallard, "To add or not to add a new treatment arm to a multiarm study: A decision-theoretic framework," *Statistics in Medicine*, vol. 38, pp. 3305–3321, 8 2019.

[62] K. Thorlund, S. Golchi, J. Haggstrom, and E. Mills, "Highly efficient clinical trials simulator (hect): Software application for planning and simulating platform adaptive trials," *Gates Open Research*, vol. 3, p. 780, 3 2019.

[63] P. F. Thall, J. K. Wathen, B. N. Bekele, R. E. Champlin, L. H. Baker, and R. S. Benjamin, "Hierarchical bayesian approaches to phase ii trials in diseases with multiple subtypes," 3 2003.

[64] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," 1972.

[65] R. E. Kass and D. Steffey, "Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models)," 1989.

[66] W. J. Skene AM, "Hierarchical models for multicentre binary response studies.," *Statistics in Medicine*, vol. 9, p. 919–929., 1990.

[67] A. Gelman and J. Hill., "Data analysis using regression and multilevel/hierarchical models.," *Cambridge university press.*, 2006.

[68] S. R., "Optimal two-stage designs for phase ii clinical trials.," *Control Clin Trials.*, vol. 10, no. 1, pp. 1–10, 1989.

[69] B. P. Hobbs and R. Landin, "Bayesian basket trial design with exchangeability monitoring," *Statistics in Medicine*, vol. 37, no. 25, pp. 3557–3572, 2018.

[70] M. Pohl, J. Krisam, and M. Kieser, "Categories, components, and techniques in a modular construction of basket trials for application and further research," 8 2021.

[71] L. Billingham, K. Malottki, and N. Steven, "Research methods to change clinical practice for patients with rare cancers," 2 2016.

[72] H. Zheng and J. M. Wason, "Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy," *Biostatistics*, vol. 23, pp. 120–135, 1 2022.

[73] A. Gelman, "Prior distributions for variance parameters in hierarchical models(comment on article by browne and draper)," 2006.

[74] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[75] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[76] H. Zheng, L. V. Hampson, and S. Wandel, "A robust bayesian meta-analytic approach to incorporate animal data into phase i oncology trials," *Statistical Methods in Medical Research*, vol. 29, pp. 94–110, 1 2020.

[77] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, pp. 457–511, 1992.

[78] M. A. Psioda, J. Xu, Q. Jiang, C. Ke, Z. Yang, and J. G. Ibrahim, "Bayesian adaptive basket trial design using model averaging," *Biostatistics*, vol. 22, pp. 19–34, 1 2021.

[79] A. M. Kaizer, J. S. Koopmeiners, and B. P. Hobbs, "Bayesian hierarchical modeling based on multisource exchangeability," *Biostatistics*, vol. 19, pp. 169–184, 4 2018.

[80] C. Jennison and B. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC., 1999.

[81] J. Herson, "Predictive probability early termination plans for phase ii clinical trials," 1979.

[82] D. L. Demets and K. K. G. Lan, "Interim analysis: The alpha spending function approach," 1994.

[83] S. Pampallona, A. A. Tsiatis, and K. Kim, "Interim monitoring of group sequential trials using spending functions for the type i and type 11 error probabilities," 2001.

[84] S. J. Pocock, "Group sequential methods in the design and analysis of clinical trials.," *Biometrika*, vol. 64, 1977.

[85] O. PC and F. TR., "A multiple testing procedure for clinical trials. biometrics.," *Biometrika*, vol. 35, no. 3, pp. 549–56, 1979.

[86] K. K. G. Lan and D. L. Demets, "Discrete sequential boundaries for clinical trials," 1983.

[87] H. Zhou, J. J. Lee, and Y. Yuan, "Bop2: Bayesian optimal design for phase ii clinical trials with simple and complex endpoints," *Statistics in Medicine*, vol. 36, pp. 3302–3314, 9 2017.

[88] B. R. Saville, J. T. Connor, G. D. Ayers, and J. Alvarez, "The utility of bayesian predictive probabilities for interim monitoring of clinical trials," in *Clinical Trials*, vol. 11, pp. 485–493, SAGE Publications Ltd, 2014.

[89] J. J. Lee and D. D. Liu, "A predictive probability design for phase ii cancer clinical trials," *Clinical Trials*, vol. 5, pp. 93–106, 2008.

[90] T. Zhou and Y. Ji, "On bayesian sequential clinical trial designs," *The New England Journal of Statistics in Data Science*, pp. 1–16, 2023.

[91] P. Gallo, L. Mao, and V. H. Shih, "Alternative views on setting clinical trial futility criteria," in *Journal of Biopharmaceutical Statistics*, vol. 24, pp. 976–993, Taylor and Francis Inc., 9 2014.

[92] M. Ghadessi, R. Tang, J. Zhou, R. Liu, C. Wang, K. Toyoizumi, C. Mei, L. Zhang, C. Q. Deng, and R. A. Beckman, "A roadmap to using historical controls in clinical trials - by drug information association adaptive design scientific working group (dia-adswg)," 3 2020.

[93] C. X, Z. J, J. Q, and Y. F., "Borrowing historical information to improve phase i clinical trials using meta-analytic-predictive priors.," *J Biopharm Stat.*, vol. 32, no. 1, pp. 34–52, 2022.

[94] K. Kim and D. L. Demets, "Design and analysis of group sequential tests based on the type i error spending rate function," 1987.

[95] Y. WJ., "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[96] H. Zheng, T. Jaki, and J. M. Wason, "Bayesian sample size determination using commensurate priors to leverage preexperimental data," *Biometrics*, vol. 79, pp. 669–683, 6 2023.

[97] D. Ferreira, P. O. Ludes, P. Diemunsch, E. Noll, K. D. Torp, and N. Meyer, "Bayesian predictive probabilities: a good way to monitor clinical trials," *British Journal of Anaesthesia*, vol. 126, pp. 550–555, 2 2021.

[98] V. Sambucini, "Bayesian sequential monitoring of single-arm trials: A comparison of futility rules based on binary data," *International Journal of Environmental Research and Public Health*, vol. 18, 8 2021.

[99] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC., 2013.

[100] B. Thomas and R. Price, *An Essay towards Solving a Problem in the Doctrine of Chances*, vol. 3. Philosophical Transactions of the Royal Society of London., 1763.

[101] R. Christian and G. Casella, *Monte Carlo Statistical Methods*. Springer Texts in Statistics, 2 ed., 2004.

[102] M. Plummer, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," *Proceedings of the 3rd international workshop on Distributed Statistical Computing*, 2003.

[103] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC., 2013.

[104] S. Sturtz, U. Ligges, and A. Gelman., "R2winbugs: A package for running winbugs from r," *Journal of Statistical Software*, vol. 3, no. 12, pp. 1–6, 2005.

[105] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 6, no. 6, p. 721–741, 1984.

[106] W. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, p. 97–109, 1970.