Universität Bremen

MEDICAL UNIVERSITY OF VIENNA

Master Thesis

Design issues in multi-arm trials

by

Cora Burgwinkel

Student number: 6013146

First Supervisor: Assoc. Prof. Dr. Franz König
Medical University of Vienna

Second Supervisor: Prof. Dr. Werner Brannath
Competence Center for Clinical Trials Bremen,
University of Bremen

University of Bremen
Faculty 3: Mathematics and Computer Science
Medical Biometry/ Biostatistics

Bremen, 26^{th} July 2023

Declaration

The work for this thesis was carried out at the Center for Medical Data Science of the Medical University of Vienna. This work was part of the EU-Pearl (EU Patient-cEntric clinicAl tRial pLatforms) project which has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Untertaking under grant agreement No 853966. This Joint Untertaking received support from the European Union´s Horizon 2020 research and innovation program and EFPIA and Children´s Tumor Foundation, Global Alliance for TB Drug Development non-profit organization, Springworks Therapeutcs Inc. This thesis reflects the author‘s point of view. Neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained herein.

Acknowledgements

This work would not have been possible without the advice and support of several people who in one way or another contributed their valuable time and expertise during the course of this work.

First and foremost, I want to express my gratitude especially to Professor Franz König from the Medical University of Vienna for his supervision, continuous support and constructive feedback throughout this work. He has contributed significantly to the creation of this work through his suggestions and assistance. Thank you for the opportunity to write my thesis in Vienna!

Secondly, I would like to thank Professor Werner Brannath from the University of Bremen for his willingness to take on the co-supervision of my master thesis and providing thoughtful and helpful feedback to the thesis.

Many thanks to Elias Laurin Meyer, Marta Bofill Roig, Mariella Gregorich and Pavla Krotka who offered their unending support in the most crucial times and always found time to provide valuable thoughts and knowledge during all stages of the work.

I am deeply grateful to have been given the opportunity to write my master thesis within EU-Pearl at the Medical University of Vienna. My sincerest thanks to the whole Institute of Medical Statistics for the warm welcome, the academic environment, the lunch and coffee breaks and the constant support from the whole team. This experience has equipped me with valuable skills and knowledge that will undoubtedly influence my future endeavors. Special thanks to my office partner, Luzia Berchtold, for going with me through all the ups and downs and consistently brightening my day.

I also want to thank my classmate, Annika Swenne, for the teamwork, support and friendship throughout the master´s programme. Last but not least, I want to thank my parents and sisters for their encouragement, love and belief in my abilities throughout my studies and this thesis. Thank you for always being there for me!

Abstract

Multi-arm trials enhance drug development by offering increased flexibility and efficiency compared to traditional randomized clinical trials. The treatment efficacy in multi-arm trials is often assessed by comparing multiple treatment arms against a shared control arm. In traditional multi-arm studies, it is generally necessary for all enrolled patients to be eligible for all treatments in the trial. However, there are situations where this requirement may not be feasible, for example, treatments may not be available at all study centres. Selective exclusion of treatment arms can be considered as a solution in such cases, allowing clinicians and patients to exclude an unsuitable treatment arm. It is important to carefully consider the implications of the selective exclusion on the overall design and analysis of the trial. A key issue is which control data can be utilized and how to address the selective subgroups. To utilize different patient populations, it has been suggested to use randomization procedures in the trial which are capable of randomizing the patients between limited subsets of interventions according to the patient background, patients preference or treatment options at the study centres.

This thesis aims to enhance the methodology for optimally utilizing different patient subgroups in the analysis of a multi-arm trial while considering different compositions of control data. It was of further interest to evaluate the analysis strategy, the randomization strategy and the distribution of the different patient populations. The results from a simulation study are presented, where the performance of the proposed approaches in terms of the type I error rate and statistical power was evaluated under a wide range of scenarios. The results from the simulation study indicate that the analyses where the different patient subgroups were not adjusted for can lead to a substantial power loss and type I error inflation as bias in the effect estimates is introduced. Therefore, the usage of adjusted analyses is recommended. Furthermore, the results show that the preferred randomization strategy implies an equal ratio for treatment arm vs. control arm within each subgroup. Regarding the different composition of control data, the preferred analysis strategy is to base the comparisons on the patients who could have been directly randomized to one of the arms which are of interest for the comparison.

Kurzfassung

Mehrarmige Studien verbessern die Arzneimittelentwicklung, da sie im Vergleich zu herkömmlichen randomisierten klinischen Studien mehr Flexibilität und Effizienz bieten. Die Wirksamkeit der Behandlung in mehrarmigen Studien wird in der Regel durch den Vergleich mehrerer Behandlungsarme mit einer gemeinsamen Kontrolle bewertet. Bei herkömmlichen mehrarmigen Studien ist es im Allgemeinen erforderlich, dass alle eingeschlossenen Patient*innen für alle Behandlungen in der Studie in Frage kommen. Es gibt jedoch Situationen, in denen diese Anforderung nicht erfüllt werden kann, z.B. wenn die Behandlungen nicht in allen Studienzentren zur Verfügung stehen. In solchen Fällen kann ein selektiver Ausschluss von Behandlungsarmen als Lösung in Betracht gezogen werden, der es Ärzt*innen und Patient*innen ermöglicht, einen ungeeigneten Behandlungsarm auszuschließen. Es ist wichtig die Auswirkungen des selektiven Ausschlusses auf das Gesamtdesign und die Analyse der Studie sorgfältig abzuwägen. Eine zentrale Frage ist, welche Kontrolldaten verwendet werden können und wie mit den selektiven Untergruppen umgegangen werden soll. Um unterschiedliche Patient*innenpopulationen zu nutzen, wurde vorgeschlagen in der Studie Randomisierungsverfahren einzusetzen, mit denen die Patient*innen je nach Patient*innenhintergrund, Patient*innenpräferenz oder Behandlungsmöglichkeiten in den Studienzentren zu Untergruppen von Interventionen randomisiert werden können.

Ziel dieser Arbeit ist es die Methodik zur optimalen Nutzung unterschiedlicher Patient*innenpopulationen in der Studienanalyse unter Berücksichtigung unterschiedlicher Zusammensetzungen von Kontrolldaten zu verbessern. Darüber hinaus war es von Interesse die Analysestrategie, die Randomisierungsstrategie und die Verteilung der verschiedenen Patient*innenpopulationen auf die Untergruppen auszuwerten. Schließlich werden die Ergebnisse einer Simulationsstudie vorgestellt, in der das Verhalten der vorgeschlagenen Ansätze im Hinblick auf den Fehler 1. Art und die Teststärke unter einer Vielzahl von Szenarien bewertet wurde. Die Ergebnisse der Simulationsstudien deuten darauf hin, dass die Verwendung adjustierter Analysen empfohlen wird, da das nicht adjustieren der Untergruppen

zu einem erheblichen Verlust der Teststärke und einer Aufblähung des Fehler 1. Art führen kann, da eine Verzerrung der Effektschätzungen entsteht. Außerdem zeigen die Ergebnisse, dass man für die bevorzugte Randomisierungsstrategie ein gleiches Verhältnis von Behandlungsarm und Kontrollarm innerhalb jeder Untergruppe annimmt. Im Hinblick auf die unterschiedliche Zusammensetzung der Kontrolldaten ist die bevorzugte Analysestrategie für die Vergleiche nur die Patient*innen einzubeziehen, die direkt in einen der für den Vergleich interessanten Arme randomisiert worden sind.

Contents

List of Tables

List of Figures

1 Introduction

The goal of a clinical trial is to gain evidence for efficacy and safety of a medical intervention by extrapolating from a sample to a defined population. In a standard clinical trial one compares a treatment (e.g. a new drug, a device or a therapy) to the standard of care (SOC) and/or placebo (see Figure 1.1). Individuals receiving the same type of treatment form an arm within a trial and patients enrolled in the trial are randomized to one of the arms such that all influencing factors have the same distribution in all treatment arms [1, p. 7]. The recruitment to the two arms is done in parallel (see Figure 1.1). A key drawback of the standard clinical trial design is that only one intervention is investigated per trial for which the participant enrolment might be slow since the subjects must meet the inclusion criteria. As result, the average duration of a trial is long and the costs are high [2–4]. Further reasons for the long duration are, for example, the development of a study protocol [5], site selection and management [6] and the ethical and regulatory approval.

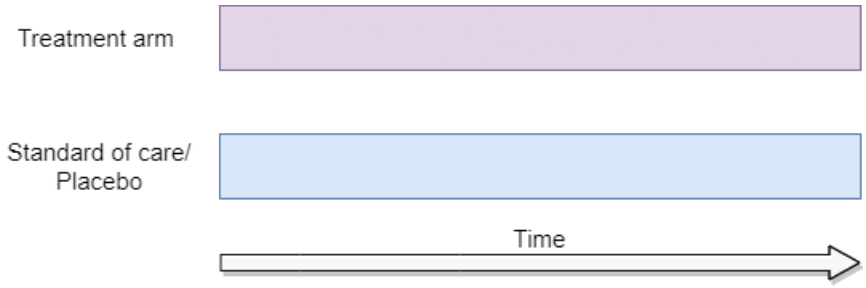

Figure 1.1: A parallel randomized controlled trial with one treatment arm and one control arm.

Consequently, trial designs are required to address the need to compare multiple treatments at the same time to reduce the logistical burden (e.g. patient recruitment and site selection [6]) of the standard clinical trial design.

Multi-arm trials are an example for such a trial design. The idea of multi-arm trials is to evaluate multiple experimental arms concurrently against a shared

control arm within one trial (see Figure 1.2). The shared control arm reduces the required sample size compared to multiple independent randomized trials [7, 8]. Multi-arm trials can be extended to incorporate multiple stages of analysis. The use of interim analyses, i.e. planned analyses conducted during the course of the trial before the final analysis is performed, in multi-arm multi-stage (MAMS) trials allows to stop arms early for futility or efficacy [9].

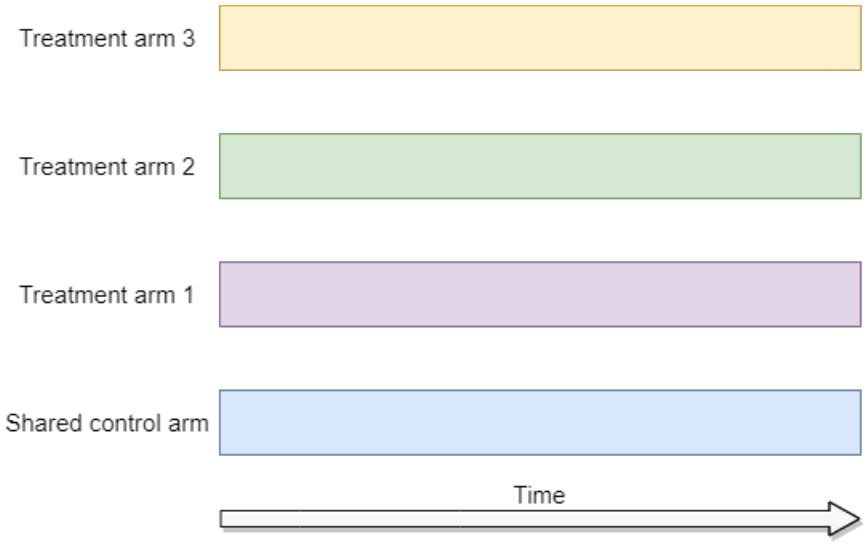

Figure 1.2: Illustration of a multi-arm trial: parallel evaluation of three treatment arms against a shared control arm.

Moreover, there has been an increased interest in trial designs that test multiple drugs and/or multiple subpopulations concurrently under a single protocol [10], which outlines, among other things, the objective, design, methodology, and statistical analysis plan for conducting a specific clinical trial [11]. Thus, there is no need to develop new protocols for each trial. The demand is met by the concept of the master protocol. In contrast to protocols for the standard trial design, master protocols use a single infrastructure, trial design and protocol to simultaneously investigate multiple drugs and/or diseases in multiple substudies [12, 13]. As a result, master protocols allow an efficient and accelerated drug development [13]. Trial designs that are covered by master protocols are basket trials, umbrella trials and platform trials [13].

In classical multi-arm trials, it is usually required that all enrolled patients are eligible for all treatments in the trial. However, in many contexts that may not be possible. For example, not all treatments are available at all participating study sites, patient preference or, depending on the context, patients are often not

eligible to all interventions because treatments regularly have contraindications [14]. Selective exclusion of treatment arms can be a solution when clinicians and patients think that one of the treatment arms is unsuitable which, in turn, no longer prevents these patients from being recruited to the trial. As a result, patient recruitment to the trial is easier and faster. The question arises how to analyse studies which allow selective exclusion of treatment arms and on which data to base the treatment vs control comparisons in the analysis. However, one must consider the implications of selective exclusion of treatment arms for the design and analysis [15].

For the analysis, Law et al. [15] recommended to base the primary comparison of two treatments only on the patients who were directly randomized between those two treatment arms. This implies that in a multi-arm trial which allows selective exclusion not all patients will be included in all comparisons and that the treatment comparisons will be conducted in a pairwise manner. For this reason, the randomization procedure used in the trial should have the capability to randomize patients between limited subsets of interventions according to the patient background or patients preference [15]. Additionally, it is crucial to ensure that patients' allocation to the treatments in the specific subgroups is clearly labeled to facilitate subsequent analysis of the trial data.

There are several contexts where some patients may not be eligible to all interventions and multiple trials encountered the challenge how best to utilize comparable patients in the analysis [16–18]. However, to our knowledge, the statistical analysis of such studies has not received much attention. For this purpose, a comprehensive simulation study was carried out in this thesis in which a trial design comprising a multi-arm trial with two treatment arms and one control arm is simulated. The advantage of such a trial design is that a shared control arm can be used rather than two control arms which would be needed in two separate trials comparing treatment 1 with control and treatment 2 with control. The patients are randomized to the treatment arm(s) or to the control arm by complete or block randomization and different randomization strategies to the three arms are considered. Since the trial design allows the selective exclusion of treatment arms, the randomization methods used in the simulation are able to randomize the patients between limited subsets of interventions depending on the patients background. The main goal of this master thesis is to investigate how comparable patients can be utilized optimally in the analysis of the trial data for different compositions of control data. Further distinctions were made between adjusting and

not adjusting for the patient background. To investigate whether the composition of the control data has an impact in a multi-arm trial with selective exclusion, operating characteristics, e.g. the power and type 1 error, are taken into account and evaluated under different simulation parameters such as the total sample size, the randomization strategy and the treatment effects.

The structure of this thesis is organized as follows: Chapter 2 delves into essential background information regarding MAMS trials and master protocols. Additionally, various design issues that may arise during the clinical trial design process, such as the randomization method and the allocation ratio, are comprehensively elucidated. Moving forward, Chapter 3 encompasses an in-depth examination of the methods employed in this master thesis. The analysis methods, e.g. the analysis of variance model, are expounded upon. Chapter 4 explains the design of the simulation and its implementation, providing a comprehensive understanding of the methodologies employed. In Chapter 5, the statistical analysis results from one simulated data set are thoroughly presented, alongside the comprehensive findings stemming from the simulation study. Lastly, Chapter 6 serves as a comprehensive summary of the thesis' findings, effectively consolidating the overall outcomes. Furthermore, the limitations of the simulation study are discussed, adding valuable insights and implications to the research.

2 Advancements in clinical trial development

This chapter comprises crucial background information concerning the development of randomized controlled trials. Initially, a comprehensive exploration of randomized controlled trials will be presented, encompassing their fundamental principles and methodologies. Subsequently, the focus will shift towards multi-arm trials, followed by an expansion to multi-arm multi-stage (MAMS) trials (see Section 2.2). Furthermore, the concept of master protocols will be introduced and the different types of master protocols will be delineated (see Section 2.3). Lastly, a detailed examination of various critical aspects relevant to the design of clinical trials will be conducted (see Section 2.4), highlighting their significance and implications.

2.1 Randomized controlled trials

The objective of a randomized clinical trial is to obtain evidence regarding the safety and effectiveness of a new medical intervention for a defined population by extrapolating from a sample of subjects who receive the intervention. It has been considered the gold-standard for demonstrating benefit of a new intervention for the last decades [7, 12, 19]. In a standard clinical trial one compares patients who receive the intervention to a second group which consists of patients who do not receive the intervention (see Figure 2.1 **A)**). The latter group is called control. In so-called placebo-controlled clinical trials, the control arm receives the placebo therapy. However, in situations where it is unethical to withhold the therapy, the patients in the control arm receive the SOC. Patients are randomized to one of the arms such that all influencing factors have the same distribution in all treatment arms [1, p. 7]. The goal of randomization is to avoid bias by facilitating the groups' comparability of baseline patient characteristics [3].

One of the first clinical trials using randomization for patient allocation was conducted in the 1940s for the study of tuberculosis and directed by Sir Bradford

Hill [1, p. 8]. Compared to randomized experiments in other disciplines, clinical trials in humans involve complex ethical issues [1, p. 9]. For example, patients who are by chance randomized to placebo are denied of a therapy that might later prove to be benefiting. Or on the other hand, patients could get randomized to a therapy which might later prove to be highly toxic.

The course of developing a new drug can usually be divided into four phases [20, p. 4]:

1. **Phase I** determines possible toxic effects and establishes the maximally tolerated dose in small groups of roughly 20-80 subjects [21, p. 3].

2. **Phase II** analyses the therapeutic effects of the drug and investigates the dose finding for phase III in small collectives of approximately 50-300 subjects.

3. **Phase III** assesses the efficacy and tolerability of the drug by comparing it to the current SOC in a controlled trial setting with larger sample sizes (about 1000 - 3000 subjects). The results from phase III provide the basis for approval by the regulatory agencies.

4. **Phase IV** is conducted once the drug is approved with the aim to investigate long-term effects and rare side effects in large collectives of several thousand patients [22, p. 75].

The **randomization** to the different arms takes place in phase III and patients are randomized to the arms such that the only difference between the patients randomized to either arm is the treatment itself. Knowledge of treatment assignment by the investigator, physician and/or patient can introduce bias at several stages of the trial. For example, due to selection bias of the physician, the distribution of patient characteristics between the arms might, on average, no longer be the same because knowledge of the next assignment could influence whether a patient is included or excluded based on the perceived prognosis [3, 23]. Patient-reported bias is presented by the fact that patients who know that they are assigned to the experimental intervention might tend to believe that the intervention is helping them while patients who know that they get placebo might believe that they are more likely to worsen than to improve. One of the procedures employed to avoid conscious or subconscious influence of staff members, physicians or patients is **concealment** [19, 23]. Concealment of the allocation sequence ensures that

the allocation sequence can be implemented without knowledge of which patient will receive which treatment [24]. As result, concealment seeks to prevent selection bias [19, 25]. Another procedure which is used to avoid bias is **blinding**. Clinical trials are often either single-blinded, meaning that the patient is blinded with respect to the treatment assignment but not the physician (or vice versa), or double-blinded, meaning that the treatment assignment is concealed from both patient and physician/investigator [26, pp. 26 sq.].

In addition to the above mentioned challenges of a standard clinical trial, it has been realized that patients who seem to have the same disease might require different therapies and that one needs to account for heterogeneous patient populations [27]. However, this is not covered by the design of the randomized controlled trials. The question arises how the challenges of a standard clinical trial can be overcome. For that purpose new trial designs are needed [27, 28] which reduce the costs and sample size, and make the trials more appealing to patients. Innovative designs can help to combine trials investigating distinct research questions into a single protocol and thereby shorten the duration of the trial and accelerate the development of an intervention.

2.2 Multi-arm and multi-arm multi-stage trials

One of the new trial designs that were introduced are **multi-arm trials** which have been recommended for many different diseases such as oncology [29], stroke [9] and tuberculosis [30]. The idea of multi-arm trials is to evaluate multiple experimental arms against a common control within one trial (see Figure 2.1 **B)**). The direct comparison of treatments in one trial not only minimises the bias that could potentially be introduced by making comparisons between treatments tested in separate trials [31], but also reduces the required sample size compared to independent randomized trials by sharing the common control arm (see Figure 2.1). For example, a multi-arm trial with three arms reduces the sample size that would be required for two separate trials by 25% due to the shared control arm when no multiplicity adjustment is performed (see Figure 2.1) [7]. Another advantage of the parallel evaluation of multiple experimental arms in multi-arm trials is that it results in an increased chance of a patient to meet the inclusion criteria of at least one experimental arm. Furthermore, there is a higher probability that a patient will be randomized to one of the experimental arms, assuming an equal allocation ratio, and therefore, the multi-arm trial might be more attractive for patient participation [32, 33] compared to a standard clinical trial. In some multi-

arm trials, fewer patients are randomized to the control arm to further increase the probability to receive an experimental treatment even though a larger overall sample size is required for this approach [7]. On the other side, a large number of patients can be allocated to the control arm without having to increase the maximum sample size considerably. This may be of interest in a scenario where the control treatment is cheaper than the experimental treatments or expected to have a better safety profile than the experimental treatments [31]. However, as consequence one might have a reduced recruitment to the trial as there is some evidence in placebo controlled trials that the willingness of the patients to participate in a trial reduces as the allocation to the control arm increases [34].

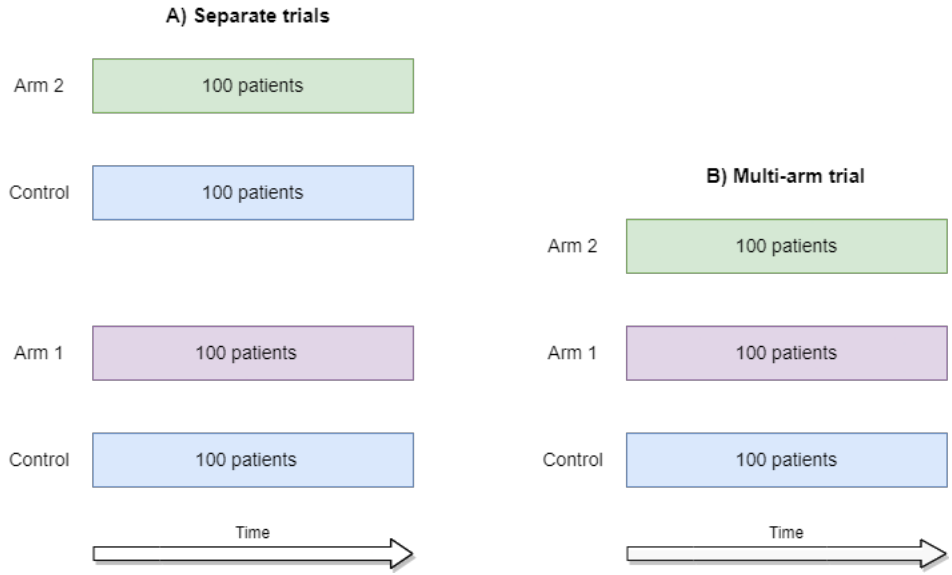

Figure 2.1: Illustrative comparison of the standard clinical trial design with the multi-arm trial design. **A)** Two clinical trials with separate control arms and **B)** a multi-arm trial with a shared control arm.

By incorporating interim analyses, multi-arm trials can be extended to **multi-arm multi-stage trials**. This has the advantage that it is ensured that only promising treatments continue to the next stage of the trial as the study design allows to drop ineffective treatments before the end of the trial [9]. Furthermore, interim analyses can be performed at multiple time points and allow to stop the trial early if sufficient evidence of a treatment being superior to control is found [8]. Clinical trials that incorporate the ability to stop early after an interim analysis are referred to as *group sequential trials* [35]. The ability to test several treatments

at once and to drop ineffective interventions reduces the duration of a trial and efficiency is gained [9]. Blinding in multi-arm randomized controlled clinical trials can be challenging since it must be ensured that none of the arms can be distinguished [7]. To ensure that none of the arms can be distinguished, techniques like double-dummy designs might be needed. In a double-dummy design each participant receives a combination of two treatments, one active treatment and one placebo and the purpose is to maintain blinding in the trial, e.g. when its challenging to blind the treatments completely, especially in trials comparing different modes of administrations (e.g. pill vs. injection), the double-dummy design can ensure that the appearance and administration of the treatments is indistinguishable [36]. However, double-dummy designs can become infeasible to implement. As result, multi-arm trials, like the TelmisArtan and InsuLin Resistance in HIV (TAILoR) trial, which will be introduced in the following, are commonly conducted as open-label studies, i.e. both the participants and researchers know which treatment is administered [37].

One example for a MAMS trial is the TAILoR trial. The trial evaluated whether telmisartan reduces the insulin resistance in HIV-positive participants on antiretrovirals. The TAILoR trial was a multi-center, randomized, open-label and dose-ranging controlled trial of telmisartan [37]. HIV-positive participants were randomized by block randomization with an equal allocation ratio to either control (no intervention) or 20, 40 or 80mg telmisartan administered once daily. The primary endpoint was the reduction in insulin resistance in the groups treated with telmisartan compared to the control arm (measured at 24 weeks by the homeostasis model assessment of insulin resistance (HOMA-IR)) [37]. The design of the trial allowed to test all of the doses of telmisartan in stage I and take the promising doses to stage II. The design further permitted to stop the study at the interim analysis in case the required benefit was not identified in any of the doses. No significant effect of telmisartan after 24 weeks was demonstrated on the primary endpoint HOMA-IR. Additional research in this population is justified to explore novel approaches for preventing cardiovascular morbidity and mortality [37].

2.3 Master protocol

Clinical trial development entails the writing of a protocol which, among other things, outlines the objective, study design and methodology. The protocol serves as a guide for the study personnel involved in the trial to ensure consistency and

standardization throughout the trial [11]. Over the past few decades there has been an increased interest in trial designs that test multiple drugs and/or multiple subpopulations in parallel under a single protocol without the need to develop new protocols for each trial [10]. In contrast to standard clinical trial designs, where a single drug is tested in a single disease population, master protocols use a single infrastructure, trial design, and protocol to simultaneously evaluate multiple drugs and/or disease populations in multiple substudies. Master protocols are defined as one overall protocol designed to answer multiple questions [13]. Hence, they enable efficient and flexible drug development by providing an opportunity to incorporate efficient approaches, such as a shared control arm and the use of centralized data capture systems. In the field of oncology many master protocols are initiated [12, 13, 38]. As those trial designs are often employed with the intention to support regulatory approval, it is important that such trials are well designed and conducted and meet the regulatory standards. Due to their complexity, master protocols require increased planning efforts. Master protocols are often classified into three different study designs: basket trials, umbrella trials and platform trials [13] (see Table 2.1) which will be introduced in the following.

Table 2.1: Types of master protocols defined by Woodcock et al. [13]

Type of trial design	**Objective**
Umbrella trial	To study multiple targeted therapies in the context of a single disease
Basket trial	To study a single targeted therapy in the context of multiple diseases or disease subtypes
Platform trial	To study multiple targeted therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm

An **umbrella trial** is a master protocol that is designed to evaluate multiple investigational drugs administered as single drugs or as drug combinations in a single disease population [39] (see Figure 2.2 **A)**). An example of a master protocol with umbrella trial design is the original version of the lung master protocol (Lung-MAP) trial which is a multi-drug, multi-substudy, biomarker-driven trial in patients with advanced metastatic squamous cell carcinoma of the lung [40]. Eligible patients were assigned to the different substudies in the trial based on their biomarkers or to a non-match therapy substudy for patients not eligible for the biomarker-specific substudies. Within the substudies, the patients were randomized to a biomarker-driven target or to the SOC therapy. Each substudy functions

autonomously, opens and closes independently, and is analyzed independently of the other substudies. The idea behind the trial design is that a drug that is found to be effective in phase II will move directly into the phase III registration setting, incorporating the patients from phase II in order to reduce time, resources and patient numbers needed to accomplish the ultimate goal of bringing novel agents to the clinic [40].

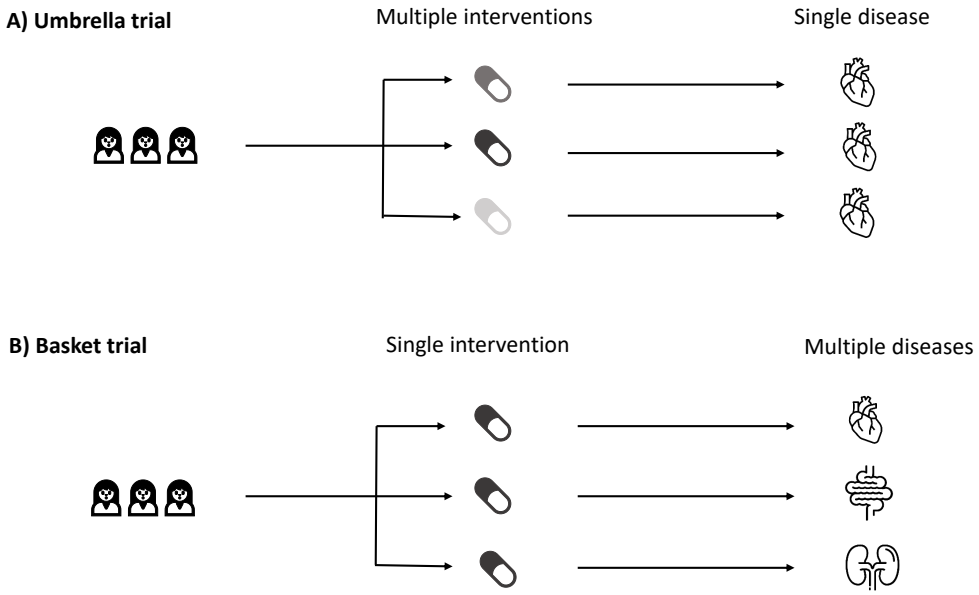

Figure 2.2: Depiction of an umbrella trial and a basket trial. **A)** shows an umbrella trial which evaluates multiple interventions in a single population and **B)** depicts a basket trial which investigates a single intervention in multiple diseases.

A master protocol that is designed to test a single investigational drug or drug combination in different populations defined by e.g. different cancers or disease stages for a specific cancer is commonly referred to as a **basket trial** [39] (see Figure 2.2 **B)**). An example of a master protocol with basket trial design is the phase II trial evaluating vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations [41]. The goal of the study was to distinguish signals of activity in the different tumor types that could be further pursued in subsequent studies or through protocol amendments in the current study. As the study design allows the enrollment of patients with different types of cancers, eligible patients were assigned to the six cohorts based on their prespecified cancers or to an all-others

cohort which included enrollment of patients with other BRAD V600 mutation-positive cancer [41]. In case enrollment in the cohorts was not sufficient for the analysis, the cohorts were closed and the patients were included in the all-others cohort. The flexible design of the study also allowed the inclusion of a cohort of patients with any tumor types that were not prespecified which facilitated identification of modest activity on certain orphan cancers. To summarize, the basket study design made it possible that different tumor types with the identical molecular biomarker can differ in their sensitivity to the given therapy [41].

2.3.1 Platform trials

Platform trials are a type of study design where clinical trials investigate multiple treatments or treatment combinations in the context of a single disease. The idea is to combine clinical trials with related questions to improve the efficiency [42]. As a consequence, one has a direct comparison between several treatments within one trial. In addition, more patients are eligible for the trial because of the broader inclusion criteria for the several treatments that are tested. Platform trials provide great benefit for sponsors and trial participants in terms of accelerating drug development. They are very flexible because the number of experimental treatments is not being fixed in advance. Arms can be added or removed during the trial [10, 13, 39] (see Figure 2.3), so new investigational treatments can easily be incorporated into the ongoing trial. This allows the trial to run, in principle, infinitely.

Platform trials are *adaptive* since changes are allowed to be made during the trial based on results from interim analyses, e.g. endpoint selection, randomization ratio or sample size reassessment. Compared to *group sequential designs*, flexibility is gained because data from the interim analysis can freely modify the course of the trial [43]. A flexible number of interim analyses is possible and interim analyses are possible with the option to stop the treatment for efficacy or futility. Thus, the required sample size is reduced. A further advantage is that control data can be shared among the treatment arms which also leads to a reduction of the required sample size [13]. In case the sharing of a common control is inappropriate, a platform will probably lose some of its inferential advantages but might still have operational advantages compared to a classical RCT [42]. Besides, the change of control within the trial is also possible in case a treatment is found to significantly outperform the existing control [10] (see Figure 2.3). There can be a staggered entry of a flexible number of treatments over time and for a treatment entering

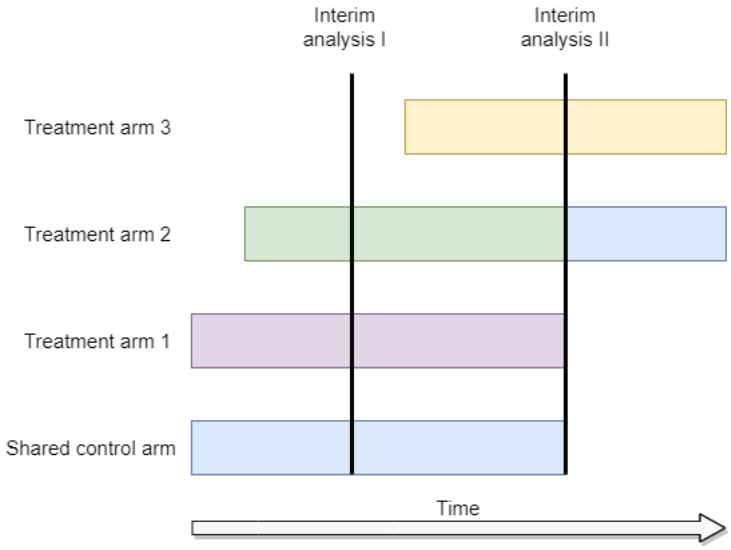

Figure 2.3: Example for a platform trial: investigating multiple experimental treatments which enter and leave the trial at different time points. Besides, the shared control arm is outperformed and a change of control takes place.

the trial at a later time point, the control arm is divided into concurrent and non-concurrent controls [44]. Concurrent controls are the patients who are recruited in parallel to the treatment arm and non-concurrent controls are the patients who have been recruited from the start of the trial until the beginning of the treatment of interest (see Figure 2.4). Using non-concurrent controls in addition to concurrent controls can improve the efficiency of the trial by increasing the power and reducing the required sample size [44].

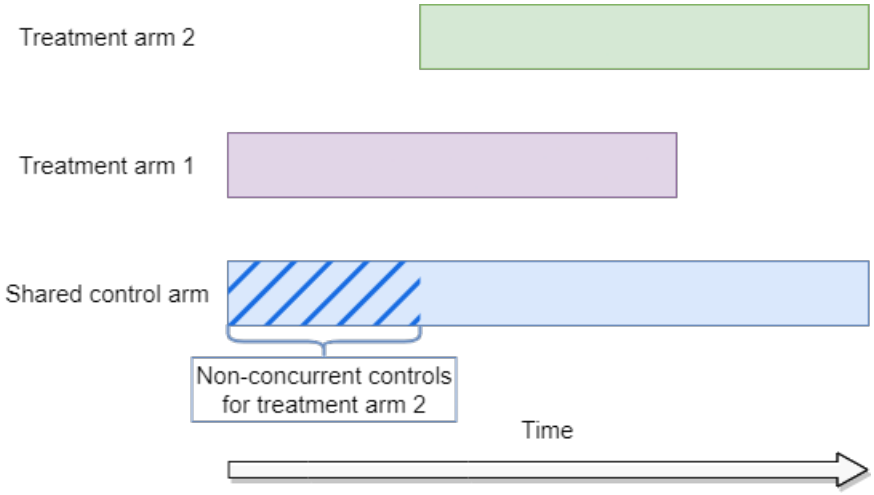

Figure 2.4: Schematic illustration for a platform trial with non-concurrent control data.

Challenges in platform trials

Platform trials are very efficient but also raise logistical issues (e.g. cost, planning) in addition to several statistical issues (e.g. the timing of adding new interventions and the error rate control). The complexity of platform trials results in higher costs and new challenges regarding trial implementation and planning. For example, platform trials may require more complex collaborations across sponsors and well-timed communication between all participating sites and data coordinating units. Besides, the shared costs of running the trial need to be determined.

One of the statistical issues arises from the fact that multiple experimental treatments can be investigated within a platform trial, which may additionally be tested with regard to multiple endpoints or in multiple subgroups [45, 46]. It is still an open issue whether to adjust for multiplicity in multi-arm trials and must be decided for each trial separately while considering, e.g. the relatedness of the tested hypotheses and the sharing of a control arm [47] which will be discussed in more detail in Section 2.4.5. Besides, adding new treatments may require modified eligibility criteria and avoiding operational bias can also be a challenge [42].

Another statistical issue in platform trials is whether to include non-concurrent control data or not. As mentioned above, the inclusion of non-concurrent control data can improve the efficiency of the trial, however, bias may be introduced due to time trends [32, 38, 48–50]. Time trends may be caused by external factors such as a change in the SOC, a change in the patient population, a pandemic or due to internal factors (e.g. a change in the recruiting centers, a change in the recruitment strategies or the inclusion and exclusion criteria) [51]. There are several different analysis methods to incorporate non-concurrent controls in the trial, for example,:

(i) *Simple approaches:* the separate approach, which analyzes only concurrent controls, and the pooled approach, which analyzes both concurrent and non-concurrent controls [52], but leads to a type I error rate inflation and biased treatment effect estimates, if time trends are present in the trial [44, 53],

(ii) *Frequentist methods:* test-then-pool [52], dynamic pooling (i.e. a weight parameter is assigned to the historical data, which controls the proportion of the historical information that will be used in the analysis [54]) and propensity score methods (i.e. balance the differences between historical and concurrent controls [55]),

(iii) *Bayesian approaches:* power prior (i.e. down-weight the historical information by introducing a power parameter [56]), commensurate power prior [57] and meta-analytic-predictive (MAP) prior [58, 59],

(iv) *Frequentist and Bayesian model-based approaches:* include time as a covariate to avoid bias due to time trends [44, 53, 60] (it it has been shown that the type I error rate is controlled and the power increased when the time trend is equal across the arms and additive in the model scale [44]).

To conclude, it can be said that it must be decided for each platform trial individually whether to use non-concurrent controls or not as the regulatory guidance for the use of non-concurrent control is currently limited. In addition, there is a lack of clear guidance on appropriate statistical methods for the incorporation of non-concurrent controls [61].

Examples for platform trials

An example of a master protocol with platform study design is the individualized screening trial of innovative glioblastoma therapy (INSIGhT) which is a Bayesian adaptive platform trial to develop precision medicines for patients with glioblastoma [62, 63]. It is a randomized, adaptive phase II trial which is still ongoing. The trial started with three different experimental arms and different therapies can be added or dropped throughout the trial. At the beginning, there was an equal randomization ratio among the arms while during the trial a Bayesian adaptive randomization ratio is used. For the adaptive randomization ratio the biomarker-specific probability of the treatment impact on the progression-free survival is estimated. The experimental arms are compared to a common control of standard chemoradiotherapy and the primary endpoint is overall survival. The goal of the trial is to shorten the execution of the trial while increasing the number of discovered biomarkers with the help of adaptive randomization [62]. The choice of study design, the adaptive platform trial, allows for the addition of new treatment arms to the overall trial structure which is anticipated to improve the overall efficiency of the trial.

The I-SPY2 (investigation of serial studies to predict your therapeutic response with imaging and molecular analysis 2) trial is another example for a platform trial. It is an ongoing phase II, response-adaptive, randomized, multi-center trial which evaluates new drugs or drug combinations in the context of neoadjuvant

treatment for women with locally advanced breast cancer [64, 65]. The trial focuses on high-risk breast cancer patients, including those with larger tumors and the primary endpoint is pathologic complete response. For assigning the patients to the drugs, adaptive randomization will be used to achieve a higher probability of efficacy. The trial incorporates the use of biomarkers, such as gene expression profiles, to match patients with specific treatments that are most likely to benefit them. Within ten predefined clinically relevant biomarker signatures efficacy is evaluated and when promising treatments are identified in the signatures, the treatments will be evaluated for further phase III testing [66]. By using an adaptive design, biomarker-driven patient selection, and neoadjuvant therapy, the I-SPY2 trial aims to accelerate the evaluation of new treatments for high-risk breast cancer patients, improve treatment outcomes, and facilitate personalized medicine approaches in breast cancer treatment.

And finally, the STAMPEDE (Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy) trial, is a multi-center, MAMS, randomized, open-entry clinical trial which assesses novel therapies in men with high-risk localised or metastatic prostate cancer who are being treated for the first time with long-term hormone therapy [29, 67]. The trial opened in 2005 with six arms and five stages and is the first MAMS trial to implement multiple arms and multiple stages synchronously [68]. The primary endpoint is overall survival. Over time, some experimental arms have finished accrual, and new experimental arms have entered the trial [69]. The STAMPEDE trial seeks to improve the understanding of optimal treatment strategies for advanced or metastatic prostate cancer. By comparing different treatment approaches and utilizing an adaptive design, the trial aims to identify treatment combinations that offer better outcomes, longer survival, and improved quality of life for patients with advanced prostate cancer.

2.4 Issues in designing clinical trials

The design issues refer to various considerations and decisions that need to be addressed when planning and conducting a trial. Issues in trial conduct and analyses should be anticipated during trial design and thoughtfully addressed. Fundamental clinical trial design issues for the study design such as the randomization method, the allocation ratio and the selective exclusion of treatment arms are discussed in the next subsections. Besides, operating characteristics are an

essential aspect of clinical trial design. Some key operating characteristics that can be considered in the study design when simulating a clinical trial will be presented. Lastly, the concept of error control and the problem of multiplicity will be explained.

2.4.1 Randomization methods

One of the most important design techniques to avoid bias in a clinical trial is **randomization** [23]. Randomization is the process of assigning participants to different treatment arms, assuming that every participant has an equal chance of being assigned to each arm independent of their baseline characteristics. The idea behind randomization in a clinical trial is to neutralize the inevitable variation by chance. If randomization is properly implemented, then per definition there is no confounding factor. Besides randomization, key issues to achieve that are **blinding** and **concealment** [19, 24]. The patient characteristics may differ between the treatment arms but just by chance [3]. As result, the heterogeneity is not canceled out completely but all influencing factors should have a similar distribution in the treatment arms [1, p. 7]. Therefore, standard statistical methods are applicable [48]. Randomization procedures can be divided into restricted and unrestricted procedures. Unrestricted randomization procedures impose no constraints on the random allocation of treatments while restricted randomization procedures apply restrictions on the probability of treatment allocation. For example, balancing restrictions are imposed to have equal numbers of patients per treatment [19]. In general, balancing restrictions can be differentiated into balancing of the covariates between the treatment arms, e.g. a balance of the covariate gender implies an equal distribution of males and females across the treatment arms, and balancing on the assignment of treatment, e.g. the treatment arms consist of an equal number of patients [70].

The simplest randomization procedure is the **complete randomization** which is an unrestricted randomization procedure where for each subject the allocation is chosen randomly, e.g. by tossing a fair coin [71]. For example, a patient is assigned to treatment if the result of tossing the coin is head, and to control if tossing the coin produces tail.

In general, let $T_i \in \{1, ..., J\}$, $i = 1, ..., n$, denote the assignment of patient i to one of J treatments in a study cohort of size n. For a trial with two arms, e.g. $J = 2$, then $T_i \in \{1, 2\}$. $T_1, ..., T_n$ are independent and identically distributed Bernoulli

random variables with $P(T_i = 1) = \frac{1}{2}$, $i = 1, ..., n$ for an equal randomization ratio [1, p. 35].

As mentioned above, one of the reasons to employ randomization is to avoid bias in a clinical trial. Blackwell et al. [72] remarked that experiments comparing two or more treatments may yield biased results if the patients are selected with knowledge of the treatments they are to receive. For example, the investigator could try to guess the next treatment assignment based on the knowledge of the previous assignments when the patients are enrolled sequentially [72]. Thus, the investigator might guess that a treatment which until then has not been allocated often is likely to be allocated next and based on that might select a patient with a higher or lower expected response to the guessed treatment to be included next in the trial. One advantage of complete randomization is that the investigator is unable to infer which treatment the patient will be assigned to because the probability for the treatment assignment is always the same [3]. Thus, selection bias can be avoided since all patients have the same probability of being assigned to either treatment arm, regardless of the imbalance in numbers or previous treatment assignments. On the other hand, a possible disadvantage of complete randomization is that the treatment allocation can lead to a high imbalance in the resulting group sizes [1, p. 36]. For example, for two treatments, A and B, and an equal allocation ratio, then for 20 patients, one would assume that 10 patients were allocated to each treatment. But since there are no constraints on the random allocation of the treatments, there might be an imbalance in the actual randomization ratio compared to the initially planned ratio, e.g. 12 patients are allocated to treatment A and 8 patients are allocated to treatment B (see Figure 2.5). The imbalance can lead to a substantial decrease in power as one has the highest precision in studies with equally sized treatment arms [1, p. 36].

Block randomization, which is a restricted randomization procedure, can be employed to control the imbalance of treatment assignments by making treatment assignments at random in blocks [3]. For a block randomization, the patients are divided into M blocks or groups of equal size containing $m = \frac{n}{M}$ patients [1, p. 42]. Within each block, e.g. the treatments are allocated equally. For example, for two treatments A and B, $\frac{m}{2}$ patients would be assigned to each treatment. In order to do so all possible balanced combinations of assignment within a block must be calculated after the block size has been determined (see Table 2.2). Table 2.2 shows six possible combination sequences for a block size of four. An equal randomization ratio is assumed within each block such that the treatments are allocated

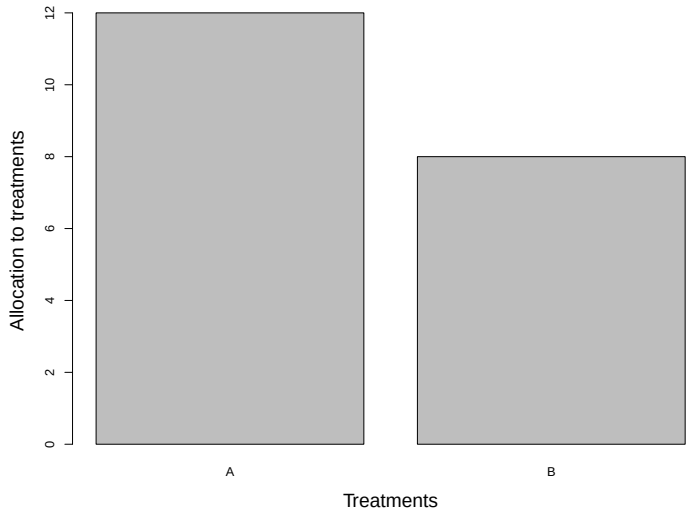

Figure 2.5: Visualization of the treatment assignment to treatments A and B for 20 patients for complete randomization assuming an equal allocation ratio. As there are no constraints on the random allocation of the treatments, the result is a high imbalance in the resulting group sizes.

equally. Each combination sequence has a probability of $\frac{1}{6}$ under random allocation. The grouping to the blocks is done in the chronological order of patient entry [73] and the block length must be confidential and investigators should be blinded to the block length [48]. Using block randomization produces the same amount of covariate imbalance as complete randomization since the randomization technique focuses on balancing the total number of subjects assigned to the treatment arms and not on balancing the prognostic factors between the treatment arms [70]. A covariate imbalance could introduce bias in the statistical analysis and reduce the power of the study. When the goal is to control the covariate imbalance, one can use covariate adaptive randomization. Covariate adaptive randomization balances both, the covariates between the treatment arms and the treatment assignments. Besides, covariate adaptive randomization may be used when several covariates are known a priori to influence the outcome. A disadvantage of covariate adaptive randomization is the complexity introduced into the analysis [70].

Besides controlling the imbalance throughout the trial, another advantage of block randomization is that it improves the comparability of the treatment arms, especially in situations where subject characteristics change over time [48]. However, that is only guaranteed if the block size is not chosen too large to avoid a

Table 2.2: Possible combination sequences for two treatments for a block size of 4

Block	1	2	3	4	5	6
Combination sequence	A	B	A	B	A	B
	A	B	B	A	B	A
	B	A	B	A	A	B
	B	A	A	B	B	A

time drift in important covariates. As a larger block size increases the allocation randomness, the block size should neither be chosen too small to avoid that the allocation is deterministic and becomes predictable towards the end of a block which could result in selection bias [74]. As explained above, the investigator might guess the next treatment assignment based on its frequency for the previous assignments knowing that the treatment arms are expected to be balanced. If, for example, the block size is four for two treatments, A and B, and the first three patients were assigned to A, B, B, then the investigator can guess that, assuming an equal allocation ratio, the fourth patient is assigned to treatment A. Selection bias may be reduced by making use of a variable block design where the block size itself is randomly selected [70]. The selection bias in multi-arm trials has been investigated by its impact on the type I error probability [75]. It was shown that selection bias leads to an inflation of the type I error rate when it was not accounted for in the analysis and that selection bias poses a serious risk even when the number of treatment arms or the sample size is large. It was recommended to use a randomization procedure with very few restrictions and to only use the permuted block design with large block sizes [75]. To conclude, it is of importance to choose the block size thoughtfully by considering the extent of the selection bias for different block sizes.

2.4.2 Allocation ratios

When using randomization, it is also important to consider the **allocation ratio** between the intervention and control arm(s) as the power depends on the allocation of the sample size to the different arms. In a standard randomized controlled trial where one control arm is compared to one experimental arm, there is rarely any reason to deviate from a *1:1* allocation. Pocock [74] has shown that, in many cases, there is only a minor loss of power employing allocation ratios as unbalanced as *3:1*, however, the power declines rapidly with ratios greater than *3:1*.

In a multi-arm trial the control arm is used in each comparison and assuming an equal allocation ratio in a multi-arm trial implies that there is a higher probability that a patient will be randomized to one of the experimental arms. When multiple experimental arms are being compared against a shared control arm, the optimal allocation is no longer *1:1:...:1*. The sample size reduction and efficiency gain of using a shared control arm is maximised when one allows to randomize more patients to the shared control arm than to each of the experimental arms [76, p. 96]. In a multi-arm trial where the pairwise comparisons against a common control are of interest, the allocation for pairwise comparisons is optimized with a 1:...:1:$\sqrt{k}$ randomization ($\sqrt{k}$ to control) with k being the number of experimental treatments [77, 78]. While allocating too few participants to the control arm may result in reduced power, this can be a worthwhile trade-off between power, precision and improved ethics since more patients would be allocated to promising experimental interventions [38]. As mentioned in Section 2.2, there is some evidence that the patient participation in a trial reduces as the allocation to the control arm increases [34]. When the trial design allows early stopping, the optimal allocation ratio is likely to be closer to one [8], as treatments can be dropped at interim analyses which reduces the number of treatments at each stage making the optimal allocation ratio closer to the situation of a RCT. Besides, in trials that employ response-adaptive randomization, the allocation ratios can be adapted during the trials based on data from interim analyses to allow more patients to be allocated to the superior arm(s) which provides major ethical benefits [79, 80].

Although efficiency (in terms of maximum sample size) can be gained by deviating from an equal allocation ratio in a multi-arm trial, the gain is generally fairly small [7, 78]. According to Freidlin et al. [7], the deviation from an equal allocation ratio does not justify the negative aspect of reducing the probability to get randomized to one of the experimental arms and the complexity of an unequal randomization. Wason et al. [31] also showed that by choosing the optimal allocation ratio the maximum sample size is reduced by only 2.5% compared to an equal allocation ratio. They further pointed out that one can allocate a large proportion of patients to the control arm without increasing the maximum sample size considerably. This may be of interest in cases where, for example, the control treatment is cheaper than the interventions or thought to have a better safety profile than the interventions [31].

In a platform trial, the number of arms may change over time as arms enter and

leave the trial. Determining an optimal allocation ratio to allocate patients to the treatment and control arms in a platform trial is challenging because the change in treatment arms implies that the optimal allocation ratio also changes when treatments enter or leave the platform. As result, the above introduced optimal allocation ratio for multi-arm trials may be hard to implement for platform trials. It may further be the case that the ratio is influenced by external factors [81], for example, a pharmaceutical company may choose to allocate extra patients to the drug they think is the most promising. Roig et al. [82] have shown that the optimal allocation ratio depends on the entry time of the arms in the trial and, in general, does not correspond to the $\sqrt{k}$ allocation rule used in the classical multi-arm trials. In a platform trial with four experimental arms, the allocation ratio of *2:1:1:1:1* might be ideal in order to maximise power with the allocation to the control arm being $\sqrt{4} = 2$ times higher than the allocation to the experimental arms [38, 68]. An example where this was implemented is the STAMPEDE trial, see Section 2.3.1. It had been decided to randomise two patients to the control arm for every patient randomised to each experimental arm and the rationale for this was that the control arm is used in every pairwise comparison, so higher precision provides greater power [68]. To conclude it can be said that the optimal allocation ratio for multi-arm trials remains a topic for future research and that the allocation ratio for each trial is chosen individually as the decision is made based on ethical reasons and depends on factors such as the study costs, the objective of the trial and the study design, e.g. interim analyses, the randomization method and the type of comparisons.

2.4.3 Selective exclusion of treatment arms

In classical multi-arm trials the eligibility of all enrolled patients for all treatments is usually required. However, it might be possible that some patients are not eligible for all interventions or can decide the subgroup to be randomized to. For example, for some patients the risk of complications can be high such that adverse events can be expected to exceed the presumed benefits [83]. In trials that involve several different experimental interventions, the safety profiles could possibly reduce the amount of potential patients which affects the recruitment and generalizability of the trial [84]. Patients could further potentially refuse randomization if they have a preference for one of the treatments being investigated or may reject certain aspects of the trial or its conduct. A possible scenario for that is a trial with three treatment arms, one with an aggressive treatment, one with the

minimal treatment and one with SOC. Clinicians and patients might be worried of side-effects of the aggressive treatment or think that the minimal treatment is not suitable because the SOC is available which could result in the patient not being recruited to the trial [15]. **Selective exclusion** is a possible solution for that because patients and clinicians can selectively exclude one of the randomized treatment arms, e.g. patients would then be eligible to get randomized to only one of the treatment arms or to SOC. This has the advantage that patients who might have not been recruited before would now be enrolled.

Regarding the implications of allowing selection exclusion of treatment arms for the analysis, it was recommended to base the comparison of two arms only on the patients who were directly randomized to either one of the arms [15]. That implies that not all patients who have been recruited to the trial are included in all comparisons and that the number of patients who are included in the different treatment comparisons is not equal. In a trial with three arms and continuous endpoints, where the comparisons are made in a pairwise fashion, the third arm, which is not directly compared, would indirectly contribute to the comparisons. The third arm would indirectly contribute to the comparisons through a common estimate of the variance when using analysis of variance models for the analysis. Law et al. [15] concluded that by including the patients randomized to the third arm, the variance estimates would be slightly smaller which would lead to greater power. The authors further pointed out that it is of interest that a rather large percentage of the patients enrolled in the trial is willing to get randomized to all arms so that the trial does not become two separate RCTs since then all advantages of multi-arm trials (see Section 2.2) are lost. Besides, it was highlighted that the selective exclusion of treatment arms could change the generalizability of the results of each treatment comparison [15, 84]. According to Law et al. [15], it is probably recommendable to stratify the analyses for all treatment comparisons for the subgroups since the different subgroups may have recruited patients with different prognosis. Selective exclusion can also be applied in trials with more than three treatment arms, however, the available options lead to more complex and complicated scenarios. Whether to allow for selective exclusion of treatment arms in a multi-arm trial setting must be decided individually. It is important to consider that there is a trade-off between increased sample size and decreased trial duration. Selective exclusion of treatment arms allows for faster recruitment of patients but increases the number of required patients compared to a standard multi-arm trial [15].

Two examples from the literature for trials which allow selective exclusion of treatment arms are the following: in a trial with three arms surgery, drug therapy and control, not all enrolled patients might be suitable for the type of surgery but could contribute to the evaluation of a drug. As result, patients could then be randomized to the following three subgroups: control vs surgery vs drug, control vs surgery (if the patient is not suitable for the drug therapy) or control vs drug therapy (if the patient is not suitable for the surgical procedure) [84]. In the second example, given by Molenberghs et al. [14], also a trial with three arms A, B and SOC is considered. A randomization ratio of 1:2:1 is assumed. Of interest are the treatment vs control comparisons, i.e. treatment A compared to SOC and treatment B compared to SOC. They expected that 10% of the patients are eligible for treatment A (subpopulation 1), 30% are eligible for treatment B only (subpopulation 2) and 60% are eligible for both treatments (subpopulation 3) [14]. In each of the subpopulations, the randomization is performed in such a way that the overall ratio of 1:2:1 is met, i.e. 1:1 for subpopulation 1, 2:1 for subpopulation 2 and a 1:2:1 ratio for subpopulation 3. For a comparison of treatment A vs. SOC, patients from subpopulation 1 and 3 were included. The problem arises that in subpopulation 3 half of the patients received treatment B. In order to be able to estimate the effect one would estimate in a placebo controlled trial of treatment A vs SOC, one needs to make sure that the subpopulations 1 and 3 are balanced. To achieve that Molenberghs et al. [14] suggested to weight subpopulation 3 by a factor of 2 to make sure that subpopulation 3 is not underrepresented in the comparison.

In multi-arm trials that allow selective exclusion of treatment arms, it is important that the trials' randomization system is capable of randomizing the patients between limited subsets of interventions according to, e.g. the patient background [51]. In addition, treatment allocation and assignment to the specific subgroups should be labelled for the later analysis of the data. The RECOVERY trial is an example for such a trial. RECOVERY is short for randomized evaluation of COVID-19 therapy and it is an ongoing open-label multi-centre trial that started recruitment in 2020. The trial investigates whether several treatments prevent death in patients with COVID-19 and is sponsored by the University of Oxford [85]. All eligible patients are randomized between multiple treatment arms and treatments can be added and removed during the trial. The study is divided into several parts depending on the age of the patients (adults or children) and

the geographic area. For example, in part K patients are randomised between no additional treatment and molnupiravir and in part L patients get randomised between no additional treatment and paxlovid. For patients who are not eligible for all treatment arms or at locations where not all treatment arms are available, randomisation will be between fewer arms [16]. Thereby, the RECOVERY trial design allows for selective exclusion of treatment arms by randomizing the patients between limited subsets of treatments depending on the patient background and availability due to the geographic area.

As mentioned in Section 2.4.2, the allocation ratio depends on the study design of the trial and is chosen for each trial individually. When allowing selective exclusion of treatment arms in the trial, the randomization ratio can be complex as multiple different eligibility profiles may be present. Viele [81] focused in his paper on the allocation ratios in platform trials in order to maintain comparability over time and over patients with different eligibility criteria. For that he considered platform trials with patient restrictions, e.g. due to non-availability of certain arms at certain sites or explicit exclusion criteria. For example, it might be the case that one arm has different exclusion criteria compared to the other arms and as result, several different eligibility profiles may be present in the platform trial. In an analysis based on eligibility, the randomization ratio might not balance the active and control arms. Viele showed that in case the allocation ratio is not maintained throughout the whole platform trial, which could lead to certain patient groups being more represented in the treatment arm(s) than in the control arm, bias can be introduced. This is also the case when restricting to concurrent controls. Viele proposed an algorithm that guarantees the comparability between active and control arms in both time and eligibility [81]. The idea is to set weights for each arm which are constant in time across the eligibility groups. The trial is divided into different cohorts and each cohort consists of an active treatment and a control. Patients are first randomized to a cohort and then in the second stage within the cohort either randomized to active or control. Allocation in platform trials is often performed in two stages to facilitate blinding and consent processes [86]. According to Viele, the most simple selection of weights is a weight of one for all active treatments in the platform. This would result in the first stage in an equal allocation to the cohorts and in the second stage in a ratio of *K:1* for active vs. control where K is the number of treatments for which the patient is eligible. The guaranteed comparability results in robustness in the analyses but excludes analyses that might be more efficient in certain scenarios, such as includ-

ing non-concurrent controls or response-adaptive randomization. Viele's suggested approach has the limitation that when comparing two treatments with each other, the analysis must be restricted to times where both treatments are enrolling and where the patients are eligible to both arms [81].

SOLIDARITY is one of the trials where the analysis was restricted to eligible controls to minimize the bias when using concurrent controls. It was a multi-country open-label randomized trial in patients hospitalized with COVID-19 that started in March 2020 and was conducted by the WHO [87]. The goal of the trial was to investigate whether any of the four re-purposed anti-viral drugs could affect the in-hospital mortality. Unpromising treatments could be dropped during the trial and new treatments could be added. The patients were equally randomized between the locally available treatments and open control. For analysis one was interested in the comparison of each treatment vs. control. As mentioned above, the analysis was restricted to eligible controls to minimize the bias when using concurrent controls which means that for the treatment vs. control comparisons the controls for one specific treatment were those who had the chance of being randomized to the treatment of interest [87].

2.4.4 Definition of operating characteristics

Operating characteristics refer to the statistical properties of a clinical trial, specifically related to its ability to correctly detect or identify treatment effects, control error rates, and provide valid and reliable results. In order to evaluate different trial designs, operating characteristics need to be chosen that take into account the special features of the trial design at hand. In a classical RCT, which evaluates one treatment against a control, power and type 1 error rate are used to judge the design under consideration. For multi-arm trials, the choice of operating characteristics is less obvious since multiple hypotheses are tested at the same time. Besides evaluating the operating characteristics for the individual comparisons, one could also examine the operating characteristics for the whole multi-arm trial, for which, for example, the conjunctive and disjunctive power are regarded. Table 2.3 summarizes different operating characteristics to consider.

2.4.5 Error control and multiplicity

Controlling errors during hypothesis testing involves effectively managing the risk of drawing incorrect conclusions or making errors in the process. When test-

Table 2.3: Operating characteristics and their definitions

Name	**Level of measurement**	**Definition**
Bias	treatment level	expected value of the difference between the expected mean and the true mean
Confidence Interval	treatment level	probability that parameter falls between a set of values a certain proportion of times
Marginal Power	treatment level	probability to correctly reject H_0 in case the alternative is true
Type I Error Rate	treatment level	mistaken rejection of an actually true null hypothesis
Conjunctive Power	trial level	probability to reject all false H_0
Disjunctive Power	trial level	probability to reject at least one false H_0
Family-wise Error Rate	trial level	probability to make one or more type I errors when testing multiple hypotheses

ing a single null hypothesis, H_0, against an alternative hypothesis, H_1, there are two types of errors to consider, the type I error and the type II error (see Table 2.4). The type I error describes the probability to falsely reject a true H_0, denoted by α [88], and the type II error, β, represents the probability of failing to reject a H_0 that is actually false. For a single hypothesis test, one aims to control the type I error rate at level α, if $P(\text{reject } H_0 \mid H_0) \leq \alpha$ [47]. Depending on the situation, controlling a specific error rate might be of higher relevance than another. For example, in a phase III setting a type I error could be serious since then an ineffective treatment is used in practice. On the other hand, however, the power, $1 - \beta$, i.e. the likelihood of a hypothesis test to detect a true effect if there is one, is also important since a type II error, could lead to an effective treatment not making it to practice. The type I and type II error rates influence each other and there is an important trade-off between the type I and type II error, i.e. setting a lower significance level decreases a type I error risk, but increases a type II error risk [89].

Table 2.4: Type I and type II error for testing a single hypothesis

Null hypothesis is ...	**True**	**False**
Not rejected	Correct decision probability $= 1 - \alpha$	Type II error probability $= \beta$
Rejected	Type I error probability $= \alpha$	Correct decision probability $= 1 - \beta$

When testing, for example, several hypotheses, multiple endpoints or several separate standard clinical trials, the problem of multiplicity arises, i.e. the more hypotheses are tested, the higher the probability to obtain type I errors. For

testing multiple hypotheses, denoted by H_{01}, H_{02}, ..., H_{0m}, one could test each of the m hypotheses separately with the same pre-specified significance level α and then the per-comparison error rate (PCER), which is the expected proportion of type I errors among all m decisions, is controlled at level α. However, one could also consider the hypotheses together, as a family of hypotheses, and then the global null hypothesis consists of the intersections of all null hypotheses one is interested in:

$$H_0 : H_{01} \bigcap H_{02} \bigcap ... \bigcap H_{0m} \text{ for } l = 1, ..., m.$$

Table 2.5: Possible outcomes when testing m hypotheses

	True H_0	False H_0	Total
Rejected	V	S	R
Not rejected	U	T	m-R
Total	m_0	m-m_0	m

When the hypotheses are all individually planned with the same one-sided type I error rate α, the probability of making one or more false discoveries or type I errors when testing multiple hypotheses, is called family-wise error rate (FWER) [47]. The FWER can be described by using a binomial distribution since each null hypothesis has a binary outcome (i.e. rejected, not rejected). For independent comparisons, the probability of exactly V type I errors across m comparisons can be expressed as (following the notation in 2.5):

$$P(V = v) = \binom{m}{v} \alpha^v (1 - \alpha)^{m-v},$$

since the FWER is the probability of at least one type I error [47],

$$\text{FWER} = P(V > 0) = 1 - P(V = 0) = 1 - (1 - \alpha)^m.$$

The FWER increases quickly with the number of hypotheses being tested as one can see in Figure 2.6. Figure 2.6 depicts the FWER in a multi-arm trial with a shared control arm for $\alpha = 0.025$ and $\alpha = 0.05$ (see Appendix A.2 for the considered simulation setup). While the FWER for one treatment arm is equal to the significance levels of $\alpha = 0.025$ and $\alpha = 0.05$, the FWER for 10 treatment arms is already larger than 0.14 for $\alpha = 0.025$ and larger than 0.25 for $\alpha = 0.05$. In a multi-arm trial with a shared control arm, the comparisons are not independent but correlated based on the shared control arm. For comparison, Figure 2.7

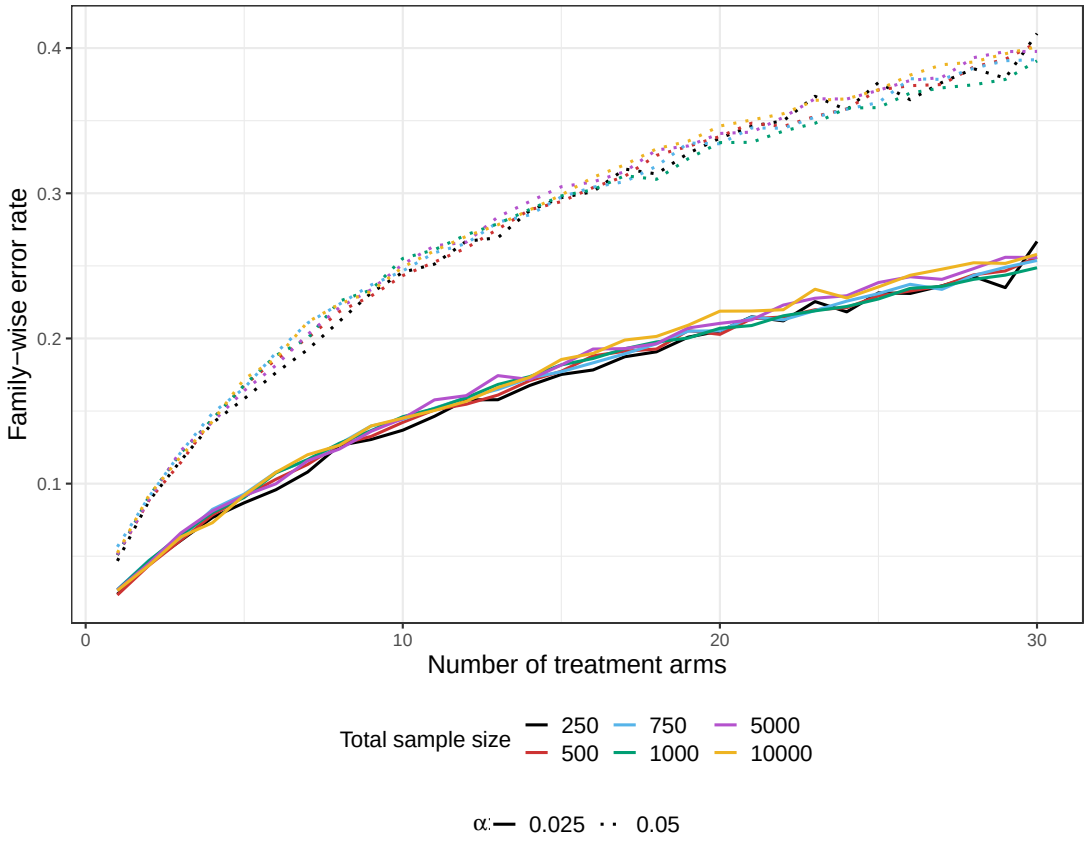

Figure 2.6: FWER over an increasing number of treatment arms in a multi-arm trial with a common control for $\alpha = 0.025$ and $\alpha = 0.05$. The different colours depict the sample size from 250 to 10000. 10000 iterations of each scenario were replicated to estimate the operating characteristic of interest. The range of the y-axis is restricted to [0, 0.4].

depicts the type I error of separate trials where each treatment has its own control arm. One can see that the FWER is larger for independent trials compared to the equivalent error performing a multi-arm trial with a shared control arm. The reason for this is that the correlation between the test statistics, due to the shared control arm, reduces the overall probability of a type I error [47, 90, 91]. The correlation of the test statistics can be decreased, for example, by allocating more patients to the control arm or by delaying the recruitment to one or more treatment arms [91]. The more interventions are included in the multi-arm trial, the higher the gain of the FWER reduction in multi-arm trials compared to independent trials.

The goal is to assure that FWER $\leq \alpha$, i.e. the probability of making one or more

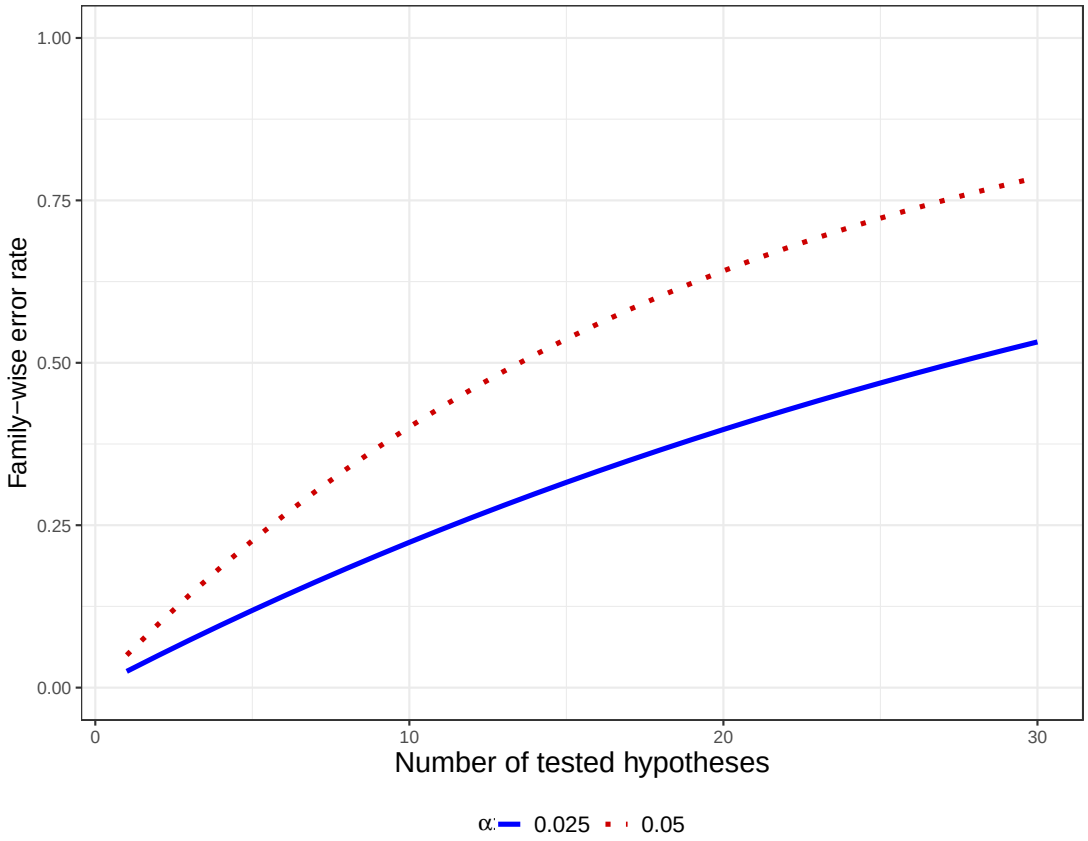

Figure 2.7: FWER over an increasing number of tested hypotheses for $\alpha = 0.025$ and $\alpha = 0.05$. Since separate trials are considered, each treatment has its own control arm.

type I errors in the family is controlled at level α because multiplicity can have an influence on the rate of false positive conclusions which may affect approval and labelling of an investigational drug [92]. Procedures that control the FWER also control the PCER but not vice versa [93]. One can control the FWER at level α in a weak sense if

$$P(\text{reject any } \mathrm{H}_{0l} | \mathrm{H}_{01}, ..., \mathrm{H}_{0m}) \leq \alpha,$$

meaning that FWER control at level α is guaranteed only when all null hypotheses are true (i.e. the global null hypothesis is true), and in a strong sense if

$$P(\text{reject any true } \mathrm{H}_{0l}) \leq \alpha,$$

meaning FWER control at level α is guaranteed for any configuration of true and false null hypotheses independent of whether the global null hypothesis is true or not. Strong error rate control implies weak error rate control but not the other

way around [93]. According to the guidance on multiplicity issues in clinical trials from the European Medicines Agency, controlling the FWER in the strong sense is required for confirmatory trials [94] as the cost of a false-positive finding is high since the treatment will probably be used in practice. For explanatory multi-arm trials on the other hand, it has been suggested that FWER control is not required [88, 89].

Several statistical tests have been developed to address the problem of an increasing FWER when testing multiple hypotheses, usually by requiring a stricter significance threshold for the individual comparisons, so as to compensate for the number of inferences being made. The Bonferroni procedure is one of the procedures which controls the FWER. It is the most conservative procedure and has no distributional or dependency assumptions. For m hypotheses H_{0l} $(l = 1, ..., m)$ with p-values p_l, the decision rule is to reject H_{0l} if $p_l \leq \alpha/m$ [95]. It requires that the number of hypotheses is fixed in advance. Other procedures are, for example, the Bonferroni-Hochberg test, the Sidak test and the Dunnett correction. A description of these procedures would go beyond the scope of this master thesis. For details on the procedures, see, for example, Chen et al. [96] and Dmitrienko et al. [97].

Table 2.6: Error rates for multiple testing

Error Rate and Probability	**Notation**	**Description**
Family-Wise Error Rate	$P(V \geq 1)$	probability to make one or more type I errors when testing multiple hypotheses
False Discovery Rate	$E(V/R)$	expected proportion of incorrect rejections among all rejected hypotheses
Per-Comparison Error Rate	$E(V/m)$	expected proportion of type I errors among all m decisions

A different point of view on the problem of multiplicity is to consider the number of incorrect rejections and not only the question whether any error was made. From this point of view, a desirable error rate to control may be the expected proportion of incorrect rejected hypotheses among all rejected hypotheses which is called the false discovery rate (FDR). The FDR is given with $E(V/R)$ where V is the number of true rejected null hypotheses and R is the number of rejected null hypotheses (assuming the notation introduced in Table 2.5) [98]. In the context of clinical trials, the FDR describes the proportion of recommended treatments that are actually ineffective. In case all null hypotheses are true, the FDR is equivalent to the FWER and otherwise it is smaller than the FWER [99]. One procedure

that controls the FDR at level α is the Benjamini-Hochberg procedure. The Benjamini-Hochberg procedure is a step-down procedure where the hypotheses are first ordered from smallest to largest p-value [98]. Then the hypotheses are either rejected or not based on their p-values using a modified Bonferroni correction defined as:

$$\text{let } m \text{ be the largest l for which } p_{(l)} \leq \frac{l}{k}q*, \text{ then reject all } H_{0l}\ l = 1, ..., m$$

where l is the rank of the p-value, k the total number of tests and q the chosen false discovery rate [98]. This implies that all p-values need to be available at the time point of decision. Procedures controlling the FDR have a greater power and control the type I error less strictly in comparison to procedures controlling the FWER. In case several treatments are successful, the FDR has the advantage that it represents the expected proportion of inefficient treatments among the efficacious treatments.

In general, it can be said that it becomes difficult to control the type I error in a platform trial because one usually does not know the number of treatments arms in advance since treatment arms enter and leave the trial at different time points [42, 45] and not all p-values are available at the time point of test decision. As result, conventional methods such as the Bonferroni method to control the FWER or the Benjamini-Hochberg method to control the FDR are no appropriate methods for platform trials and instead other methods have been suggested in the last years, for example, the online control of the FDR. The (online) control of the FDR can be seen as a compromise between no adjustment and the conservative FWER adjustment [99].

Multiplicity adjustment in multi-arm trials

In multi-arm trials there are multiple sources of structural multiplicity, e.g. multiple endpoints, multiple subgroups, multiple control groups and multiple interventions. Whether to correct for multiplicity or not is still an open issue in multi-arm trials. When considering to control for the FWER when testing multiple hypotheses, the question arises how to a define a family of hypotheses [89]. One could, for example, define a family of hypotheses by saying that all treatment to control comparisons form a single family and then control the FWER across all treatment arms or subgroups. Alternatively, one could say that each treatment to control comparison is its own family and then no control for multiplicity is needed. Traditionally, one adjusts for multiplicity when testing multiple confirmatory hypotheses in a trial but not when they are conducted in separate trials

[89]. Some argue that a multi-arm trial is a collection of independent trials and aims and therefore, no correction is recommended since adjustments would not be necessary when the interventions were compared in separate trials [7, 47, 88, 100].

The need for multiplicity control in a multi-arm trial depends on the relatedness of the hypotheses being tested. When the hypotheses are being tested independently, e.g. in different substudies, no further multiplicity correction is necessary due to simply sharing a protocol [45, 47]. In case the use of each treatment is restricted to a single subgroup, the treatments for randomization depend on the subgroup so that a patient is only randomized once their subgroup is known and the effects of the treatments are assessed in each of the subgroups [101]. However, when testing different doses of the same intervention, correcting for multiplicity is more appropriate compared to testing multiple arms with unrelated interventions [101]. The potential dependencies between the hypotheses can, among other criteria, be taken into account when defining a family of hypotheses. Further scenarios where adjustment of the FWER is recommended have been mentioned, for example,: if there is an increased chance of making a single claim of effectiveness by testing multiple hypotheses in a multi-arm trial [47] and when testing of multiple hypotheses increases the chance of erroneously declaring a given ineffective treatment to be effective in the population [101].

To illustrate the different arguments in the discussion about multiplicity adjustment, Odutayo et al. [102] included a decision tool in their paper to determine the need for multiplicity adjustment in multi-arm trials. Their decision tool focused on multiplicity issues due to the number of interventions. The tool has three decision points (simplified) which summarize the above mentioned arguments:

1. Is it a confirmatory or exploratory trial?

2. What type of confirmatory trial is it, e.g. multiple interventions vs a common control?

3. Is independence of the interventions given?

In the statistical analysis plan of the study MND-SMART it was also discussed whether multiplicity adjustment is necessary for the design at hand [103]. MND-SMART, which stands for Motor Neuron Disease Systematic Multi-Arm Adaptive Randomised Trial, is a multi-arm, multi-stage, multi-centre platform randomized

controlled trial which recruits people with motor neuron disease which is an incurable neurodegenerative disorder. Even though Parker et al. [103] had stated in earlier versions of their protocol that at the moment, there is no consensus whether control of the FWER is required in MAMS trials, the plan was initially to control the FWER with respect to the multiple treatment comparisons. However, over time they had the feeling that the majority of the experts discussing multiplicity correction in the literature was against it (see for example Odutayo et al. [102], Parker and Weir [89], Collignon et al. [45] and Molloy et al. [104]). Besides, they planned to test multiple hypotheses on different treatments. As result, they decided to not adjust for multiplicity and changed their trial design to pairwise error control which had the advantage that it increased the power considerably. The authors pointed out that in each situation the decision must be made in consideration of the legal environment [103].

As the question whether to correct for multiplicity or not is still not answered for multi-arm trials, Molloy et al. [104] asked in their commentary letter for clearer guidance from stakeholders, such as regulators and scientific journals, on the specific settings for which multiplicity adjustment is required.

3 Methods for the statistical analysis

In this chapter, the methods for the statistical analysis will be presented. To evaluate the type I error rate and power of the randomization procedures, analysis of variance (ANOVA) models can be considered. In the following sections, first the theoretical background of ANOVA models will be introduced and the relationship to linear regression will be highlighted (see Section 3.1). Then, contrasts are discussed (see Section 3.1.2) as well as the possibility to adjust for covariates (see Section 3.1.3).

3.1 Analysis of variance

The ANOVA is used to analyze differences between means and in its simplest form, it can be seen as an extension of the t-test. In a t-test one investigates whether two population means are equal and the ANOVA extends this to more than two groups. The one-way ANOVA with two factor levels is equivalent to the two-sample t-test [105, p. 525].

The standard one-way ANOVA is the appropriate choice for the analysis of a single factor, e.g. treatment, which has more than two factor levels, e.g. treatment arm 1, treatment arm 2 and treatment arm 3. A statistical model for the one-way ANOVA for an outcome Y_{ij} with i denoting the factor level, e.g. the treatment arm, and j representing the number of the outcome in the factor level is the following:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \text{ for } i = 1, ..., I, \text{ and } j = 1, ..., n_i, \text{ where}$$

Y_{ij} is the observation of the i-th treatment arm and j-th observation unit in the treatment arm,

μ is the overall expected value, i.e. $\mu = \frac{1}{n}\sum_{i=1}^{I} n_i\mu_i$ where μ_i is the expected value of treatment arm i,
α_i is the effect of the i-th treatment arm with $\alpha_i = \mu_i - \mu$,
ϵ_{ij} are the residuals [105, p. 521].
I denotes the number of treatment arms and $n_1, ..., n_I$ the sample size in each treatment arm.

As it is assumed that the dependent variable in the treatment arms is normally distributed, one can further assume that the residuals are normally distributed with $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ [105, p. 521]. This assumption means that the residuals balance out on average and that the variability is the same in all factor levels. Note, the factor levels, e.g. the treatment arms, do not need to be of the same size [106, p. 522].

The estimators for the overall mean and the treatment specific effects, $\hat{\mu}$ and $\hat{\alpha}_i$ are given as [105, p. 522]

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i\cdot} \text{ and } \hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} \text{ with } \overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}.$$

The dots indicate that the subscript has been summed over and the bar shows that the mean has been taken. The question of interest is whether the treatment arms differently affect the dependent variable. Under the null hypothesis there are no varying effects on the dependent variable due to the treatment arm. Under the alternative hypothesis, such an effect is present.

$$\mathrm{H}_0\text{: } \alpha_1 = \alpha_2 = ... = \alpha_I = 0 \text{ vs. } \mathrm{H}_1\text{: at least two } \alpha_i \neq 0.$$

In order to quantify the variation between and within the treatment arms, the total sum of squares is computed which measures the total variability. It is an unscaled measure of dispersion and calculates the sum of the squared deviations of each observation from the overall mean or the treatment means. The total sum of squares can be divided into the sum of squares between and within the treatment arms [107, p. 127]:

$$SS_{Total} = SS_{Between} + SS_{Within}$$
$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$$

The total sum of squares (SS_{Total}) represents the total variability in the outcome variable, while the sum of squares between ($SS_{Between}$) represents the variability between different arms and the sum of squares within (SS_{Within}) measures the

variation within each treatment arm. When scaled for the number of degrees of freedom, one gets the mean square. By dividing the sum of squares by the corresponding degrees of freedom, the sum of squares is adjusted so that it does not increase as the sample size increases. Again, the idea is to divide the total variability into variability within the treatment arms (MS_{Within}) and variability between the treatment arms ($MS_{Between}$). The mean square within the treatment arms is given as

$$\text{MS}_{Within} = \frac{SS_{Within}}{n-I} \; \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2.$$

and the mean square between the treatment arms is given as

$$\text{MS}_{Between} = \frac{SS_{Between}}{I-1} = \frac{1}{I-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

By dividing the sum of squares by its degrees of freedom, the mean square provides an average measure of variation that can be used for hypothesis testing and comparing differences in the outcome variable between treatment arms. Table 3.1 provides an overview of the quantification of variation.

The test statistic is then defined as

$$\text{F} = \frac{MS_{Between}}{MS_{Within}} = \frac{\frac{1}{I-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{\frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2}.$$

Table 3.1: Analysis of variance: quantification of variation

Variation	**Degrees of freedom**	**Mean squared error**	**Test statistic**
$SS_{Between}$	I - 1	$MS_{Between} = \frac{SS_{Between}}{(I-1)}$	$\text{F} = \frac{MS_{Between}}{MS_{Within}}$
SS_{Within}	n - I	$MS_{Within} = \frac{SS_{Within}}{(n-1)}$	
SS_{Total}	n - 1		

The null hypothesis is rejected if the variability between the treatment arms is substantially larger than the variability within the treatment arms: $MS_{Between} > MS_{Within}$. In other words: the null hypothesis at significance level α is rejected if $\text{F} > \text{F}_{1-\alpha}(I-1, n-I)$ [105, p. 525]. To determine if the F statistic is statistically significant, it is compared to a critical value from an F-distribution table or calculated using the degrees of freedom associated with the ANOVA. In the case of a significant result, one knows that there is a significant difference between the

treatment arms but not between which arms. In order to answer that question one can follow one of the subsequent approaches. One can either calculate multiple t-tests, where one must be aware of the multiple testing problem (explained in Section 2.4.5), or calculate post-hoc tests, for example, the Tukey's range test or the Dunnett's test when performing pairwise comparisons against a common control [77]. Another approach is to calculate so-called contrasts (see Section 3.1.2). For the post-hoc tests all pairwise comparisons are calculated which implies that pairwise comparisons are evaluated which might not be of interest for the research question and that one needs to adjust for multiplicity for testing all comparisons. On the other hand, when calculating contrasts one can only test the pairwise comparisons of interest. Therefore, one needs to have the hypotheses about the differences between the factor levels defined before calculating the ANOVA.

It is important to keep in mind that the ANOVA model is based on a series of model assumptions that must be examined or at least critically questioned in each individual case. The assumptions to consider are the following:

(i) the observations are normally distributed on the dependent variable in each treatment arm,

(ii) independence of the observations [105, p. 528],

(iii) homogeneity of the variance in the treatment arms [106, p. 524].

The statistical model of the ANOVA can also be written as a linear model:

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, ..., n, \text{ with } \epsilon_i \sim N(\mu, \sigma^2) \text{ where}$$

Y_i is the response for the i-th observation,
α is the intercept,
β is the regression coefficient,
x_i are the predictor variables,
ϵ_i are the residuals [106, p. 539].

The variance-analytical model given above can be extended in order to satisfy various problems, e.g. for repeated measurements or more than one outcome variable. It can also be extended to examine the influence of two factors on one continuous dependent variable as described in the following.

3.1.1 Two-way analysis of variance

When looking at the influence of two factors on a response variable, the question arises whether the factors have a different effect when considered together than when each of the factors is considered individually. With the two-way ANOVA one can assess the main effects of two factors and also if there is an interaction between the independent variables. For that the statistical model from the one-way ANOVA can be extended to the following model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \text{ for } i = 1, ..., I,$$
$$j = 1, ..., J \text{ and } k = 1, ..., K, \text{ where}$$

Y_{ijk} is the dependent variable,
μ is the overall expected value,
α_i is the main effect of the i-th level from factor A,
β_j is the main effect of the j-th level from factor B,
$(\alpha\beta)_{ij}$ is the interaction effect of the i-th level from factor A and the j-th level from factor B,
ϵ_{ijk} are the residuals [105, p. 531].

I denotes the number of factor levels for factor A, J represents the number of factor levels for factor B and K is the number of observations within each factor combination. For simplicity it is assumed that for each factor combination an equal number of observations is given.

As mentioned above, it is assumed that the dependent variable in the factor levels is normally distributed and one can further assume that the residuals are normally distributed with $\epsilon_{ijk} \sim N(0, \sigma^2)$ [105, p. 531]. The two-way ANOVA makes the same assumptions as the one-way ANOVA which have been stated above. The partition of the sum of squares work similarly to the one-way ANOVA but the mean square between the factor levels is divided for different effects: main effect of the first factor, main effect of the second factor and the interaction between both factors [105, p. 538]. These effects are used to calculate the F-statistics and p-values for the main effects and interactions. Besides, the two-way ANOVA can also be written as a linear model.

3.1.2 Contrasts

In order to maintain the type I error rate a common approach when analysing multi-arm randomized controlled trials is to first perform an overall heterogeneity test of the arms and only perform pairwise comparisons of treatment arms in case the first test is significant [21, p. 230], as explained above. The pairwise comparisons, so-called contrasts, are used to test specific hypotheses and perform comparisons within the framework of the analysis of variance models [106, p. 526]. Contrasts are linear combinations of the treatment means used to perform specific comparisons of interest and allow to focus on specific comparisons within the overall ANOVA framework. In order to perform the specific comparisons of interest, the hypotheses about the differences between the factor levels, e.g. treatment arms, need to be defined before calculating the ANOVA.

Any comparison between treatment means can be represented as a linear contrast in the form of a weighted sum (linear combination) of the means. Let $\mu = (\mu_1, ..., \mu_m)$ be a set of parameters or statistics and let c $= (c_1, ..., c_m)$ be known constants [106, p. 526]. Then, the linear combination of the treatments means μ_i is the following:

$$\Lambda = c_1 * \mu_1 + c_2 * \mu_2 + ... + c_m * \mu_m = \sum_{i=1}^{m} c_i * \mu_i \text{ with } \sum_{i=1}^{m} c_i = 0$$

The known constants c_i, so-called weights, are chosen such that they express the desired comparison. The sign of the weights must correspond to the direction in the hypothesis and the weight for the treatment arm which is not included in the comparison is set to zero.

Two contrasts, $\sum_{i=1}^{m} c_i * \mu_i$ and $\sum_{i=1}^{m} d_i * \mu_i$, are called orthogonal if

$$\sum_{i=1}^{m} c_i * d_i = 0.$$

Linear contrasts can be converted into sum of squares (introduced above) with one degree of freedom and n representing the number of observations per treatment arm:

$$SS_{contrasts} = \frac{n(\sum_{i=1}^{m} c_i * \mu_i)^2}{\sum_{i=1}^{m} c_i^2}.$$

For the implementation consider the following example: it is of interest to compare three different learning methods A, B and C. An ANOVA was calculated for the comparison and the result was significant. Now, the question is of interest between which learning methods the differences are. In order to answer that

question, contrasts for comparing the three learning methods can be defined. For example, the contrast for comparing two learning methods, A and B, out of the three learning methods, can be defined as:

$$\Lambda_1 = 1 * \mu_A + (-1) * \mu_B + 0 * \mu_C$$

Since learning method C is not included in the comparison, its weight is set to zero. It is of interest to test whether learning method A is better than learning method B and the sign of the weights match to the direction of the hypothesis. A more common way is to display the weights in a vector [108]:

$$\Lambda_1 = [1\ \text{-}1\ 0].$$

It is often of interest whether the contrast is significantly different from zero. The hypotheses for that could, for example, be the following:

- two-sided: H_0: $\Lambda_1 = 0$ and H_1:$\Lambda_1 \neq 0$,
- one-sided: H_0: $\Lambda_1 \leq 0$ and H_1:$\Lambda_1 > 0$.

In words: learning method A is different from learning method B (two-sided) and learning method A is better than learning method B (one-sided).

To summarize, a contrast can be calculated after a significant overall heterogeneity test to perform the specific comparisons of interest [21, p. 230].

3.1.3 Covariate adjustment

In addition to the primary outcome, there are often covariates which are also important for the result as the primary outcome is often not only related to the treatment. A baseline covariate in the context of this master thesis is defined as a variable or factor which is measured or observed before randomization and which is expected to influence the primary outcome. Covariates that potentially influence the primary outcome depend on the context of the study and are, for example, demographic factors such as gender, age and weight or diagnostic factors such as the disease state. Baseline covariates can be considered at two stages in a clinical trial: at the randomization process or in the analysis. It is not advisable to adjust the main analysis for covariates measured after randomization because they may be affected by the treatments [48]. There are many different techniques to account for baseline covariates. One of them is the analysis of covariance which

is often used in clinical trials that make use of pre-treatment baseline measurements of outcome variables [109]. The ANCOVA can be seen as an extension of the ANOVA. It can be used to answer the question whether there is a statistically significant difference between more than two independent groups after accounting for one or more covariates. According to Senn [110, p. 105], the ANCOVA is an appropriate measure for clinical trials to produce an unbiased estimate of the causal effect of treatment. The treatment effect describes the difference between what happens when the treatment is given and what would have happened had the treatment been denied [110, p. 101]. For that the covariates which are expected to have an important influence on the primary outcome are identified by, for example, literature research and practical experience. The number of covariates that can be accounted for statistically depends on the sample size. Covariates must be taken into account in the planning and the study design and it is important to adjust for the influence of baseline covariates in order to improve the precision and compensate for any lack of balance between treatment arms [48]. It is a way of controlling for initial individual differences that could not be randomized as potential baseline differences between treatment groups in prognostic factors reduce the power of significance tests. That is because, when considering the F statistic introduced above (see Table 3.1): $F = \frac{MS_{Between}}{MS_{Within}}$, the influence of the covariates is part of the unexplained variance which is grouped in the denominator. When adjusting for the influence of covariates, the unexplained variance no longer includes the influence of the covariates. Therefore, F gets larger and as result, the power is increased.

4 Simulation and implementation

This chapter focuses on describing the design of the multi-arm trial considered in the simulation study and presents the chosen design parameters for the different simulation scenarios. Then, the hypotheses of interest (see Section 4.3) and the data selection process are explained (see Section 4.4). Finally, the implementation in R is described (see Section 4.6).

4.1 Motivation

Since the statistical analysis of multi-arm trials that allow the selective exclusion of treatment arms has not received much attention, the following simulation study enhances the methodology for the optimal usage of different patient populations in the analysis of the trial. It was decided to consider a multi-arm trial with concurrent experimental arms since it is the most efficient design [82]. Besides, the implications of allowing different patient populations are largest for the concurrent comparisons of treatment arm vs. control arm. The objectives of the simulation study are to consider a setup where not all patients can be randomized to all available open treatment arms and to scrutinize how to utilize the information from those patients when making inference. Besides, different randomization strategies to the treatment arms will be investigated. The impact on the type I error and the statistical power is of interest.

4.2 Simulation setup

For the simulation study a multi-arm trial with two experimental treatment arms that enter the trial at the beginning of the trial and a shared control arm is considered (see Figure 4.1). The recruitment of the control arm and the two treatment arms start at the beginning of the trial and patients are recruited during the whole observation period. The considered multi-arm trial allows the selective exclusion of treatment arms. Since a trial with three arms is considered, this

results in three different patient subgroups X, Y and Z, as stated in Table 4.1 and visualized in Figure A.1,:

- subgroup X: patients eligible to only treatment 1,
- subgroup Y: patients eligible to only treatment 2,
- subgroup Z: patients eligible to both treatments.

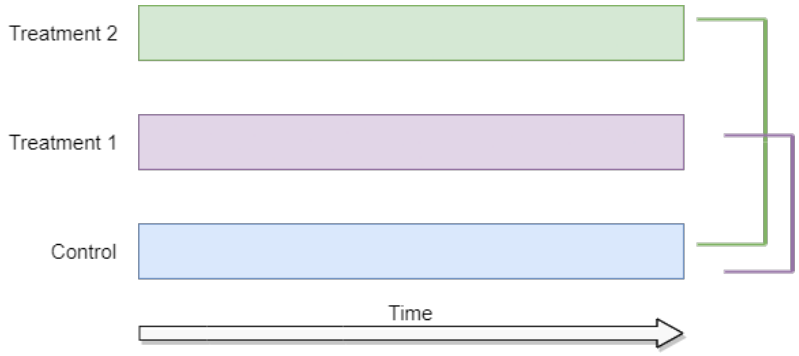

Figure 4.1: Visualization of the considered simulation setup: a multi-arm trial comprising two treatment arms and one control arm is considered. All arms start at the beginning of the trial and the patients are recruited during the whole observation period. Of interest are the pairwise comparisons of treatment 1 vs. control and treatment 2 vs. control.

Table 4.1: Different subgroups considered for the simulation

		Subgroups		
		X	Y	Z
Treatment arms	Treatment 1	Yes	No	Yes
	Treatment 2	No	Yes	Yes
	Control	Yes	Yes	Yes

In this thesis, the distribution of patients in the three subgroups X, Y and Z is referred to with the word *prevalence*. For the distribution to the subgroups, a random prevalence distribution as well as a deterministic prevalence distribution was tested (see 5.2 for explanations). It was distinguished between three settings for the prevalence to the subgroups in which the design is varied according to the objectives of the simulation study:

- **Setting 1**: a prevalence of 0:0:1 to the subgroups X, Y and Z is considered. It illustrates the extreme case where all patients are in subgroup Z. As subgroup Z represents the subgroup in which patients can get randomized

to all three arms, this setting is equivalent to a study design that does not allow for selective exclusion of treatment arms, i.e. the setting does not consider different patient populations. The setting corresponds to a traditional multi-arm design (see Figure 4.2).

- **Setting 2**: a prevalence of 0.5:0.5:0 to the subgroups X, Y, Z is regarded. It illustrates the second extreme case where patients get recruited from subgroups X and Y. This setting corresponds to two completely independent substudies.

- **Setting 3**: patients get recruited from all three subgroups X, Y and Z. For that different prevalence to the three subgroups are considered:
  - an equal prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X:Y:Z
  - an unequal prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X:Y:Z
  - a high prevalence to subgroups X and Y with $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X:Y:Z
  - a low prevalence to subgroups X and Y with $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X:Y:Z

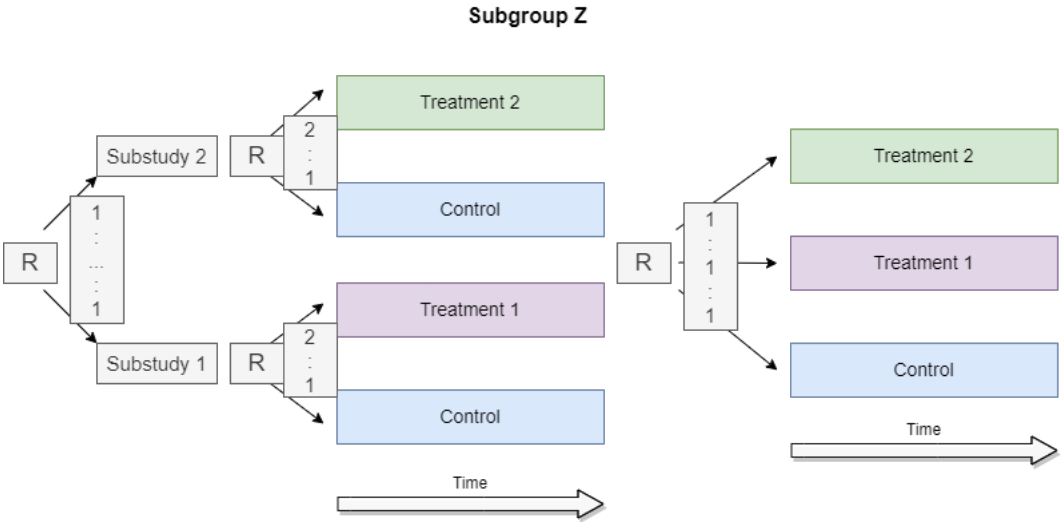

Figure 4.2: Illustration of subgroup Z for the considered simulation study: for setting 1, the case where all patients are only in subgroup Z, there is no difference between a trial with two substudies and a traditional multi-arm design. R stands for randomization, e.g. for a trial with two substudies one is first randomized to one of the two substudies and then within each substudy to either treatment arm or control and for the multi-arm trial one is directly randomized to one of the arms.

Table 4.2: Randomization strategies to the three subgroups X, Y and Z

Randomization strategy/ Subgroup	**R-I**	**R-II**	**R-III**
Subgroup X	2:1 ratio for T1 vs. C	1:1 ratio for T1 vs. C	1:1 ratio for T1 vs. C
Subgroup Y	2:1 ratio for T2 vs. C	1:1 ratio for T2 vs. C	1:1 ratio for T2 vs. C
Subgroup Z	1:1:1 ratio for T1 vs. T2 vs. C	1:1:2 ratio for T1 vs. T2 vs. C	1:1:1 ratio for T1 vs. T2 vs. C

Figure 4.2 stresses the fact that for setting 1, where all patients are recruited from subgroup Z, there is no difference between a trial with two substudies and a traditional multi-arm design. For the remainder of this thesis the following abbreviations will be used to refer to the three treatment arms: T1 for treatment 1, T2 for treatment 2 and C for control. For the three prevalence settings, scenarios with different randomization strategies to the three arms were considered (see Table 4.2 and Figure 4.3):

- **R-I**: for subgroups X and Y the ratio is 2:1 for T1 vs. C and T2 vs. C and for subgroup Z the ratio is 1:1:1 for T1 vs. T2 vs. C (see Figure 4.3 **A)**).
- **R-II**: for subgroups X and Y a ratio of 1:1 within the substudies for T1 vs. C and T2 vs. C is assumed while for subgroup Z a ratio of 1:1:2 for T1 vs. T2 vs. C (see Figure 4.3 **B)**).
- **R-III**: an equal ratio within each subgroup is assumed, meaning that for subgroup Z the ratio is 1:1:1 for T1 vs. T2 vs. C and for subgroups X and Y the ratio is 1:1 for T1 vs. C and T2 vs. C (see Figure 4.3 **C)**).

Within the substudies, patients are randomized to the different arms with either complete randomization or block randomization. The block size depends on the randomization strategy: for **R-I** a block size of 3 was chosen for all three subgroups. The idea was to thereby allocate more subjects to the treatment arm in subgroups X and Y. For **R-II** a block size of 2 for subgroups X and Y was chosen and for subgroup Z a block size of 4. As result, overall more patients are randomized to the control arm. In **R-III** a block size of 2 was chosen for subgroups X and Y and a block size of 3 for subgroup Z to achieve an equal ratio within each subgroup. For simplicity the smallest block size that meets the desired ratio for each randomization strategy was chosen for the simulation.

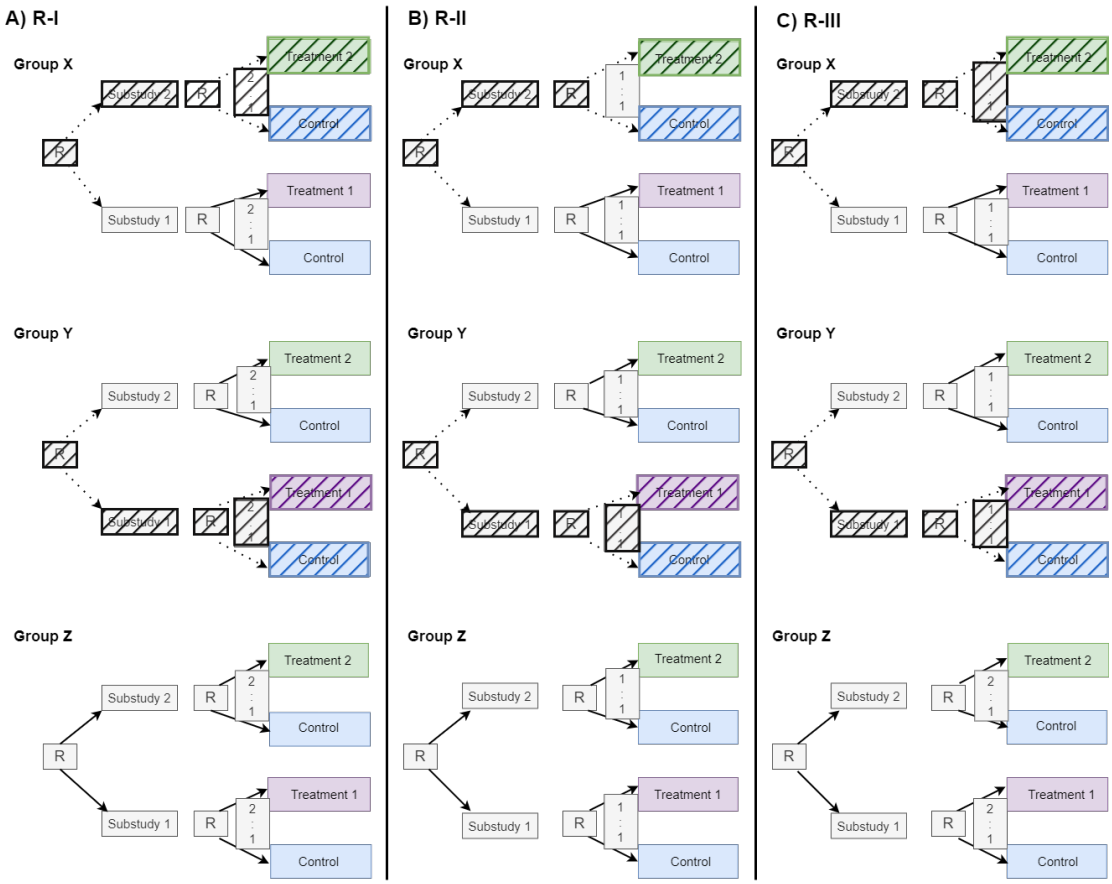

Figure 4.3: Visualization of the randomization strategies to the three subgroups X, Y and Z for the simulation study: **A) R-I** a 2:1 ratio for T1 vs. C and T2 vs. C for subgroups X and Y and a ratio of 1:1:1 for T1 vs. T2 vs. C for subgroup Z. **B) R-II** a 1:1 ratio for T1 vs. C and T2 vs. C for subgroups X and Y and a 1:1:2 ratio for T1 vs. T2 vs. C for subgroup Z. **C) R-III** a 1:1 ratio for T1 vs. C and T2 vs. C for subgroups X and Y and a 1:1:1 ratio for T1 vs. T2 vs. C for subgroup Z. R stands again for randomization.

Figure 4.3 provides an overview of the considered randomization strategies and the implications for the different subgroups. The figure emphasizes that for subgroup Z in order to achieve a 1:1:1 ratio for T1 vs. T2 vs. C, within the substudies a ratio of 2:1 for T1 vs. C and T2 vs. C is employed. That this is possible has been stressed by Figure 4.2 which shows that for subgroup Z there is no difference between a trial with two substudies and a traditional multi-arm design.

To generate the trial data, different total sample sizes are considered (n = 30, 75, 150, 300). The sample size per arm is influenced by the different randomization strategies introduced above. For example, for **R-I** the sample size per arm for a total sample size of 150 is 50 while for **R-II** the sample size per arm for a total sample size of 150 is, for example, 37:38:75 for T1 vs. T2 vs. C. The stopping criterion after which the analysis is conducted is that the total number of patients is enrolled in the trial.

The continuous outcome Y_{ij} for the patients in treatment arm i and subgroup j is drawn from a normal distribution according to:

$$Y_{ij} \sim N(\mu_j^i, \sigma^2) \text{ where } i \in \{T1, T2, C\} \text{ and } j \in \{X, Y, Z\},$$

with

$$\mu_j^i = \Delta_j^i + \mu_j^C \text{ for } i \in \{T1, T2\}$$

and

$$\sigma^2 = 1,$$

where μ_j^C is the response in the control arm and Δ_j^i the effect of treatment i in subgroup j. For the sake of simplicity the variances are assumed to be equal to 1. Moreover, different cases for the treatment effects were considered in the simulation study:

- **Case 1:** the treatment effects in the subgroups are the same: $\Delta_X^{T1} = \Delta_Z^{T1}$ and $\Delta_Y^{T2} = \Delta_Z^{T2}$
- **Case 2:** the treatment effects in the subgroups are different: $\Delta_X^{T1} \neq \Delta_Z^{T1}$ and $\Delta_Y^{T2} \neq \Delta_Z^{T2}$

Table 4.3 summarizes the considered simulation scenarios and parameters. For one prevalence setting approximately 15000 different simulation scenarios were considered. 10000 replicates of each scenario were generated to estimate the operating characteristics of interest. The empirical proportion of times when the null hypothesis is rejected provides us with an estimate of the type I error rate (when the null hypothesis is true) or the power (when the alternative hypothesis is true).

Table 4.3: Simulation setup overview

Name	**Investigated values**	**Description**
Number of treatment arms	2	Number of treatment arms that get compared to the shared control arm.
Sample size n	n = 30, 75, 150, 300	Total sample size after which the analysis is conducted. The sample size in the subgroups is influenced by the prevalence and in the arms by the randomization strategy.
Randomization method	Complete randomization, block randomization	Patients are randomized by complete or block randomization to the different arms.
Randomization strategy	R-I, R-II, R-III	Randomization strategy by which the patients get randomized to the three arms T1, T2, C.
Prevalence	Setting 1: 0:0:1, Setting 2: 0.5:0.5:0, Setting 3: - $\frac{1}{3}:\frac{1}{3}:\frac{1}{3}$ - $\frac{1}{4}:\frac{1}{4}:\frac{1}{2}$ - $\frac{2}{5}:\frac{2}{5}:\frac{1}{5}$ - $\frac{1}{10}:\frac{1}{10}:\frac{4}{5}$	Distribution of patients in the three different subgroups X, Y and Z.
Standard deviation σ	$\sigma = 1$	For simplicity is the standard deviation for the normally distributed random outcome variables the same for all subgroups and treatment arms.
Treatment effect T1 Δ^{T1}	Case 1: $\Delta_X^{T1} = \Delta_Z^{T1}$ Case 2: $\Delta_X^{T1} \neq \Delta_Z^{T1}$	Treatment effect for T1 vs. C in both subgroups X and Z.
Treatment effect T2 Δ^{T2}	Case 1: $\Delta_Y^{T2} = \Delta_Z^{T2}$ Case 2: $\Delta_Y^{T2} \neq \Delta_Z^{T2}$	Treatment effect for T2 vs. C in both subgroups Y and Z.
Significance level α	$\alpha = 2.5\%$	Significance level for one-sided testing of the pairwise comparisons.

4.3 Hypotheses of interest

The above mentioned approach of only performing pairwise comparisons of treatment arms after a significant overall heterogeneity test would imply for the design at hand, when following Law‘s recommendation [15] and basing the comparisons of treatment vs. control only on the patients who are directly randomized between the two arms, that only the patients who are enrolled in subgroup Z (the overall randomization) would be included in the overall heterogeneity test. That is not efficient and not the appropriate approach for analysing the study design of the simulation. Therefore, pairwise comparisons are performed as not all recruited patients are included in all comparisons. For the pairwise comparison of T1 vs. C and T2 vs. C, contrasts for each considered scenario are calculated based on the linear model which would be estimated for the ANOVA. For the case where one does not adjust for the subgroups, it is of interest how the factor treatment arm (T1, T2, C) affects the response variable outcome. For the two-way ANOVA the two factors treatment arm (T1, T2, C) and type of subgroup (X, Y, Z) are taken into account and thereby, one adjusts for the subgroups. The interaction of the two factors was not considered.

The following one-sided hypotheses are considered (rejecting at significance level $\alpha = 0.025$) for the pairwise comparisons of interest, where μ_{T1} is the mean in T1, μ_{T2} the mean in T2 and μ_C the mean in C.

$$\begin{aligned} &\text{H}_0\text{: } \mu_{T1} = \mu_C \text{ and } \text{H}_1\text{:} \mu_{T1} > \mu_C, \\ &\text{H}_0\text{: } \mu_{T2} = \mu_C \text{ and } \text{H}_1\text{:} \mu_{T2} > \mu_C. \end{aligned}$$

Since the hypothesis is that the mean of the treatment arm is higher than the mean of the control arm, the contrast for T1 vs. C is determined as follows:

$$\Lambda_1 = 1 * \mu_{T1} + 0 * \mu_{T2} - 1 * \mu_C$$

The sign of the weights must correspond to the direction in the hypothesis and the weight for the arm which is not included in the comparison is set to 0. The contrast for the comparison of T2 vs. C is respectively:

$$\Lambda_2 = 0 * \mu_{T1} + 1 * \mu_{T2} - 1 * \mu_C$$

The chosen weights for the statistics are 1 and -1 and the vectors displaying the weights are the following:

$$\Lambda_1 = [1\ 0\ \text{-}1] \text{ and } \Lambda_2 = [0\ 1\ \text{-}1]$$

The matrix which follows from combining the two vectors is given with:

$$L = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

4.4 Data selection

In the analysis, the impact of the composition of the control data is of interest and the above defined contrasts are calculated for the three different control data compositions.

First, *all data* are considered for the analysis, depicted in Figure 4.4 **A)**. That means that the full data set is regarded, e.g. regardless whether the patients were in the subgroup or arm of interest for the comparison they are included for the comparisons.

Next, *all control data* are regarded, see Figure 4.4 **B)**, which means that the control data of all patients are used for the analysis independent of the subgroup they were recruited from. For example, for a comparison of T1 vs. C, one would also include the control data of patients in subgroup Y who never had the chance of getting treated with T1.

Finally, *restricted data* are taken into account, represented in Figure 4.4 **C)**, which refers to the patients who had the chance of getting treated with the treatment of interest. For a comparison of T1 vs. C, one would only use the data of patients in subgroups X and Z who were randomized to either T1 or C.

In order to calculate the different control data compositions, the data set is split accordingly. Tables 4.4, 4.5, 4.6 show the distribution of patients per arm and subgroup for the different data sets for the comparison of T1 vs. C for the three considered randomization strategies for a single simulation run. The fixed design parameters were the following: a total sample size of 150, complete randomization and an equal prevalence to the three subgroups X, Y and Z. Tables 4.5 and 4.6 show that for *all data* and *all control data*, one has more patients in the control arm than in the treatment arm for **R-II** and **R-III**. Besides, one can see that for **R-I** (see Table 4.4) one has less controls for the comparisons based on *restricted data*. In comparison, for **R-II** and **R-III** (see Tables 4.5 and 4.6) the distribution of patients to T1 and C is almost equal for comparisons based on *restricted data*.

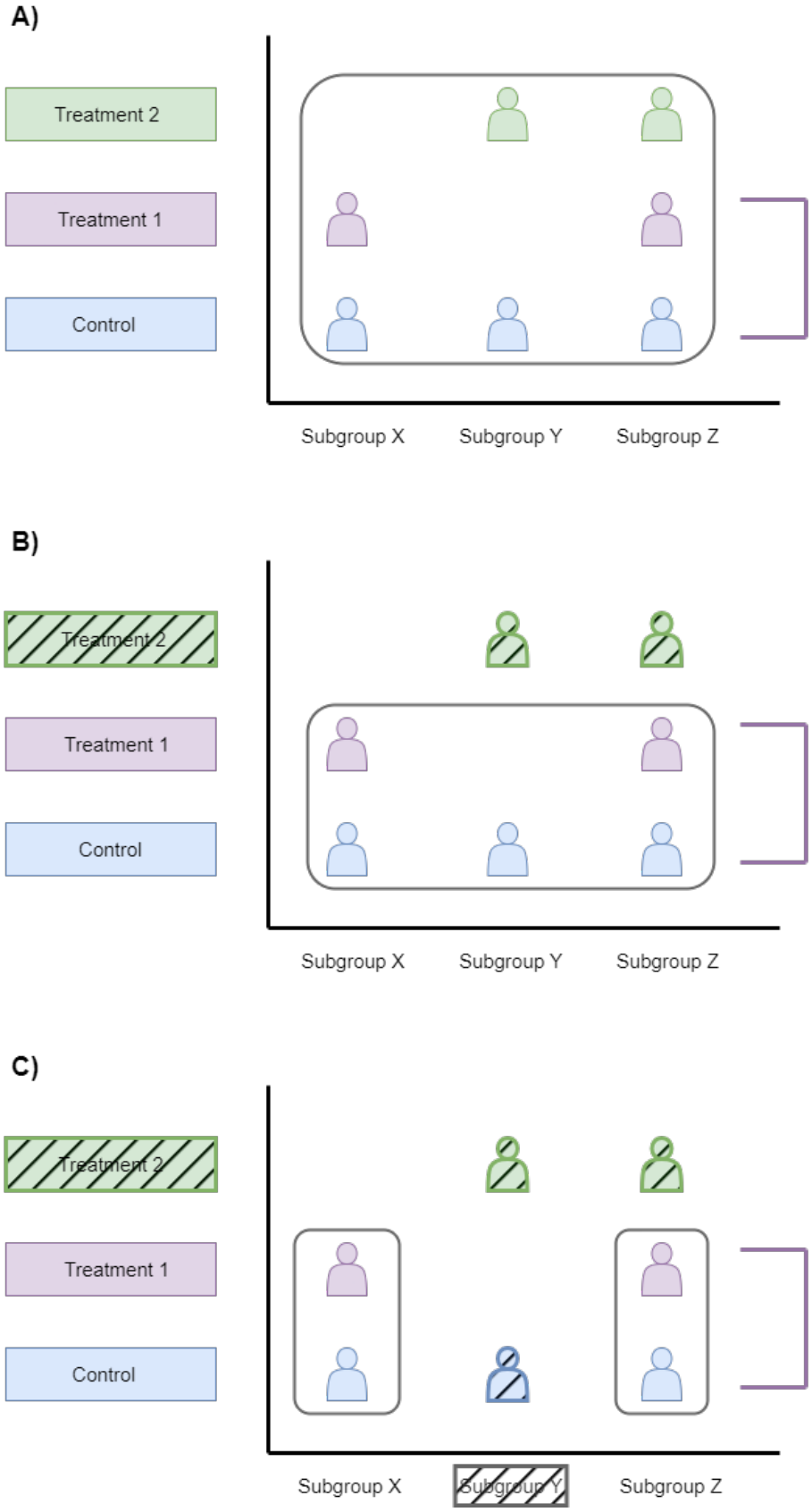

Figure 4.4: Illustration of the different compositions of control data for the comparison of T1 vs. C. **A)** *All data* are used for the comparison. **B)** *All control data* are used for the comparison. **C)** *Restricted data* are used for the comparison.

Table 4.4: **R-I:** Distribution of patients for the comparison of T1 vs. C for complete randomization and a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X:Y:Z

Data set		*All data*				*All control data*				*Restricted data*			
Treatment arms		T1	T2	C	$\sum$	T1	T2	C	$\sum$	T1	T2	C	$\sum$
Subgroups	X	30	0	20	50	30		20	50	30		20	50
	Y	0	32	18	50			18	18				
	Z	19	18	13	50	19		13	32	19		13	32
Total per arm		49	50	51	150	49		51	100	49		33	82

Table 4.5: **R-II:** Distribution of patients for the comparison of T1 vs. C for complete randomization and a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X:Y:Z

Data set		*All data*				*All control data*				*Restricted data*			
Treatment arms		T1	T2	C	$\sum$	T1	T2	C	$\sum$	T1	T2	C	$\sum$
Subgroups	X	24	0	26	50	24		26	50	24		26	50
	Y	0	24	26	50			26	26				
	Z	17	14	19	50	17		19	36	17		19	36
Total per arm		41	38	71	150	41		71	112	41		45	86

Table 4.6: **R-III:** Distribution of patients for the comparison of T1 vs. C for complete randomization and a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X:Y:Z

Data set		*All data*				*All control data*				*Restricted data*			
Treatment arm		T1	T2	C	$\sum$	T1	T2	C	$\sum$	T1	T2	C	$\sum$
Subgroups	X	24	0	26	50	24		26	50	24		26	50
	Y	0	24	26	50			26	26				
	Z	19	18	13	50	19		13	32	19		13	32
Total per arm		43	42	65	150	43		65	108	43		39	82

4.5 Multiplicity adjustment

The gain of using the design of a multi-stage trial is that patients in the control group can be used twice, for the comparisons of T1 vs. C and T2 vs. C. However, when one has a trial design with a shared use of the control data, the comparisons are no longer independent but correlated based on the shared control arm. The question arises whether it is necessary to control for multiplicity or not for which the current consensus in the literature has been reviewed (see Section 2.4.5). In

the considered trial design, the patients are randomized to different experimental treatment options and the hypotheses are being tested independently which does not require multiplicity correction [47, 101, 104]. Besides, as discussed above, the shared control arm itself does not necessitate adjustment of the error rate, however, the chance of multiple simultaneous false decisions might increase due to the shared control arm [45, 47]. In the simulation study at hand, the hypotheses are being tested independently and as result, no adjustment for the level-α-test for the multiple hypotheses testing was performed. This does not contradict ICH E9 [48] which states that adjustment should always be taken into account and an explanation of why adjustment is not considered necessary should be included. Following the decision tool provided by Odutayo et al. [102], the result for the study design at hand is that multiple testing procedures to control the error rate are not necessary.

4.6 Implementation in R

The simulations and analyses are conducted on a RStudio Server 2022.12.0 with R version 4.2.2 (2022-10-31) on a 64-bit Linux platform with 42 Cores. For the implementation in R, the linear models are fitted with the lm function from the *stats* package:

$$\text{for the unadjusted scenarios: } lm(Outcome \sim TreatmentArm),$$
$$\text{for the adjusted scenarios: } lm(Outcome \sim TreatmentArm + Typeofgroup).$$

Then, the pairwise contrasts are calculated based on the lm models using the lsmeans and contrast function from the R package *emmeans* (formerly provided by the R package *lsmeans*). The method in the contrast function is set to "trt.vs.ctrl" since the pairwise comparison of T1 vs. T2 are not of interest. No multiplicity adjustment for the contrasts was performed for the reasons explained in Section 4.5. In order to test the contrasts one-sided, the test function from the *emmeans* R package was used. The entire R code can be found in the Appendix A.4.

5 Results

The present chapter focuses on interpreting the results from the statistical analysis of one simulated data set and assessing the performance of the proposed methods in a simulation study. First, the results from the statistical analysis of one simulated data set are explained (see Section 5.1) and then, the results from the simulation study are presented (see Section 5.2). Besides, the properties of the examined methods and the influence of certain design parameters on the operating characteristics are discussed. For the statistical analysis the results from one simulated data set are interpreted while for the results from the simulation study 10000 replicates are considered.

The results from the statistical analysis and the simulation study are presented in three subsections, corresponding to the three different settings of the prevalence in the simulation study. First, a prevalence of 0:0:1 to the three subgroups X, Y and Z is considered which illustrates the extreme case where all patients are in subgroup Z. This corresponds to a classical multi-arm design comparing two treatment arms against a common control. Next, a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z is taken into account which illustrates the second extreme case for the prevalence and corresponds to two completely independent substudies. Finally, the effect of recruiting patients from all three subgroups is shown.

5.1 Statistical analysis

In the statistical analysis the focus was on comparing the different compositions of control data for the three settings from only one simulated data set. It was of interest to visualize the distribution of patients to the different arms for the comparisons based on all three data sets. For better comparability and easier legibility, it was decided to only plot the arms of interest for the comparison in the histograms depicting the treatment arms. As result, T2 is not depicted for the comparison of T1 vs. C and T1 is not depicted for the comparison of T2 vs. C for the comparisons based on *all data*. Besides, within the arms the distribution of the patients to the subgroups was investigated by considering different randomization

strategies. For each figure a table is provided which summarizes the statistical measures for each considered scenario. The considered statistical measures are the following: the standard error of the contrast (SE), the p-value of the contrast (p), the difference of the means of the treatment arms ($\widehat{\mu^{T1}} - \widehat{\mu^{C}}$ and $\widehat{\mu^{T2}} - \widehat{\mu^{C}}$) and the two-sided confidence interval of the contrast (CI). Besides, the number of patients per arm is provided (n^C, n^{T1} and n^{T2}). For all three prevalence settings, the above explained simulation setup comprising a multi-arm trial with two treatment arms and a shared control arm (see Chapter 4) was considered.

5.1.1 Setting 1: all patients are recruited from subgroup Z

In setting 1 the extreme case of the prevalence is investigated where all patients are in subgroup Z which corresponds to a traditional multi-arm trial.

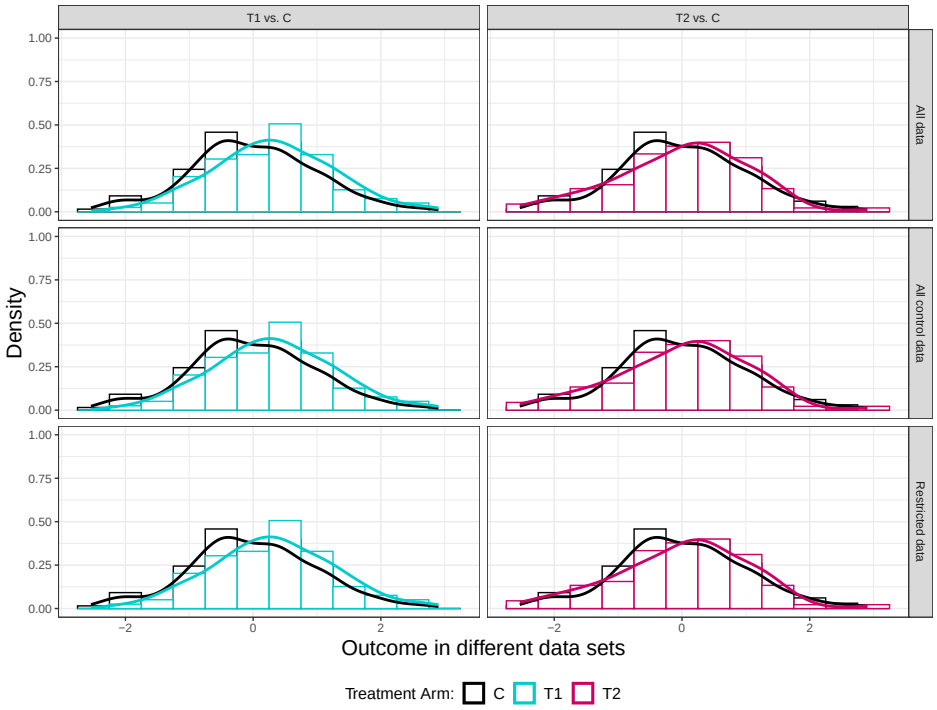

Figure 5.1: Distribution of patients for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of 0:0:1 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-II, $\mu_Z^C = 0$, $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

Figure 5.1 depicts the distribution of patients to the treatment arms for the comparison of T1 vs. C and T2 vs. C for the different compositions of control

Table 5.1: **Setting 1:** Results statistical analysis for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 131$, $n^{T1} = 79$, $SE = 0.142$, $p = 0.017$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.297$, CI = [0.022, 0.573]	$n^C = 131$, $n^{T2} = 90$, $SE = 0.135$, $p = 0.485$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.005$, CI = [-0.260, 0.270]
All control data	$n^C = 131$, $n^{T1} = 79$, $SE = 0.143$, $p = 0.016$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.297$, CI = [0.026, 0.569]	$n^C = 131$, $n^{T2} = 90$, $SE = 0.136$, $p = 0.485$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.005$, CI = [-0.263, 0.274]
Restricted data	$n^C = 131$, $n^{T1} = 79$, $SE = 0.143$, $p = 0.016$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.297$, CI = [0.026, 0.569]	$n^C = 131$, $n^{T2} = 90$, $SE = 0.136$, $p = 0.485$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.005$, CI = [-0.263, 0.274]

data. The data set has been simulated under the scenario assumptions that the total sample size is 300, $\Delta^{T1} = 0.5$, $\Delta^{T2} = 0$ and R-II which means that for subgroup Z the ratio is 1:1:2 for T1 vs. T2 vs. C. It was decided to only consider one randomization strategy, i.e. R-II, for setting 1 as there is no difference between a traditional multi-arm design (which corresponds to R-I and R-III) and a trial with two substudies (which correlates to R-II) (see Figure 4.2).

Figure 5.1 stresses the fact that in case all patients are recruited from subgroup Z there are no differences between the different compositions of the control data. That is further emphasized by the statistical measures summarized in Table 5.1. One can see that the p-value and the difference in means are the same for the three data sets. Besides, one can see in Table 5.1 that the comparisons of T1 vs. C are significant when testing one-sided at significant level $\alpha = 0.025$ while the comparisons of T2 vs. C are not significant. That is because of the scenario assumptions for the treatment effects for T1 and T2, i.e. $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$. Besides, Table 5.1 stresses that for R-II, one has more patients in the control arm than in each of the two treatment arms when all patients are recruited from substudy Z. The table further shows that the SE slightly changes from *all data* to *all control data*. The reason for that is the following: for the comparisons based on *all control data* one uses all control data and the data from the respective

treatment arm. In comparison to *all data* the SE changes because the treatment arm which is not used for the comparison has been excluded from the data set for the comparisons based on *all control data*. For *all control data* and *restricted data*, the SE is the same per definition.

5.1.2 Setting 2: all patients are recruited from subgroups X and Y

For setting 2 the second extreme case of the prevalence was considered where patients were only recruited from the subgroups X and Y. It was of interest how the simulation parameters influence the composition of the control data and the statistical measures. Homogeneous as well as heterogeneous controls were considered and the focus was on comparing R-I with R-II. For the randomization strategies only R-I and R-II are considered since when all patients are recruited from only subgroups X and Y the ratio in the subgroups is the same for R-II and R-III, i.e. for both strategies a ratio of 1:1 is considered for the subgroups X and Y (see Table 4.2 and Figure 4.3).

Table 5.2: **Setting 2:** Results statistical analysis for R-I for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 108$, $n^{T1} = 98$, $SE = 0.137$, $p = 6.7\text{e-}07$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.677$, $CI = [0.407, 0.947]$	$n^C = 108$, $n^{T2} = 94$, $SE = 0.138$, $p = 0.251$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.093$, $CI = [-0.180, 0.366]$
All control data	$n^C = 108$, $n^{T1} = 98$, $SE = 0.146$, $p = 3.2\text{e-}06$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.677$, $CI = [0.389, 0.966]$	$n^C = 108$, $n^{T2} = 94$, $SE = 0.134$, $p = 0.243$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.093$, $CI = [-0.171, 0.357]$
Restricted data	$n^C = 52$, $n^{T1} = 98$, $SE = 0.180$, $p = 5.6\text{e-}05$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.714$, $CI = [0.359, 1.071]$	$n^C = 56$, $n^{T2} = 94$, $SE = 0.155$, $p = 0.353$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.059$, $CI = [-0.247, 0.365]$

Figure 5.2 shows the distribution of patients to the treatment arms for the dif-

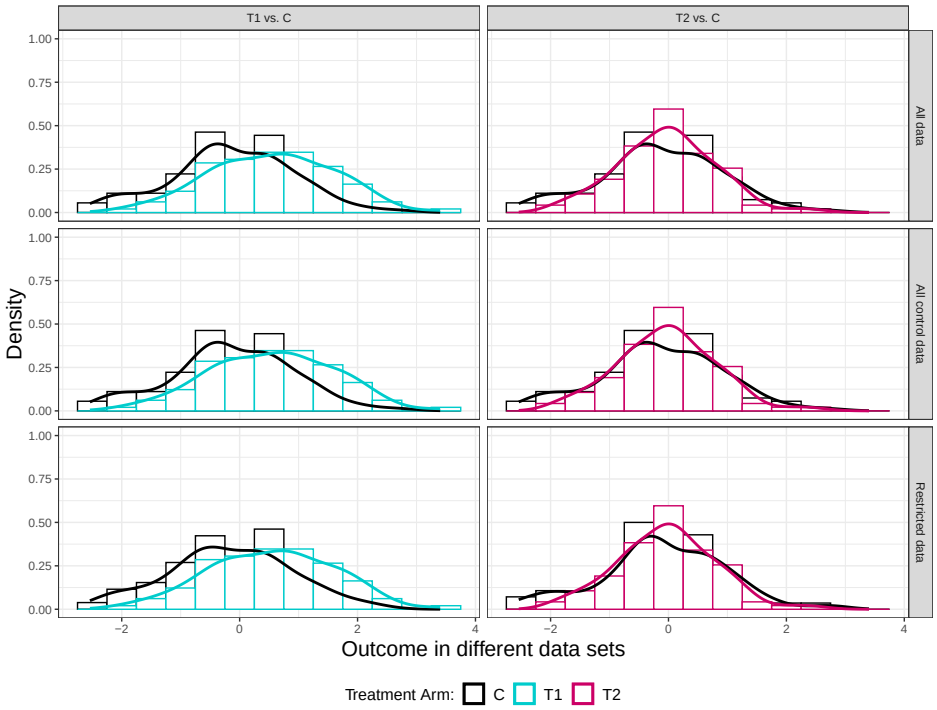

Figure 5.2: Distribution of patients per treatment arm for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-I, $\mu_X^C = \mu_Y^C = 0$ and $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

ferent compositions of control data for R-I. The data set has been simulated under the same scenario assumptions as the one for setting 1, meaning that the total sample size is 300, homogeneous controls, $\Delta^{T1} = 0.5$, $\Delta^{T2} = 0$ and R-I, i.e. the ratio in subgroups X and Y is 2:1 for the respective treatment arm vs. control. The figure stresses that recruiting patients from only subgroups X and Y (the prevalence is 0.5:0.5:0 to the subgroups X, Y and Z) results in differences between the different compositions of control data. One can see that the distribution of patients is the same for *all data* and *all control data* but different for *restricted data*. That is because one has less controls for the comparisons based on *restricted data* which is stressed in Table 5.2. One can see that for R-II one has almost twice as many patients in the treatment arms than in the control arm for the comparisons based on *restricted data* because all patients are recruited from subgroups X and Y with a ratio of 2:1 for the respective treatment arm vs. control. The position of the density lines depict the chosen treatment effects. For the comparison of T1 vs. C one can see that the distribution in the arms is shifted because $\Delta^{T1} = 0.5$ is as-

sumed. Besides, Table 5.2 summaries the statistical measures for the comparisons illustrated in Figure 5.2. One can see slight differences in the statistical measures across the different data sets, e.g. the p-value and SE are larger the less data available for the comparisons. Furthermore, one can see that the comparisons of T1 vs. C are significant when testing one-sided at significance level $\alpha = 0.025$ for all three data sets while for the comparisons of T2 vs. C are not significant. Again, the reason for that are the simulation assumptions, i.e. $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

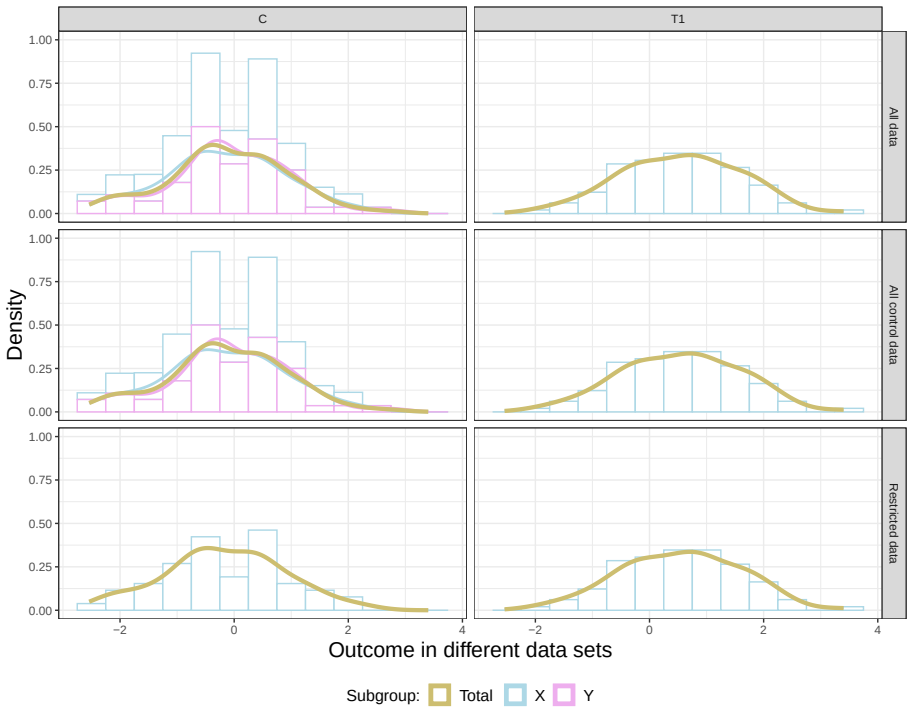

Figure 5.3: Distribution of patients in the subgroups for different compositions of control data for the comparison of T1 vs. C for a prevalence of 0.5:0.5:0 for the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-I, $\mu_X^C = \mu_Y^C = 0$ and $\Delta^{T1} = 0.5$.

It was of further interest to investigate the distribution of patients in subgroups X and Y for the treatment arms for the different compositions of control data. Figure 5.3 depicts the distribution of patients in subgroups X and Y for the comparison of T1 vs. C. The data set is simulated under the same scenario assumptions as Figure 5.2. Figure 5.3 emphasizes that for *all data* and *all control data* one uses the control data from subgroups X and Y and for *restricted data* the comparisons are based only on the control data from subgroup X. As consequence, one has less

controls for the comparison based on *restricted data*. The golden line depicts the total distribution of patients in the control arm when not differentiating between the subgroups. For homogeneous controls one cannot see big differences between the density lines when comparing the total distribution with the separate distributions in the subgroups.

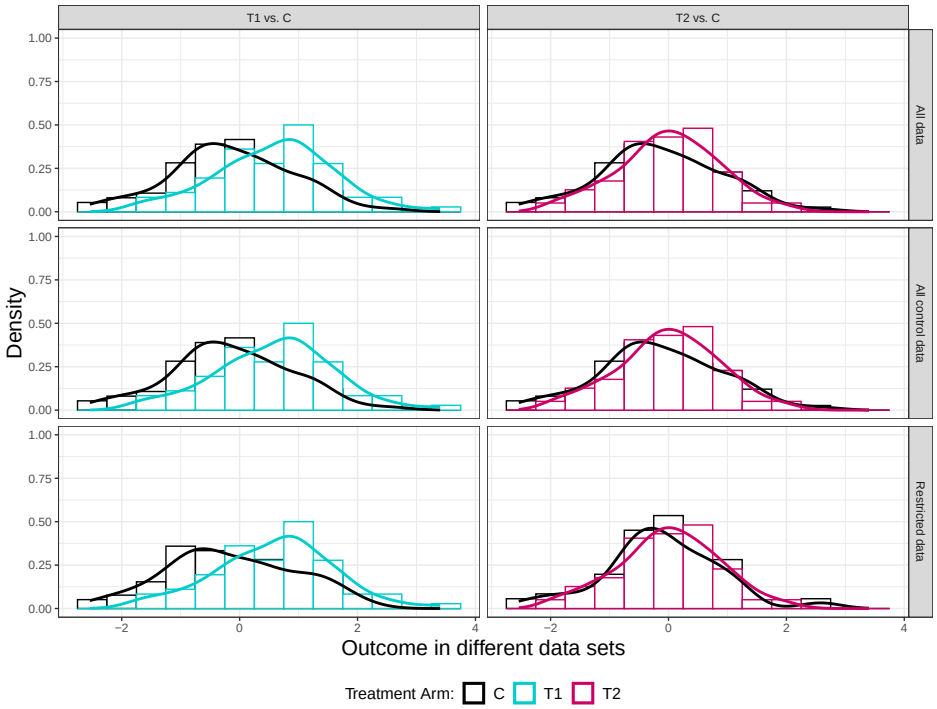

Figure 5.4: Distribution of patients per treatment arm for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-II, $\mu_X^C = \mu_Y^C = 0$, $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

Figure 5.4 shows the distribution of patients to the treatment arms for different compositions of control data for R-II, i.e. a ratio of 1:1 for the respective treatment arm vs. control in the subgroups X and Y is considered. When comparing Figure 5.2 with Figure 5.4, the only assumption that was varied for simulating the data set is the randomization strategy. For Figure 5.2 one assumed R-I and for Figure 5.4 R-II. One can see that for R-II the density lines for the comparison of T1 vs. C are more distinct than for R-I. Besides, a comparison of Table 5.2 and Table 5.3 shows only slight differences in the statistical measures. For example, the SEs and p-values are slightly higher for R-I. Furthermore, the tables stress the effect of the randomization strategies on the sample size per arm. R-II results in an

Table 5.3: **Setting 2:** Results statistical analysis for R-II for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 149$, $n^{T1} = 72$, $SE = 0.141$, $p = 2.2\text{e-}07$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.678$, $CI = [0.451, 1.006]$	$n^C = 149$, $n^{T2} = 79$, $SE = 0.136$, $p = 0.161$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.005$, $CI = [-0.133, 0.405]$
All control data	$n^C = 149$, $n^{T1} = 72$, $SE = 0.147$, $p = 7.1\text{e-}07$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.728$, $CI = [0.439, 1.021]$	$n^C = 149$, $n^{T2} = 79$, $SE = 0.136$, $p = 0.159$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.135$, $CI = [-0.132, 0.403]$
Restricted data	$n^C = 78$, $n^{T1} = 72$, $SE = 0.171$, $p = 1.6\text{e-}05$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.735$, $CI = [0.396, 1.071]$	$n^C = 71$, $n^{T2} = 79$, $SE = 0.149$, $p = 0.196$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.128$, $CI = [-0.167, 0.424]$

unequal ratio of treatment vs. control for comparisons based on *all data* and *all control data* and in an equal ratio for *restricted data*. For R-I it is the other way around, e.g. R-I results in an equal ratio of treatment vs. control for the comparisons based on *all data* and *all control data* and in unequal ratio for *restricted data*.

Figure 5.5 illustrates the distribution of patients per subgroup for T1 vs. C for the different compositions of control data simulated under the same scenario assumptions as Figure 5.4. It stresses the fact that for *all data* and *all control data*, one uses the control data from subgroups X and Y and for *restricted data* the comparison is based only on the control data from subgroup X which are the patients who could have get randomized to the treatment of interest. As result, one has less control data for the comparisons based on the *restricted data*. The golden line depicts the overall density when one does not differentiate between the two subgroups for the control arm. Compared to Figure 5.3 the only assumption that was varied for simulating the data set is the randomization strategy. In Figure 5.5 one can see a larger difference between the controls from subgroups X and Y for *all data* and *all control data* compared to Figure 5.5 because the difference in the sample size of the control arm between the subgroups is larger for R-II.

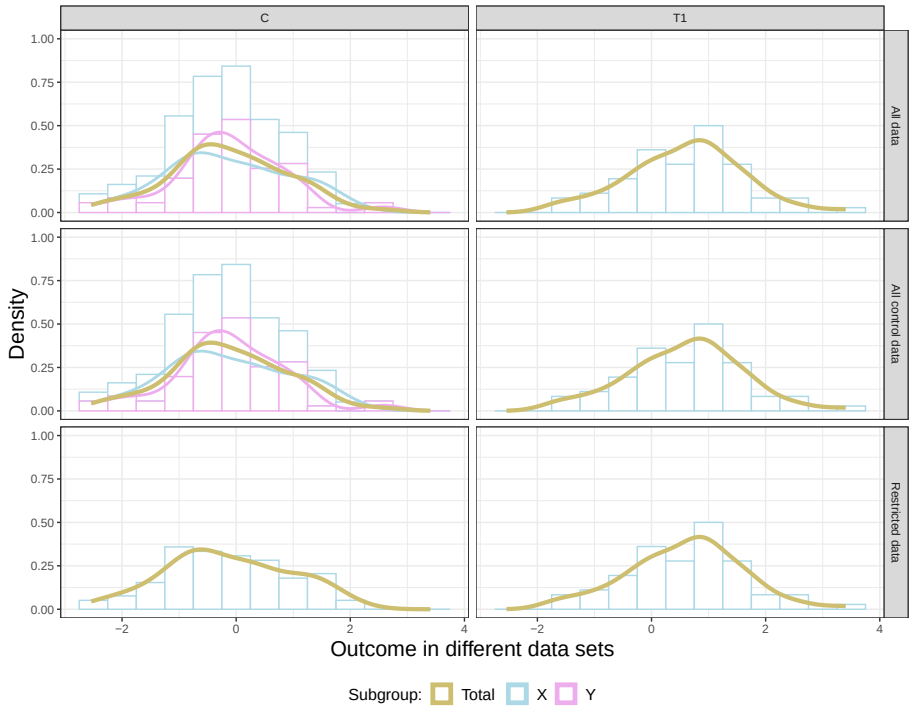

Figure 5.5: Distribution of patients in the subgroups for different compositions of control data for the comparison of T1 vs. C for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-II, $\mu_X^C = \mu_Y^C = 0$ and $\Delta^{T1} = 0.5$.

Next, it was also of interest to investigate the effect of heterogeneous controls. A negative control mean for subgroup X ($\mu_X^C = -0.5$) and a positive control mean for subgroup Y ($\mu_Y^C = 0.5$) was chosen. Except for the means in the control arm nothing was changed in the scenario assumptions to simulate the data set, meaning a total sample size of 300, $\Delta^{T1} = 0.5, \Delta^{T2} = 0$ and R-II, i.e. a ratio of 1:1 for the respective treatment arm vs. control for the subgroups X and Y is considered. Figure 5.6 shows the distribution of patients to the treatment arms for heterogeneous controls. When comparing Figures 5.4 and 5.6, one can see the difference between simulating homogeneous and heterogeneous controls. The figures show that the distribution of patients is different for the comparisons based on *all data* and *all control data* for the different means in the control arm but the same for the comparisons based on *restricted data*. That is further stressed in Tables 5.3 and 5.4 as the statistical measures differ for *all data* and *all control data* but not for *restricted data*. The reason is as follows: for *all data* and *all control data* the controls from subgroups X and Y are pooled while for *restricted data* only the controls from the respective subgroup are used for the comparison. For heterogeneous controls the controls to be pooled are further away from each other (see

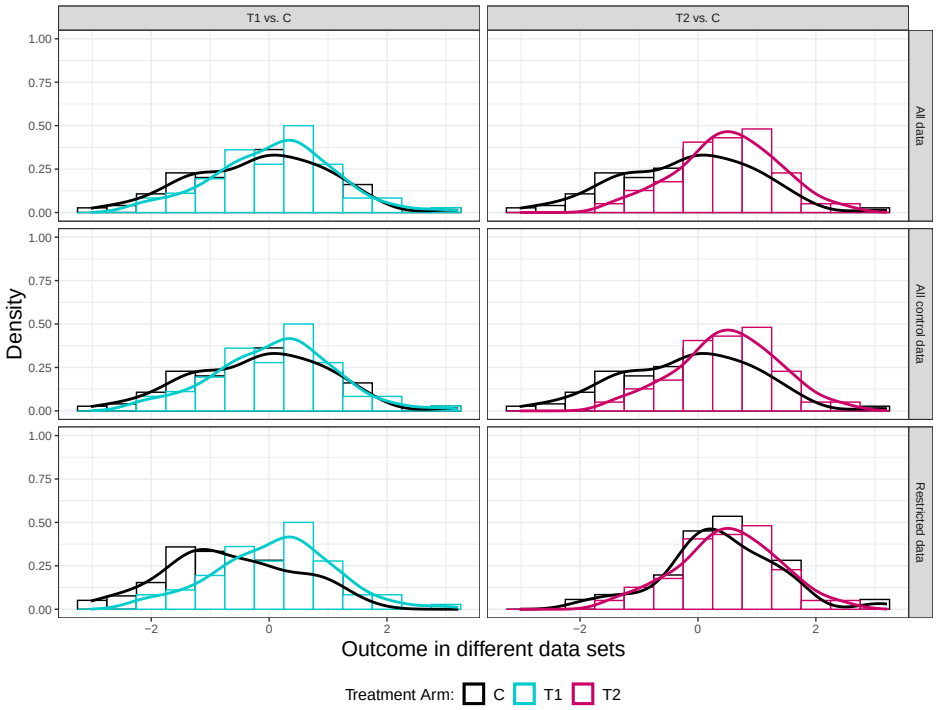

Figure 5.6: Distribution of patients per treatment arm for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-II, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

Figure 5.7) and as result, bias in the effect estimates is introduced. That is not the case for the *restricted data* because only the correct controls are used for the comparisons.

Figure 5.7 visualizes the distribution of patients per subgroup for heterogeneous controls for T1 vs. C for the different compositions of control data. Compared to Figure 5.5 where homogeneous controls were assumed, one can see that the density lines for subgroups X and Y are shifted in the direction of the mean in the control arm. For example, for subgroup X a negative mean in the control arm was assumed and the density line is shifted to the left. As result, one can clearly see the bimodal distribution of the subgroups for the control arm. The bimodal distribution of the subgroups for the control arm is the reason for the above explained bias in the effect estimates that is introduced when pooling the controls from subgroups X and Y for *all data* and *all control data* because it is not adjusted for.

Table 5.4: **Setting 2:** Results statistical analysis for heterogeneous controls for R-II for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 149$, $n^{T1} = 72$, $SE = 0.150$, $p = 4.7\text{e-}02$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.252$, $CI = [-0.043, 0.547]$	$n^C = 149$, $n^{T2} = 79$, $SE = 0.146$, $p = 4.3\text{e-}06$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.659$, $CI = [0.373, 0.946]$
All control data	$n^C = 149$, $n^{T1} = 72$, $SE = 0.252$, $p = 0.057$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.252$, $CI = [-0.061, 0.565]$	$n^C = 149$, $n^{T2} = 79$, $SE = 0.147$, $p = 6.1\text{e-}06$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.659$, $CI = [0.369, 0.951]$
Restricted data	$n^C = 78$, $n^{T1} = 72$, $SE = 0.171$, $p = 1.6\text{e-}05$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.735$, $CI = [0.396, 1.071]$	$n^C = 71$, $n^{T2} = 79$, $SE = 0.149$, $p = 0.196$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.128$, $CI = [-0.167, 0.424]$

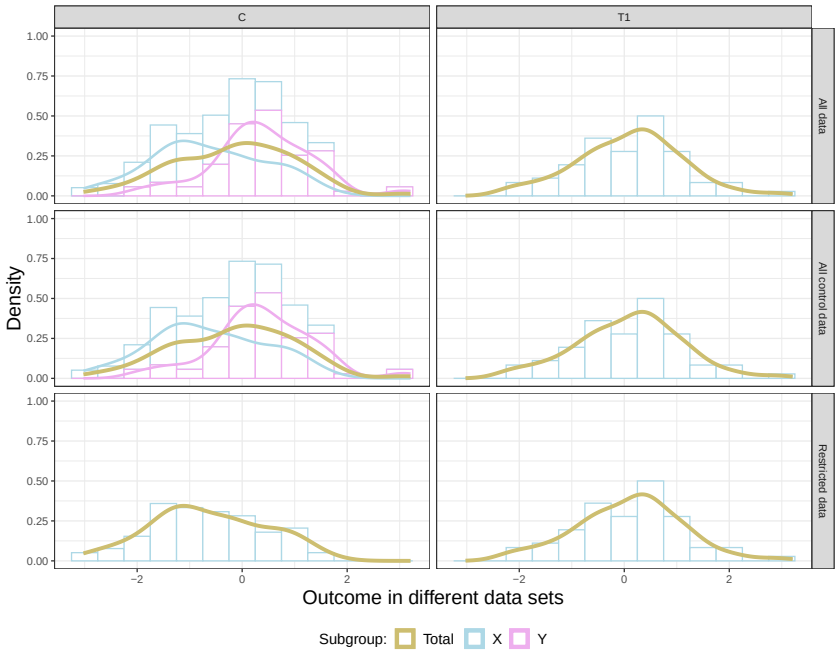

Figure 5.7: Distribution of patients in the subgroups for different compositions of control data for the comparison of T1 vs. C for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, R-II, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\Delta^{T1} = 0.5$.

5.1.3 Setting 3: patients are recruited from all three subgroups X, Y and Z

Finally, in setting 3, patients are recruited from the three subgroups X, Y and Z. The influence of the simulation parameters on the composition of control data is investigated. Heterogeneous controls were assumed to simulate the data sets as it was shown in setting 2 that the chosen heterogeneous controls lead to recognizable distinct bimodal distributions of the subgroups for the control arm. The scenario assumptions for simulating the data sets in this subsection are the following: a total sample size of 300, complete randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$, $\Delta^{T1} = 0.5$, $\Delta^{T2} = 0$ and either R-I or R-II.

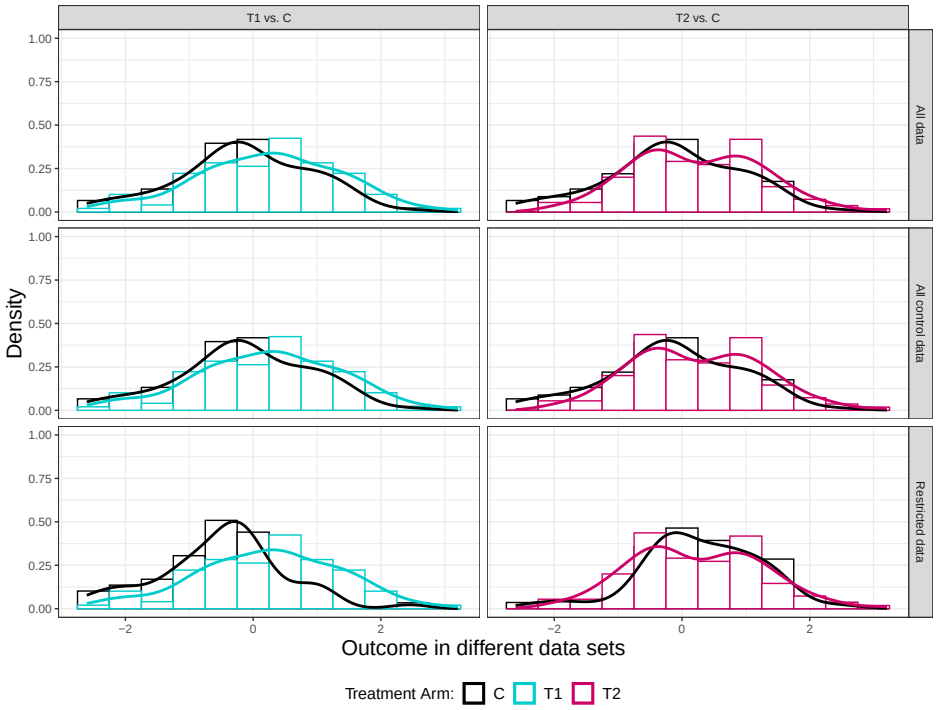

Figure 5.8: Distribution of patients for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$, R-I, $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

Figure 5.8 shows the distribution of patients to the different arms for R-I, i.e. for subgroups X and Y the ratio is 2:1 for the respective treatment arm vs. control and for subgroup Z the ratio is 1:1:1 for T1 vs. T2 vs. C. One can see that the distribution of patients is the same for *all data* and *all control data* but differs

for *restricted data*. For setting 3, the statistical measures for the unadjusted as well as adjusted analysis scenarios are provided. Tables 5.5 and 5.6 show that the SE, p-values and CIs depend on the analysis method. While according to the p-values the comparisons of T2 vs. C are significant for *all data* and *all control data* of the unadjusted analysis scenarios (see Table 5.5), none of the comparisons is significant for the adjusted analysis scenarios (see Table 5.5). The statistical measures for *restricted data* are different to the other two compositions of control data, e.g. the SEs are higher and the p-values are lower.

Table 5.5: **Setting 3:** Results statistical analysis for R-I for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 91$, $n^{T1} = 99$, $SE = 0.154$, $p = 0.011$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.353$, $CI = [0.050, 0.655]$	$n^C = 91$, $n^{T2} = 110$, $SE = 0.150$, $p = 0.008$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.361$, $CI = [0.066, 0.656]$
All control data	$n^C = 91$, $n^{T1} = 99$, $SE = 0.157$, $p = 0.013$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.353$, $CI = [0.042, 0.663]$	$n^C = 91$, $n^{T2} = 110$, $SE = 0.146$, $p = 0.007$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.361$, $CI = [0.073, 0.649]$
Restricted data	$n^C = 59$, $n^{T1} = 99$, $SE = 0.176$, $p = 4.1\text{e-}05$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.712$, $CI = [0.365, 1.060]$	$n^C = 56$, $n^{T2} = 110$ $SE = 0.161$, $p = 0.631$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = \text{-}0.054$, $CI = [\text{-}0.372, 0.264]$

Figure 5.9 pictures the distribution of patients to the subgroups X, Y and Z for the different composition of control data for the comparison of T1 vs. C for R-I. It stresses the fact that for *all data* and *all control data*, one uses the same composition of control data, namely, the control data from the three subgroups X, Y and Z. For *restricted data*, one uses only the control data from subjects assigned to subgroups X and Z. The control data are restricted to those patients who could have been treated with T1. This has the consequence that one has less control data for comparisons based on the *restricted data*. The figure further depicts the density curves for the respective patients per subgroup as well as the overall density (the golden line). The density curves for subgroups X and Y are shifted in

Table 5.6: **Setting 3:** Results statistical analysis for R-I for adjusted analyses

Data sets	**Adjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 91$, $n^{T1} = 99$, $SE = 0.154$, $p = 6.8\text{e-}07$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.353$, $CI = [0.456, 1.062]$	$n^C = 91$, $n^{T2} = 110$, $SE = 0.152$, $p = 0.616$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.361$, $CI = [-0.345, 0.255]$
All control data	$n^C = 91$, $n^{T1} = 99$, $SE = 0.165$, $p = 5.4\text{e-}06$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.353$, $CI = [0.422, 1.074]$	$n^C = 91$, $n^{T2} = 110$, $SE = 0.151$, $p = 0.716$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.361$, $CI = [-0.385, 0.212]$
Restricted data	$n^C = 59$, $n^{T1} = 99$, $SE = 0.171$, $p = 1.1\text{e-}05$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.712$, $CI = [0.409, 1.086]$	$n^C = 56$, $n^{T2} = 110$ $SE = 0.151$, $p = 0.715$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = -0.054$, $CI = [-0.386, 0.213]$

opposite directions due to the chosen means in the control arm. R-I was followed to randomize the subjects to one of the arms which implies that for subgroups X and Y more subjects are randomized to the treatment arms. That is highlighted in the histograms, as one can see higher densities for T1 than for the control for subgroup X. Besides, one can see especially for the control arm that for an equal prevalence and R-I one has less patients from subgroup Z than from subgroups X and Y.

Figure 5.10 shows the distribution of patients to the different arms for R-II, i.e. for subgroups X and Y the ratio is 1:1 for the respective treatment arm vs. control and for subgroup Z the ratio is 1:1:2 for T1 vs. T2 vs. C. One can see that the distribution of the patients is the same for *all data* and *all control data* but differs for *restricted data*. For R-II also the statistical measures for the unadjusted as well as adjusted analysis scenarios are provided. Tables 5.7 and 5.8 show that the SEs, p-values and CIs depend on the analysis method. While the comparisons of T1 vs. C are only significant for the *restricted data* for the unadjusted analysis scenarios (see Table 5.7), they are significant for all adjusted analysis scenarios (see Table 5.8).

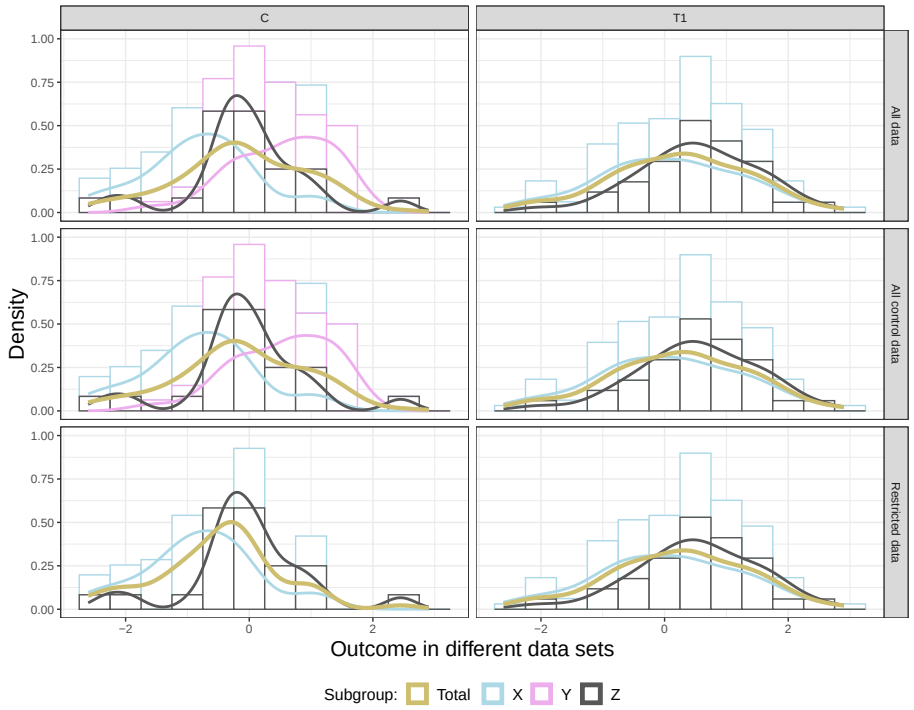

Figure 5.9: Distribution of patients per subgroup for different compositions of control data for the comparisons of the T1 vs. C for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$, R-I and $\Delta^{T1} = 0.5$

When comparing Figures 5.8 and 5.10, the only simulation parameter that was varied is the randomization strategy to the arms. In Figure 5.8 R-I is used to randomize the patients to the three arms while in Figure 5.10 R-II is employed. The figures emphasize that one has more controls for the comparisons when one aims for R-II, especially for the comparisons based on the *restricted data*.

Regarding the corresponding Tables 5.5 and 5.7 for the unadjusted analyses one can see differences for all statistical measures. For R-I all comparisons expect the comparison of T2 vs. control based on *restricted data* are significant while for R-II no comparison is significant. Besides, for the differences in means, the differences for R-II are roughly twice as large as for R-I. For the adjusted analyses, see Tables 5.6 and 5.8, for the differences in means, the differences for R-II are roughly twice as large as for R-I. Besides, the p-values for R-II are larger and the SE is smaller.

Figure 5.11 depicts the distribution of patients to the subgroups X, Y and Z for the comparison of T1 vs. C for the different compositions of control data. The figure stresses that one has less controls for the comparison based on *restricted*

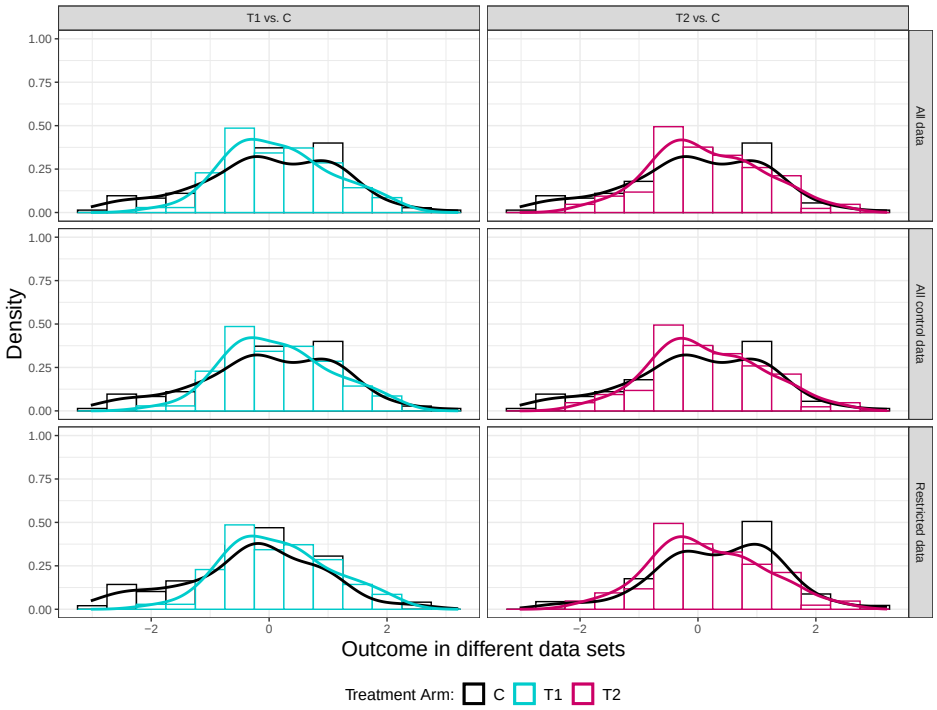

Figure 5.10: Distribution of patients for different compositions of control data for the comparisons of the treatment arms vs. the control arm for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$, R-II, $\Delta^{T1} = 0.5$ and $\Delta^{T2} = 0$.

data. When comparing Figure 5.9 with Figure 5.11, the only parameter that was varied for simulating the data sets is the randomization strategy. In Figure 5.11, R-II is used to randomized to the patients to the different arms while in Figure 5.9 R-I is applied for the randomization to the arms. One can see that in Figure 5.9 more patients were randomized to the treatment arm than to the control arm for subgroup X compared to Figure 5.11. Besides, Figure 5.11 shows that for R-II more patients were randomized to the control arm than to the treatment arm for subgroup Z compared to R-I depicted in Figure 5.9.

Table 5.7: **Setting 3:** Results statistical analysis for R-II for unadjusted analyses

Data sets	**Unadjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 145$, $n^{T1} = 70$, $SE = 0.155$, $p = 0.148$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.162$, CI = [-0.143, 0.468]	$n^C = 145$, $n^{T2} = 85$, $SE = 0.146$, $p = 0.137$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.159$, CI = [-0.127, 0.446]
All control data	$n^C = 145$, $n^{T1} = 70$, $SE = 0.162$, $p = 0.158$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.162$, CI = [-0.156, 0.481]	$n^C = 145$, $n^{T2} = 85$, $SE = 0.159$, $p = 0.295$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.159$, CI = [-0.141, 0.459]
Restricted data	$n^C = 98$, $n^{T1} = 70$, $SE = 0.167$, $p = 0.005$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.441$, CI = [0.111, 0.769]	$n^C = 91$, $n^{T2} = 85$, $SE = 0.149$, $p = 0.911$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.506$, CI = [-0.496, 0.093]

Table 5.8: **Setting 3:** Results statistical analysis for R-II for an adjusted analyses

Data sets	**Adjusted analysis scenarios**	
	T1 vs. C	**T2 vs. C**
All data	$n^C = 145$, $n^{T1} = 70$, $SE = 0.152$, $p = 0.001$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.162$, CI = [0.241, 0.839]	$n^C = 145$, $n^{T2} = 85$, $SE = 0.146$, $p = 0.955$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.159$, CI = [-0.536, 0.039]
All control data	$n^C = 145$, $n^{T1} = 70$, $SE = 0.163$, $p = 0.001$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.162$, CI = [0.183, 0.824]	$n^C = 145$, $n^{T2} = 85$, $SE = 0.154$, $p = 0.961$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.159$, CI = [-0.575, 0.029]
Restricted data	$n^C = 98$, $n^{T1} = 70$, $SE = 0.162$, $p = 0.001$, $\widehat{\mu^{T1}} - \widehat{\mu^C} = 0.441$, CI = [0.184, 0.822]	$n^C = 91$, $n^{T2} = 85$. $SE = 0.142$, $p = 0.972$, $\widehat{\mu^{T2}} - \widehat{\mu^C} = 0.506$, CI = [-0.553, 0.008]

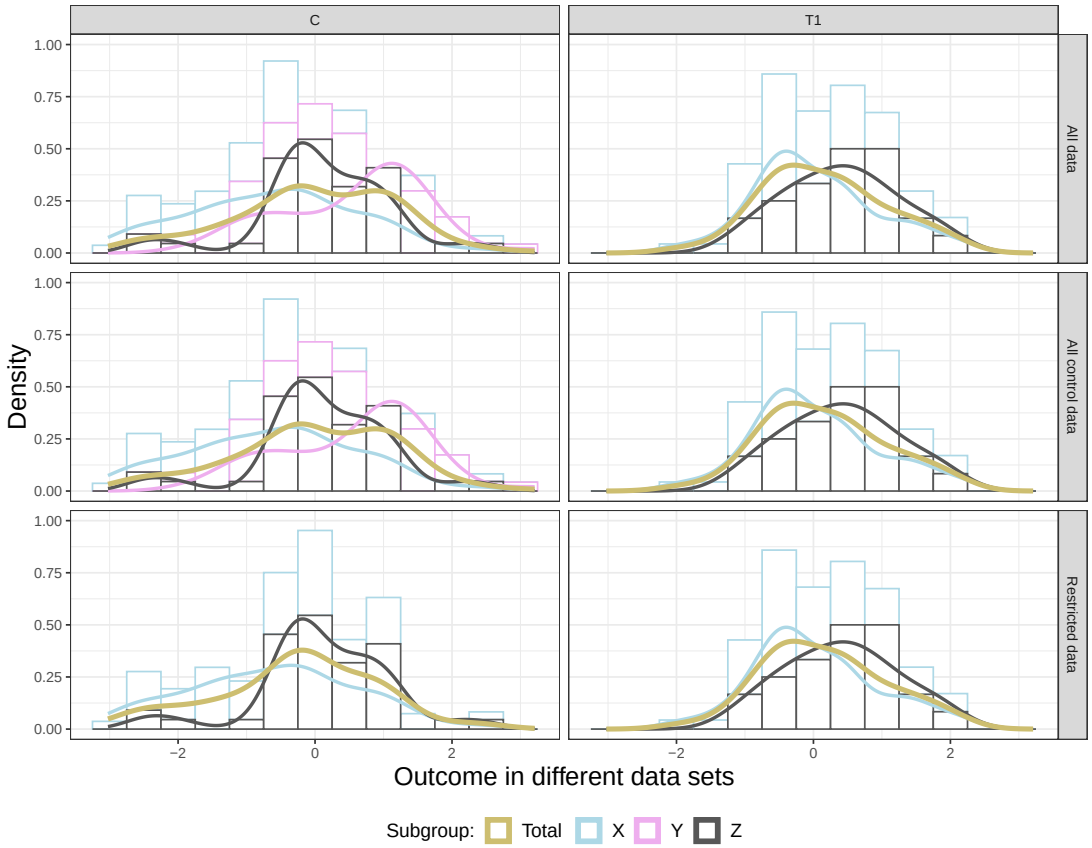

Figure 5.11: Distribution of patients to the subgroups for different compositions of control data for the comparisons of the T1 vs. C for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z. Fixed parameters: n = 300, complete randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$, R-II and $\Delta^{T1} = 0.5$.

5.2 Simulation results

This section presents the results from the simulation study described in Chapter 4. For all settings the above explained simulation setup comprising a multi-arm trial with two treatment arms and a shared control arm was considered. Table 5.9 provides an overview for the different data sets, analysis scenarios and operating characteristics which were investigated for the three settings. For each setting, different simulation scenarios were considered, e.g different total sample sizes and randomization strategies (see Table 4.3 for an extensive overview of the simulation parameters).

Table 5.9: Overview of investigated settings

Settings	**Data sets**	**Analysis scenarios**	**Operating characteristics**
Setting 1: all patients are in subgroup Z	*All data,* *All control data,* *Restricted data*	Unadjusted analysis scenarios	Type I error, Marginal power
Setting 2: all patients are in subgroups X and Y	*All data,* *All control data,* *Restricted data*	Unadjusted analysis scenarios	Type I error, Marginal power
Setting 3: patients are in all three subgroups X, Y and Z	*All data,* *All control data,* *Restricted data*	Unadjusted and adjusted analysis scenarios	Type I error, Marginal power

For better legibility and easier comparison, the range of the y-axis is restricted to [0, 0.1] in the plots showing the estimated type I error rate unless otherwise stated. Besides, the plots, which depict the type I error rate, include the confidence interval around the nominal significance level of 0.025 for 10000 simulation replications as a grey box. For the plots different colours and shapes are chosen to differentiate between the analysis scenarios. The line type is varied to either distinguish between the treatments or the randomization strategies.

As mentioned in Chapter 4, two different prevalence distribution strategies to the three subgroups X, Y and Z were tested. In one simulation the proportion of subgroups X, Y and Z is deterministic and corresponds to the true prevalence and in the other simulation the distribution is random. For a deterministic prevalence distribution, a total sample size of 300 and an equal prevalence to the subgroups X, Y and Z exactly 100 patients would be distributed to each subgroup. For a random prevalence distribution, a total sample size of 300 and an equal prevalence to the subgroups X, Y and Z, the result could be, e.g., 104:106:90 for X:Y:Z. Figure 5.12 shows the rejection probability for the different prevalence distributions for

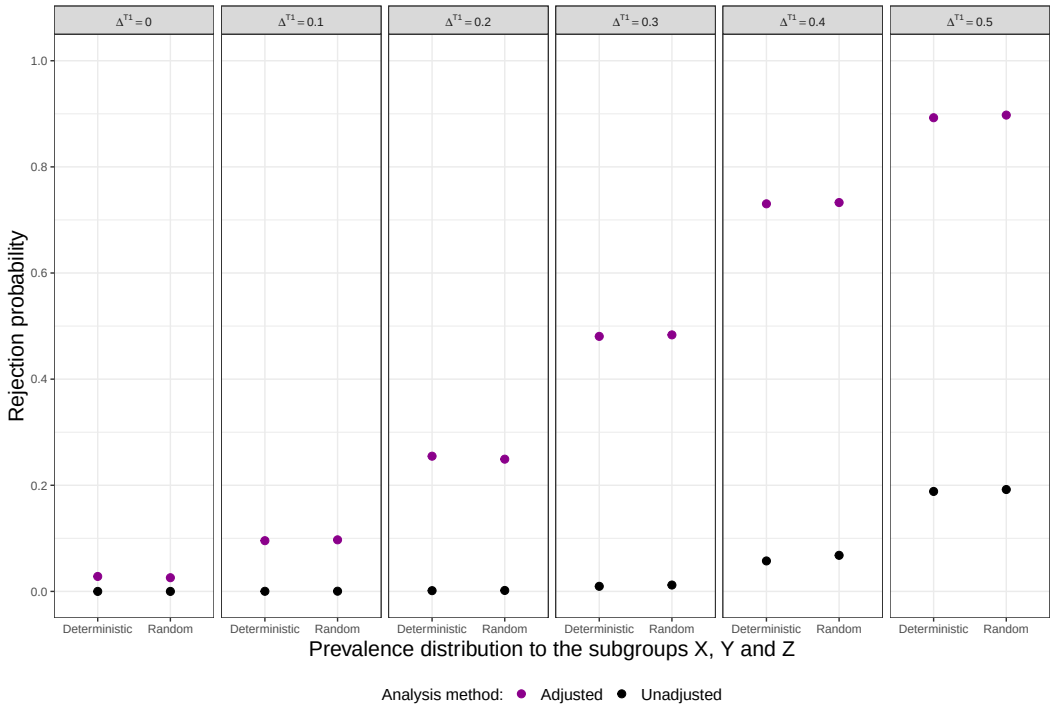

Figure 5.12: Rejection probability over different prevalence distribution strategies for increasing Δ^{T1}. Fixed parameters: block randomization, R-I, prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z, n = 300, *all data*, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\mu_Z^C = 0$ and $\Delta^{T2} = 0$.

10000 simulation replications. The figure stresses that one cannot see any decisive differences for the two distribution strategies for both unadjusted and adjusted analyses. Therefore, it was decided to only distribute the patients to subgroups X, Y and Z deterministically to ensure that the distribution of patients in the three subgroups corresponds to the true prevalence.

Besides, two randomization methods, complete and block randomization, to the three arms, i.e. T1, T2 and C, were considered in the simulation study. Figure 5.13 depicts the two randomization methods for the three randomization strategies for an equal prevalence to the subgroups X, Y and Z. The other design parameters that were considered are the following: a total sample size of 300, $\Delta^{T2} = \Delta^{T2} = 0.5$ and heterogeneous controls ($\mu_X^C = -0.5$, $\mu_Y^C = -0.5$ and $\mu_X^C = 0$). Besides, only the comparisons for T1 vs. C are shown. Figure 5.13 shows that for there are hardly any visible differences between complete and block randomization for the considered randomization strategies and analysis scenarios. One can see that the power for block randomization for R-II is slightly higher than

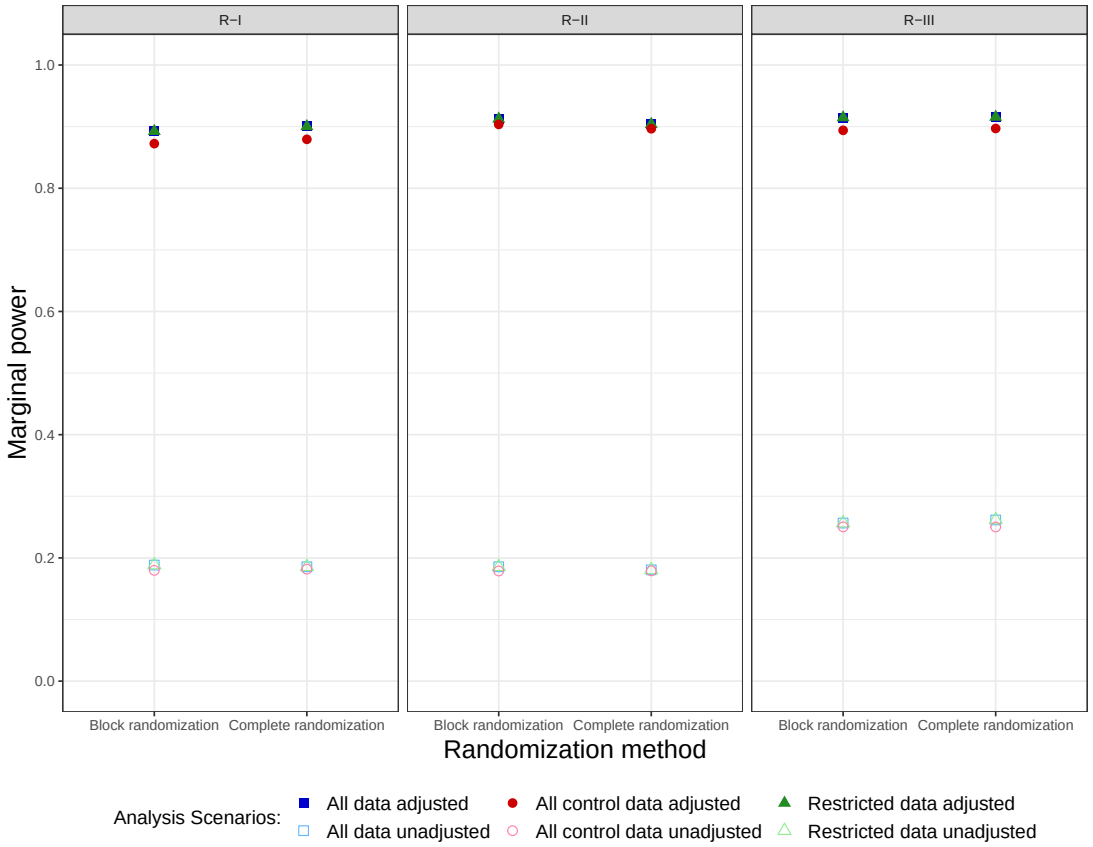

Figure 5.13: Power for complete and block randomization for all considered randomization strategies. Fixed parameters: prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z, n = 300, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\mu_Z^C = 0$ and $\Delta^{T1} = \Delta^{T2} = 0.5$.

for complete randomization. Therefore, the decision was made to employ block randomization for randomizing the patients to the three treatment arms, as it is commonly favored, e.g. for its ability to enhance statistical efficiency by reducing variability and improving the precision of treatment effect estimates (see Section 2.4.1 for more information).

5.2.1 Setting 1: all patients are recruited from subgroup Z

The total sample size and randomization strategies are varied to evaluate the effect of the prevalence on the type I error rate and statistical power. The considered prevalence in this subsection is 0:0:1 to the subgroups X, Y and Z which corresponds to a traditional multi-arm design. Since it was shown above that there are no differences for setting 1 for the different compositions of control data (see Figure 5.1), only one analysis scenario, i.e. *all data*, is plotted in the following.

Besides, when recruiting all patients from only subgroup Z, only the contrasts based on the one-way models were of interest since the factor type of group only has one level, i.e., subgroup Z. A solid line is chosen to represent T1 and a dotted line is chosen to visualize T2.

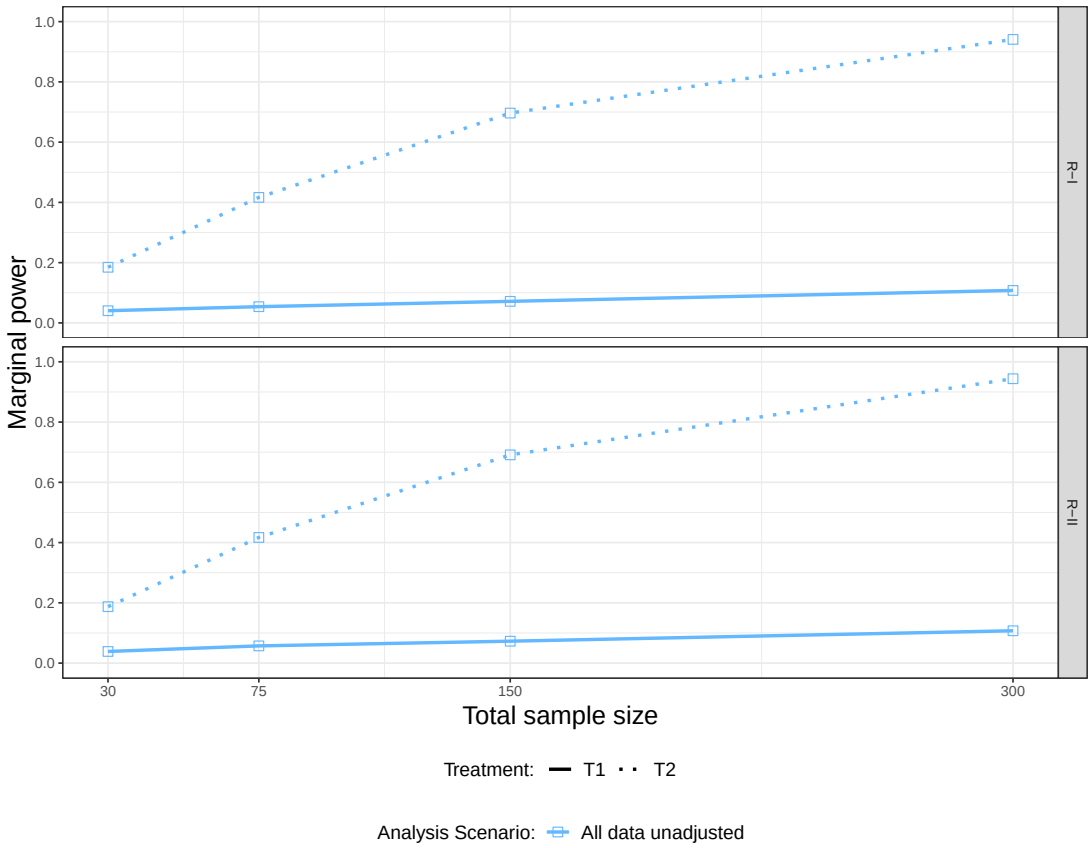

Figure 5.14: Power over increasing sample size for a prevalence of 0:0:1 to the subgroups X, Y and Z for R-I and R-II. Fixed simulation parameters: block randomization, $\mu_Z^C = 1$, $\Delta^{T1} = 0.1$ and $\Delta^{T2} = 0.5$.

Figures 5.14 and 5.15 show the impact of the total sample size on the operating characteristics. Figure 5.14 shows that the power increases for an increasing sample size. One can see that for a higher treatment effect, i.e. $\Delta^{T2} = 0.5$, the power increases more for an increasing sample size. While for $\Delta^{T1} = 0.1$ the power increase is only 5% from a total sample size of 30 to a total sample size of 300, the power increase for $\Delta^{T2} = 0.5$ is approximately 70%, e.g. the power is 20% for a total sample size of 30, and it increases up to 95% for a total sample size of 300. Figure 5.14 also shows that the different randomization strategies do not affect the power. For R-I the ratio for subgroup Z is 1:1:1 for T1 vs. T2 vs. C

and for R-II the ratio is 1:1:2 for T1 vs. T2 vs. C. The randomization strategies do not affect the power because, as shown in Figure 4.2, for setting 1 there is no difference between a traditional multi-arm design (which corresponds to R-I) and a trial with two substudies (which correlates to R-II). R-III was not investigated in this setting because when only recruiting patients from subgroup Z, the ratio for subgroup Z is the same for R-I and R-III (see Table 4.2). Figure 5.15 demonstrates that the type I error rate is controlled for all considered sample sizes and both randomization strategies.

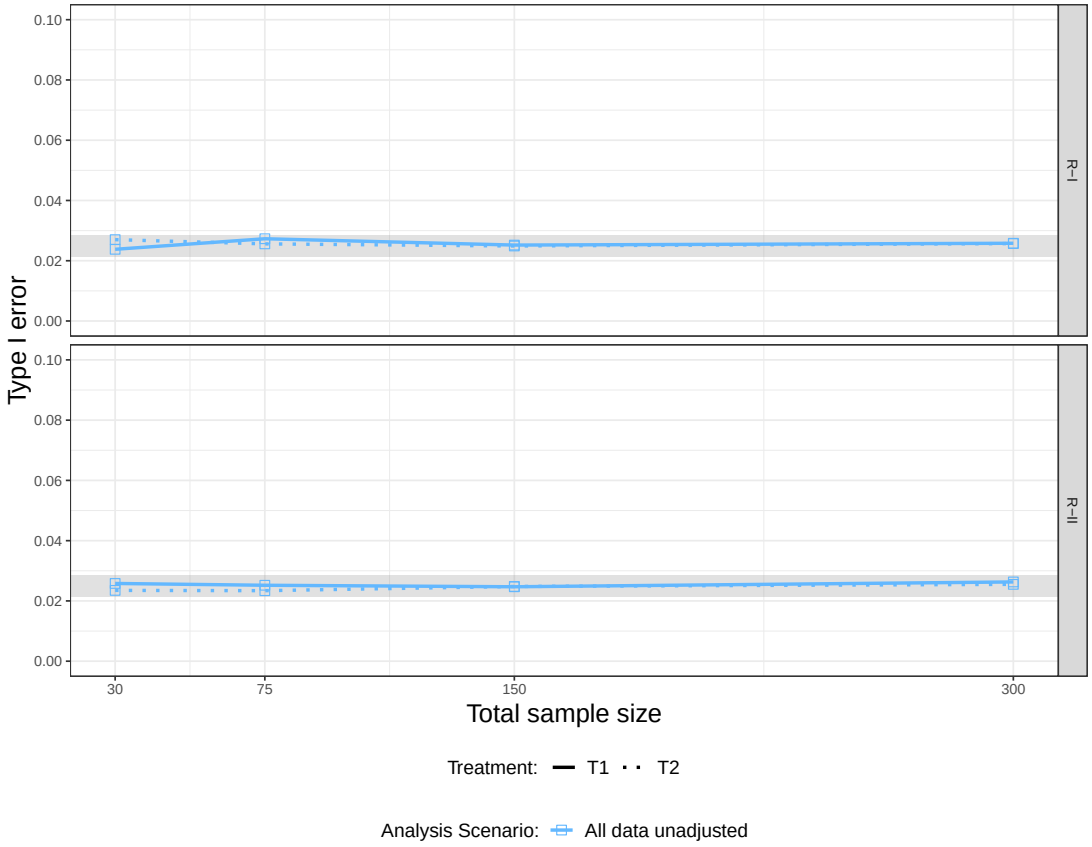

Figure 5.15: Type I error rate over increasing sample size for a prevalence of 0:0:1 to the subgroups X, Y and Z for R-I and R-II. Fixed simulation parameters: block randomization, $\mu_Z^C = 1$ and $\Delta^{T1} = \Delta^{T2} = 0$.

5.2.2 Setting 2: all patients are recruited from subgroups X and Y

The total sample size and randomization strategies as well as the treatment effects are varied to evaluate the effect of the prevalence on the type I error rate

and statistical power. In this subsection a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z is considered which corresponds to two independent substudies. For the randomization strategies only R-I and R-II are considered since when all patients are recruited from only subgroups X and Y the ratio to the subgroups is the same for R-II and R-III, i.e. for both strategies a ratio of 1:1 is considered for the subgroups X and Y (see Table 4.2 and Figure 4.3). Besides, when all patients are recruited from subgroups X and Y, the focus was also on the contrasts based on the one-way models. Therefore, only results for the unadjusted analyses are presented for setting 2. For better legibility it was decided to add different shapes and colours for the different compositions of control data.

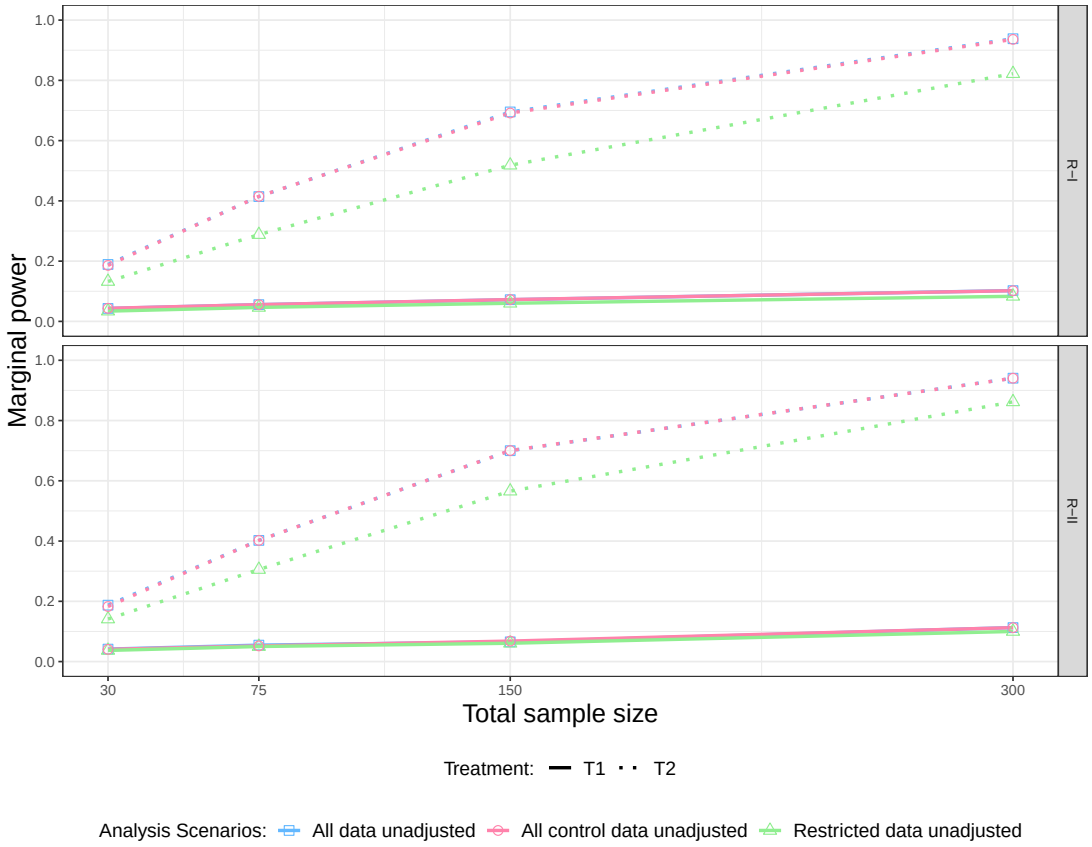

Figure 5.16: Power over increasing sample size for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z for R-I and R-II. Fixed design parameters: block randomization, $\mu_X^C = \mu_Y^C = 0.5$, $\Delta^{T1} = 0.1$ and $\Delta^{T2} = 0.5$.

Figure 5.16 provides us with an idea of how the power increases as the sample size increases. One can see that the power increases for increasing sample size for all considered analysis scenarios. Note, the lines for *all data* are masked by

all control data. It has already been shown in Section 5.1.2 that there are no visible differences between the comparisons based on *all data* and *all control data*. Different line types are chosen for the two treatments, e.g. a solid line is used to visualize T1 and dotted line to represent T2. One can see that the power increase is higher for the larger treatment effect, e.g. the power for T2 is higher than for T1. Besides, one can see that the power is higher for *all control data* and *all data* than for the *restricted data*. The reason for that was illustrated in Figure 5.5, namely, one has less control data for the comparisons based on *restricted data* and as result, the power is lower. While the power for *all data* and *all control data* is the same for both randomization strategies, the power for the *restricted data* is lower for R-I than for the R-II.

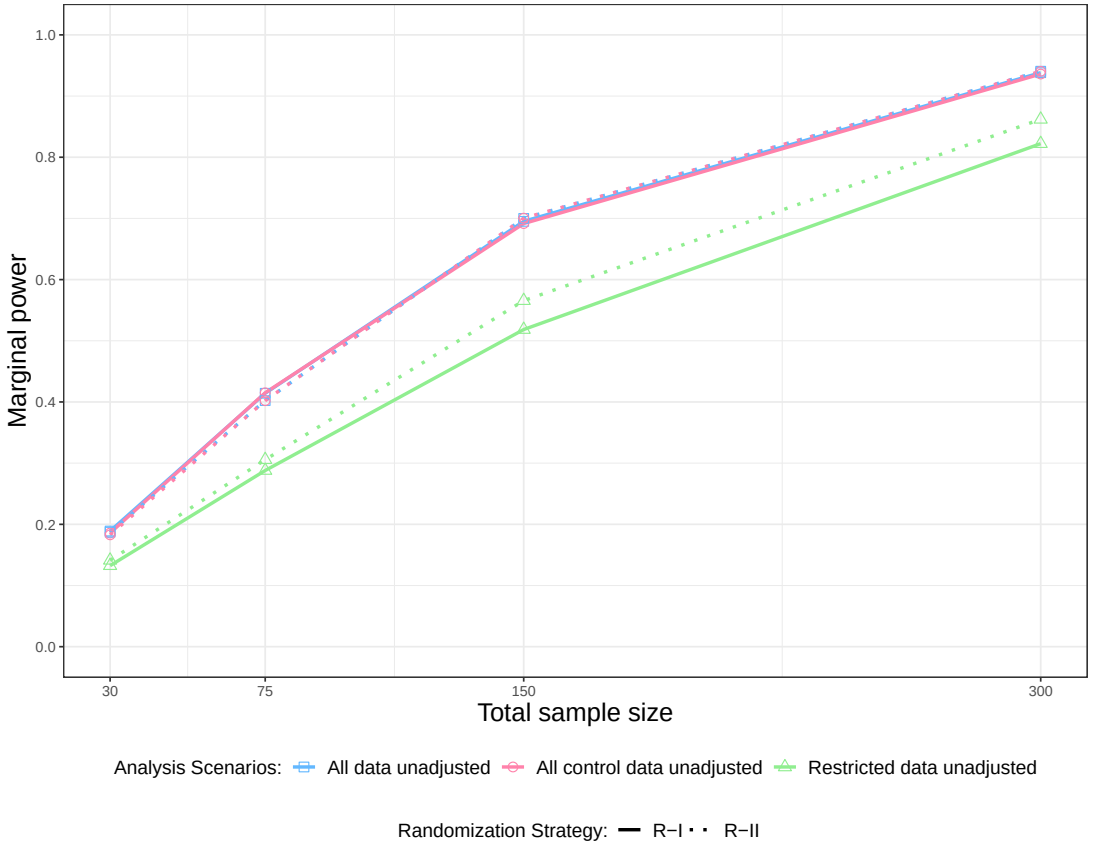

Figure 5.17: Power over increasing sample size for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z for treatment 2. Fixed design parameters: block randomization, $\mu_X^C = \mu_Y^C = 0.5$, $\Delta^{T2} = 0.5$ and n = 150.

That the power for the comparisons based on the *restricted data* is lower for R-I than for R-II is emphasized in Figure 5.17. Different line types are chosen

to distinguish between the two randomization strategies for T2, e.g. the solid line represents R-I and the dotted line is chosen for R-II. While the power is the same for *all data* and *all control data* for both randomization strategies, it declines for the *restricted data* from 57.6% to 52.2% when comparing R-II with R-I. The reason for that is the following: for R-I one has a ratio of 2:1 for the respective treatment vs. control in the subgroups X and Y (see Figure 4.3 **A)**) while for R-II one has a ratio of 1:1 for the respective treatment vs. control in the subgroups X and Y (see Figure 4.3 **B)**). As result, one has less controls for the comparisons when aiming for R-I and therefore, the power declines. In Table 5.2 one can see that for the *restricted data* for one simulated data set, one has almost twice as many patients in the treatment arms than in the control arm. Note, the lines for *all data* are masked by *all control data*. Figure 5.18 shows that the considered total sample sizes and randomization strategies have no effect on the type I error, which is guaranteed for all analysis scenarios.

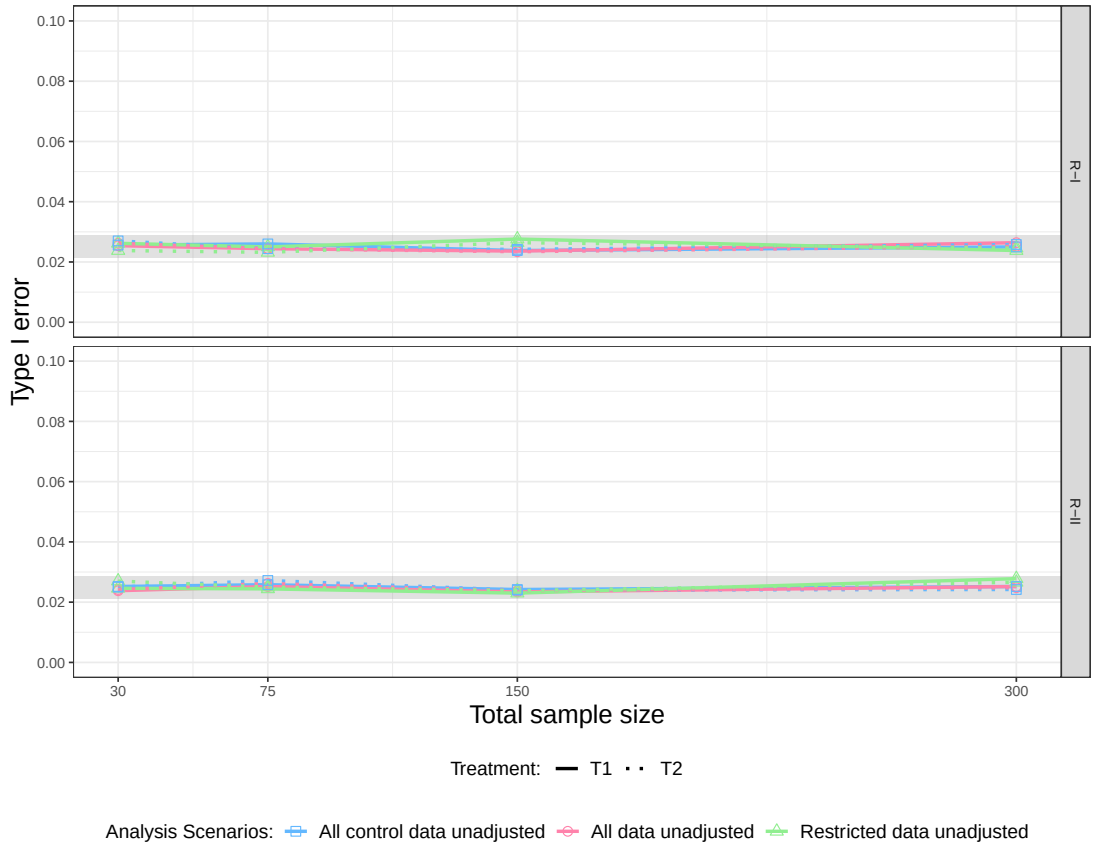

Figure 5.18: Type I error over increasing sample size for a prevalence of 0.5:0.5:0 to the subgroups X, Y and Z for R-I and R-II. Fixed design parameters: block randomization, $\mu_X^C = \mu_Y^C = 0.5$ and $\Delta^{T1} = \Delta^{T2} = 0$.

5.2.3 Setting 3: patients are recruited from all three subgroups X, Y and Z

In order to investigate the impact of allowing selective exclusion of treatment arms, the more complex setting where patients come from all three subgroups X, Y and Z is now investigated. Among other scenario assumptions, the differences between adjusting and not adjusting for the subgroups X, Y and Z will be examined. For better visibility it was decided to use darker colors (dark blue, dark red and dark green) to represent the adjusted analysis scenarios and lighter versions of the colors (light blue, pink and light green) for the unadjusted analysis scenarios. Besides, points were added for better comparability. Three different shapes were chosen to differentiate between the three different compositions of control data. The adjusted analysis scenarios are represented by the filled shapes and for the unadjusted analysis scenarios the corresponding unfilled shapes were chosen.

In the first subsection, the effect of an equal prevalence in the subgroups X, Y and Z and case 1, meaning $\Delta_X^{T1} = \Delta_Z^{T1}$ and $\Delta_Y^{T2} = \Delta_Z^{T2}$, will be investigated. In the second subsection, the effect of different prevalence in the subgroups will be taken into account and both cases for the treatment effects are considered, i.e. equal as well as unequal treatment effects in the different subgroups X, Y and Z. As a reminder, the four prevalence in the three subgroups X, Y and Z that were taken into account are the following:

- an equal prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X:Y:Z
- an unequal prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X:Y:Z
- a high prevalence of $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X:Y:Z
- a low prevalence of $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X:Y:Z

Hence, a high prevalence refers to a prevalence distribution where the majority of patients is recruited from subgroups X and Y and a low prevalence describes that less patients are recruited from subgroups X and Y than from subgroup Z.

Equal prevalence in the subgroups X, Y and Z

The total sample size, randomization strategy and means in the control arm were varied in order to investigate whether the type I error is controlled. Figure 5.19 shows that the increasing sample size and randomization strategies have no

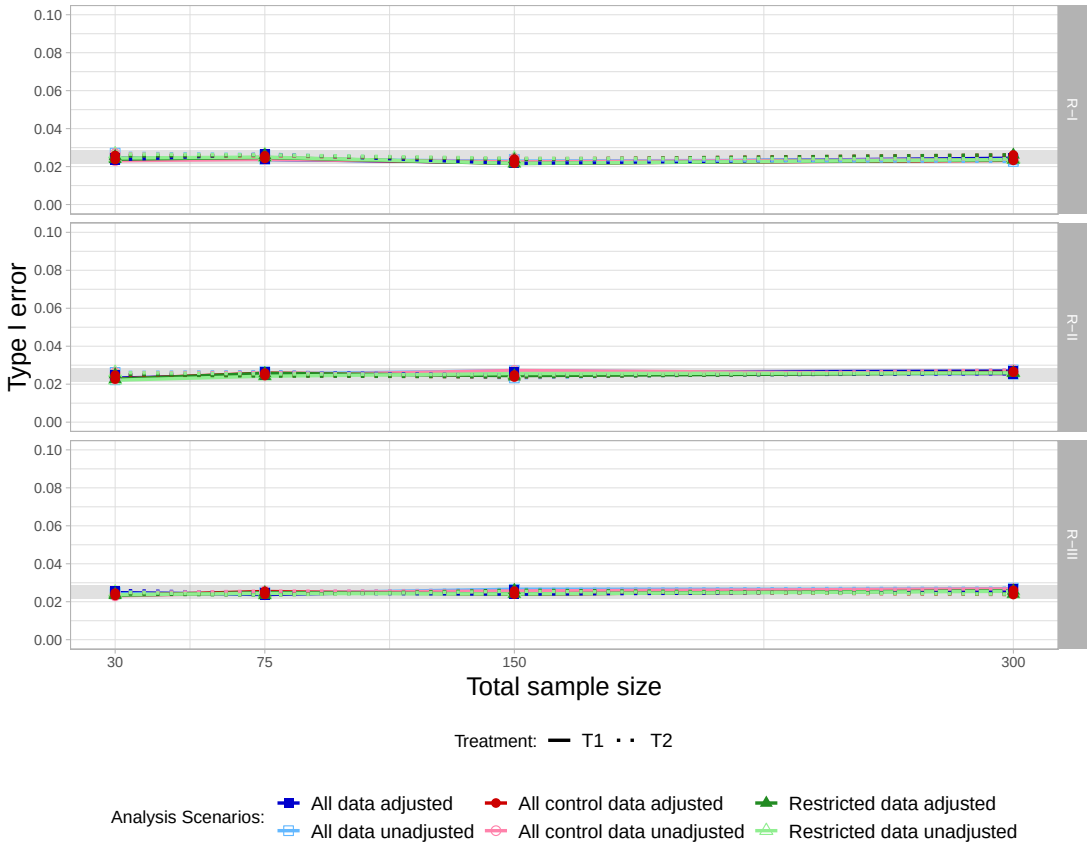

Figure 5.19: Type I error over increasing sample size for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z for all considered randomization strategies. Fixed design parameters: block randomization, $\mu_X^C = \mu_Y^C = \mu_Z^C = 0$ and $\Delta^{T1} = \Delta^{T2} = 0$.

effect on the type I error for $\Delta^{T1} = \Delta^{T2} = 0$ and $\mu_X^C = \mu_Y^C = \mu_Z^C = 0$. Under the null hypothesis all considered analysis scenarios maintain the type I error rate at the nominal level of 0.025 for homogeneous controls. Note, the lines for the analysis scenarios mask each other.

Next, the means in the control arm were varied as follows: $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$. Figure 5.20 shows that the type I error is no longer controlled for all analysis scenarios. For heterogeneous controls, the type I error is controlled for all adjusted analyses. Note, for the adjusted analyses the *restricted data* mask *all data* and *all control data*. For T2, where $\mu_Y^C = 0.5$ was assumed, the type I error is significantly inflated for the unadjusted analyses based on *all data* and *all control data* for all randomization strategies. Note, for the unadjusted analysis scenarios *all data* and *all control data* overlap each other. The type I error for *all data*

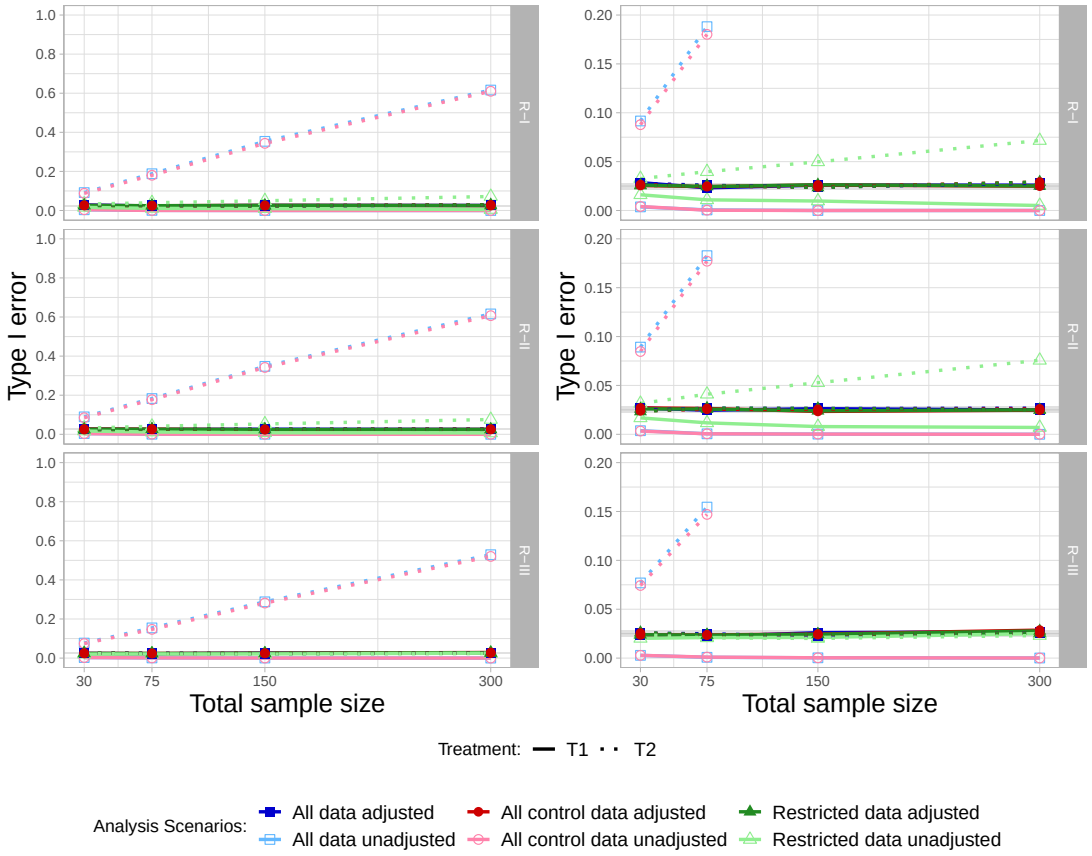

Figure 5.20: Type I error over increasing sample size for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z for R-I, R-II and R-III. Fixed design parameters: block randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\mu_Z^C = 0$ and $\Delta^{T1} = \Delta^{T2} = 0$. The right plot represents a zoom in on the y-axis of the left plot and the y-axis is restricted to [0, 0.2].

and *all control data* is less inflated for R-III than for the other two randomization strategies (e.g. the type I error is 0.6 for a total sample size of 300 for R-I and R-II and 0.5 for a total sample size of 300 for R-III). Furthermore, increasing the sample size did not help to reduce the type I error inflation since the type I error inflation increases for increasing sample size. Besides, for the unadjusted analysis scenarios the inflation of the type I error is higher for *all data* and *all control data* than for *restricted data*. For *restricted data*, the type I error is controlled for R-III while for R-I and R-II it is controlled for a total sample size of 30 but not for larger sample sizes. For T1, where one assumed $\mu_X^C = -0.5$, one can see in the zoom in of the y-axis that for the unadjusted analysis scenarios for *all data* and *all control data* the type I error is lower than the nominal significance level of 0.025. That is the case for all randomization strategies. The negative mean in

the control arm results in a stricter control over false positives. Note, *all control data* mask *all data*. For the unadjusted analysis based on the *restricted data* the type I error is lower than the significance level for R-I and R-II. For R-III the type I error is controlled for the unadjusted analysis based on the *restricted data*. To summarize, for heterogeneous controls the type I error is only controlled for the adjusted analysis scenarios for all considered randomization strategies and sample sizes.

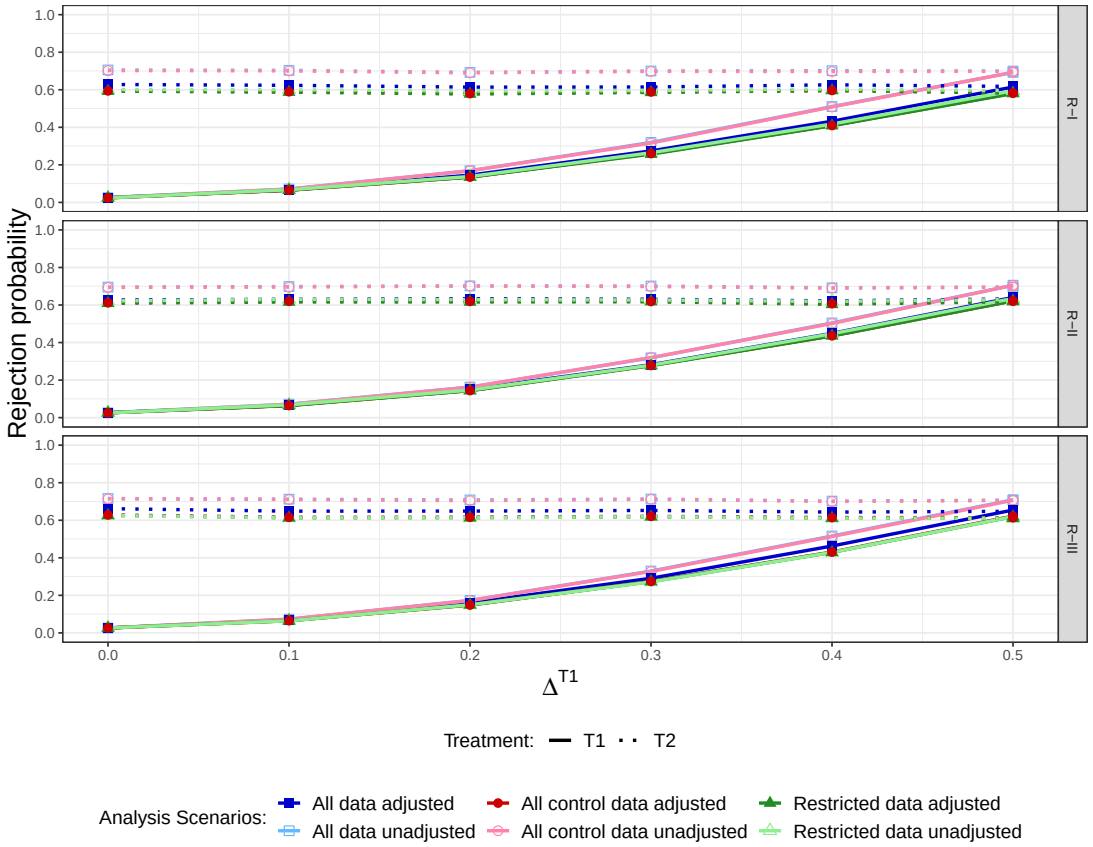

Figure 5.21: Rejection probability over increasing Δ^{T1} for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z for R-I, R-II and R-III. Fixed design parameters: block randomization, $\mu_X^C = \mu_Y^C = \mu_Z^C = 0$, n = 150 and $\Delta^{T2} = 0.5$.

In Figures 5.21 and 5.22 it was investigated how the randomization strategies and means in the control arm affect the rejection probability. The simulation parameters that were varied between Figures 5.21 and 5.22 are the means in the control arm. Since $\Delta^{T1} = 0$ corresponds to the null hypothesis, the term "rejection probability" is employed for the next plots instead of "power".

For homogeneous controls, see Figure 5.21, one can see that the rejection prob-

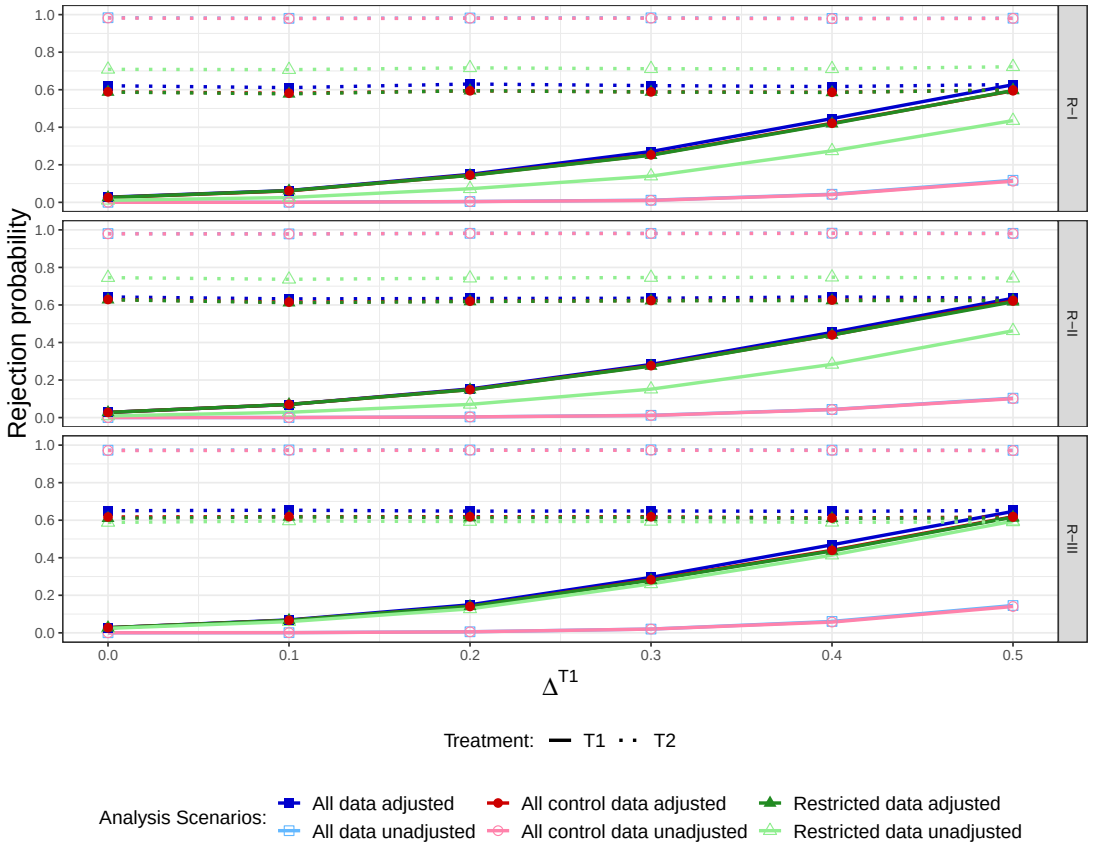

Figure 5.22: Rejection probability over increasing Δ^{T1} for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z for R-I, R-II and R-III. Fixed design parameters: block randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\mu_Z^C = 0$, n = 150 and $\Delta^{T2} = 0.5$.

ability increases for T1 for an increasing treatment effect for all randomization strategies. Besides, the adjusted analysis scenarios have a higher rejection probability than the unadjusted analysis scenarios. An exception are the unadjusted analyses based on the *restricted data*. One can see that the rejection probability of the unadjusted *restricted data* is close to the rejection probability of the adjusted analysis scenarios. For the adjusted analysis scenarios the rejection probability is slightly higher for R-III. When aiming for R-III one has a ratio of 1:1 in each subgroup (see Figure 4.3) which results in a higher rejection probability. Regarding the adjusted analysis scenarios, the comparisons based on *all data* have a slightly higher rejection probability compared to *all control data* and *restricted data* in all three considered randomization strategies. Note, for treatment 1 for the unadjusted analysis scenarios, *all control data* overlaps *all data* and for the adjusted analysis scenarios *all control data* and *restricted data* are masked by the

unadjusted *restricted data*. For T2 one can see that the rejection probability for the unadjusted scenarios based on *all data* and *all control data* is the highest and slightly higher for R-III. Besides, the rejection probability for the unadjusted analysis scenario based on *restricted data* is also the highest for R-III and the lowest for R-I. The rejection probability of the unadjusted analysis based on the *restricted data* is close to that of the adjusted analysis scenarios. The adjusted analyses scenarios also have the highest rejection probability for R-III. Note, for T2 for the unadjusted analysis scenarios, *all control data* overlap *all data* and for the adjusted analysis scenarios *all control data* and *restricted data* are masked by the unadjusted *restricted data*. For the adjusted analysis scenarios, the comparisons based on *all data* have the highest rejection probability. By basing the analysis on the full data set, the variance estimates are slightly smaller which leads to a greater power.

For Figure 5.22 heterogeneous controls were considered: $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$. One can see that the rejection probability for the unadjusted analysis scenarios changed compared to Figure 5.21. The reason for this is that a bias in the effect estimates is introduced for the unadjusted analysis scenarios based on *all data* and *all control data*. That is because one pools the control data of the three subgroups X, Y and Z for the unadjusted scenarios, which have different means, and does not adjust for the subgroups. By pooling the control data of the other subgroups, a bias in the effect estimates is introduced. For T1 the bias in the effect estimate is negative, $\mu_X^C = -0.5$, and for T2 the bias in the effect estimate is positive, $\mu_Y^C = 0.5$. As result, the rejection probability for the unadjusted scenarios for T1 is lower and for T2 higher in comparison to the rejection probability of the adjusted analysis scenarios. An exception is the unadjusted analysis of the *restricted data* for R-III since the rejection probability is the same as for the adjusted analysis scenarios. When comparing the adjusted analysis scenarios there are slight differences between the randomization strategies, e.g. the rejection probability is slightly lower for R-I and slightly higher for R-III. Besides, one can see that the comparisons based *all data* have the highest rejection probability for all randomization strategies. That is because the variance can be estimated more efficiently, assuming variance homogeneity, as more data are available for the estimation. As result, the variance estimates are slightly smaller and lead to a greater power. To conclude, R-III is the preferred randomization strategy since it results in the highest rejection probability.

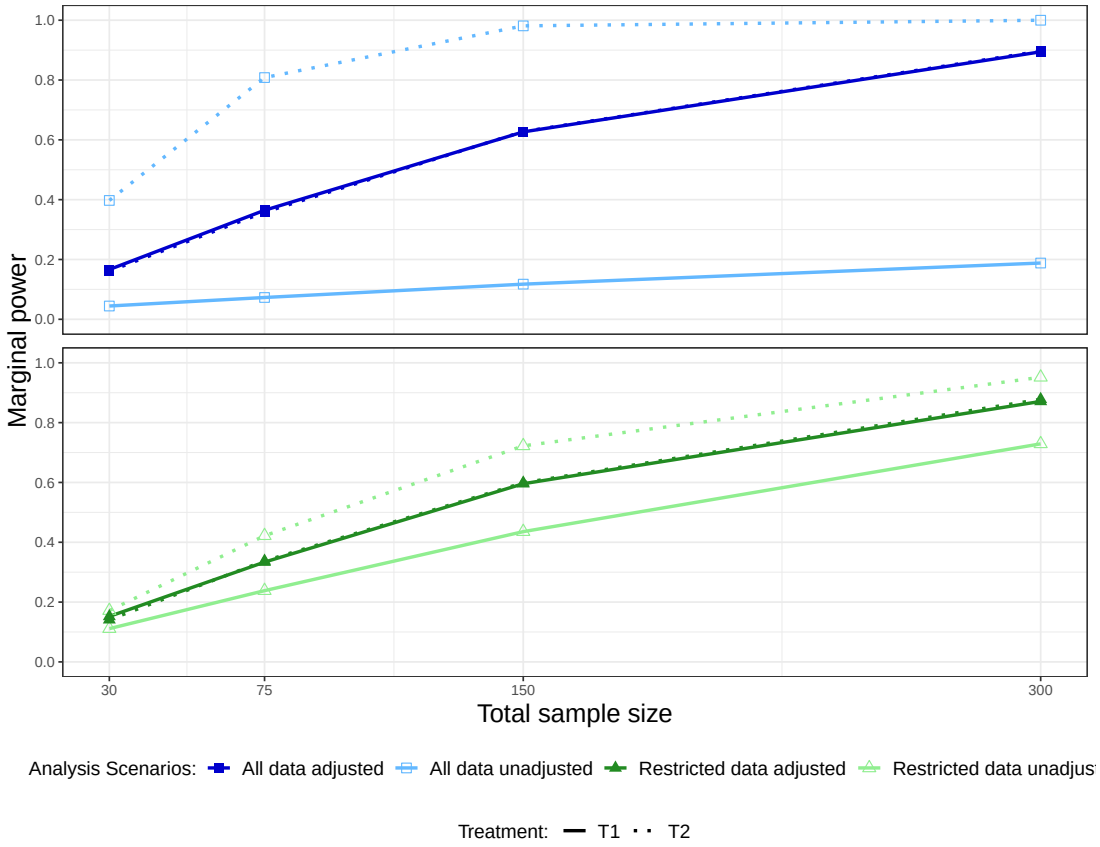

Figure 5.23: Power over increasing sample size for a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z for analyses based on *all data* and *restricted data*. Fixed design parameters: block randomization, $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$, $\mu_Z^C = 0$, R-I and $\Delta^{T1} = \Delta^{T2} = 0.5$.

Figure 5.23 stresses the difference between adjusting and not adjusting for heterogeneous controls. For that one considers the comparisons based on *all data* and *restricted data* for the same heterogeneous controls as above, i.e. $\mu_X^C = -0.5$, $\mu_Y^C = 0.5$ and $\mu_Z^C = 0$. The upper row depicts *all data* and the *restricted data* are displayed in the lower row. For the unadjusted analysis scenarios one can see a big difference between the power for T1 and T2. For *all data* the difference between the analysis scenarios is due to the bias in the effect estimates as explained above: the control data are pooled over the three subgroups X, Y and Z without adjusting for the subgroups. One can see that by adjusting for the different subgroups the variability between the subgroups is removed and the power for both treatments results in the same values as expected because the same treatment effects were assumed. For *restricted data* the difference between the power of the treatments is not due to the bias in the effect estimates, as the correct data have been used

for the comparisons, but due to the bimodal distribution of the subgroups which is not accounted for (see Figure 5.8 for a visualization of the bimodal distribution for one simulated data set). For the *restricted data* the power for the adjusted analysis scenario also results in the same values as expected since the same treatment effect was assumed. Figure 5.23 stresses the fact that it is recommendable to adjust for the subgroups in the analysis.

Different prevalence in the subgroups X, Y and Z

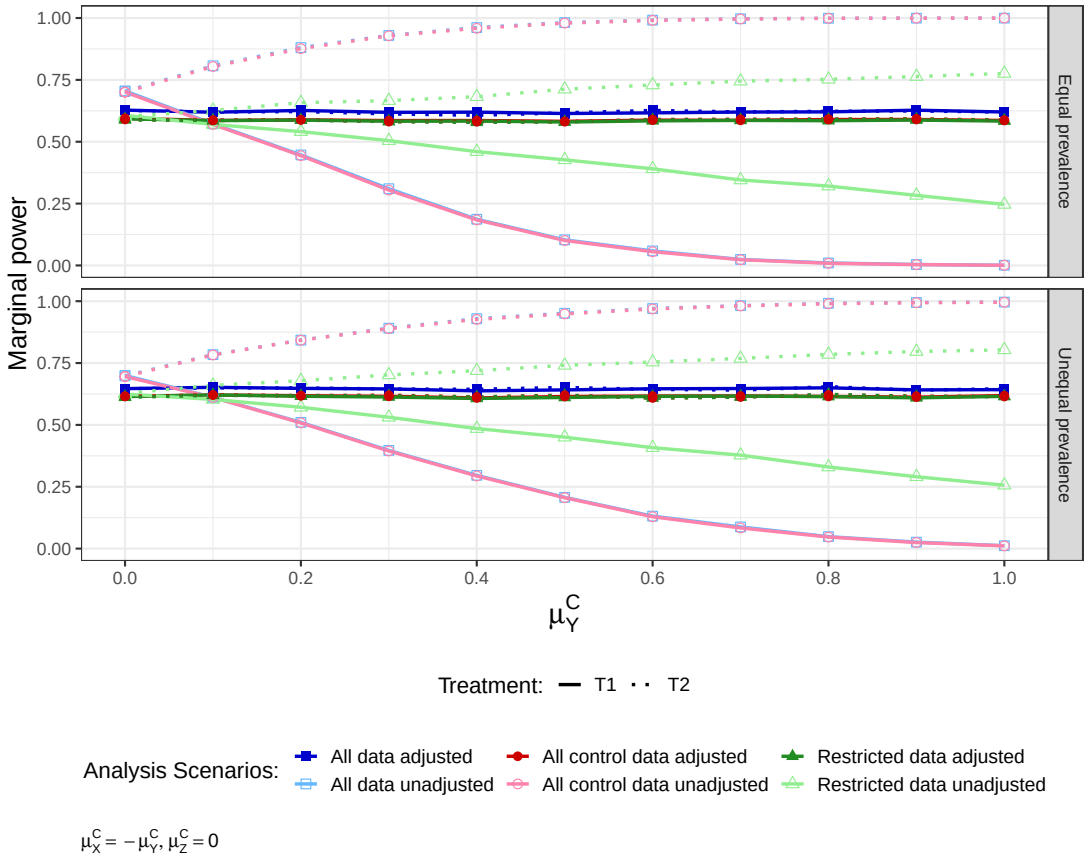

Figure 5.24: Power for heterogeneous controls for all considered analysis scenarios. The upper row shows a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z while the lower row depicts a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, R-I, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

Since heterogeneous controls introduce bias in the effect estimates and increase the variance, it was of interest to investigate the effect further. Therefore, the

means in the control arm were varied as follows: one subgroup has a mean of 0, one subgroup a positive mean between 0 and 1 and one subgroup a negative mean between 0 and -1. For the exact values see Tables A.2 and A.3 which provide an overview over the means in the control arm and treatment arms for the investigated simulation scenarios.

Figure 5.24 compares the power for all analysis scenarios for an equal prevalence and an unequal prevalence to the subgroups X, Y and Z. For the means per treatment arm in the subgroups see Table A.2. One can see that for the two considered prevalence to the subgroups X, Y and Z there are hardly any differences. The power for all adjusted analysis scenarios and unadjusted analysis scenario based on *restricted data* is slightly higher for an unequal prevalence. The reason for that is for an unequal prevalence more patients are recruited from subgroup Z. Figure 5.24 shows that the power for the adjusted analysis scenarios is constant over the different means. Note, for the adjusted analysis scenarios the *restricted data* mask *all control data*. However, one can see that for the unadjusted analysis scenarios the power increases over increasing μ_Y^C and decreasing μ_Z^C for T2 and decreases for T1. The power for the unadjusted analysis scenarios based on *all data* and *all control data* for T2 is the highest and lowest. Note, for the unadjusted analysis scenarios *all data* and *all control data* overlap each other.

Figure 5.25 is a simplification of Figure 5.24 by focusing on an unequal prevalence to the subgroups X, Y and Z and the unadjusted analysis scenarios. As mentioned above, the power loss for T1 as well as the power increase for T2 for comparisons based on *all data* and *all control data* is due to the bias in the effect estimates. Note, *all data* and *all control data* overlap each other. The bias is introduced by pooling the control data of the subgroups X, Y and Z and not accounting for the subgroups. For the comparisons based on the *restricted data* no bias in the effect estimates is introduced since the comparisons are based on the correct data. But the means in subgroups X and Y are further and further away from the mean in subgroup in Z. The bimodal distribution is not adjusted for and the variance increases. Besides, for R-I there are less controls in subgroup Z and one has less patients in the treatment arms in subgroup Z compared to subgroup X (see Table 4.4). The bias in the effect estimates as well as the increase in variance is removed when adjusting for the subgroups, as shown in Figure 5.24.

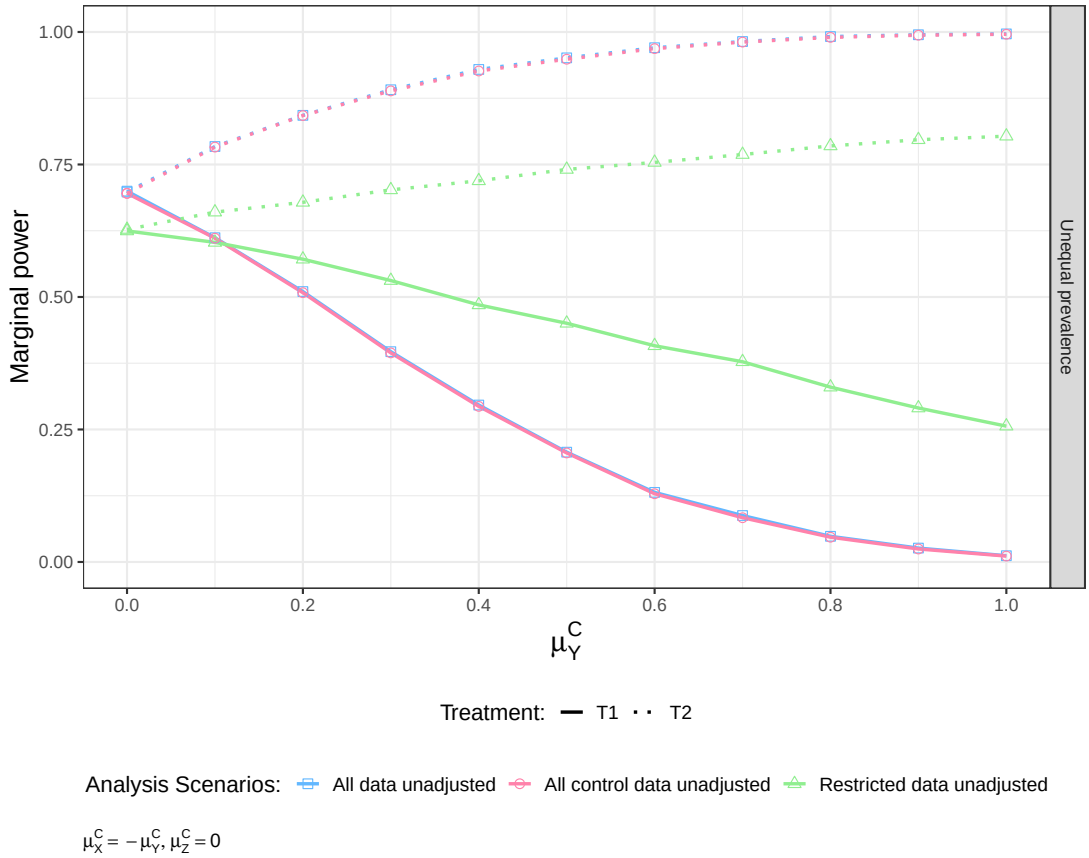

Figure 5.25: Power for heterogeneous controls for the unadjusted analysis scenarios for a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z for R-I, R-II and R-III. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

To summarize, unadjusted analyses may result in an inflation of the type I error, particularly in the presence of heterogeneous controls. While there are instances where unadjusted analyses lead to an increased power, there are also cases where the power is reduced. Given the possibility of a type I error inflation for the unadjusted analyses, they are not the primary focus of interest. Consequently, the subsequent investigations will primarily concentrate on the adjusted analyses to determine the randomization strategy and control data composition that yield the highest power. For the sake of completeness the plots for the unadjusted analysis scenarios for different prevalence to the subgroups X, Y and Z can be found in the appendix (see Section A.3.1).

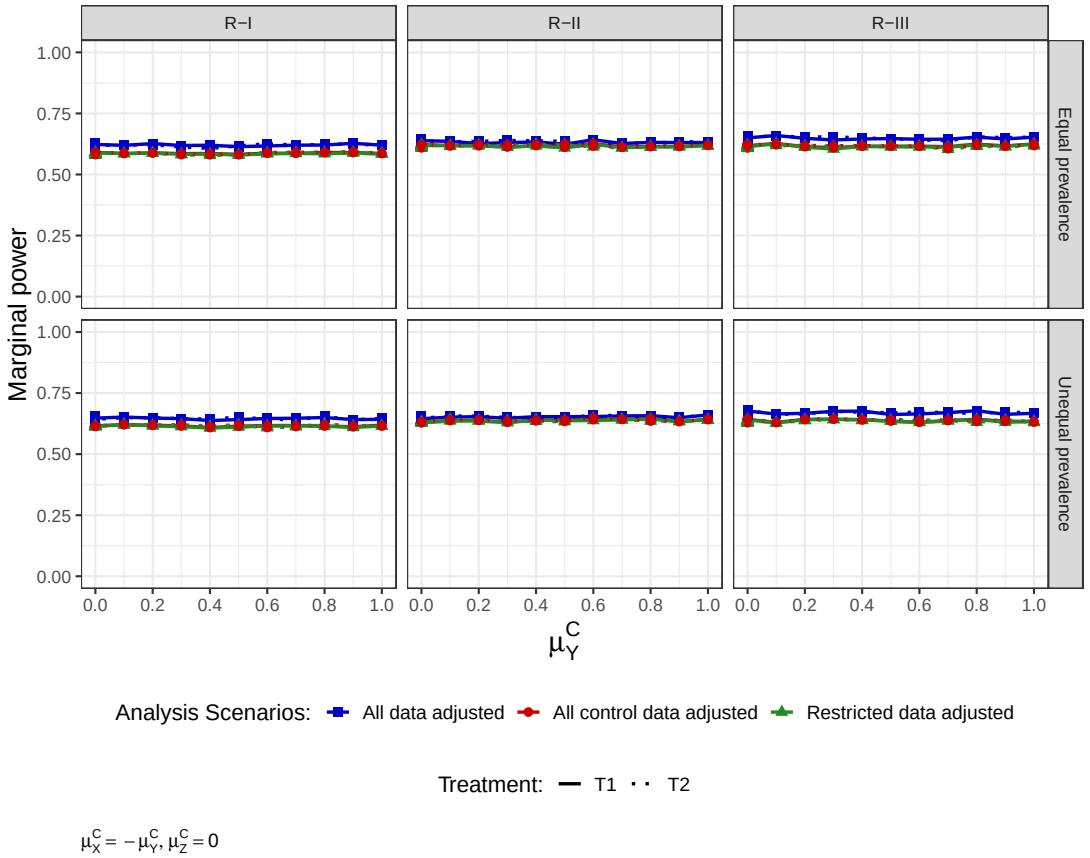

Figure 5.26: Power for heterogeneous controls for the adjusted analysis scenarios. The upper row shows a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z while the lower row shows a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

The following figures depict the randomization strategies for all adjusted analysis scenarios for different prevalence to the subgroups X, Y and Z. Figure 5.26 depicts the adjusted analysis scenarios for the three randomization strategies for an unequal and equal prevalence to the subgroups X, Y and Z. For the exact means for the control and treatment arms in the subgroups see Table A.2. One can see slight differences for the composition of the control data for the randomization strategies. The power for *all data* is higher than the power for *all control data* and *restricted data* for R-I and R-III for both considered prevalence. Note, *all control data* and *restricted data* overlap each other. The power for *all data* is the highest for R-III. Variance homogeneity is assumed and the variance estimates for *all data* can be estimated more efficiently compared to the other two data sets because more data are available for the estimation. As result, the variance estimates are

slightly smaller which yields a greater power. For *all control data* and *restricted data* the power is the lowest for R-I and the same for R-II and R-III. For R-II there are the least differences between the different compositions of control data. Comparing the prevalence, the power is slightly higher for an unequal prevalence because more patients are recruited from subgroup Z. The same results can be found when swapping the means for the control arm in subgroups Y and Z (see Figure A.3).

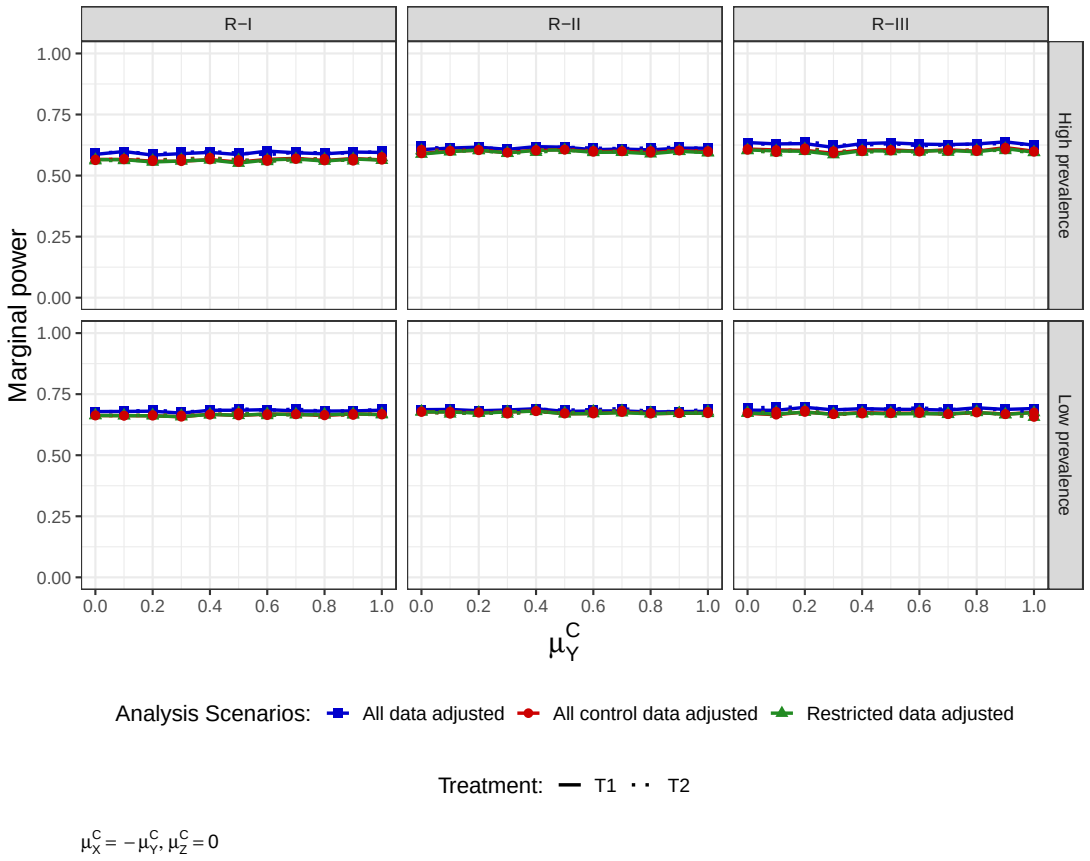

Figure 5.27: Power for heterogeneous controls for the adjusted analysis scenarios. The upper row shows a high prevalence of $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X:Y:Z and the lower row shows a low prevalence of $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X:Y:Z. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

Compared to the plots before, the only simulation parameter that was varied for Figure 5.27 is the prevalence to the subgroups X, Y and Z. Now a high and low prevalence to the three subgroups is considered. For a high prevalence more patients come from subgroups X and Y than from subgroup Z and for a low preva-

lence less patients are recruited from subgroups X and Y compared to subgroup Z. For the exact means for the control and treatment arms in the subgroups see Table A.2. When comparing the high prevalence in the subgroups with the low prevalence in the subgroups, one can see that the power for the low prevalence is higher for all randomization strategies. The reason behind this is that in the case of a low prevalence, a larger proportion of patients is recruited from subgroup Z. For a high prevalence the power for *all data* is highest for R-III. Besides, for a high prevalence the power for *all data* is higher compared to the other two compositions of control data for R-I and R-III while for R-II the three data sets overlap each other. Since variance homogeneity is assumed, the estimation of the variance for the entire data set can be performed more efficiently than for the other two data sets, primarily due to the larger amount of available data for estimation purposes. The power is lowest for R-I because for a high prevalence most patients are recruited from subgroups X and Y and for R-I one has a ratio of 2:1 for the respective treatment arm vs. control in subgroups X and Y and as result, one has less controls for the comparisons. For a low prevalence one can see less differences between the randomization strategies. For R-I and R-III the power for *all data* is again slightly higher than for the other two data sets. The reason for the hardly visible differences between the randomization strategies for a low prevalence is that most patients (80%) are recruited from subgroup Z. For subgroup Z it was shown that there is no difference between a multi-arm trial and a trial with two substudies (see Figure 4.2). As result, the randomization ratios considered in subgroup Z for the three randomization strategies, i.e. either 1:1:1 for T1 vs. T2 vs. C (R-I and R-III) or 1:1:2 for T1 vs. T2 vs. C (R-II) have no effect on the power. Even though one considers different ratios for subgroups X and Y for the randomization strategies, the effect is hardly visible because only a minority of the patients (20%) are recruited from subgroups X and Y.

Compared to Figure 5.27, the means for subgroups Y and Z were swapped for Figure 5.28 (see Table A.3). Figure 5.28 shows that the power for a high prevalence to the subgroups is lower than for a low prevalence to the subgroups. Besides, when the majority of patients is recruited from subgroup Z there are hardly any differences between the randomization strategies. For a high prevalence, the power is highest for R-III, i.e. an equal ratio within each subgroups leads to the highest power for all considered compositions of control data. The power is lowest for R-I, i.e. an ratio of 2:1 for the respective treatment arm vs. control in subgroups X and Y results in less controls and thus, in a lower power. Figure 5.28 further

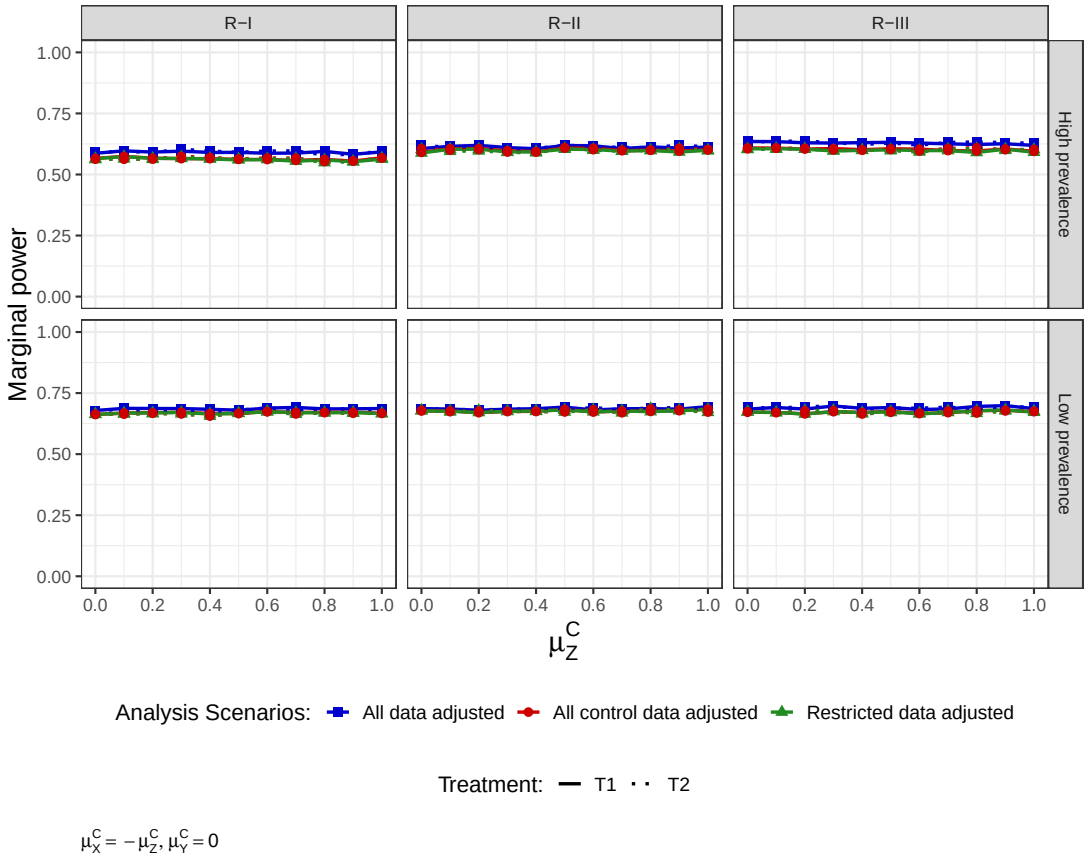

Figure 5.28: Power over heterogeneous controls for the adjusted analysis scenarios. The upper row shows a high prevalence of $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X:Y:Z and the lower row shows a low prevalence of $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X:Y:Z. The following heterogeneous controls were considered: μ_Z^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Y^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

shows that the smaller control arm for *restricted data* compared to *all control data* does not result in a smaller power. As expected, one cannot see any differences between Figures 5.27 and 5.28 because the heterogeneous controls in the subgroups are adjusted for.

Case 2: unequal treatment effects in the subgroups X, Y and Z

In the following, it was of interest to vary the treatment effects in the subgroups X, Y and Z. For that the following two cases were investigated:

- case 1: $\Delta_Z^{T1} = \Delta_X^{T1} * 2$ and $\Delta_Z^{T2} = \Delta_Y^{T2} * 2$
- case 2: $\Delta_Z^{T1} = \Delta_X^{T1}/2$ and $\Delta_Z^{T2} = \Delta_Y^{T2}/2$.

The values that were chosen for Δ_X^{T1} and Δ_Y^{T2} are 0.25 and 0.5. For better legibility only the treatment effects for T1 are written in the facets, since T1 and T2 have the same values in the subgroups, e.g. for case 1 $\Delta_Z^{T1} = \Delta_Z^{T2} = 0.5$ and $\Delta_X^{T1} = \Delta_Y^{T2} = 0.25$. The following figures depict the randomization strategies for all adjusted analysis scenarios for different prevalence to the subgroups X, Y and Z while taking different treatment effects in the subgroups into account.

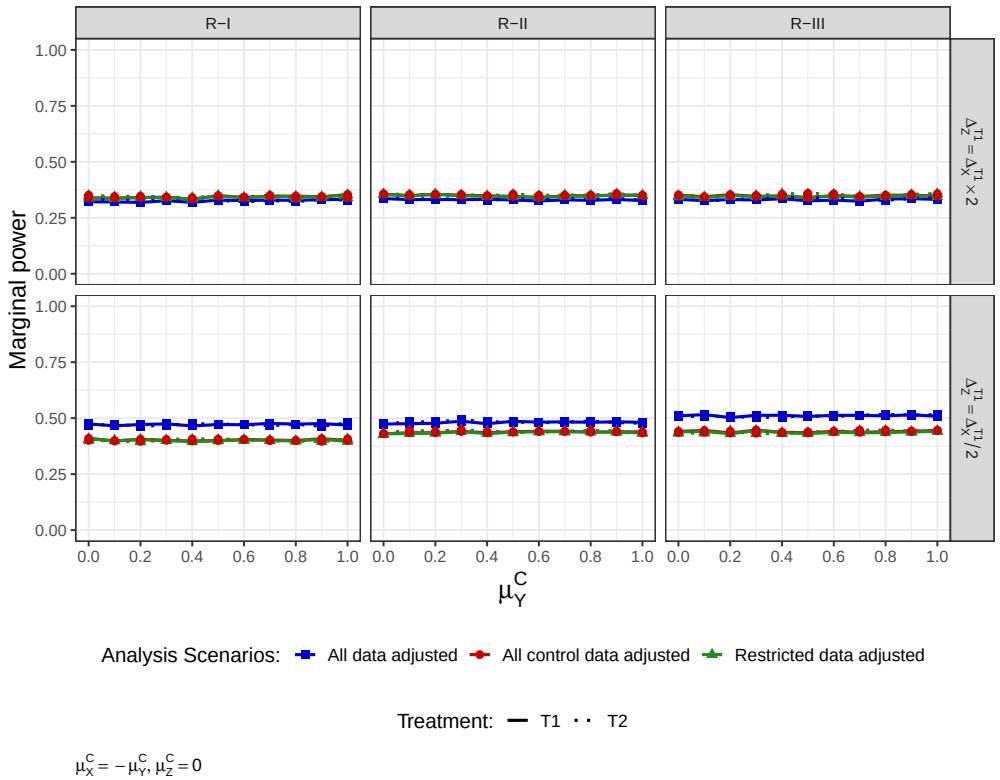

Figure 5.29: Power for all adjusted analysis scenarios and randomization strategies for different treatment effects in the subgroups. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed parameters: n = 150, block randomization and a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z.

Figure 5.29 shows the effect of varying the treatment effects in the subgroups for all adjusted analysis scenarios for an equal prevalence to the subgroups X, Y and Z. The upper plots depict the case where the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y. One cannot see any differences between the randomization strategies or the composition of the control data. The lower row shows the case where the treatment effects in subgroup Z are half the treatment effects in subgroups X and Y. One can see differences between the randomization strategies, i.e. the power is the highest for R-III. Besides, one can see that the power for *all data* is higher than for *all control data* and *restricted data*. Note, *all control data* and *restricted data* overlap each other. As mentioned before, the power for *all data* is higher because by including the patients of the third treatment arm the variance estimates are slightly smaller which results in a higher power. In comparison to Figure 5.26, where the same treatment effects in the subgroups were considered, e.g. $\Delta^{T1} = \Delta^{T2} = 0.5$, one can see that for an equal prevalence the power is lower when a third or two thirds of the patients are recruited from a subgroup with a lower treatment effect. $\Delta^{T1} = \Delta^{T2} = 0.5$ in Figure 5.26 resulted in a power of approximately 60%, while for $\Delta_Z^{T1} = \Delta_Z^{T2} = 0.5$ and $\Delta_X^{T1} = \Delta_Y^{T2} = 0.25$ the power is only approximately 43%.

Figure 5.30 shows the effect of different treatment effects when considering an unequal prevalence to the subgroups X, Y and Z. For an unequal prevalence, the power is almost the same for both considered treatment effect cases. Again, one cannot see any visible differences between the randomization strategies for the upper row, where the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y. Compared to Figure 5.29, the unequal prevalence in the subgroups results in a larger power for the case where the treatment effects in subgroup Z are larger. This is due to the unequal prevalence, where half of the patients are recruited from subgroup Z. Furthermore, in subgroup Z, the treatment effects are twice as large as those in the other half of the patients. When the treatment effects are larger in subgroups X and Y one can see that the power for *all data* is the largest for all randomization strategies. Besides, the power is largest for R-III. Compared to 5.29, the unequal prevalence in the subgroups results in a lower power for the case where the treatment effects in subgroup Z are smaller. The reasons for that is that for an unequal prevalence half of the patients are recruited from subgroup Z with treatment effects half the size than those of the other half of patients. In Figure 5.26, where $\Delta^{T1} = \Delta^{T2} = 0.5$, the power was approximately at 60%. One can see that for an unequal prevalence to the subgroups

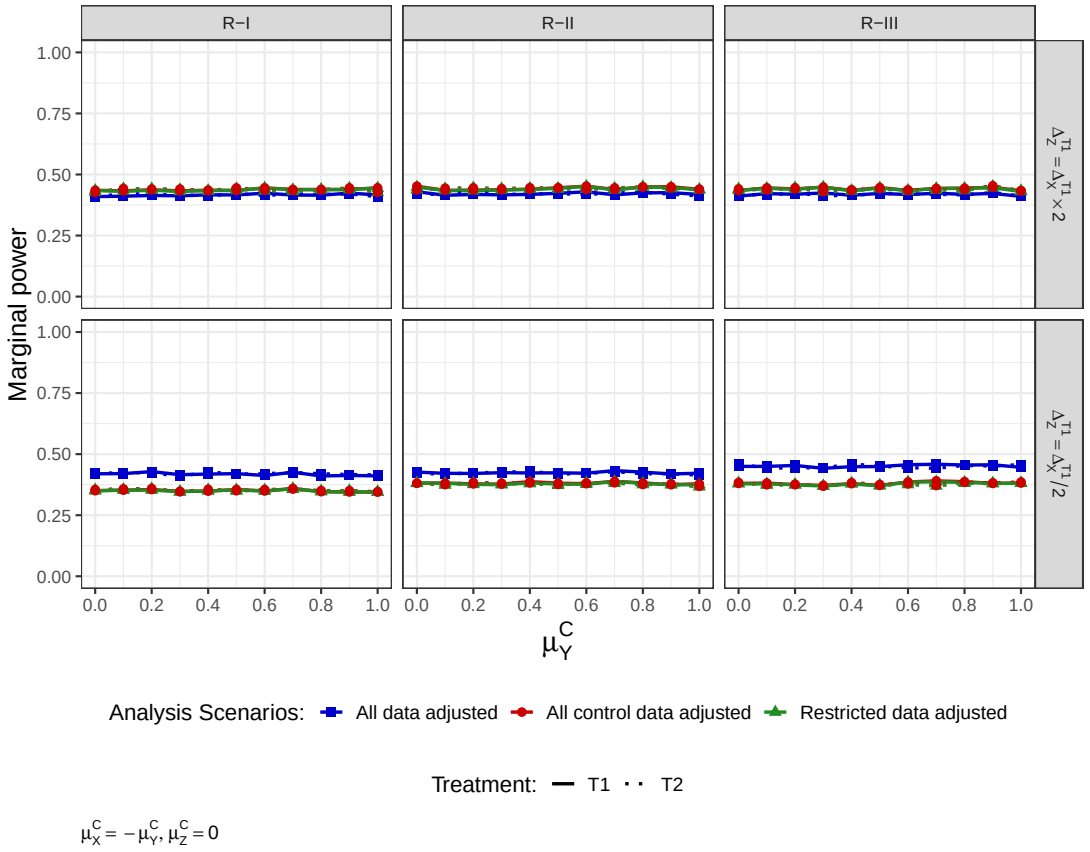

Figure 5.30: Power for all adjusted analysis scenarios and randomization strategies. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed parameters: n = 150, block randomization and a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z.

X, Y and Z, both considered cases for the treatment effects result in a lower power.

One can see the effect of the considered treatment effects even stronger for a high and low prevalence to the subgroups X, Y and Z. Figure 5.31 illustrates that when the treatment effects in subgroups X and Y are larger than in subgroup Z, the power is higher for a high prevalence. That is because for a high prevalence the majority of patients are recruited from subgroups X and Y and a higher treatment effect in those subgroups results in a higher power. Accordingly, the power is lower then the treatment effects in subgroups X and Y are smaller than the treatment effects in subgroup Z. Besides, when the treatment effects are larger in subgroups X and Y one can see that the power for *all data* is the largest for all randomization strategies. The figure shows that for both cases the power is slightly larger for

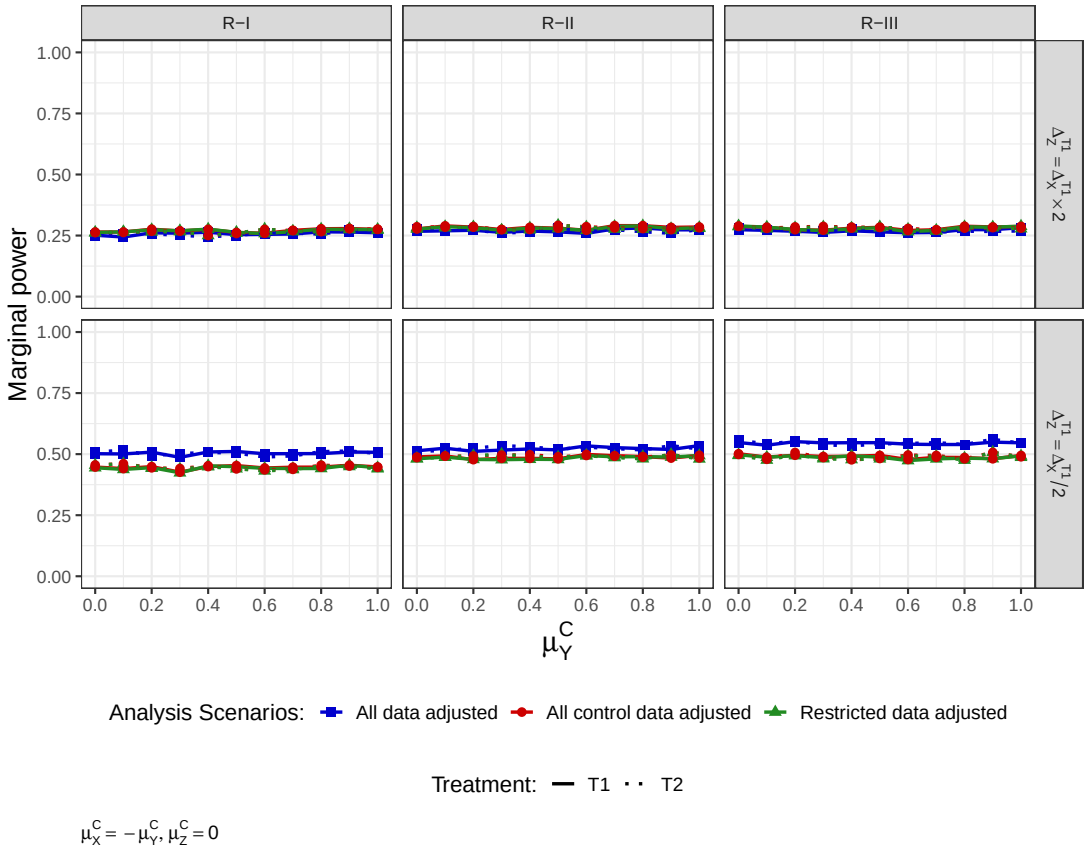

Figure 5.31: Power for all adjusted analysis scenarios and all randomization strategies. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed parameters: n = 150, block randomization and a prevalence of $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X, Y and Z.

R-III compared to the other two considered randomization strategies.

Figure 5.32 shows that for a low prevalence the power is larger when the treatment effects are larger in subgroup Z. The reason for that is that for a low prevalence most patients are recruited from subgroup Z and higher treatment effects in subgroup Z result in a larger power. In Figure 5.27, where $\Delta^{T1} = \Delta^{T2} = 0.5$, the power was approximately at 60%. One can see that for $\Delta_Z^{T1} = \Delta_Z^{T2} = 0.5$ and $\Delta_X^{T1} = \Delta_Y^{T2} = 0.25$, the power is also approximately 60% for a low prevalence. It can be observed that achieving the same power is possible by recruiting 80% of the patients from the subgroup with the larger treatment effects, as compared to the scenario where all patients had uniformly high treatment effects. In comparison, the unequal prevalence, where 50% of the patients were recruited from the

subgroup with the larger treatment effects, resulted in a power loss (see Figure 5.30).

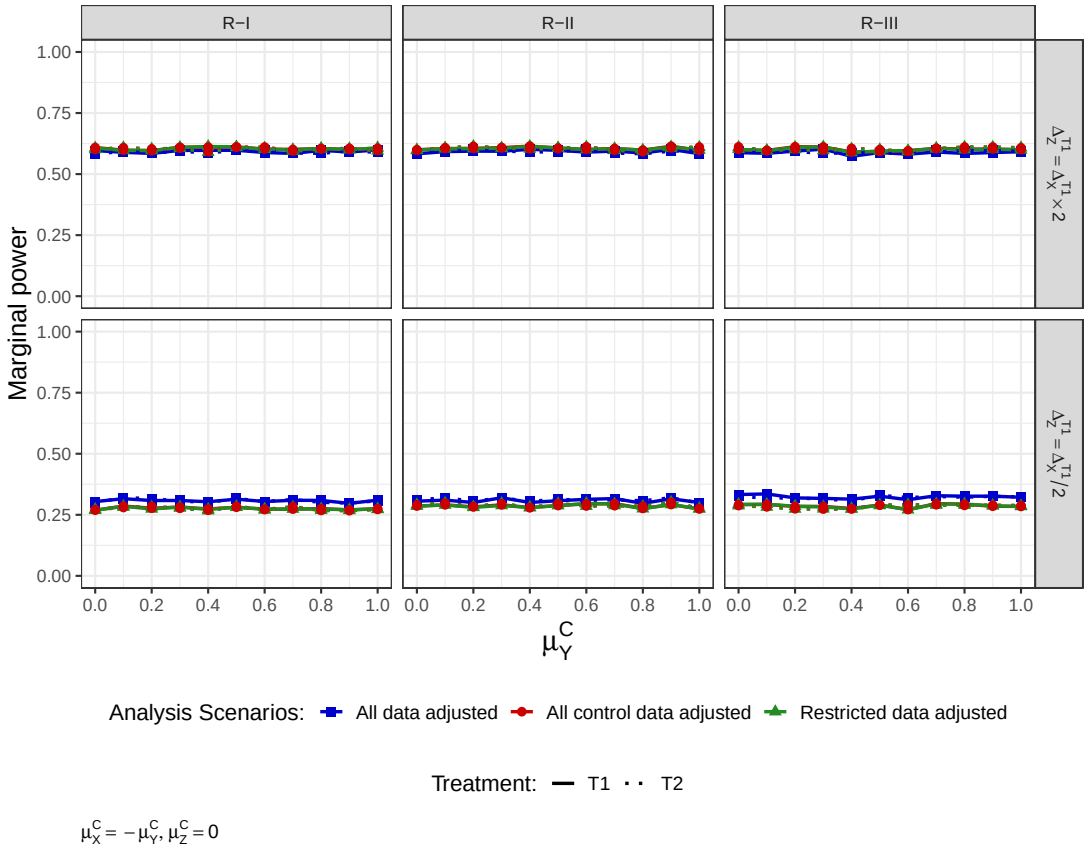

Figure 5.32: Power for all adjusted analysis scenarios and all randomization strategies. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed parameters: n = 150, block randomization and a prevalence of $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X, Y and Z.

In conclusion, when the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y, no noticeable differences are observed for the power for the adjusted analysis scenarios among the randomization strategies. However, in the case where the treatment effects in subgroups X and Y are twice the treatment effects in subgroup Z, following R-III yields the highest power for the adjusted analysis scenarios. Additionally, for this latter case, the comparisons based on *all data* demonstrate a higher power compared to the comparisons based on *all control data* and *restricted data*.

6 Conclusion and discussion

6.1 Summary

While study designs with selective exclusion of treatment arms have been described in the literature, the implementation and statistical analysis of such study design has not received much attention. Therefore, a multi-arm trial with selective exclusion of treatment arms was considered in this thesis to enhance the methodology for the optimal usage of different patient populations in the analysis of the trial. In this thesis, the performance of the proposed methods in terms of the type I error rate and statistical power was evaluated in a simulation study, considering a wide range of scenarios. The sample size to the three arms was varied, different prevalence to the subgroups were considered and multiple randomization strategies were investigated. Besides, it was of interest to explore the effect of different treatment effects in the subgroups and heterogeneous means in the control arm. Finally, the comparisons were based on different compositions of control data and the difference between adjusted and unadjusted analyses was investigated.

The results from the simulation study show that not adjusting for the different subgroups may result in an inflation of the type I error, particularly in the presence of heterogeneous controls. The type I inflation increases for increasing sample size for the unadjusted analysis scenarios. Given the possibility of a type I error inflation for the unadjusted analyses, the usage of adjusted analyses is recommended since by adjusting for the subgroups the variability between the subgroups is removed.

Regarding the different compositions of control data, it was shown that the treatment effects in the subgroups have an effect on whether the compositions of control data result in differences. When the treatment effects are the same in the subgroups, the comparisons based on the full data set yield the highest power. The full data set also includes the patients who are randomized to the third arm, which is not used for the comparisons. Thus, the variance estimates are slightly

smaller which leads to a greater power [15]. However, as basing the comparisons on the full data set relies on further assumptions, the preferred analysis strategy is to adjust for the subgroups and base the comparisons on the restricted data set, i.e. only include the patients who have been directly randomized to one of the arms which are of interest for the comparisons. This is line with the recommendations by Law et al. [15].

It was of further interest to examine the influence of the randomization strategies to the three treatment arms. It has been shown that the type I error is the least inflated for R-III, i.e. assuming an equal ratio for the respective treatment arm(s) vs. control within each subgroup. Besides, it was shown that the power is lowest for R-I which is the randomization strategy where a ratio of 2:1 for the respective treatment arm vs. control in the subgroups X and Y and a ratio of 1:1:1 for treatment 1 vs. treatment 2 vs. control for subgroup Z was used. Especially when less or no patients are recruited from subgroup Z, R-I results in a power loss because the comparisons are based on less controls. Assuming an equal ratio within each subgroup, i.e. R-III, resulted in the highest power. Furthermore, it has been shown that for R-III the unadjusted comparisons based on the restricted data set lead to results which are close to the results one obtains from the adjusted analysis scenarios. In summary, it can be said that R-III, i.e. an equal ratio within each subgroup, is the favored randomization strategy.

Besides, different prevalence in the three subgroups X, Y and Z have been considered. It has been shown that the prevalence in subgroup Z has an effect on the power. For equal treatment effects in the subgroups, the power is higher when a large number of subjects is recruited from subgroup Z. For unequal treatment effects in the subgroups it has been shown that when a large number of subjects is recruited from subgroup Z the power is larger when the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y. When a low number of patients is recruited from subgroup Z the power is larger when the treatment effects in subgroup Z are half the treatment effects in subgroups X and Y. Additionally, it has been indicated that when the majority of patients are recruited from subgroup Z, there are only minimal differences observed among the randomization strategies. The reason for that is the following: for subgroup Z there is no difference between a traditional multi-arm trial and a trial with two substudies. Consequently, when the majority of patients are recruited from subgroup Z, the power is unaffected by the randomization strategies.

6.2 Limitations

Despite the various scenarios that were considered in this thesis, the results presented in this thesis are limited by the assumptions made about the considered trial design and by the examined scenarios which do not reflect the full complexity of a multi-arm trial run in practice. In particular, a predefined number of experimental treatment arms was considered in the trial. This work has been restricted to a multi-arm trial evaluating two treatment arms against one shared control arm with a continuous primary endpoint. It has been assumed that the control treatment does not change over time. Selective exclusion can also be applied in trials with more than two treatment arms, however it results in more complex and complicated scenarios [15]. Besides, it was assumed that all arms start and end at the same time. It was further assumed that the trial stops after the predetermined overall sample size. However, it was shown that when the prevalence in subgroups X and Y is high, this can lead to a reduced power since less control data of subgroup Z can be shared. To mitigate this issue, one potential approach is to establish a predetermined sample size based on realistic and/or worst-case assumptions regarding the prevalence in subgroups X and Y. An alternative approach would involve terminating the trial once a predetermined number of patients have been administered the treatment. However, this implies that the time points for analyzing treatment 1 and treatment 2 could potentially differ.

Throughout this thesis, complete or block randomization was used, however, other strategies can be considered in practice. Extending the work to cover more advanced randomization methods, such as response adaptive randomization, is left open for further research. Moreover, the simulation in this thesis considers a situation where the endpoints are normally distributed with known and equal variance. The assumption of only continuous endpoints is somewhat restrictive, with MAMS trials being potentially far broader than this. Some trials being carried out use different endpoints at each stage. For example, the STAMPEDE trial, which has been introduced before, evaluates different intermediate and definite outcome measures: progression-free survival was considered at earlier stages and overall survival at later stages [18, 68]. The simulation presented in this thesis could be extended to different endpoints for the different arms and/or groups.

In the simulation study at hand, it was decided to not adjust for multiplicity and only assess the power for the individual treatment vs control comparisons.

When and how to account for multiplicity in multi-arm trials is still under debate, see Section 2.4.5.

6.3 Future research

As study designs involving selective exclusion of treatment arms have been discussed in the literature but not often been put into practice, numerous questions regarding the statistical analysis of such designs remain unanswered. In the following open questions are outlined which are outside the scope of this thesis but are potential topics for future research.

Multi-arm trials with interim analyses and changes in the control arm
In practice, most of the multi-arm trials take advantage of the possibility to perform an interim analysis and stop the trial earlier for efficacy or futility or to modify the trial design based on interim data. If the sample sizes for the treatment arms are modified based on the results of the interim analyses, the effect estimates are no longer guaranteed to be unbiased. Further research is needed to assess under which conditions unbiasedness and type I error rate control can be guaranteed when considering a multi-arm trial which allows the selective exclusion of treatment arms. Besides, multi-arm trials in practice frequently include multiple control arms, and it is worth noting that cases where the control arm changes over the trial's duration have not been addressed. However, this situation is quite common, particularly when an existing treatment proves to be effective and subsequently becomes the new SOC. For a multi-arm trial which allows the selective exclusion of treatment arms, this necessitates careful consideration of the evolving control arm dynamics and their impact on the trial's design and analysis.

Extension to platform trials
In this thesis the focus was on a study design where all arms started and finished at the same time. As a further step it would be of interest to consider a platform trial that allows the selective exclusion of treatment arms. Then the number of experimental arms does not need to be fixed in advance and arms could be added during the trial. It would be of interest to investigate the effect of different patient populations when using non-concurrent controls. The use of non-concurrent controls can improve the efficiency of a trial by increasing the power and reducing the required sample size, however, bias may be introduced due to time trends [32, 38, 48]. Further research could assess under which conditions unbiasedness and

type I error rate control can be guaranteed. When extending the study at hand to a platform trial, Viele´s [81] suggested allocation method to obtain comparable treatment and control arms could be considered. For a platform trial which considers different eligibility profiles, the idea is to set weights for each arm which are constant in time across the eligibility groups. The allocation method has the limitation that it is restricted to concurrent controls and to patients eligible for both arms [81].

Covariates

Incorporating covariates into the trials is often an important consideration. Baseline covariates can help to identify subgroups of participants who may respond differently to the treatment. By exploring interactions between treatment and baseline covariates, one can assess treatment heterogeneity and explore differential treatment effects across these subgroups. That would extend the study design at hand since now the differences between the subgroups are due to the heterogeneous means in the control arm. Besides, by incorporating baseline covariates, the simulation study can mimic the real-world scenario more closely. This improves the external generalizability of the findings, as the simulation study can account for the heterogeneity in the study population that may be associated with the baseline covariates.

Weighting of the underrepresented subgroup

Molenberghs et al [14] gave the example of a multi-arm trial with two treatment arms and a shared control arm that allows the selective exclusion of treatment arms. As result, one has three different subgroups. The authors assume an unequal prevalence in the three subgroups and only considered concurrent controls for the pairwise comparisons. The unequal prevalence for the three subgroups results in an imbalance between the subgroups for the comparisons. That is because the subgroup, which included the patients who are eligible for all three arms, is underrepresented in the comparison since half of the patients did not receive the treatment of interest. Therefore, Molenbergh et al. [14] suggested to weight the underrepresented subgroup with a factor of two to restore the balance between the subgroups. However, they did not implement their suggested approach. The simulation study at hand could be extended to implement Molenbergh´s approach of weighing the underrepresented subgroup, i.e. subgroup Z.

Bibliography

[1] W. F. Rosenberger and J. M. Lachin. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons, 2002.

[2] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. "Innovation in the pharmaceutical industry: new estimates of R&D costs". *Journal of Health Economics*, vol. 47 (2016), pp. 20–33.

[3] V. Berger, L. J. Bour, K. Carter, J. J. Chipman, C. C. Everett, N. Heussen, C. Hewitt, R.-D. Hilgers, Y. A. Luo, J. Renteria, et al. "A roadmap to using randomization in clinical trials". *BMC Medical Research Methodology*, vol. 21, no. 1 (2021), pp. 1–24.

[4] S. Treweek, P. Lockhart, M. Pitkethly, J. A. Cook, M. Kjeldstrøm, M. Johansen, T. K. Taskila, F. M. Sullivan, S. Wilson, C. Jackson, et al. "Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis". *BMJ open*, vol. 3, no. 2 (2013).

[5] M. G. Kahn, C. A. Broverman, N. Wu, W. J. Farnsworth, and L. Manlapaz-Espiritu. "Improving protocol quality". *Applied Clinical Trials*, vol. 11 (2002), pp. 40–50.

[6] A. Hurtado-Chong, A. Joeris, D. Hess, and M. Blauth. "Improving site selection in clinical studies: a standardised, objective, multistep method and first experience results". *BMJ open*, vol. 7, no. 7 (2017).

[7] B. Freidlin, E. Korn, R. Gray, and A. Martin. "Multi-arm clinical trials of new agents: some design considerations". *Clinical Cancer Research*, vol. 14, no. 14 (2008), pp. 4368–4371.

[8] J. Wason and T. Jaki. "Optimal design of multi-arm multi-stage trials". *Statistics in Medicine*, vol. 31, no. 30 (2012), pp. 4269–4279.

[9] T. Jaki and J. Wason. "Multi-arm multi-stage trials can improve the efficiency of finding effective treatments for stroke: a case study". *BMC Cardiovascular Disorders*, vol. 18, no. 1 (2018), pp. 1–8.

[10] 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". *Clinical Therapeutics*, vol. 42, no. 7 (2020), pp. 1330–1360.

[11] J. M. Tetzlaff, A.-W. Chan, J. Kitchen, M. Sampson, A. C. Tricco, and D. Moher. "Guidelines for randomized clinical trial protocol content: a systematic review". *Systematic Reviews*, vol. 1, no. 1 (2012), pp. 1–11.

[12] M. Redman and C. Allegra. "The master protocol concept". *Seminars in Oncology*, vol. 42, no. 5 (2015), pp. 724–730.

[13] J. Woodcock and L. LaVange. "Master protocols to study multiple therapies, multiple diseases, or both". *New England Journal of Medicine*, vol. 377, no. 1 (2017), pp. 62–70.

[14] G. Molenberghs, M. Buyse, S. Abrams, N. Hens, P. Beutels, C. Faes, G. Verbeke, P. Van Damme, H. Goossens, T. Neyens, et al. "Infectious diseases epidemiology, quantitative methodology, and clinical research in the midst of the COVID-19 pandemic: Perspective from a European country". *Contemporary Clinical Trials*, vol. 99 (2020).

[15] M. Law and S. Emery. "Selective exclusion of treatment arms in multi-arm randomized clinical trials". *Statistics in Medicine*, vol. 22, no. 1 (2003), pp. 19–30.

[16] *Randomised Evaluation of COVID-19 Therapy (RECOVERY).* [Accessed: 2023-04-12]. URL: `https://clinicaltrials.gov/ct2/show/record/NCT04381936?view=record`.

[17] *RECOVERY Respiratory Support: Respiratory Strategies in patients with coronavirus COVID-19 – CPAP, high-flow nasal oxygen, and standard care.* [Accessed: 2023-04-15]. URL: `https://www.isrctn.com/ISRCTN16912075`.

[18] M. Sydes, M. K. Parmar, M. D. Mason, N. W. Clarke, C. Amos, J. Anderson, J. de Bono, D. P. Dearnaley, J. Dwyer, C. Green, et al. "Flexible trial design in practice-stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial". *Trials*, vol. 13 (2012), pp. 1–14.

[19] O. D. Flecha, D. W. D. de Oliveira, L. S. Marques, and P. F. Gonçalves. "A commentary on randomized clinical trials: How to produce them with a good level of evidence". *Perspectives in Clinical Research*, vol. 7, no. 2 (2016).

[20] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of Clinical Trials*. Springer, 2015.

[21] S. J. Pocock. *Clinical Trials: A Practical Approach*. John Wiley & Sons, 1983.

[22] T. D. Cook and D. L. DeMets. *Introduction to Statistical Methods for Clinical Trials*. CRC Press, 2007.

[23] L. L. Gluud. "Bias in clinical intervention research". *American Journal of Epidemiology*, vol. 163, no. 6 (2006), pp. 493–501.

[24] J. Dettori. "The random allocation process: two things you need to know". *Evidence-based Spine Care Journal*, vol. 1, no. 3 (2010), pp. 7–9.

[25] K. F. Schulz. "Randomised trials, human nature, and reporting guidelines". *The Lancet*, vol. 348, no. 9027 (1996), pp. 596–598.

[26] L. A. Moyé. *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. Springer, 2003.

[27] D. A. Berry. "Emerging innovations in clinical trial design". *Clinical Pharmacology & Therapeutics*, vol. 99, no. 1 (2016), pp. 82–91.

[28] D. A. Berry. "The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research". *Molecular Oncology*, vol. 9, no. 5 (2015), pp. 951–959.

[29] M. K. Parmar, F. M.-S. Barthel, M. Sydes, R. Langley, R. Kaplan, E. Eisenhauer, M. Brady, N. James, M. A. Bookman, A.-M. Swart, et al. "Speeding up the evaluation of new agents in cancer". *Journal of the National Cancer Institute*, vol. 100, no. 17 (2008), pp. 1204–1214.

[30] P. P. Phillips, S. H. Gillespie, M. Boeree, N. Heinrich, R. Aarnoutse, T. McHugh, M. Pletschette, C. Lienhardt, R. Hafner, C. Mgone, et al. "Innovative trial designs are practical solutions for improving the treatment of tuberculosis". *Journal of Infectious Diseases*, vol. 205 (2012), S250–S257.

[31] J. Wason, D. Magirr, M. Law, and T. Jaki. "Some recommendations for multi-arm multi-stage trials". *Statistical Methods in Medical Research*, vol. 25, no. 2 (2016), pp. 716–727.

[32] European Medicines Agency. *ICH E10 Choice of Control Group in Clinical Trials*. 2001.

[33] A. Avins. "Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials." *Journal of Medical Ethics*, vol. 24, no. 6 (1998), pp. 401–408.

[34] S. Halpern, J. H. Karlawish, D. Casarett, J. A. Berlin, R. R. Townsend, and D. A. Asch. "Hypertensive patients' willingness to participate in placebo-controlled trials: implications for recruitment efficiency". *American Heart Journal*, vol. 146, no. 6 (2003), pp. 985–992.

[35] S. J. Pocock. "Group sequential methods in the design and analysis of clinical trials". *Biometrika*, vol. 64, no. 2 (1977), pp. 191–199.

[36] A. Marušić and S. F. Ferenčić. "Adoption of the double dummy trial design to reduce observer bias in testing treatments". *Journal of the Royal Society of Medicine*, vol. 106, no. 5 (2013), pp. 196–198.

[37] S. Pushpakom, R. Kolamunnage-Dona, C. Taylor, T. Foster, C. Spowart, M. Garcıa-Fiñana, G. J. Kemp, T. Jaki, S. Khoo, P. Williamson, et al. "TAILoR (TelmisArtan and InsuLin Resistance in Human Immunodeficiency Virus [HIV]): an adaptive-design, dose-ranging phase IIb randomized trial of telmisartan for the reduction of insulin resistance in HIV-positive individuals on combination antiretroviral therapy". *Clinical Infectious Diseases*, vol. 70, no. 10 (2020), pp. 2062–2072.

[38] 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". *Journal of Clinical Epidemiology*, vol. 125 (2020), pp. 1–8.

[39] FDA. *Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics. Guidance for Industry.* 2022.

[40] R. Herbst, D. R. Gandara, F. R. Hirsch, M. W. Redman, M. LeBlanc, P. C. Mack, L. H. Schwartz, E. Vokes, S. S. Ramalingam, J. D. Bradley, et al. "Lung Master Protocol (Lung-MAP)—A Biomarker-Driven Protocol for Accelerating Development of Therapies for Squamous Cell Lung Cancer: SWOG S1400Lung-MAP: A Protocol for Accelerating Drug Development". *Clinical Cancer Research*, vol. 21, no. 7 (2015), pp. 1514–1524.

[41] D. Hyman, I. Puzanov, V. Subbiah, J. E. Faris, I. Chau, J.-Y. Blay, J. Wolf, N. S. Raje, E. L. Diamond, A. Hollebecque, et al. "Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations". *New England Journal of Medicine*, vol. 373, no. 8 (2015), pp. 726–736.

[42] B. Saville and S. Berry. "Efficiencies of platform clinical trials: a vision of the future". *Clinical Trials*, vol. 13, no. 3 (2016), pp. 358–366.

[43] F. Bretz, F. Koenig, W. Brannath, E. Glimm, and M. Posch. "Adaptive designs for confirmatory clinical trials". *Statistics in Medicine*, vol. 28, no. 8 (2009), pp. 1181–1217.

[44] M. B. Roig, P. Krotka, C.-F. Burman, E. Glimm, S. M. Gold, K. Hees, P. Jacko, F. Koenig, D. Magirr, P. Mesenbrink, et al. "On model-based time trend adjustments in platform trials with non-concurrent controls". *BMC Medical Research Methodology*, vol. 22, no. 1 (2022), pp. 1–16.

[45] O. Collignon, C. Gartner, A.-B. Haidich, R. James Hemmings, B. Hofner, F. Pétavy, M. Posch, K. Rantell, K. Roes, and A. Schiel. "Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials". *Clinical Pharmacology & Therapeutics*, vol. 107, no. 5 (2020), pp. 1059–1067.

[46] D. S. Robertson, J. M. Wason, F. König, M. Posch, and T. Jaki. "Online error rate control for platform trials". *Statistics in Medicine* (2023).

[47] D. Howard, J. Brown, S. Todd, and W. Gregory. "Recommendations on multiple testing adjustment in multi-arm trials with a shared control group". *Statistical Methods in Medical Research*, vol. 27, no. 5 (2018), pp. 1513–1530.

[48] European Medicines Agency. *ICH E9 Statistical Principles for Clinical Trials*. 1998.

[49] O. Collignon, C.-F. Burman, M. Posch, and A. Schiel. "Collaborative platform trials to fight COVID-19: methodological and regulatory considerations for a better societal outcome". *Clinical Pharmacology & Therapeutics*, vol. 110, no. 2 (2021), pp. 311–320.

[50] L. Dodd, B. Freidlin, and E. Korn. "Platform trials—beware the noncomparable control group". *New England Journal of Medicine*, vol. 384, no. 16 (2021), pp. 1572–1573.

[51] 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". *Trials*, vol. 22 (2021), pp. 1–10.

[52] K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. "Use of historical control data for assessing treatment effects in clinical trials". *Pharmaceutical Statistics*, vol. 13, no. 1 (2014), pp. 41–54.

[53] 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 (2020), pp. 1–12.

[54] F. Jiao, W. Tu, S. Jimenez, V. Crentsil, and Y.-F. Chen. "Utilizing shared internal control arms and historical information in small-sized platform clinical trials". *Journal of Biopharmaceutical Statistics*, vol. 29, no. 5 (2019), pp. 845–859.

[55] H. Schmidli, D. A. Häring, M. Thomas, A. Cassidy, S. Weber, and F. Bretz. "Beyond randomized clinical trials: use of external controls". *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4 (2020), pp. 806–816.

[56] A. Banbeta, J. Van Rosmalen, D. Dejardin, and E. Lesaffre. "Modified power prior with multiple historical trials for binary endpoints". *Statistics in Medicine*, vol. 38, no. 7 (2019), pp. 1147–1169.

[57] 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, no. 3 (2011), pp. 1047–1056.

[58] 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, no. 4 (2014), pp. 1023–1032.

[59] S. Weber, Y. Li, J. Seaman, T. Kakizume, and H. Schmidli. "Applying meta-analytic-predictive priors with the R Bayesian evidence synthesis tools". *arXiv preprint arXiv:1907.00603* (2019).

[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, no. 5 (2022), pp. 490–501.

[61] M. B. 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". *Trials*, vol. 24 (2023).

[62] B. M. Alexander, L. Trippa, S. Gaffey, I. C. Arrillaga-Romany, E. Q. Lee, M. L. Rinne, M. S. Ahluwalia, H. Colman, G. Fell, E. Galanis, et al. "Individualized screening trial of innovative glioblastoma therapy (INSIGhT): a Bayesian adaptive platform trial to develop precision medicines for patients with glioblastoma". *JCO Precision Oncology*, vol. 3 (2019), pp. 1–13.

[63] B. M. Alexander and T. F. Cloughesy. "Platform trials arrive on time for glioblastoma". *Neuro-Oncology*, vol. 20, no. 6 (2018), pp. 723–725.

[64] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and L. J. Esserman. "I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy". *Clinical Pharmacology & Therapeutics*, vol. 86, no. 1 (2009), pp. 97–100.

[65] L. J. Esserman and J. Woodcock. "Accelerating identification and regulatory approval of investigational cancer drugs". *Jama*, vol. 306, no. 23 (2011), pp. 2608–2609.

[66] S. M. Berry, J. T. Connor, and R. J. Lewis. "The platform trial: an efficient strategy for evaluating multiple treatments". *Jama*, vol. 313, no. 16 (2015), pp. 1619–1620.

[67] N. D. James, M. R. Sydes, N. W. Clarke, M. D. Mason, D. P. Dearnaley, J. Anderson, R. J. Popert, K. Sanders, R. C. Morgan, J. Stansfeld, et al. "Systemic therapy for advancing or metastatic prostate cancer (STAMPEDE): a multi-arm, multistage randomized controlled trial". *BJU international*, vol. 103, no. 4 (2009), pp. 464–469.

[68] M. Sydes, M. Parmar, N. James, N. Clarke, D. Dearnaley, M. Mason, R. Morgan, K. Sanders, and P. Royston. "Issues in applying multi-arm multistage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial". *Trials*, vol. 10, no. 1 (2009), pp. 1–16.

[69] C. Gilson, S. Chowdhury, M. K. Parmar, M. R. Sydes, S. Investigators, et al. "Incorporating biomarker stratification into STAMPEDE: an adaptive multi-arm, multi-stage trial platform". *Clinical Oncology*, vol. 29, no. 12 (2017), pp. 778–786.

[70] S. Hedden, R. Woolson, and R. Malcolm. "Randomization in substance abuse clinical trials". *Substance Abuse Treatment, Prevention, and Policy*, vol. 1, no. 1 (2006), pp. 1–17.

[71] J. Lachin. "Properties of simple randomization in clinical trials". *Controlled Clinical Trials*, vol. 9, no. 4 (1988), pp. 312–326.

[72] D. Blackwell and J. L. Hodges Jr. "Design for the control of selection bias". *The Annals of Mathematical Statistics*, vol. 28, no. 2 (1957), pp. 449–460.

[73] M. Zelen. "The randomization and stratification of patients to clinical trials". *J Chron Dis*, vol. 27 (1974), pp. 365–375.

[74] S. J. Pocock. "Allocation of patients to treatment in clinical trials". *Biometrics* (1979), pp. 183–197.

[75] D. Uschner, R.-D. Hilgers, and N. Heussen. "The impact of selection bias in randomized multi-arm parallel group clinical trials". *PLoS One*, vol. 13, no. 1 (2018).

[76] J. Fleiss. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, 1986.

[77] C. W. Dunnett. "A multiple comparison procedure for comparing several treatments with a control". *Journal of the American Statistical Association*, vol. 50, no. 272 (1955), pp. 1096–1121.

[78] G. Wassmer. "On sample size determination in multi-armed confirmatory adaptive designs". *Journal of Biopharmaceutical Statistics*, vol. 21, no. 4 (2011), pp. 802–817.

[79] J. J. Park, K. Thorlund, and E. J. Mills. "Critical concepts in adaptive clinical trials". *Clinical Epidemiology* (2018), pp. 343–351.

[80] K. Thorlund, J. Haggstrom, J. J. Park, and E. J. Mills. "Key design considerations for adaptive clinical trials: a primer for clinicians". *BMJ*, vol. 360 (2018).

[81] K. Viele. "Allocation in platform trials to maintain comparability across time and eligibility". *Statistics in Medicine* (2023).

[82] M. B. Roig, E. Glimm, T. Mielke, and M. Posch. "Optimal allocation strategies in platform trials". *arXiv preprint arXiv:2304.03035* (2023).

[83] A. Britton, M. McKee, N. Black, K. McPherson, C. Sanderson, and C. Bain. "Threats to applicability of randomised trials: exclusions and selective participation". *SageJournals*, vol. 4, no. 2 (1999), pp. 112–121.

[84] E. Juszczak, D. G. Altman, S. Hopewell, and K. Schulz. "Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement". *Jama*, vol. 321, no. 16 (2019), pp. 1610–1620.

[85] *Randomised Evaluation of COVID-19 Therapy (RECOVERY).* [Accessed: 2023-04-12]. URL: `https://www.recoverytrial.net/`.

[86] S. Paganoni, J. D. Berry, M. Quintana, E. Macklin, B. R. Saville, M. A. Detry, M. Chase, A. V. Sherman, H. Yu, K. Drake, et al. "Adaptive platform trials to transform amyotrophic lateral sclerosis therapy development". *Annals of Neurology*, vol. 91, no. 2 (2022), pp. 165–175.

[87] WHO Solidarity Trial Consortium. "Repurposed antiviral drugs for Covid-19-interim WHO solidarity trial results". *New England Journal of Medicine*, vol. 384, no. 6 (2021), pp. 497–511.

[88] J. Wason, L. Stecher, and A. Mander. "Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?" *Trials*, vol. 15 (2014), pp. 1–7.

[89] R. Parker and C. Weir. "Non-adjustment for multiple testing in multi-arm trials of distinct treatments: rationale and justification". *Clinical Trials*, vol. 17, no. 5 (2020), pp. 562–566.

[90] X. Bai, Q. Deng, and D. Liu. "Multiplicity issues for platform trials with a shared control arm". *Journal of Biopharmaceutical Statistics*, vol. 30, no. 6 (2020), pp. 1077–1090.

[91] Q. Nguyen, K. Hees, and B. Hofner. "Platform Trials: the Impact of common Controls on Type One Error and Power". *arXiv preprint arXiv:2302.04713* (2023).

[92] European Medicines Agency. *Guideline on multiplicity issues in clinical trials*. 2017.

[93] F. Bretz and F. Koenig. "Commentary on Parker and Weir". *Clinical Trials*, vol. 17, no. 5 (2020), pp. 567–569.

[94] Committee for Medicinal Products for Human Use and others. *Points to consider on multiplicity issues in clinical trials*. 2002.

[95] M. Aickin and H. Gensler. "Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods." *American Journal of Public Health*, vol. 86, no. 5 (1996), pp. 726–728.

[96] S.-Y. Chen, Z. Feng, and X. Yi. "A general introduction to adjustment for multiple comparisons". *Journal of Thoracic Disease*, vol. 9, no. 6 (2017).

[97] A. Dmitrienko and R. D'Agostino Sr. "Traditional multiplicity adjustment methods in clinical trials". *Statistics in Medicine*, vol. 32, no. 29 (2013), pp. 5172–5218.

[98] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1 (1995), pp. 289–300.

[99] S. Zehetmayer, M. Posch, and F. Koenig. "Online control of the False Discovery Rate in group-sequential platform trials". *Statistical Methods in Medical Research*, vol. 31, no. 12 (2022), pp. 2470–2485.

[100] K. Schulz and D. Grimes. "Multiplicity in randomised trials I: endpoints and treatments". *The Lancet*, vol. 365, no. 9470 (2005), pp. 1591–1595.

[101] N. Stallard, S. Todd, D. Parashar, P. K. Kimani, and L. A. Renfro. "On the need to adjust for multiplicity in confirmatory clinical trials with master protocols". *Annals of Oncology*, vol. 30, no. 4 (2019), pp. 506–509.

[102] A. Odutayo, D. Gryaznov, B. Copsey, P. Monk, B. Speich, C. Roberts, K. Vadher, P. Dutton, M. Briel, S. Hopewell, et al. "Design, analysis and reporting of multi-arm trials and strategies to address multiple testing". *International Journal of Epidemiology*, vol. 49, no. 3 (2020), pp. 968–978.

[103] R. A. Parker, C. J. Weir, T. M. Pham, I. R. White, N. Stallard, M. K. Parmar, R. J. Swingler, R. S. Dakin, S. Pal, and S. Chandran. "Statistical analysis plan for the motor neuron disease systematic multi-arm adaptive randomised trial (MND-SMART)". *Trials*, vol. 24, no. 1 (2023), pp. 1–15.

[104] S. F. Molloy, I. R. White, A. J. Nunn, R. Hayes, D. Wang, and T. S. Harrison. "Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed". *Contemporary Clinical Trials*, vol. 113 (2022).

[105] L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, and G. Tutz. *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag, 2009.

[106] G. Casella and R. L. Berger. *Statistical Inference*. Cengage Learning, 2002.

[107] B. R. Clarke. *Linear Models: The Theory and Application of Analysis of Variance*. John Wiley & Sons, 2008.

[108] S. M. van den Berg. "Analysing data using linear models." (2022).

[109] M. Crager. "Analysis of covariance in parallel-group clinical trials with pretreatment baselines". *Biometrics* (1987), pp. 895–901.

[110] S. S. Senn. *Statistical Issues in Drug Development*. John Wiley & Sons, 2008.

A. Appendix

In Section A.1 of this appendix, additional information for the setup of the conducted simulation study is presented. Next, in Section A.2 the simulation setup for the FWER is described in more detail. Then, in Section A.3, additional results of the simulation study are explained. Finally, Section A.4, includes the R code for the simulation study which has been presented in Chapters 4 and 5.

A.1 Simulation setup

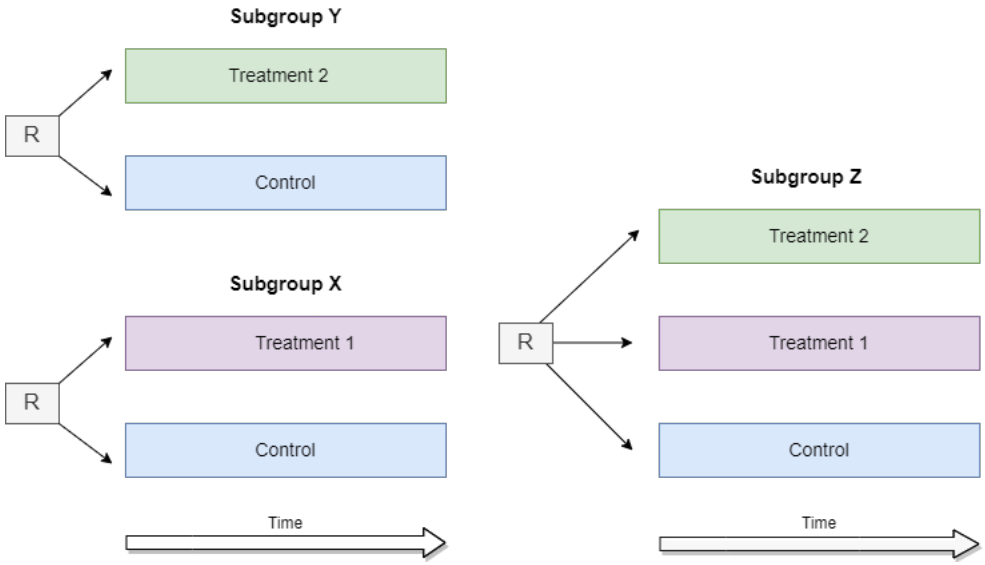

Figure A.1: Schematic illustration of the different subgroups: patients in subgroup X are eligible for treatment 1, patients in subgroup Y are eligible for treatment 2 and patients in subgroup Z are eligible for both treatments.

Figure A.1 illustrates the different subgroups that were considered for the simulation study. A patient in subgroup X is randomized to either treatment 1 or control while a patient in subgroup Y is randomized to either treatment 2 or control. Patients in subgroup Z can get randomized to all three arms: treatment 1, treatment 2 or control.

A.2 Considered design for simulating the family-wise error rate

The present subsection focuses on assessing the performance of the FWER in a simulation study. First, the design of the considered study is described. Then, the question whether multiplicity adjustment is required in the setup at hand is discussed by reviewing the current consensus in the literature. The results can be found in Section 2.4.5.

A multi-arm trial with k experimental treatment arms that enter the trial at trial start is considered (see Figure A.2). The recruitment of the control arm also starts at the beginning of the trial and runs in parallel to the treatment arms. Several treatment arms are compared to the shared control arm.

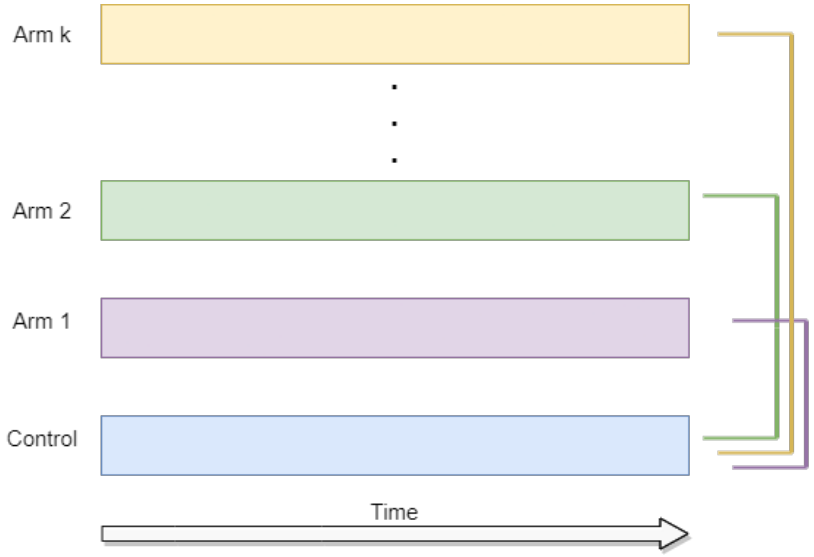

Figure A.2: Visualization of the study design considered for the simulation: a multi-arm trial is considered with a varying number of treatment arms k. The treatment arms all start at the same time and run in parallel to the control arm. Since the start of the trial patients are recruited to the control arm. The pairwise comparisons of each treatment arm vs. the shared control arm are of interest.

To generate the trial data, equal sample sizes in all arms (n_{total} = 250, 500, 750, 1000, 5000, 10000) were assumed. Patients are assigned to the arms following complete randomization with an equal randomization ratio of *1: 1: k + 1*. The continuous outcome Y_j for patient j is drawn from a normal distribution according to:

$$Y_j \sim N(\mu, \sigma^2) \text{ for } j = 1, ..., n_{total}$$

with

$$\mu = \Delta_k + \mu_C,$$

where μ_C is the response in the control arm and Δ_k the effect of treatment k. For the sake of simplicity it was assumed that the variances are equal to 1. The comparisons are tested under the global null hypothesis and a treatment effect of $\Delta_k = 0$ is assumed for the treatment arms. For the pairwise comparisons one-sided t-tests are calculated at a significance level $\alpha = 0.025$ and $\alpha = 0.05$ once the total number of patients is enrolled in the trial. Table A.1 summaries the considered simulation settings and parameters. 10000 replicates of each scenario were simulated to estimate the FWER.

Table A.1: Simulation setup overview family-wise error rate

Name	**Investigated values**	**Description**
Number of treatment arms k	k = 1 - 30	Number of treatment arms that independently get compared to the shared control arm.
Sample size n_{total}	n_{total} = 250, 500, 750, 1000, 5000, 10000	Total sample size after which the analysis is conducted.
Randomization method	Complete randomization	Patients are randomized by complete randomization to the different arms.
Allocation ratio	*1: 1: k + 1*	Allocation ratio by which the patients are randomized to one of the arms with k being the number of treatment arms.
Standard deviation σ	$\sigma = 1$	Standard deviation for the normally distributed random outcome variables.
Treatment effect Δ	$\Delta = 0$	Treatment effect for the treatment versus control comparisons.
Mean control arm μ_C	$\mu_C = 0$	Mean for the normally distributed outcome in the control arm.
Mean treatment arms μ_k	$\mu_k = 0$	Mean for the normally distributed outcome in the treatment arms.
Significance level α	$\alpha = 2.5\%$, $\alpha = 5\%$	Significance level for one-sided testing of the pairwise comparisons.

The question arises whether it is necessary to control for multiplicity or not due to the multiple comparisons in the considered simulation setup. In the considered trial design, the patients are randomized to different experimental treatment options and the hypotheses are being tested independently which does not require multiplicity correction [47, 101, 104]. Besides, as discussed above, the shared control arm itself does not necessitate FWER adjustment, however, the chance

of multiple simultaneous false decisions might increase due to the shared control arm [45, 47]. To conclude, it was decided to perform no adjustment for the level-α-test for the multiple hypotheses testing since the hypotheses are being tested independently.

A.3 Additional results

Additional results of the simulation study are presented for setting 3. Besides, Tables A.2 and A.3 summarize the means in the three arms in the different subgroups that were considered in the simulation study. The difference between Tables A.2 and A.3 is that the means in the control arm for subgroups Y and Z are swapped. The tables represent the scenarios where equal treatment effects of 0.5 were assumed. The means for the control arm are simulation parameters and the means in the treatment arms were calculated by adding the respective mean in the control arm to the respective treatment effect, e.g. $\mu_X^{T1} = \mu_X^C + \Delta^{T1}$.

Table A.2: Means in the different arms, T1, T2 or C, in the subgroups X, Y and Z for $\Delta^{T1} = \Delta^{T2} = 0.5$

μ_X^C	μ_Y^C	μ_Z^C	μ_X^{T1}	μ_Y^{T2}	$\mu_Z^{T1} = \mu_Z^{T2}$
0	0	0	0.5	0.5	0.5
-0.1	0.1	0	0.4	0.6	0.5
-0.2	0.2	0	0.3	0.7	0.5
-0.3	0.3	0	0.2	0.8	0.5
-0.4	0.4	0	0.1	0.9	0.5
-0.5	0.5	0	0	1	0.5
-0.6	0.6	0	-0.1	1.1	0.5
-0.7	0.7	0	-0.2	1.2	0.5
-0.8	0.8	0	-0.3	1.3	0.5
-0.9	0.9	0	-0.4	1.4	0.5
-1	1	0	-0.5	1.5	0.5

A.3.1 Setting 3: patients are in all three subgroups X, Y and Z

Figure A.3 shows the power for all adjusted analysis scenarios for an equal and unequal prevalence to the subgroups X, Y and Z. Table A.1 depicts the choice of means. One can see that there are hardly any differences between an equal and unequal prevalence, e.g. for *all data* and *all control data* the power is slightly

Table A.3: Means in the different arms, T1, T2 or C, in the subgroups X, Y and Z for $\Delta^{T1} = \Delta^{T2} = 0.5$

μ_X^C	μ_Y^C	μ_Z^C	μ_X^{T1}	μ_Y^{T2}	$\mu_Z^{T1} = \mu_Z^{T2}$
0	0	0	0.5	0.5	0.5
-0.1	0	0.1	0.4	0.5	0.6
-0.2	0	0.2	0.3	0.5	0.7
-0.3	0	0.3	0.2	0.5	0.8
-0.4	0	0.4	0.1	0.5	0.9
-0.5	0	0.5	0	0.5	1
-0.6	0	0.6	-0.1	0.5	1.1
-0.7	0	0.7	-0.2	0.5	1.2
-0.8	0	0.8	-0.3	0.5	1.3
-0.9	0	0.9	-0.4	0.5	1.4
-1	0	1	-0.5	0.5	1.5

higher for an equal prevalence. Note, *all control data* is masked by *restricted data*. Besides, when comparing Figures A.3 and 5.26 the means in the control are swapped. The figures show that the power for the adjusted analysis scenarios is not affected by the different means in the control arm as one adjusts for them.

Figure A.4 shows the power for all unadjusted analysis scenarios for an equal and unequal prevalence to the subgroups X, Y and Z. Table A.2 depicts the choice of means. One can see that the power increases over increasing mean for T2 and decreases over increasing mean for T1. The power loss/ increase for *all data* and *all control data* is due to the bias in the effect estimates which is introduced by pooling the controls of the different subgroups and not accounting for it. For the comparisons based on *restricted data* no bias is introduced since the comparisons are based on the correct data. However, the means in subgroups X and Y are further and further away from the means in subgroup Z. The bimodal distribution is not adjusted for and the variance increases. The figure further shows that there are hardly any visible differences between an equal and unequal prevalence to the subgroups, e.g. for *restricted data* the power is slightly higher for an unequal prevalence. That is due to the reason that for an unequal prevalence more patients are recruited from subgroup Z. Note, *all control data* mask *all data*. Besides, Figure A.4 illustrates the differences for the randomization strategies. One can see that the power is slightly lower for R-I. For R-III no bias for the comparisons based on *restricted data* is introduced, e.g. the power is the same for T1 and T2.

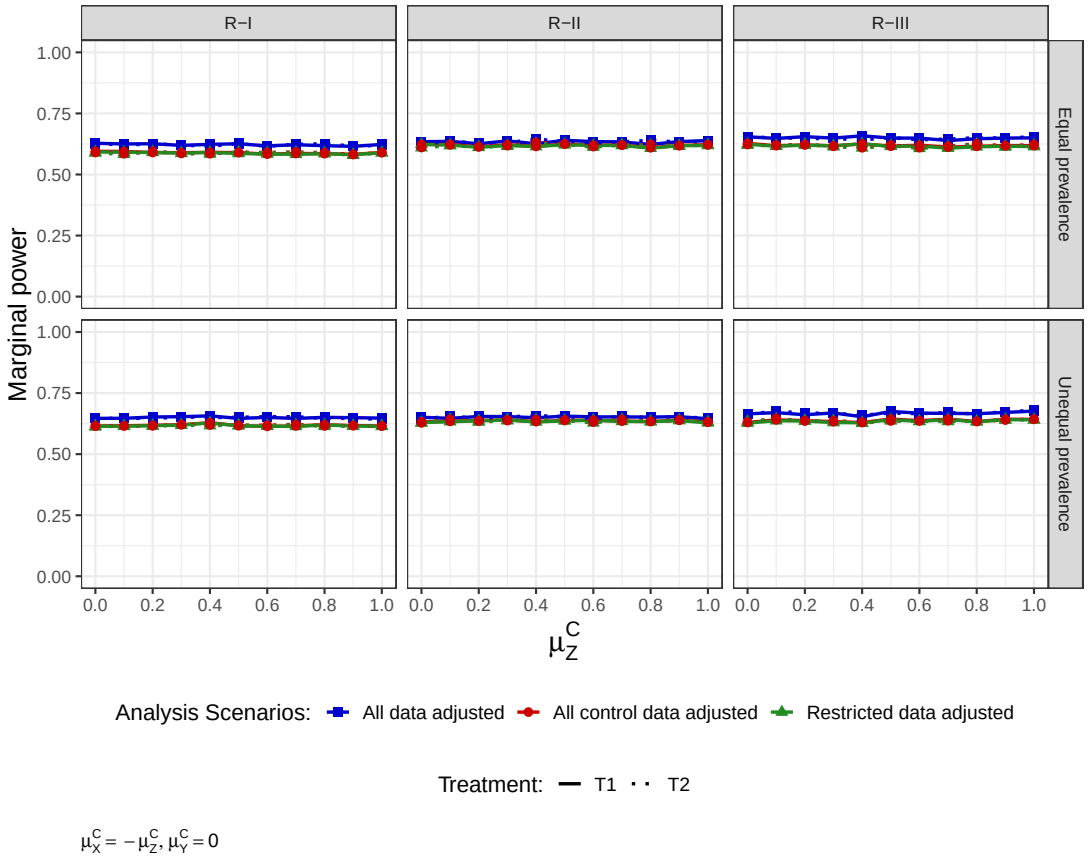

Figure A.3: Power over heterogeneous controls for the adjusted analysis scenarios. The upper row shows a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z while the lower row shows a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Z^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Z^C) and $\mu_Y^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

That is because only the correct data are used for the comparisons and an equal ratio within each subgroup is assumed.

Figure A.5 shows the power for all unadjusted analysis scenarios for an equal and unequal prevalence to the subgroups X, Y and Z. Table A.3 depicts the choice of means. The only difference between Figures A.4 and A.5 is that the means for the control arm in the subgroups Y and Z were swapped. For the adjusted analysis scenarios it was shown that the different means do not result in differences, however, one can see that the power for the unadjusted analysis scenarios is affected. One can see the biggest change for the comparisons based on *restricted data*. While the power for T2 for the *restricted data* was increasing before, it now also decreases.

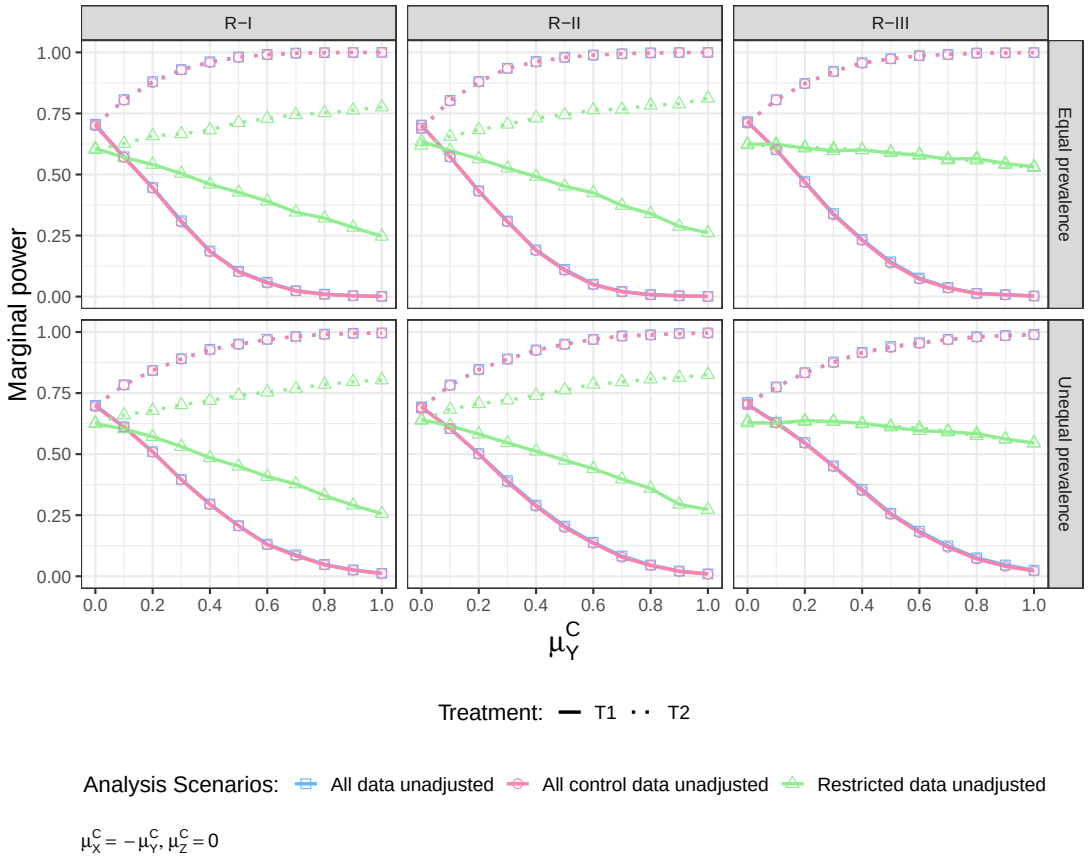

Figure A.4: Power over heterogeneous controls for the unadjusted analysis scenarios. The upper row shows a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z while the lower row shows a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

Even though it was shown that the power for the unadjusted analysis scenarios is affected by the means in the control arm, due to space limitations and for simplicity the following plots depict only one choice of means, namely μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$.

Figure A.6 shows the power for all unadjusted analysis scenarios for a high and low prevalence to the subgroups X, Y and Z. Table A.2 depicts the choice of means. The only simulation parameter that was varied between Figures A.4 and A.6 is the prevalence in the subgroups X, Y and Z. One can see that for the low prevalence the power is less increased/ decreased than for the other considered prevalence.

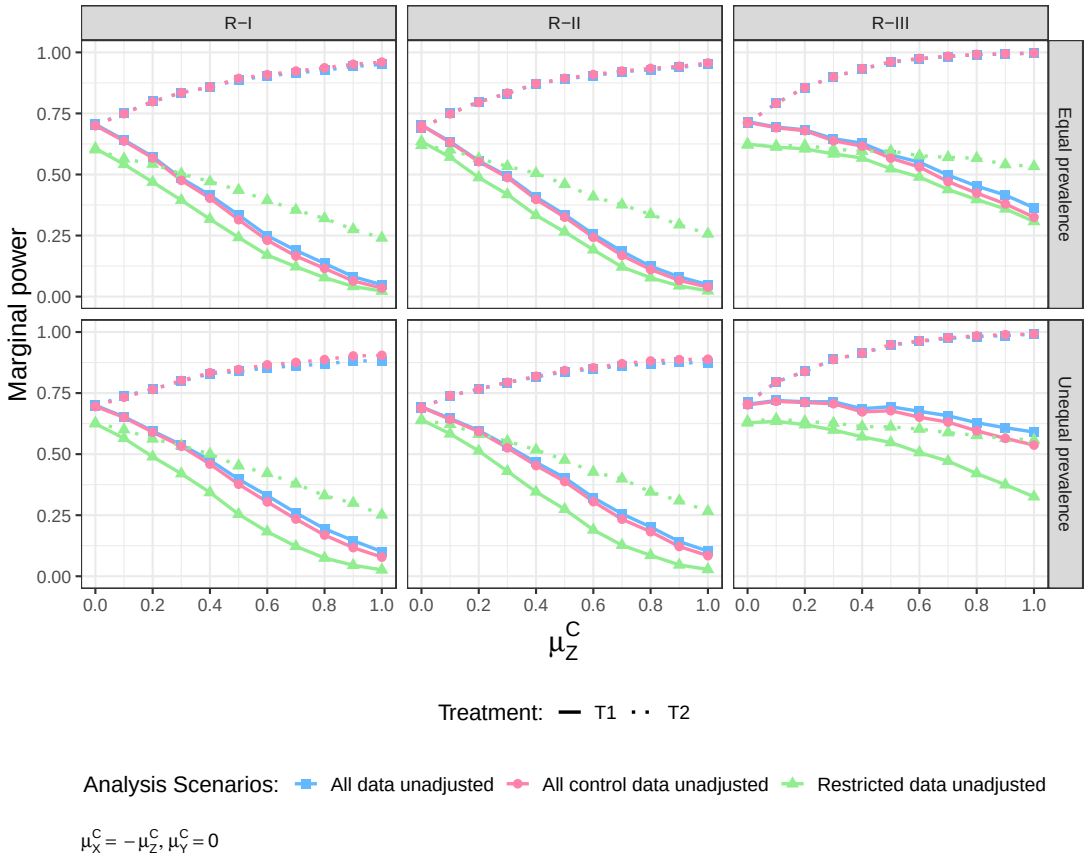

Figure A.5: Power over heterogeneous controls for the unadjusted analysis scenarios. The upper row shows a prevalence of $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z while the lower row shows a prevalence of $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Z^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Z^C) and $\mu_Y^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

The reason for that is that for a low prevalence the majority of patients is recruited from subgroup Z. As result, the bias in the effect estimates, which is introduced by pooling the control data of the subgroups, is smaller. In comparison, the bias for the high prevalence, where only a minority of patients is recruited from subgroup Z, is the largest. Regarding the randomization strategies, one can see that the power for the *restricted data* for R-III is the same for T1 and T2. Comparing it with the power for the adjusted analysis scenarios for the same simulation scenario, see Figure 5.27, one can see slight differences. While the power for the adjusted analysis scenarios is constant over the heterogeneous controls, the power for *restricted data* slightly decreases the further away the means in the subgroups are from each other. Besides, Figure A.6 shows that the bias which is introduced

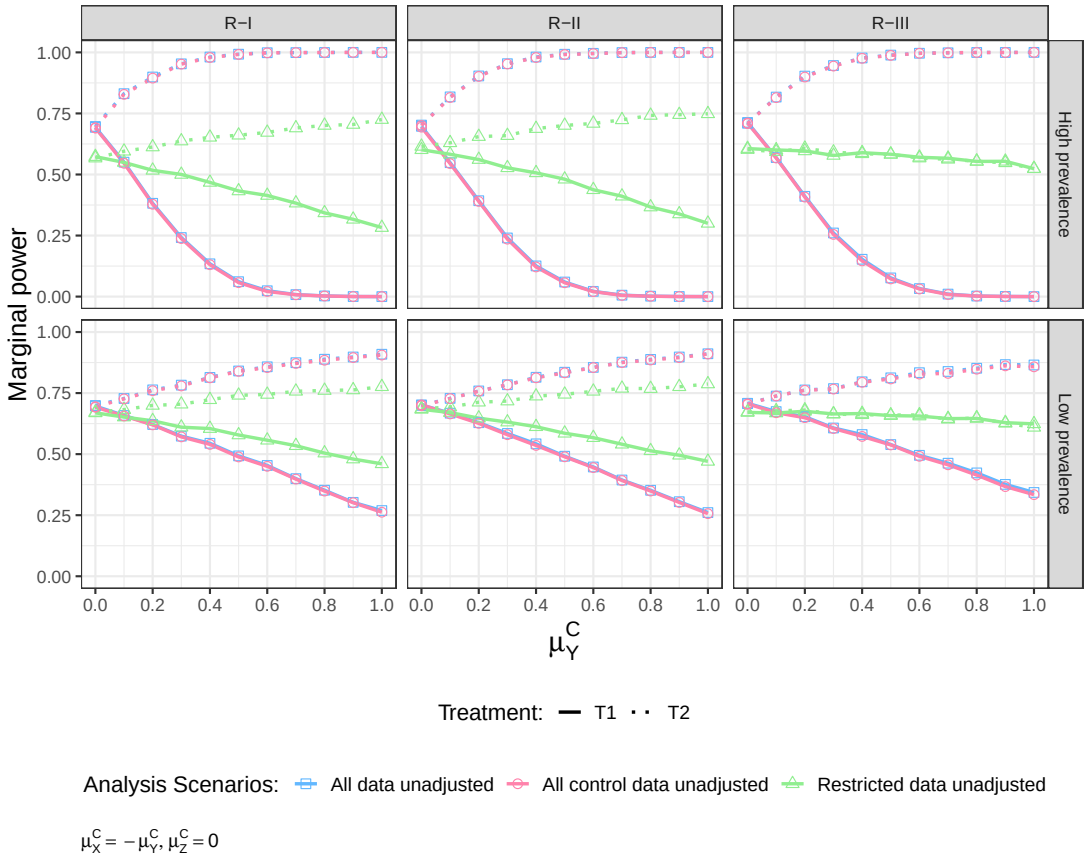

Figure A.6: Power over heterogeneous controls for the unadjusted analysis scenarios. The upper row shows a prevalence of $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X, Y and Z while the lower row shows a prevalence of $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X, Y and Z. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150 and $\Delta^{T1} = \Delta^{T2} = 0.5$.

for *all data* and *all control data* is smaller for R-III. Note, *all data* and *all control data* overlap each other.

Case 2: unequal treatment effects in the subgroups X, Y and Z

The next figures investigate the effect of unequal treatment effects in the subgroups X, Y and Z for different prevalence for the unadjusted analysis scenarios.

Figure A.7 depicts the power for the considered randomization strategies for an equal prevalence to the subgroups X, Y and Z for all unadjusted analysis scenarios. One can see that the power is smaller when the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y. Note, *all data* and *all control*

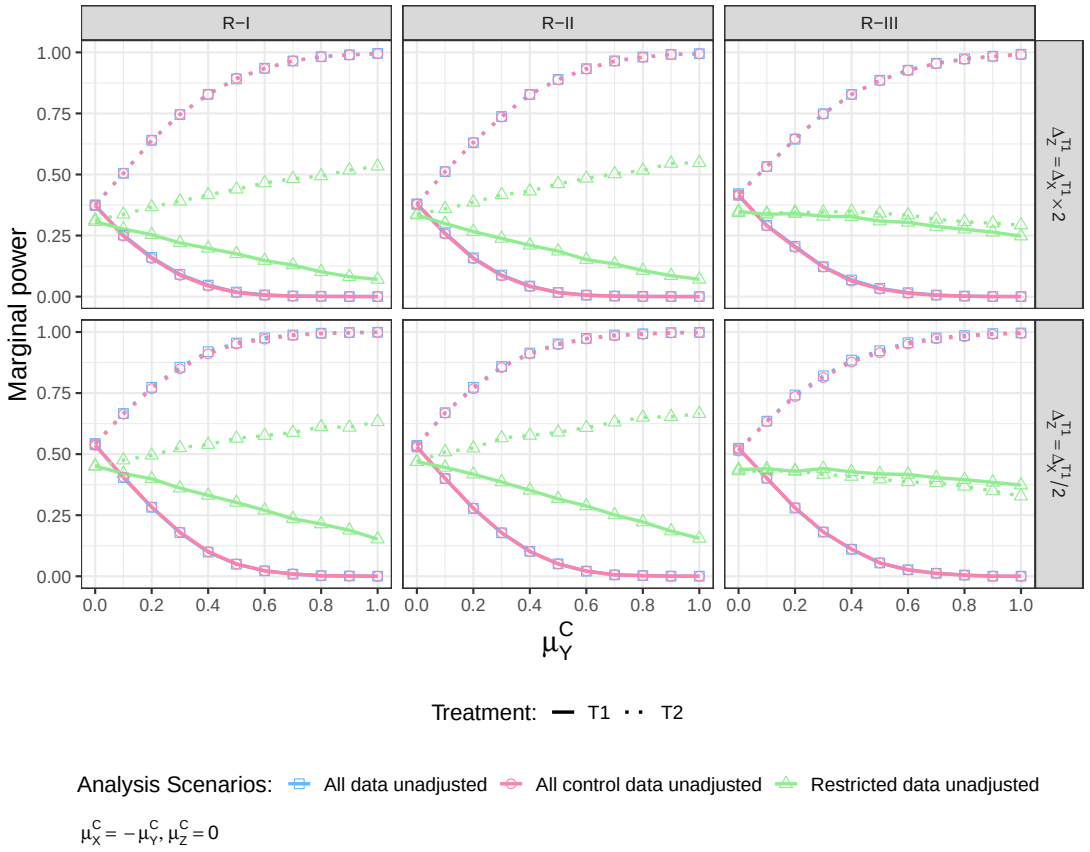

Figure A.7: Power for all unadjusted analysis scenarios and randomization strategies for different treatment effects in the subgroups. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150, and a prevalence $\frac{1}{3}$:$\frac{1}{3}$:$\frac{1}{3}$ to the subgroups X, Y and Z.

data overlap each other. Concerning the different randomization strategies one cannot see any differences for *all data* and *all control data*. However, for *restricted data* one can see that for R-III the power for T1 and T2 is almost the same. The power for the *restricted data* in R-III is close to that of the adjusted analyses, see Figure 5.29. One difference is, however, that the power for the unadjusted analysis of the *restricted data* is not constant over the different means but slightly decreases.

Figure A.8 shows the power for all considered randomization strategies for an unequal prevalence in the subgroups X, Y and Z. In comparison to Figure A.7 the only simulation parameter that was changed is the prevalence in the subgroups. For an unequal prevalence in the subgroups one can see less differences between the different treatment effects. Note, *all data* and *all control data* overlap each

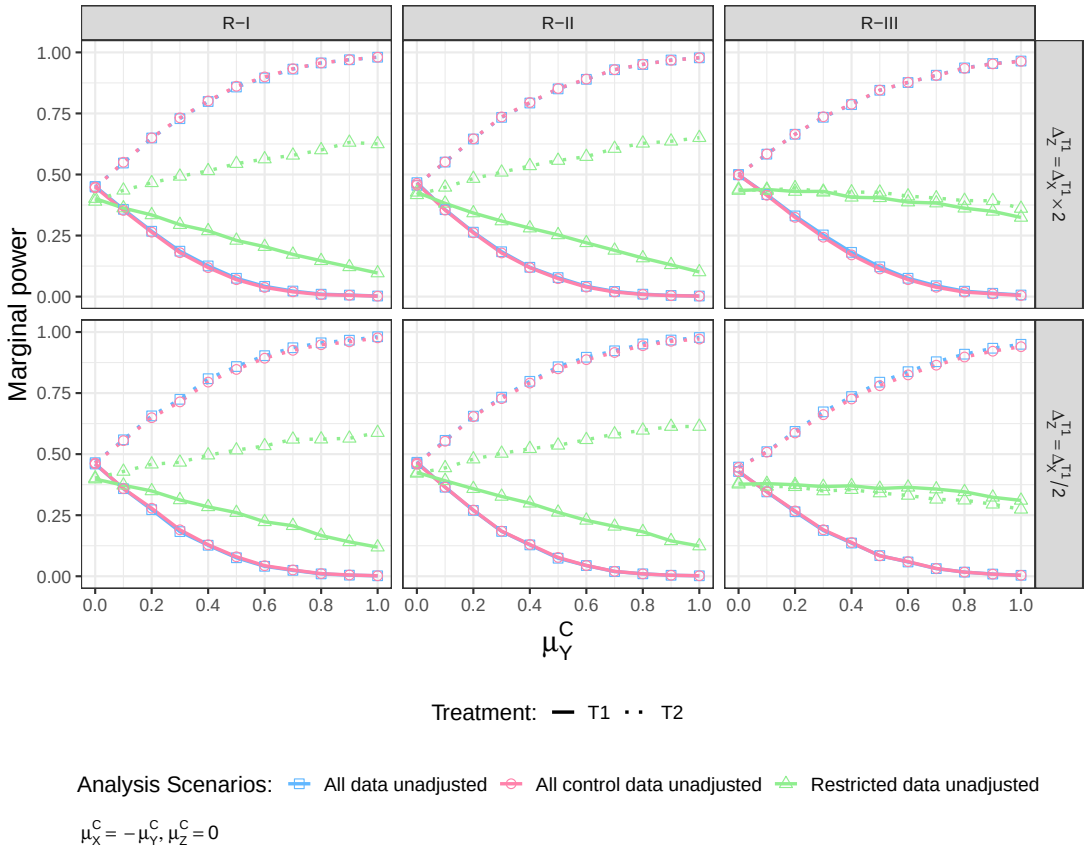

Figure A.8: Power for all unadjusted analysis scenarios and randomization strategies for different treatment effects in the subgroups. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150, and a prevalence $\frac{1}{4}$:$\frac{1}{4}$:$\frac{1}{2}$ to the subgroups X, Y and Z.

other. When comparing the different compositions of control data over the randomization strategies, one can see the biggest difference for the *restricted data* for R-III. For R-III the power for T1 and T2 is almost the same for the *restricted data*. One can see that in comparison to Figure A.4, where equal treatment effects in the subgroups were simulated, the unequal treatment effects resulted in a lower power.

Figure A.9 displays the power the considered randomization strategies for a high prevalence to the subgroups X, Y and Z. One can see that the power is higher when the treatment effects in subgroup Z are half the treatment effects in subgroups X and Y. That is because for a high prevalence the majority of patients is recruited from subgroups X and Y. When those two subgroups have higher treatments, the power is higher. Note, *all data* and *all control data* overlap each other. Regarding

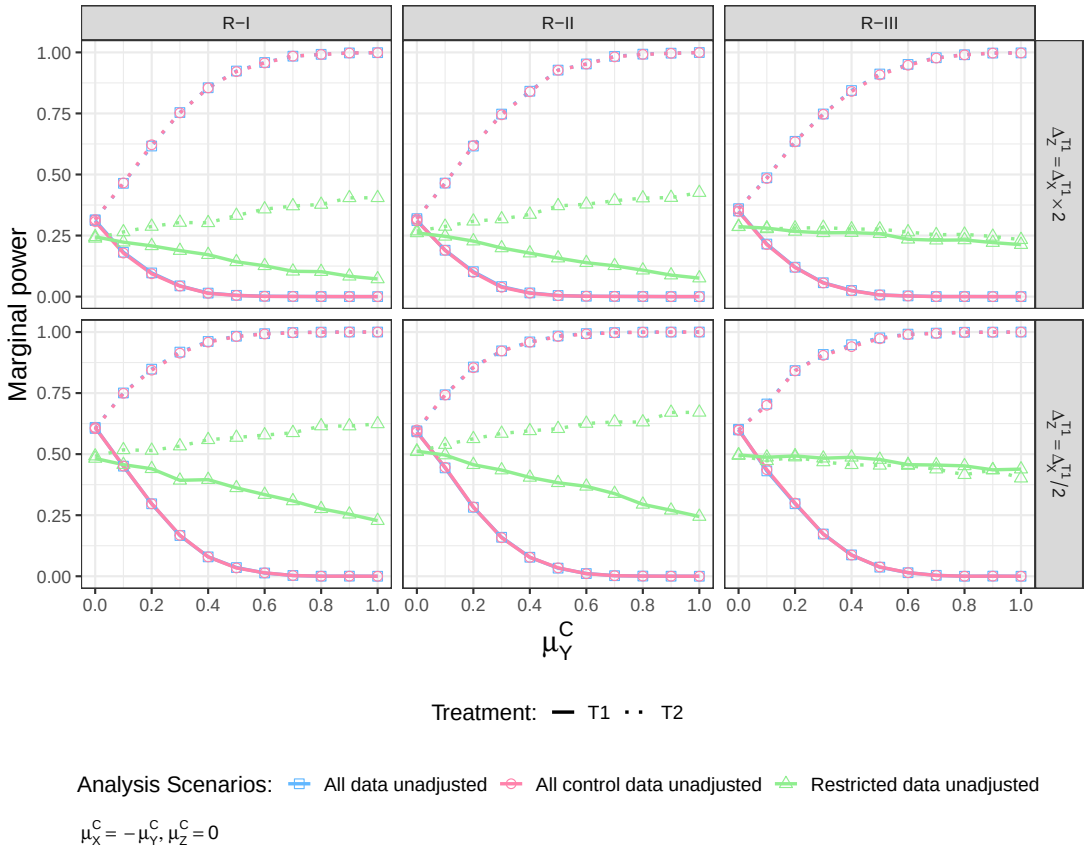

Figure A.9: Power for all unadjusted analysis scenarios and randomization strategies for different treatment effects in the subgroups. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150, and a prevalence $\frac{2}{5}$:$\frac{2}{5}$:$\frac{1}{5}$ to the subgroups X, Y and Z.

the randomization strategies one cannot see any differences for *all data* and *all control data*. For *restricted data* the power for T1 and T2 is almost the same for R-III. Compared to Figure A.6, where equal treatment effects in the subgroups were simulated, one can see that the power is lower for unequal treatment effects.

Figure A.10 shows the power for the randomization strategies for a low prevalence to the subgroups, i.e. the majority of patients is recruited from subgroup Z. One can see that the power is larger when the treatment effects in subgroup Z are twice the treatment effects in subgroups X and Y. The reason for this is that the majority of patients (80%) is recruited from subgroup Z. Besides, one can see that power is almost the same as in Figure A.6, where equal treatment effects were assumed. Note, *all data* and *all control data* overlap each other. One can see that

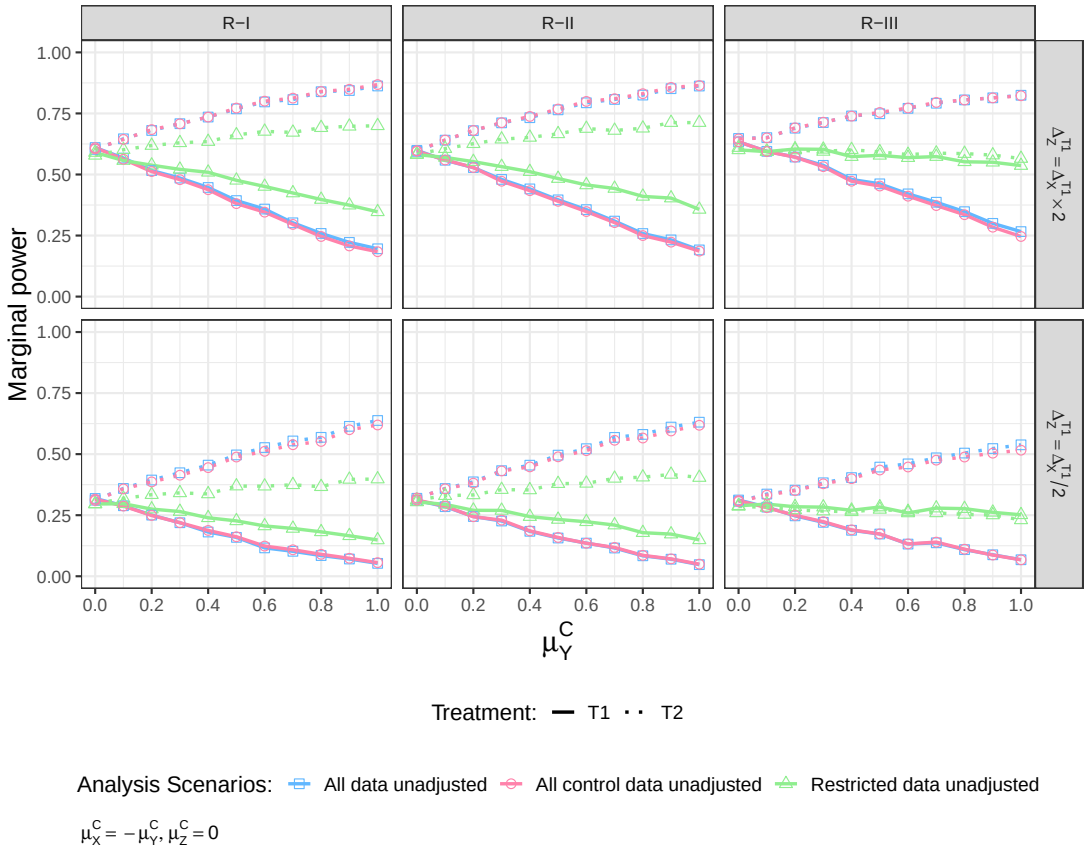

Figure A.10: Power for all unadjusted analysis scenarios and randomization strategies for different treatment effects in the subgroups. The following heterogeneous controls were considered: μ_Y^C with values from 0 to 1 (see x-axis), μ_X^C with values from -1 to 0 (μ_X^C = - μ_Y^C) and $\mu_Z^C = 0$. Fixed design parameters: block randomization, n = 150, and a prevalence $\frac{1}{10}$:$\frac{1}{10}$:$\frac{4}{5}$ to the subgroups X, Y and Z.

the bias in the effect estimates which is introduced for *all data* and *all control data* by pooling the control data is the smallest for R-III. Furthermore, the power for T1 and T2 for the *restricted data* for R-III is almost the same.

To conclude, for the unadjusted analysis scenarios the preferred randomization strategy is R-III, i.e. an equal ratio within each subgroup. The recommended composition of control data is the *restricted data*, i.e. to only base the comparisons on the patients who had the chance of getting randomized to one of the arm of interest. Besides, the unequal treatment effects and different prevalence in the subgroups affect the power for the unadjusted analysis scenarios.

A.4 Code

The code for the simulation study described in Chapter 4 is structured as follows: the first function includes the assignment to the three different subgroups X, Y and Z, the randomization to the three arms T1, T2 and C, the splitting of the data sets into *all data, all control data and restricted* and the adjusted and unadjusted analyses. In the second function, the estimates and p-values over the iterations are collected and the operating characteristics such as power, bias and confidence intervals are calculated. Besides, the iterations are parallelized to accelerate the simulating process. Finally, in the third function a grid is build which consists of all simulation parameters of interest. Then, for each row of the grid one iterates over the first two functions. In the end, the data are saved as csv file and one row in the csv file corresponds to one row in the grid, i.e. one simulation scenario.

A.4.1 Function for generating the data set

```
1
2  fnSimulation <-
3    function(
4      t, # number of time periods
5      n, # number of patients for each time period
6      armsNumb_vec = 1:3, # number of arms
7      allocProb_vec, # allocation probability to the three groups X, Y and Z
8      deltaT1, # treatment effect for treatment 1
9      deltaT2, # treatment effect for treatment 2
10     meanCX, # mean for control in group X
11     meanCY, # mean for control in group Y
12     meanCZ, # mean for control in group Z
13     sdT1X = 1, # standard deviation for treatment 1 in group X
14     sdT2Y = 1, # standard deviation for treatment 2 in group Y
15     sdT1Z = 1, # standard deviation for treatment 1 in group Z
16     sdT2Z = 1, # standard deviation for treatment 2 in group Z
17     sdCX = 1, # standard deviation for control in group X
18     sdCY = 1, # standard deviation for control in group Y
19     sdCZ = 1, # standard deviation for control in group
20     blocksizeXY, # block size for block randomization for groups X and Y
21     blocksizeZ, # block size for block randomization for group Z
22     alpha = 0.025, # alpha level for t-test
23     allocProbRandXY_vec, # allocation probability for groups X and Y to treatment or control
24     allocProbRandZ_vec, # allocation probability for group Z to different arms
25     complete, # boolean for randomization method
26     random, # boolean for deterministic or random allocation to groups
27     full # boolean for returning full output including input parameters or just the output
28   ) {
29
30     # calculate the overall sample size
31     sampleSize <- n * t
32
33     # create an ID vector
34     ID_vec <- c(1:sampleSize)
35
36     # set counting variable for the different groups to 0
37     nGroupX <- nGroupY <- nGroupZ <- nGroup <- 0
38
39     # create an empty data frame and vectors for future values
40     time <- arm_char <- metOutcome1_vec <- tmp_vec <- vector()
41
42     data <- data.frame(time,
43                        arm_char,
44                        tmp_vec,
45                        metOutcome1_vec)
```

```
46
47      # calculate the means for treatment 1 and treatment 2
48      # in the different subgroups X, Y and Z
49      meanT1X <- meanCX + deltaT1
50      meanT2Y <- meanCY + deltaT2
51      meanT1Z <- meanCZ + deltaT1
52      meanT2Z <- meanCZ + deltaT2
53
54
55      # assign the patients randomly to the three groups X, Y and Z
56      if(random == TRUE) {
57
58        # determine the list of allocation for all patients
59        # when getting allocated randomly
60        listAlloc_vec <- sample(x = armsNumb_vec,
61                                size = sampleSize,
62                                replace = TRUE,
63                                prob = allocProb_vec)
64
65      # assign the patients in a fixed way to the three groups X, Y and Z
66      } else {
67
68        # calculate how many patients should get allocated to each group
69        # for a fixed allocation and multiply by sample size because
70        # times in the next step only takes integers
71        allocProb_vec1 <- allocProb_vec * sampleSize
72
73        # determine the list of allocation for all patients
74        listAlloc_vec <- sample(x = c(rep(armsNumb_vec,
75                                          times = ceiling(allocProb_vec1))),
76                                size = sampleSize,
77                                replace = FALSE)
78      }
79
80      # use complete randomization to randomize the patient to one of the three arms T1, T2, C
81      if(complete == TRUE){
82
83        # the randomization list depends on the block size which is different for the different
84        # ratios: if the block size is smaller or equal to the number of treatments, there is
85        # no replacement within each block
86
87        if(length(allocProbRandXY_vec) == 2) {
88          # determine the list of complete randomization for patients in group X
89          # list has length of n*t because the most extreme case is that
90          # all patients are in one group
91          armX_char <- sample(x = c("T1", "C"),
92                              size = sampleSize,
93                              replace = TRUE,
94                              prob = allocProbRandXY_vec)
95
96          # determine the list of randomization for patients in group Y
97          # list has length of n*t because the most extreme case is that
98          # all patients are in one group
99          armY_char <- sample(x = c("T2", "C"),
100                              size = sampleSize,
101                              replace = TRUE,
102                              prob = allocProbRandXY_vec)
103       } else {
104
105         # determine the list of complete randomization for patients in group X
106         # list has length of n*t because the most extreme case is that
107         # all patients are in one group
108         armX_char <- sample(x = c("T1", "T1", "C"),
109                             size = sampleSize,
110                             replace = TRUE,
111                             prob = allocProbRandXY_vec)
112
113         # determine the list of randomization for patients in group Y
114         # list has length of n*t because the most extreme case is that
115         # all patients are in one group
116         armY_char <- sample(x = c("T2", "T2", "C"),
117                             size = sampleSize,
118                             replace = TRUE,
119                             prob = allocProbRandXY_vec)
120       }
121
```

```
122       if(length(allocProbRandZ_vec) == 3) {
123         # determine the list of randomization for patients in group Z
124         # list has length of n*t because the most extreme case is that
125         # all patients are in one group
126         armZ_char <- sample(x = c("T1", "T2", "C"),
127                             size = sampleSize,
128                             replace = TRUE,
129                             prob = allocProbRandZ_vec)
130       } else {
131         # determine the list of randomization for patients in group Z
132         # list has length of n*t because the most extreme case is that
133         # all patients are in one group
134         armZ_char <- sample(x = c("T1", "T2", "C", "C"),
135                             size = sampleSize,
136                             replace = TRUE,
137                             prob = allocProbRandZ_vec)
138       }
139
140       # use block randomization to randomize the patients to one of the three arms T1, T2, C
141     } else {
142
143       # determine the length of the randomization list
144       nBlockXY_vec <- rep(1:ceiling(sampleSize / blocksizeXY),
145                           each = blocksizeXY)
146
147       # if the block size is smaller or equal to the number of
148       # treatments, there is no replacement within each block
149       if(blocksizeXY <= 2) {
150
151         # calculate the first block for group X: each treatment occurs once and the
152         # probability to get allocated to one of the treatments is the same
153         randListX_vec <- sample(x = c("T1", "C"),
154                                 size = blocksizeXY,
155                                 replace = FALSE,
156                                 prob = allocProbRandXY_vec)
157
158         # determine the randomization list by making sure that
159         # the treatments are allocated equally
160         while(length(randListX_vec) <= length(nBlockXY_vec)) {
161
162           randListX_vec <- c(randListX_vec, sample(x = c("T1", "C"),
163                                                   size = blocksizeXY,
164                                                   replace = FALSE,
165                                                   prob = allocProbRandXY_vec))
166         }
167
168         # calculate the first block for group Y: each treatment occurs once and the
169         # probability to get allocated to one of the treatments is the same
170         randListY_vec <- sample(x = c("T2", "C"),
171                                 size = blocksizeXY,
172                                 replace = FALSE,
173                                 prob = allocProbRandXY_vec)
174
175         # determine the randomization list by making sure that
176         # the treatments are allocated equally
177         while(length(randListY_vec) <= length(nBlockXY_vec)) {
178
179           randListY_vec <- c(randListY_vec, sample(x = c("T2", "C"),
180                                                   size = blocksizeXY,
181                                                   replace = FALSE,
182                                                   prob = allocProbRandXY_vec))
183         }
184       } else {
185         # calculate the first block for group X: each treatment can occur more than once and
186         # the probability to get allocated  to one of the treatments is not the same
187         randListX_vec <- sample(x = c(rep("T1", blocksizeXY/blocksizeXY * 2),
188                                       "C"),
189                                 size = blocksizeXY,
190                                 replace = FALSE,
191                                 prob = allocProbRandXY_vec)
192
193         # determine the randomization list
194         while(length(randListX_vec) <= length(nBlockXY_vec)) {
195
196           randListX_vec <- c(randListX_vec, sample(x = c(rep("T1", blocksizeXY/blocksizeXY * 2), "C"),
197                                                   size = blocksizeXY,
```

```
                                            replace = FALSE,
                                            prob = allocProbRandXY_vec))
      }

      # calculate the first block for group Y: the treatment can occur more than once and the
      # probability to get allocated to one of the treatments is the same
      randListY_vec <- sample(x = c(rep("T2", blocksizeXY/blocksizeXY * 2),
                                    "C"),
                              size = blocksizeXY,
                              replace = FALSE,
                              prob = allocProbRandXY_vec)

      # determine the randomization list
      while(length(randListY_vec) <= length(nBlockXY_vec)) {

        randListY_vec <- c(randListY_vec, sample(x = c(rep("T2", blocksizeXY/blocksizeXY * 2),
                                                       "C"),
                                            size = blocksizeXY,
                                            replace = FALSE,
                                            prob = allocProbRandXY_vec))
      }
    }

    # determine the length of the randomization list
    nBlockZ_vec <- rep(1:ceiling(sampleSize/blocksizeZ),
                       each = blocksizeZ)

    # if the block size is smaller or equal to the number of treatments,
    # there is no replacement within each block
    if(blocksizeZ <= 3) {
      # calculate the first block for group Z: each treatment occurs once and the
      # probability to get allocated to one of the treatments is the same
      randListZ_vec <- sample(x = c("T1", "T2", "C"),
                              size = blocksizeZ,
                              replace = FALSE,
                              prob = allocProbRandZ_vec)

      # determine the randomization list by making sure that the treatments are allocated equally
      while(length(randListZ_vec) <= length(nBlockZ_vec)) {

        randListZ_vec <- c(randListZ_vec, sample(x = c("T1", "T2", "C"),
                                            size = blocksizeZ,
                                            replace = FALSE,
                                            prob = allocProbRandZ_vec))
      }
    } else{
      # calculate the first block for group Z: each treatment can
      # occur more than once and the probability to get allocated
      # to one of the treatments is not the same
      randListZ_vec <- sample(x = c("T1", "T2", "C", "C"),
                              size = blocksizeZ,
                              replace = FALSE,
                              prob = allocProbRandZ_vec)

      # determine the randomization list
      while(length(randListZ_vec) <= length(nBlockZ_vec)) {

        randListZ_vec <- c(randListZ_vec, sample(x = c("T1", "T2", "C", "C"),
                                            size = blocksizeZ,
                                            replace = FALSE,
                                            prob = allocProbRandZ_vec))
      }
    }
  }

  # repeat the simulation for a given number of time periods
  for(time in 1:t) {
    # repeat the simulation for a given number of patients
    for (i in 1:n) {

        # when considering different groups of patients
        # i.e. patients are not willing to get randomized to all three
        # arms, the prevalence can either be determined at random or fixed

          # 1: group X = treatment 1, but not treatment 2
          # 2: group Y = treatment 2, but not treatment 1
```

```
274             # 3: group Z = treatment 1 and 2
275
276           # determine the prevalence at random
277           if(random == TRUE) {
278
279             # counting variable
280             nGroup <- nGroup + 1
281
282             # iterate over all patients and assign the corresponding
283             # type of group from the allocation list
284             typeofgroup <- listAlloc_vec[nGroup]
285
286             # determine the prevalence fixed
287           } else {
288
289             # counting variable
290             nGroup <- nGroup + 1
291
292             # iterate over all patients and assign the corresponding
293             # type of group from the allocation list
294             typeofgroup <- listAlloc_vec[nGroup]
295           }
296
297           # determine the allocation to treatment or control arm depending on the type of group
298           # by complete or  block randomization
299
300           if(complete == TRUE) {
301
302             # patients in group 1 are either randomized to treatment 1 or control
303             if(typeofgroup == 1) {
304
305               # counting variable within group 1
306               nGroupX <- nGroupX + 1
307
308               # iterate over the number of patients in the group and assign the
309               # corresponding treatment from the randomization list
310               arm_char <- armX_char[nGroupX]
311
312               # patients in group 2 are either randomized to treatment 2 or control
313             } else if(typeofgroup == 2) {
314
315               # counting variable within group 2
316               nGroupY <- nGroupY + 1
317
318               # iterate over the number of patients in the group and assign the
319               # corresponding treatment from the randomization list
320               arm_char <- armY_char[nGroupY]
321
322               # patients in group 3 are either randomized to treatment 1, 2 or control
323             } else {
324
325               # counting variable in group 3
326               nGroupZ <- nGroupZ + 1
327
328               # iterate over the number of patients in the group and assign the
329               # corresponding treatment from the randomization list
330               arm_char <- armZ_char[nGroupZ]
331             }
332
333             # use block randomization
334           } else {
335
336             # for group X: patient is randomized to either treatment 1 or control
337             if(typeofgroup == 1) {
338
339               # count the patients in group 1
340               nGroupX <- nGroupX + 1
341
342               # iterate over the number of patients in the group and assign the
343               # corresponding treatment from the randomization list
344               arm_char <- randListX_vec[nGroupX]
345
346               # for group Y: patient is randomized to either treatment 2 or control
347             } else if(typeofgroup == 2) {
348
349               # count the patients in group 2
```

```
350             nGroupY <- nGroupY + 1
351
352             # iterate over the number of patients in the group and assign the
353             # corresponding treatment from the randomization list
354             arm_char <- randListY_vec[nGroupY]
355
356             # for group Z: patient is randomized to either treatment 1, treatment 2 or control
357           } else {
358
359             # count the patients in group 3
360             nGroupZ <- nGroupZ + 1
361
362             # iterate over the number of patients in the group and assign the
363             # corresponding treatment from the
364             randomization list
365             arm_char <- randListZ_vec[nGroupZ]
366           }
367         }
368
369       # generate normally distributed outcome variables
370       # depending on the type of group and treatment arm
371       if(arm_char == "T1" && typeofgroup == 1) {
372         tmp_vec <- rnorm(1,
373                          mean = meanT1X,
374                          sd = sdT1X)
375       } else if (arm_char == "T2" && typeofgroup == 2) {
376         tmp_vec <- rnorm(1,
377                          mean = meanT2Y,
378                          sd = sdT2Y)
379       } else if(arm_char == "T1" && typeofgroup == 3) {
380         tmp_vec <- rnorm(1,
381                          mean = meanT1Z,
382                          sd = sdT1Z)
383       } else if(arm_char == "T2" && typeofgroup == 3) {
384         tmp_vec <- rnorm(1,
385                          mean = meanT2Z,
386                          sd = sdT2Z)
387       } else if(arm_char == "C" && typeofgroup == 1) {
388         tmp_vec <- rnorm(1,
389                          mean = meanCX,
390                          sd = sdCX)
391       } else if(arm_char == "C" && typeofgroup == 2) {
392         tmp_vec <- rnorm(1,
393                          mean = meanCY,
394                          sd = sdCY)
395       } else {
396         tmp_vec <- rnorm(1,
397                          mean = meanCZ,
398                          sd = sdCZ)
399       }
400
401       # return a data frame with four variables: time when included,
402       # the type of group, the treatment arm and the outcome
403       data <- rbind(data,
404                     cbind(time,
405                           typeofgroup,
406                           arm_char,
407                           tmp_vec))
408
409       # save how many patients are in each arm
410       patientsPerArm <- data$arm_char
411
412     }
413   }
414   # add the ID variable to the data frame and order the rows such that
415   # the ID variable is in the first row
416   data$ID <- ID_vec
417   data <- data[, c(5, 1:4)]
418
419   # rename the rows in the data frame
420   colnames(data) <- c("ID",
421                       "Time",
422                       "Typeofgroups",
423                       "TreatmentArm",
424                       "Outcome")
425
```

```
##### Analysis #####
# depending on whether one adjusts for the different type of groups or not and the
# composition of the control data, the analysis  is different

# Analysis scenario 1
# fit one-way model for all data unadjusted for type of control
fit1 <- lm(Outcome ~ TreatmentArm,
           data = data)

# use lsmeans to compare the groups
fit1means <- lsmeans(fit1,
                     "TreatmentArm")

# calculate treatment vs control comparison
contrastsTC <- contrast(object = fit1means,
                        method = "trt.vs.ctrl",
                        adjust = "none")
# calculate the contrasts one-sided
contrastsTC <- test(contrastsTC,
                    side = ">")
# transform emmsgrid into data frame to extract values
contrastsTC <- data.frame(contrastsTC)

# extract the estimates
estimateCT1All <- contrastsTC[[2]][1]
estimateCT2All <- contrastsTC[[2]][2]
# extract the p-values
pValCT1All <- contrastsTC[[6]][1]
pValCT2All <- contrastsTC[[6]][2]

# calculate confidence interval for treatment vs control contrast
contrastsTCCI <- confint(contrast(object = fit1means,
                                  method = "trt.vs.ctrl",
                                  adjust = "none"))
# extract upper and lower bounds
CICT1All <- c(contrastsTCCI[[5]][1],
              contrastsTCCI[[6]][1])
CICT2All <- c(contrastsTCCI[[5]][2],
              contrastsTCCI[[6]][2])



# Analysis Scenario 2
# edit the data frame and delete all patients in group 2 (Y) for the comparison of T1-C with only
# the patients who had the chance of getting treated with the treatment of interest
# --> restricted data for the unadjusted comparison of treatment 1 vs control
dataXZ <- data %>%
  filter(!Typeofgroups %in% '2' & !TreatmentArm %in% "T2")

# fit one-way model
fit2 <- lm(Outcome ~ TreatmentArm,
           data = dataXZ)

# use lsmeans to compare the groups
fit2means <- lsmeans(fit2,
                     "TreatmentArm")

# calculate treatment vs control comparison
contrasts2TC <- contrast(object = fit2means,
                         method = "trt.vs.ctrl",
                         adjust = "none")
# calculate the contrasts one-sided
contrasts2TC <- test(contrasts2TC,
                     side = ">")
# transform emmsgrid into data frame to extract values
contrasts2TC <- data.frame(contrasts2TC)

# extract the estimate
estimateCT1Fit <- contrasts2TC[[2]][1]
# extract the p-value
pValCT1Fit <- contrasts2TC[[6]][1]

# calculate confidence interval for treatment vs control contrast
contrasts2TCCI <- confint(contrast(object = fit2means,
                                   method = "trt.vs.ctrl",
                                   adjust = "none"))
```

```
502     # extract upper and lower bounds
503     CICT1Fit <- c(contrasts2TCCI[[5]][1],
504                   contrasts2TCCI[[6]][1])
505
506
507     # Analysis Scenario 2
508     # edit the data frame and delete all patients in group 1 (X) for the comparison of T2-C with only
509     # with the patients who had the chance of getting treated with the treatment of interest
510     # --> restricted data for the unadjusted comparison of treatment 2 vs control
511     dataYZ <- data %>%
512       filter(!Typeofgroups %in% '1' & !TreatmentArm %in% "T1")
513
514     # fit one-way model
515     fit3 <- lm(Outcome ~ TreatmentArm,
516                data = dataYZ)
517
518     # use lsmeans to compare the groups
519     fit3means <- lsmeans(fit3,
520                          "TreatmentArm")
521
522     # calculate treatment vs control comparison
523     contrasts3TC <- contrast(object = fit3means,
524                              method = "trt.vs.ctrl",
525                              adjust = "none")
526     # calculate the contrasts one-sided
527     contrasts3TC <- test(contrasts3TC,
528                          side = ">")
529     # transform emmsgrid into data frame to extract values
530     contrasts3TC <- data.frame(contrasts3TC)
531
532     # extract the estimate
533     estimateCT2Fit <- contrasts3TC[[2]][1]
534     # extract the p-value
535     pValCT2Fit <- contrasts3TC[[6]][1]
536
537     # calculate confidence interval for treatment vs control contrast
538     contrasts3TCCI <- confint(contrast(object = fit3means,
539                                        method = "trt.vs.ctrl",
540                                        adjust = "none"))
541     # extract upper and lower bounds of CI
542     CICT2Fit <- c(contrasts3TCCI[[5]][1],
543                   contrasts3TCCI[[6]][1])
544
545
546
547     # Analysis Scenario 3
548     # edit the data frame and delete the patients who were allocated to treatment 2
549     # --> all control data for the unadjusted comparison of treatment 1 vs control
550     dataT1C <- data %>%
551       filter(!TreatmentArm %in% 'T2')
552
553     # fit one-way model
554     fit4 <- lm(Outcome ~ TreatmentArm,
555                data = dataT1C)
556
557     # use lsmeans to compare the groups
558     fit4means <- lsmeans(fit4,
559                          "TreatmentArm")
560
561     # calculate treatment vs control comparison
562     contrasts4TC <- contrast(object = fit4means,
563                              method = "trt.vs.ctrl",
564                              adjust = "none")
565     # calculate the contrasts one-sided
566     contrasts4TC <- test(contrasts4TC,
567                          side = ">")
568     # transform emmsgrid into data frame to extract values
569     contrasts4TC <- data.frame(contrasts4TC)
570
571     # extract the estimate
572     estimateCT1FitAllC <- contrasts4TC[[2]][1]
573     # extract the p-value
574     pValCT1FitAllC <- contrasts4TC[[6]][1]
575
576     # calculate confidence interval for treatment vs control contrast
577     contrasts4TCCI <- confint(contrast(object = fit4means,
```

```
                                    method = "trt.vs.ctrl",
                                    adjust = "none"))
  # extract upper and lower bounds of CI
  CICT1FitAllC<- c(contrasts4TCCI[[5]][1],
                   contrasts4TCCI[[6]][1])


  # Analysis Scenario 3
  # edit the data frame and delete the patients  who were allocated to treatment 1
  # --> all control data for the unadjusted comparison of treatment 2 vs control
  dataT2C <- data %>%
    filter(!TreatmentArm %in% 'T1')

  # fit one-way model
  fit5 <- lm(Outcome ~ TreatmentArm,
             data = dataT2C)

  # use lsmeans to compare the groups
  fit5means <- lsmeans(fit5,
                       "TreatmentArm")

  # calculate treatment vs control comparison
  contrasts5TC <- contrast(object = fit5means,
                           method = "trt.vs.ctrl",
                           adjust = "none")
  # calculate the contrasts one-sided
  contrasts5TC <- test(contrasts5TC,
                       side = ">")
  # transform emmsgrid into data frame to extract values
  contrasts5TC <- data.frame(contrasts5TC)

  # extract the estimate
  estimateCT2FitAllC <- contrasts5TC[[2]][1]
  # extract the p-value
  pValCT2FitAllC <- contrasts5TC[[6]][1]

  # calculate confidence interval for treatment vs control contrast
  contrasts5TCCI <- confint(contrast(object = fit5means,
                                     method = "trt.vs.ctrl",
                                     adjust = "none"))
  # extract upper and lower bounds of CI
  CICT2FitAllC<- c(contrasts5TCCI[[5]][1],
                   contrasts5TCCI[[6]][1])




  # Analysis Scenario 4
  # fit two-way model for all data
  # --> adjust for the types of group
  fitTwo <- lm(Outcome ~ Typeofgroups + TreatmentArm,
               data = data)

  # use lsmeans to compare the groups
  fitTwoMeans <- lsmeans(fitTwo,
                         "TreatmentArm")

  # calculate treatment vs control comparison
  contrastsTCTwo <- contrast(object = fitTwoMeans,
                             method = "trt.vs.ctrl",
                             adjust = "none")
  # calculate the contrasts one-sided
  contrastsTCTwo <- test(contrastsTCTwo,
                         side = ">")
  # transform it into a data frame to be able to extract the values
  contrastsTCTwo <- data.frame(contrastsTCTwo)

  # extract the estimates
  estimateCT1Two <- contrastsTCTwo[[2]][1]
  estimateCT2Two <- contrastsTCTwo[[2]][2]
  # extract the p-values
  pValCT1Two <- contrastsTCTwo[[6]][1]
  pValCT2Two <- contrastsTCTwo[[6]][2]

  # calculate confidence interval for treatment vs control comparison
```

```
654     contrastsTCTwoCI <- confint(contrast(object = fitTwoMeans,
655                                          method = "trt.vs.ctrl",
656                                          adjust = "none"))
657     # transform CI contrasts into data frame to extract upper and lower bounds
658     contrastsTCTwoCI <- data.frame(contrastsTCTwoCI)
659     # extract upper and lower bounds
660     CICT1Two <- c(contrastsTCTwoCI[[5]][1],
661                   contrastsTCTwoCI[[6]][1])
662     CICT2Two <- c(contrastsTCTwoCI[[5]][2],
663                   contrastsTCTwoCI[[6]][2])
664
665
666     # Analysis Scenario 5
667     # fit two-way model for only the control data of groups X and Z
668     # --> restricted data for the adjusted comparison of treatment 1 vs control
669     fitTwoXZ <- lm(Outcome ~ Typeofgroups + TreatmentArm,
670                    data = dataXZ)
671
672     # use lsmeans to compare the groups
673     fitTwoXZMeans <- lsmeans(fitTwoXZ,
674                              "TreatmentArm")
675
676     # calculate treatment vs control comparison
677     contrastsTCTwoXZ <- contrast(object = fitTwoXZMeans,
678                                  method = "trt.vs.ctrl",
679                                  adjust = "none")
680     # calculate the contrasts one-sided
681     contrastsTCTwoXZ <- test(contrastsTCTwoXZ,
682                              side = ">")
683     # transform it into a data frame to be able to extract the values
684     contrastsTCTwoXZ <- data.frame(contrastsTCTwoXZ)
685
686     # extract the estimates
687     estimateCT1XZ <- contrastsTCTwoXZ[[2]][1]
688     # extract the p-values
689     pValCT1XZ <- contrastsTCTwoXZ[[6]][1]
690
691     # calculate confidence interval for treatment vs control comparison
692     contrastsTCTwoXZCI <- confint(contrast(object = fitTwoXZMeans,
693                                            method = "trt.vs.ctrl",
694                                            adjust = "none"))
695     # transform it into a data frame to be able to extract the values
696     contrastsTCTwoXZCI <- data.frame(contrastsTCTwoXZCI)
697     # get upper and lower bounds for contrasts of interest
698     CICT1XZ <- c(contrastsTCTwoXZCI[[5]][1],
699                  contrastsTCTwoXZCI[[6]][1])
700
701
702     # Analysis Scenario 5
703     # fit two-way anova model for only the control data of groups Y and Z
704      # --> restricted data for the adjusted comparison of treatment 2 vs control
705     fitTwoYZ <- lm(Outcome ~ Typeofgroups + TreatmentArm,
706                    data = dataYZ)
707
708     # use lsmeans to compare the groups
709     fitTwoYZMeans <- lsmeans(fitTwoYZ,
710                              "TreatmentArm")
711
712     # calculate treatment vs control comparison
713     contrastsTCTwoYZ <- contrast(object = fitTwoYZMeans,
714                                  method = "trt.vs.ctrl",
715                                  adjust = "none")
716     # calculate the contrasts one-sided
717     contrastsTCTwoYZ <- test(contrastsTCTwoYZ,
718                              side = ">")
719     # transform it into a data frame to be able to extract the values
720     contrastsTCTwoYZ <- data.frame(contrastsTCTwoYZ)
721
722     # extract the estimates
723     estimateCT2YZ <- contrastsTCTwoYZ[[2]][1]
724     # extract the p-values
725     pValCT2YZ <- contrastsTCTwoYZ[[6]][1]
726
727     # calculate confidence interval for treatment vs control comparison
728     contrastsTCTwoYZCI <- confint(contrast(object = fitTwoYZMeans,
729                                            method = "trt.vs.ctrl",
```

```
730                                         adjust = "none"))
731     # transform it into a data frame to be able to extract the values
732     contrastsTCTwoYZCI <- data.frame(contrastsTCTwoYZCI)
733     # get upper and lower bounds for contrasts of interest
734     CICT2YZ <- c(contrastsTCTwoYZCI[[5]][1],
735                  contrastsTCTwoYZCI[[6]][1])
736
737
738     # Analysis Scenario 6
739     # fit two-way model for all control data of the data set which only includes T1 and C
740     # --> all control data for the adjusted comparison of treatment 1 vs control
741     fitTwoT1C <- lm(Outcome ~ Typeofgroups + TreatmentArm,
742                     data = dataT1C)
743
744     # use lsmeans to compare the groups
745     fitTwoT1CMeans <- lsmeans(fitTwoT1C,
746                               "TreatmentArm")
747
748     # calculate treatment vs control comparison
749     contrastsTCTwoT1C <- contrast(object = fitTwoT1CMeans,
750                                   method = "trt.vs.ctrl",
751                                   adjust = "none")
752     # calculate the contrasts one-sided
753     contrastsTCTwoT1C <- test(contrastsTCTwoT1C,
754                               side = ">")
755     # transform it into a data frame to be able to extract the values
756     contrastsTCTwoT1C <- data.frame(contrastsTCTwoT1C)
757
758     # extract the estimates
759     estimateCT1All1C <- contrastsTCTwoT1C[[2]][1]
760     # extract the p-values
761     pValCT1All1C <- contrastsTCTwoT1C[[6]][1]
762
763     # calculate confidence interval for treatment vs control comparison
764     contrastsTCTwoT1CCI <- confint(contrast(object = fitTwoT1CMeans,
765                                             method = "trt.vs.ctrl",
766                                             adjust = "none"))
767     # transform it into a data frame to be able to extract the values
768     contrastsTCTwoT1CCI <- data.frame(contrastsTCTwoT1CCI)
769     # get upper and lower bounds for contrasts of interest
770     CICT1All1C <- c(contrastsTCTwoT1CCI[[5]][1],
771                     contrastsTCTwoT1CCI[[6]][1])
772
773
774     # Analysis Scenario 6
775     # fit two-way model for all control data of the data set which only includes T2 and C
776     # --> restricted data for the adjusted comparison of treatment 2 vs control
777     fitTwoT2C <- lm(Outcome ~ Typeofgroups + TreatmentArm,
778                     data = dataT2C)
779
780     # use lsmeans to compare the groups
781     fitTwoT2CMeans <- lsmeans(fitTwoT2C,
782                               "TreatmentArm")
783
784     # calculate treatment vs control comparison
785     contrastsTCTwoT2C <- contrast(object = fitTwoT2CMeans,
786                                   method = "trt.vs.ctrl",
787                                   adjust = "none")
788     # calculate the contrasts one-sided
789     contrastsTCTwoT2C <- test(contrastsTCTwoT2C,
790                               side = ">")
791     # transform it into a data frame to be able to extract the values
792     contrastsTCTwoT2C <- data.frame(contrastsTCTwoT2C)
793
794     # extract the estimates
795     estimateCT2All1C <- contrastsTCTwoT2C[[2]][1]
796     # extract the p-values
797     pValCT2All1C <- contrastsTCTwoT2C[[6]][1]
798
799     # calculate confidence interval for treatment vs control comparison
800     contrastsTCTwoT2CCI <- confint(contrast(object = fitTwoT2CMeans,
801                                             method = "trt.vs.ctrl",
802                                             adjust = "none"))
803     # transform it into a data frame to be able to extract the values
804     contrastsTCTwoT2CCI <- data.frame(contrastsTCTwoT2CCI)
805     # get upper and lower bounds for contrasts of interest
```

```
806     CICT2AllC <- c(contrastsTCTwoT2CCI[[5]][1],
807                    contrastsTCTwoT2CCI[[6]][1])
808
809     # optionally give back the dataset and the input parameters otherwise just return the output
810     # parameters which are necessary for the next function
811     if(full == TRUE) {
812
813     # return a list containing the data frame, the simulation parameters, the operating
814     # characteristics and the results from the analyses in a sublist
815     results_list <- list(Dataset = data,
816                          Timepoints = t,
817                          SampleSize = sampleSize,
818                          NumbArms = armsNumb_vec,
819                          Prevalence = allocProb_vec,
820                          MeanCX = meanCX,
821                          MeanCY = meanCY,
822                          MeanCZ = meanCZ,
823                          MeanT1X = meanT1X,
824                          MeanT2Y = meanT2Y,
825                          MeanT1Z = meanT1Z,
826                          MeanT2Z = meanT2Z,
827                          TreatmentEffectT1 = deltaT1,
828                          TreatmentEffectT2 = deltaT2,
829                          blocksizeXY = blocksizeXY,
830                          blocksizeZ = blocksizeZ,
831                          StandDevT1X = sdT1X,
832                          StandDevT2Y = sdT2Y,
833                          StandDevT1Z = sdT1Z,
834                          StandDevT2Z = sdT2Z,
835                          StandDevCX = sdCX,
836                          StandDevCY = sdCY,
837                          StandDevCZ = sdCZ,
838                          AllocProbXY = allocProbRandXY_vec,
839                          AllocProbZ = allocProbRandZ_vec,
840                          CompleteRand = complete,
841                          RandomAlloc = random,
842                          list2 = list(PatientsperArm = patientsPerArm,
843                                       TrtvsCrtlAll = contrastsTC,
844                                       estimateCT1All = estimateCT1All,
845                                       estimateCT2All = estimateCT2All,
846                                       PValueControlT1All = pValCT1All,
847                                       PValueControlT2All = pValCT2All,
848                                       CITrtvsCrtlAll = contrastsTCCI,
849                                       CICT1All = CICT1All,
850                                       CICT2All = CICT2All,
851                                       TrtvsCrtlCT1 = contrasts2TC,
852                                       estimateCT1Fit = estimateCT1Fit,
853                                       pValCT1Fit = pValCT1Fit,
854                                       CITrtvsCrtlCT1Fit = contrasts2TCCI,
855                                       CICT1Fit = CICT1Fit,
856                                       TrtvsCrtlCT2 = contrasts3TC,
857                                       estimateCT2Fit = estimateCT2Fit,
858                                       pValCT2Fit = pValCT2Fit,
859                                       CITrtvsCrtlCT2Fit = contrasts3TCCI,
860                                       CICT2Fit = CICT2Fit,
861                                       TrtvsCrtlCT1AllC = contrasts4TC,
862                                       estimateCT1FitAllC = estimateCT1FitAllC,
863                                       pValCT1FitAllC = pValCT1FitAllC,
864                                       CITrtvsCrtlCT1FitAllC = contrasts4TCCI,
865                                       CICT1FitAllC = CICT1FitAllC,
866                                       TrtvsCrtlCT2AllC = contrasts5TC,
867                                       estimateCT2FitAllC = estimateCT2FitAllC,
868                                       pValCT2FitAllC = pValCT2FitAllC,
869                                       CITrtvsCrtlCT2FitAllC = contrasts5TCCI,
870                                       CICT2FitAllC = CICT2FitAllC,
871                                       TrtvsCrtlTwo = contrastsTCTwo,
872                                       estimateCT1Two = estimateCT1Two,
873                                       estimateCT2Two = estimateCT2Two,
874                                       pValCT1Two = pValCT1Two,
875                                       pValCT2Two = pValCT2Two,
876                                       CITrtvsCrtlTwo = contrastsTCTwoCI,
877                                       CICT1Two = CICT1Two,
878                                       CICT2Two = CICT2Two,
879                                       TrtvsCrtlTwoXZ = contrastsTCTwoXZ,
880                                       estimateCT1XZ = estimateCT1XZ,
881                                       pValCT1XZ = pValCT1XZ,
```

```
882                                     CITrtvsCrtlTwoXZ = contrastsTCTwoXZCI,
883                                     CICT1XZ = CICT1XZ,
884                                     TrtvsCrtlTwoYZ = contrastsTCTwoYZ,
885                                     estimateCT2YZ = estimateCT2YZ,
886                                     pValCT2YZ = pValCT2YZ,
887                                     CITrtvsCrtlTwoYZ = contrastsTCTwoYZCI,
888                                     CICT2YZ = CICT2YZ,
889                                     TrtvsCrtlTwoT1C = contrastsTCTwoT1C,
890                                     estimateCT1AllC = estimateCT1AllC,
891                                     pValCT1AllC = pValCT1AllC,
892                                     CITrtvsCrtlTwoT1C = contrastsTCTwoT1CCI,
893                                     CICT1AllC = CICT1AllC,
894                                     TrtvsCrtlTwoT2C = contrastsTCTwoT2C,
895                                     estimateCT2AllC = estimateCT2AllC,
896                                     pValCT2AllC = pValCT2AllC,
897                                     TrtvsCrtlTwoT2CCI = contrastsTCTwoT2CCI,
898                                     CICT2AllC = CICT2AllC,
899                                     AllocationList = listAlloc_vec
900                           ))
901     return(results_list)
902
903     } else {
904
905       # return a sublist containing the operating characteristics and the results from the analyses
906       results_list <- list(list2 = list(PatientsperArm = patientsPerArm,
907                                         estimateCT1All = estimateCT1All,
908                                         estimateCT2All = estimateCT2All,
909                                         PValueControlT1All = pValCT1All,
910                                         PValueControlT2All = pValCT2All,
911                                         CICT1All = CICT1All,
912                                         CICT2All = CICT2All,
913                                         estimateCT1Fit = estimateCT1Fit,
914                                         pValCT1Fit = pValCT1Fit,
915                                         CICT1Fit = CICT1Fit,
916                                         estimateCT2Fit = estimateCT2Fit,
917                                         pValCT2Fit = pValCT2Fit,
918                                         CICT2Fit = CICT2Fit,
919                                         estimateCT1FitAllC = estimateCT1FitAllC,
920                                         pValCT1FitAllC = pValCT1FitAllC,
921                                         CICT1FitAllC = CICT1FitAllC,
922                                         estimateCT2FitAllC = estimateCT2FitAllC,
923                                         pValCT2FitAllC = pValCT2FitAllC,
924                                         CICT2FitAllC = CICT2FitAllC,
925                                         estimateCT1Two = estimateCT1Two,
926                                         estimateCT2Two = estimateCT2Two,
927                                         pValCT1Two = pValCT1Two,
928                                         pValCT2Two = pValCT2Two,
929                                         CICT1Two = CICT1Two,
930                                         CICT2Two = CICT2Two,
931                                         estimateCT1XZ = estimateCT1XZ,
932                                         pValCT1XZ = pValCT1XZ,
933                                         CICT1XZ = CICT1XZ,
934                                         estimateCT2YZ = estimateCT2YZ,
935                                         pValCT2YZ = pValCT2YZ,
936                                         CICT2YZ = CICT2YZ,
937                                         estimateCT1AllC = estimateCT1AllC,
938                                         pValCT1AllC = pValCT1AllC,
939                                         CICT1AllC = CICT1AllC,
940                                         estimateCT2AllC = estimateCT2AllC,
941                                         pValCT2AllC = pValCT2AllC,
942                                         CICT2AllC = CICT2AllC,
943                                         AllocationList = listAlloc_vec
944                           ))
945       return(results_list)
946     }
947 }
```

A.4.2 Function for parallelization and calculating the operating characteristics

```
callSimulation <-
  function(
    t, # number of time periods
    n, # number of patients for each time period
    armsNumb_vec = 1:3, # number of arms
    allocProb_vec, # allocation probability to the different types of groups
    deltaT1, # treatment effect for treatment 1
    deltaT2, # treatment effect for treatment 2
    meanCX, # mean for control in group X
    meanCY, # mean for control in group Y
    meanCZ, # mean for control in group Z
    sdT1X = 1, # standard deviation for treatment 1 in group X
    sdT2Y = 1, # standard deviation for treatment 2 in group Y
    sdT1Z = 1, # standard deviation for treatment 1 in group Z
    sdT2Z = 1, # standard deviation for treatment 2 in group Z
    sdCX = 1, # standard deviation for control in group X
    sdCY = 1, # standard deviation for control in group Y
    sdCZ = 1, # standard deviation for control in group
    blocksizeXY, # block size for block randomization for groups X and Y
    blocksizeZ, # block size for block randomization for group Z
    alpha = 0.025, # alpha level for t-test
    allocProbRandXY_vec, # allocation probability for groups X & Y to treatment or control
    allocProbRandZ_vec, # allocation probability for group Z to different arms
    complete, # boolean for randomization method
    random, # boolean for deterministic or random allocation to groups
    full,  # boolean for returning input parameters or just the output
    iterations # number of iterations
  ) {

    # determine how many cores to use
    nCores <- parallel::detectCores() - 1

    #create the cluster
    myCluster <- parallel::makeCluster(nCores,
                                       type = "PSOCK")

    # register the cluster to be used by %dopar%
    doParallel::registerDoParallel(cl = myCluster)

    # create empty list
    opChar_list <- list()

    # run the simulation for a given number of iterations
    # over the grid of all combinations
    opChar_list <- foreach(i = 1:iterations,
                           .packages = c('emmeans', 'dplyr', 'stats'),
                           .export = "fnSimulation") %dopar% {

      # call the function
      opChar_list[[i]] <- fnSimulation(t = t, # number of time periods
                                       n = n, # number of patients for each time period
                                       armsNumb_vec = armsNumb_vec, # number of arms
                                       allocProb_vec = allocProb_vec, # allocation probability to the
      different groups
                                       deltaT1 = deltaT1, # treatment effect for treatment 1
                                       deltaT2 = deltaT2, # treatment effect for treatment 2
                                       meanCX = meanCX, # mean for control in group X
                                       meanCY = meanCY, # mean for control in group Y
                                       meanCZ = meanCZ, # mean for control in group Z
                                       sdT1X = 1, # standard deviation for treatment 1 in group X
                                       sdT2Y = 1, # standard deviation for treatment 2 in group Y
                                       sdT1Z = 1, # standard deviation for treatment 1 in group Z
                                       sdT2Z = 1, # standard deviation for treatment 2 in group Z
                                       sdCX = 1, # standard deviation for control in group X
                                       sdCY = 1, # standard deviation for control in group Y
                                       sdCZ = 1, # standard deviation for control in group
                                       blocksizeXY = blocksizeXY, # block size for block randomization for
      groups X, Y
                                       blocksizeZ = blocksizeZ, # block size for block randomization for
      group Z
                                       alpha = 0.025, # alpha level for t-test
```

```
69                                      allocProbRandXY_vec = allocProbRandXY_vec, # allocation prob. for
      groups X, Y
70                                      allocProbRandZ_vec, # allocation probability for group Z to different
       arms
71                                      complete = complete, # boolean for randomization method
72                                      random = random, # boolean for deterministic or random allocation to
      groups
73                                      full = full # boolean for returning input parameters or just the
      output
74       )$list2
75     }
76       # stop cluster when no longer needed
77       parallel::stopCluster(cl = myCluster)
78
79       # get the mean for the prevalence for the different types of groups over all iterations
80       listAllocAll_mat <- do.call(rbind,
81                                   lapply(opChar_list,
82                                          function(x) table(x[["AllocationList"]])))
83       groupMean <- colMeans(listAllocAll_mat)
84       groupMean <- unname(groupMean)
85
86       # save the mean for each group separately
87       groupMeanX <- groupMean[1]
88       groupMeanY <- groupMean[2]
89       groupMeanZ <- groupMean[3]
90
91       # get the mean for the number of patients per arm over all iterations
92       tablePatientsPerArm <- do.call(rbind,
93                                      lapply(opChar_list,
94                                             function(x) table(x[["PatientsperArm"]])))
95       armMean <- colMeans(tablePatientsPerArm)
96       armMean <- unname(armMean)
97
98       # save the mean for each arm separately
99       meanControl <- armMean[1]
100      meanT1 <- armMean[2]
101      meanT2 <- armMean[3]
102
103      ### calculate power, bias, type I error and CI for the different scenarios ###
104
105      ## Analysis Scenario 1
106      # one-way model for treatment 1 vs control using all control data (unadjusted analysis)
107      pValCT1All <- do.call(rbind,
108                            lapply(opChar_list,
109                                   function(x) x[["PValueControlT1All"]]))
110
111      # save which p-values are smaller than alpha
112      rejectOneCT1All_mat <- pValCT1All < alpha
113
114      # calculate the rejection probability
115      rejProbCT1All <- mean(rejectOneCT1All_mat)
116
117      # get estimates and calculate bias
118      estimateCT1All <- do.call(rbind,
119                                lapply(opChar_list,
120                                       function(x) x[["estimateCT1All"]]))
121      biasCT1All <- mean(estimateCT1All - deltaT1)
122
123      # get lower and upper bound of confidence intervals and calculate the average
124      CICT1All <- do.call(rbind,
125                          lapply(opChar_list,
126                                 function(x) x[["CICT1All"]]))
127      CICT1All <- colMeans(CICT1All)
128
129
130
131      ## Analysis Scenario 1
132      # one-way model for treatment 2 vs control using all control data (unadjusted analysis)
133      pValCT2All <- do.call(rbind,
134                            lapply(opChar_list,
135                                   function(x) x[["PValueControlT2All"]]))
136
137      # save which p-values are smaller than alpha
138      rejectOneCT2All_mat <- pValCT2All < alpha
139
140      # calculate the rejection probability
```

```
141       rejProbCT2All <- mean(rejectOneCT2All_mat)
142
143       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2
144       rejectAtLeastOneCT12All_mat <- (pValCT1All < alpha) | (pValCT2All < alpha)
145       rejectBothCT12All_mat <- (pValCT1All < alpha) & (pValCT2All < alpha)
146
147       # get estimates and calculate bias
148       estimateCT2All <- do.call(rbind,
149                                 lapply(opChar_list,
150                                        function(x) x[["estimateCT2All"]]))
151       biasCT2All <- mean(estimateCT2All - deltaT2)
152
153       # get lower and upper bound of confidence intervals and calculate the average
154       CICT2All <- do.call(rbind,
155                           lapply(opChar_list,
156                                  function(x) x[["CICT2All"]]))
157       CICT2All <- colMeans(CICT2All)
158
159
160
161       ## Analysis Scenario 2
162       # one-way model for treatment 1 vs. control using only suitable control data of groups X and Z
163       # --> restricted data for unadjusted comparison
164       pValCT1Fit <- do.call(rbind,
165                             lapply(opChar_list,
166                                    function(x) x[["pValCT1Fit"]]))
167
168       # save which p-values are smaller than alpha
169       rejectOneCT1Fit_mat <- pValCT1Fit < alpha
170
171       # calculate the rejection probability
172       rejProbCT1Fit <- mean(rejectOneCT1Fit_mat)
173
174       # get estimates and calculate bias
175       estimateCT1Fit <- do.call(rbind,
176                                 lapply(opChar_list,
177                                        function(x) x[["estimateCT1Fit"]]))
178       biasCT1Fit <- mean(estimateCT1Fit - deltaT1)
179
180       # get lower and upper bound of confidence intervals and calculate the average
181       CICT1Fit <- do.call(rbind,
182                           lapply(opChar_list,
183                                  function(x) x[["CICT1Fit"]]))
184       CICT1Fit1 <- colMeans(CICT1Fit)
185
186
187
188       ## Analysis Scenario 2
189       # one-way model for treatment 2 vs control using only suitable control data of groups Y and Z
190       # --> restricted data for unadjusted comparison
191       pValCT2Fit <- do.call(rbind,
192                             lapply(opChar_list,
193                                    function(x) x[["pValCT2Fit"]]))
194
195       # save which p-values are smaller than alpha
196       rejectOneCT2Fit_mat <- pValCT2Fit < alpha
197
198       # calculate the rejection probability
199       rejProbCT2Fit <- mean(rejectOneCT2Fit_mat)
200
201       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2
202       rejectAtLeastOneCT12Fit_mat <- (pValCT1Fit < alpha) |(pValCT2Fit < alpha)
203       rejectBothCT12Fit_mat <- (pValCT1Fit < alpha) & (pValCT2Fit < alpha)
204
205       # get estimates and calculate bias
206       estimateCT2Fit <- do.call(rbind,
207                                 lapply(opChar_list,
208                                        function(x) x[["estimateCT2Fit"]]))
209       biasCT2Fit <- mean(estimateCT2Fit - deltaT2)
210
211       # get lower and upper bound of confidence intervals and calculate the average
212       CICT2Fit <- do.call(rbind,
213                           lapply(opChar_list,
214                                  function(x) x[["CICT2Fit"]]))
215       CICT2Fit <- colMeans(CICT2Fit)
216
```

```
217
218
219
220       ## Analysis Scenario 3
221       # one-way model for treatment 1 vs control using all control data for T1 and C
222       # --> all control data for unadjusted comparison
223       pValCT1FitAllC <- do.call(rbind,
224                                 lapply(opChar_list,
225                                        function(x) x[["pValCT1FitAllC"]]))
226
227       # save which p-values are smaller than alpha
228       rejectOneCT1FitAllC_mat <- pValCT1FitAllC < alpha
229
230       # calculate the rejection probability
231       rejProbCT1FitAllC <- mean(rejectOneCT1FitAllC_mat)
232
233       # get estimates and calculate bias
234       estimateCT1FitAllC <- do.call(rbind,
235                                     lapply(opChar_list,
236                                            function(x) x[["estimateCT1FitAllC"]]))
237       biasCT1FitAllC <- mean(estimateCT1FitAllC - deltaT1)
238
239       # get lower and upper bound of confidence intervals and calculate the average
240       CICT1FitAllC <- do.call(rbind,
241                               lapply(opChar_list,
242                                      function(x) x[["CICT1FitAllC"]]))
243       CICT1FitAllC1 <- colMeans(CICT1FitAllC)
244
245
246
247       ## Analysis Scenario 3
248       # one-way model for treatment 2 vs control using all control data for T2 and C
249       # --> all control data for the unadjusted comparison
250       pValCT2FitAllC <- do.call(rbind,
251                                 lapply(opChar_list,
252                                        function(x) x[["pValCT2FitAllC"]]))
253
254       # save which p-values are smaller than alpha
255       rejectOneCT2FitAllC_mat <- pValCT2FitAllC < alpha
256
257       # calculate the rejection probability
258       rejProbCT2FitAllC <- mean(rejectOneCT2FitAllC_mat)
259
260       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2
261       rejectAtLeastOneCT12FitAllC_mat <- (pValCT1FitAllC < alpha) |(pValCT2FitAllC < alpha)
262       rejectBothCT12FitAllC_mat <- (pValCT1FitAllC < alpha) & (pValCT2FitAllC < alpha)
263
264       # get estimates and calculate bias
265       estimateCT2FitAllC <- do.call(rbind,
266                                     lapply(opChar_list,
267                                            function(x) x[["estimateCT2FitAllC"]]))
268       biasCT2FitAllC <- mean(estimateCT2FitAllC - deltaT2)
269
270       # get lower and upper bound of confidence intervals and  calculate the average
271       CICT2FitAllC <- do.call(rbind,
272                               lapply(opChar_list,
273                                      function(x) x[["CICT2FitAllC"]]))
274       CICT2FitAllC <- colMeans(CICT2FitAllC)
275
276
277       ## Analysis Scenario 4
278       # two-way model for treatment 1 vs control for all data (adjusted analysis)
279       pValCT1Two <- do.call(rbind,
280                             lapply(opChar_list,
281                                    function(x) x[["pValCT1Two"]]))
282
283       # save which p-values are smaller than alpha
284       rejectOne2CT1_mat <- pValCT1Two < alpha
285
286       # calculate the rejection probability
287       rejProbCT1Two <- mean(rejectOne2CT1_mat)
288
289       # get estimates and calculate bias
290       estimateCT1Two <- do.call(rbind,
291                                 lapply(opChar_list,
292                                        function(x) x[["estimateCT1Two"]]))
```

```
293       biasCT1Two <- mean(estimateCT1Two - deltaT1)
294
295       # get lower and upper bound of confidence intervals and calculate the average
296       CICT1Two <- do.call(rbind,
297                           lapply(opChar_list,
298                                  function(x) x[["CICT1Two"]]))
299       CICT1Two <- colMeans(CICT1Two)
300
301
302       ## Analysis Scenario 4
303       # two-way model for treatment 2 vs control for all data (adjusted analysis)
304       pValCT2Two <- do.call(rbind,
305                             lapply(opChar_list,
306                                    function(x) x[["pValCT2Two"]]))
307
308       # save which p-values are smaller than alpha
309       rejectOne2CT2_mat <- pValCT2Two < alpha
310
311       # calculate the rejection probability
312       rejProbCT2Two <- mean(rejectOne2CT2_mat)
313
314       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2 and take their mean
315       rejectAtLeastOne2CT12_mat <- (pValCT1Two < alpha) | (pValCT2Two < alpha)
316       rejectBoth2CT12_mat <- (pValCT1Two < alpha) & (pValCT2Two < alpha)
317
318       # get estimates and calculate bias
319       estimateCT2Two <- do.call(rbind,
320                                 lapply(opChar_list,
321                                        function(x) x[["estimateCT2Two"]]))
322       biasCT2Two <- mean(estimateCT2Two - deltaT2)
323
324       # get lower and upper bound of confidence intervals and calculate the average
325       CICT2Two <- do.call(rbind,
326                           lapply(opChar_list,
327                                  function(x) x[["CICT2Two"]]))
328       CICT2Two <- colMeans(CICT2Two)
329
330
331
332       ## Analysis Scenario 5
333       # two-way model for treatment 1 vs control
334       # --> restricted data for the adjusted comparison of treatment 1 vs control
335       pValCT1XZ <- do.call(rbind,
336                            lapply(opChar_list,
337                                   function(x) x[["pValCT1XZ"]]))
338
339       # save which p-vales are smaller than alpha
340       rejectOne2CT1XZ_mat <- pValCT1XZ < alpha
341
342       # calculate the rejection probability
343       rejProbCT1XZ <- mean(rejectOne2CT1XZ_mat)
344
345       # get estimates and calculate bias
346       estimateCT1XZ <- do.call(rbind,
347                                lapply(opChar_list,
348                                       function(x) x[["estimateCT1XZ"]]))
349       biasCT1XZ <- mean(estimateCT1XZ - deltaT1)
350
351       # get lower and upper bound of confidence intervals and calculate the average
352       CICT1XZ <- do.call(rbind,
353                          lapply(opChar_list,
354                                 function(x) x[["CICT1XZ"]]))
355
356       CICT1XZ <- colMeans(CICT1XZ)
357
358
359
360       ## Analysis Scenario 5
361       # two-way model for treatment 2 vs control
362       # --> restricted data for the adjusted comparison of treatment 2 vs control
363       pValCT2YZ <- do.call(rbind,
364                            lapply(opChar_list,
365                                   function(x) x[["pValCT2YZ"]]))
366
367       # save which p-vales are smaller than alpha
368       rejectOne2CT2YZ_mat <- pValCT2YZ < alpha
```

```
369
370       # calculate the rejection probability
371       rejProbCT2YZ <- mean(rejectOne2CT2YZ_mat)
372
373       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2 and take their mean
374       rejectAtLeastOne2CT12Fit_mat <- (pValCT1XZ < alpha) | (pValCT2YZ < alpha)
375       rejectBoth2CT12Fit_mat <- (pValCT1XZ < alpha) & (pValCT2YZ < alpha)
376
377       # get estimates and calculate bias
378       estimateCT2YZ <- do.call(rbind,
379                                lapply(opChar_list,
380                                       function(x) x[["estimateCT2YZ"]]))
381       biasCT2YZ <- mean(estimateCT2YZ - deltaT2)
382
383       # get lower and upper bound of confidence intervals and calculate the average
384       CICT2YZ <- do.call(rbind,
385                          lapply(opChar_list,
386                                 function(x) x[["CICT2YZ"]]))
387       CICT2YZ <- colMeans(CICT2YZ)
388
389
390
391       ## Analysis Scenario 6
392       # two-way model for treatment 1 vs control
393       # --> all control data for the adjusted comparison of treatment 1 vs control
394       pValCT1AllC <- do.call(rbind,
395                              lapply(opChar_list,
396                                     function(x) x[["pValCT1AllC"]]))
397
398       # save which p-vales are smaller than alpha
399       rejectOne2CT1AllC_mat <- pValCT1AllC < alpha
400
401       # calculate the rejection probability
402       rejProbCT1AllC <- mean(rejectOne2CT1AllC_mat)
403
404       # get estimates and calculate bias
405       estimateCT1AllC <- do.call(rbind,
406                                  lapply(opChar_list,
407                                         function(x) x[["estimateCT1AllC"]]))
408       biasCT1AllC <- mean(estimateCT1AllC - deltaT1)
409
410       # get lower and upper bound of confidence intervals and calculate the average
411       CICT1AllC <- do.call(rbind,
412                            lapply(opChar_list,
413                                   function(x) x[["CICT1AllC"]]))
414       CICT1AllC <- colMeans(CICT1AllC)
415
416
417
418       ## Analysis Scenario 6
419       # two-way model for treatment 2 vs control
420       # --> all control data for the adjusted comparison of treatment 2 vs control
421       pValCT2AllC <- do.call(rbind,
422                              lapply(opChar_list,
423                                     function(x) x[["pValCT2AllC"]]))
424
425       # save which p-vales are smaller than alpha
426       rejectOne2CT2AllC_mat <- pValCT2AllC < alpha
427
428       # calculate the rejection probability
429       rejProbCT2AllC <- mean(rejectOne2CT2AllC_mat)
430
431       # check which p-values are smaller than alpha for both T1 and T2 and T1 or T2 and take their mean
432       rejectAtLeastOne2CT12AllC_mat <- (pValCT1AllC < alpha) | (pValCT2AllC < alpha)
433       rejectBoth2CT12AllC_mat <- (pValCT1AllC < alpha) & (pValCT2AllC < alpha)
434
435       # get estimates and calculate bias
436       estimateCT2AllC <- do.call(rbind,
437                                  lapply(opChar_list,
438                                         function(x) x[["estimateCT2AllC"]]))
439       biasCT2AllC <- mean(estimateCT2AllC - deltaT2)
440
441       # get lower and upper bound of confidence intervals and calculate the average
442       CICT2AllC <- do.call(rbind,
443                            lapply(opChar_list,
444                                   function(x) x[["CICT2AllC"]]))
```

```
445       CICT2AllC <- colMeans(CICT2AllC)
446
447
448       # calculate the conjunctive and disjunctive power for all analysis scenarios, the conjunctive and
449       # disjunctive power are calculated when the treatment effects are different from 0
450       if(deltaT1 != 0 & deltaT2 != 0) {
451         conjPower1All <- mean(rejectBothCT12All_mat)
452         disjPower1All <- mean(rejectAtLeastOneCT12All_mat)
453         conjPower1Fit <- mean(rejectBothCT12Fit_mat)
454         disjPower1Fit <- mean(rejectAtLeastOneCT12Fit_mat)
455         conjPower1AllC <- mean(rejectBothCT12FitAllC_mat)
456         disjPower1AllC <- mean(rejectAtLeastOneCT12FitAllC_mat)
457
458         conjPower2All <- mean(rejectBoth2CT12_mat)
459         disjPower2All <- mean(rejectAtLeastOne2CT12_mat)
460         conjPower2Fit <- mean(rejectBoth2CT12Fit_mat)
461         disjPower2Fit <- mean(rejectAtLeastOne2CT12Fit_mat)
462         conjPower2AllC <- mean(rejectBoth2CT12AllC_mat)
463         disjPower2AllC <- mean(rejectAtLeastOne2CT12AllC_mat)
464
465       }  else {
466         conjPower1All <- NA
467         disjPower1All <- NA
468         conjPower1Fit <- NA
469         disjPower1Fit <- NA
470         conjPower1AllC <- NA
471         disjPower1AllC <- NA
472
473         conjPower2All <- NA
474         disjPower2All <- NA
475         conjPower2Fit <- NA
476         disjPower2Fit <- NA
477         conjPower2AllC <- NA
478         disjPower2AllC <- NA
479       }
480
481       # return a list containing the simulation parameters and the operating characteristics of interest
482       results_list <- list(Timepoints = t,
483                            SampleSize = n * t,
484                            NumbArms = length(armsNumb_vec),
485                            Prevalence = allocProb_vec,
486                            MeanCX = meanCX,
487                            MeanCY = meanCY,
488                            MeanCZ = meanCZ,
489                            Iterations = iterations,
490                            TreatmentEffectT1 = deltaT1,
491                            TreatmentEffectT2 = deltaT2,
492                            blocksizeXY = blocksizeXY,
493                            blocksizeZ = blocksizeZ,
494                            StandDevT1X = sdT1X,
495                            StandDevT2Y = sdT2Y,
496                            StandDevT1Z = sdT1Z,
497                            StandDevT2Z = sdT2Z,
498                            StandDevCX = sdCX,
499                            StandDevCY = sdCY,
500                            StandDevCZ = sdCZ,
501                            AllocProbXY = allocProbRandXY_vec,
502                            AllocProbZ = allocProbRandZ_vec,
503                            CompleteRand = complete,
504                            RandomAlloc = random,
505                            MeanGroupX = groupMeanX,
506                            MeanGroupY = groupMeanY,
507                            MeanGroupZ = groupMeanZ,
508                            MeanControl = meanControl,
509                            MeanT1 = meanT1,
510                            MeanT2 = meanT2,
511                            rejProbCT1All = rejProbCT1All,
512                            biasCT1All = biasCT1All,
513                            CICT1All = CICT1All,
514                            rejProbCT2All = rejProbCT2All,
515                            biasCT2All = biasCT2All,
516                            CICT2All = CICT2All,
517                            rejProbCT1Fit = rejProbCT1Fit,
518                            biasCT1Fit = biasCT1Fit,
519                            CICT1Fit = CICT1Fit1,
520                            rejProbCT2Fit = rejProbCT2Fit,
```

```
521                             biasCT2Fit = biasCT2Fit,
522                             CICT2Fit = CICT2Fit,
523                             rejProbCT1FitAllC = rejProbCT1FitAllC,
524                             biasCT1FitAllC = biasCT1FitAllC,
525                             CICT1FitAllC = CICT1FitAllC1,
526                             rejProbCT2FitAllC = rejProbCT2FitAllC,
527                             biasCT2FitAllC = biasCT2FitAllC,
528                             CICT2FitAllC = CICT2FitAllC,
529                             rejProbCT1Two = rejProbCT1Two,
530                             biasCT1Two = biasCT1Two,
531                             CICT1Two = CICT1Two,
532                             rejProbCT2Two = rejProbCT2Two,
533                             biasCT2Two = biasCT2Two,
534                             CICT2Two = CICT2Two,
535                             rejProbCT1XZ = rejProbCT1XZ,
536                             biasCT1XZ = biasCT1XZ,
537                             CICT1XZ = CICT1XZ,
538                             rejProbCT2YZ = rejProbCT2YZ,
539                             biasCT2YZ = biasCT2YZ,
540                             CICT2YZ = CICT2YZ,
541                             rejProbCT1AllC = rejProbCT1AllC,
542                             biasCT1AllC = biasCT1AllC,
543                             CICT1AllC = CICT1AllC,
544                             rejProbCT2AllC = rejProbCT2AllC,
545                             biasCT2AllC = biasCT2AllC,
546                             CICT2AllC = CICT2AllC,
547                             ConjPower1AllData = conjPower1All,
548                             DisjPower1AllData = disjPower1All,
549                             ConjPower1FittedData = conjPower1Fit,
550                             DisjPower1FittedData = disjPower1Fit,
551                             ConjPower1AllControlData = conjPower1AllC,
552                             DisjPower1AllControlData = disjPower1AllC,
553                             ConjPower2AllData = conjPower2All,
554                             DisjPower2AllData = disjPower2All,
555                             ConjPower2FittedData = conjPower2Fit,
556                             DisjPower2FittedData = disjPower2Fit,
557                             ConjPower2AllControlData = conjPower2AllC,
558                             DisjPower2AllControlData = disjPower2AllC
559       )
560
561     return(results_list)
562 }
```

A.4.3 Function for iterating over the simulation parameters

```
# set seed for reproducibility
set.seed(234)

runSimulation <-
    function(t, # number of time periods
             n, # number of patients for each time period
             armsNumb_vec = 1:3, # number of arms
             prevalence, # allocation probability to the different types of groups
             deltaT1, # treatment effect for treatment 1
             deltaT2, # treatment effect for treatment 2
             meanCX, # mean for control in group X
             meanCY, # mean for control in group Y
             meanCZ, # mean for control in group Z
             sdT1X, # standard deviation for treatment 1 in group X
             sdT2Y, # standard deviation for treatment 2 in group Y
             sdT1Z, # standard deviation for treatment 1 in group Z
             sdT2Z, # standard deviation for treatment 2 in group Z
             sdCX, # standard deviation for control in group X
             sdCY, # standard deviation for control in group Y
             sdCZ, # standard deviation for control in group
             blocksizeXY, # block size for block randomization for groups X and Y
             blocksizeZ, # block size for block randomization for group Z
             alpha, # alpha level for t-test
             allocProbRandXY_vec, # allocation probability for groups X and Y to treatment or control
             allocProbRandZ_vec, # allocation probability for group Z to different arms
             complete, # boolean for randomization method
             random, # boolean for deterministic or random allocation to groups
             full,  # boolean for returning input parameters or just the output
             iterations # number of iterations
) {

  # calculate the total sample size
  n = (n * 3) / t

  # determine the grid over all combinations
  grid1_mat <- expand.grid("Treatment effect T1" = deltaT1,
                           "Treatment effect T2" = deltaT2,
                           "Mean Control X" = round(meanCX, 1),
                           "Mean Control Y" = round(meanCY, 1),
                           "Mean Control Z" = round(meanCZ, 1),
                           "Std Control X" = sdCX,
                           "Std Control Y" = sdCY,
                           "Std Control Z" = sdCZ,
                           "Std T1 X" = sdT1X,
                           "Std T1 Z" = sdT1Z,
                           "Std T2 Y" = sdT2Y,
                           "Std T2 Z" = sdT2Z,
                           "Time periods" = t,
                           "Sample Size" = n,
                           "Prevalence" = prevalence,
                           "Alloc Prob Rand XY" = allocProbRandXY_vec,
                           "Alloc Prob Rand Z" = allocProbRandZ_vec,
                           "Blocksize XY" = blocksizeXY,
                           "Blocksize Z" = blocksizeZ,
                           "Number of Iterations" = iterations,
                           "Signficance level" = alpha,
                           "Complete Randomization" = complete,
                           "Random Allocation" = random,
                           "Full Output" = full)

  # only keep combinations of possible allocation ratios
  grid1_mat <- grid1_mat %>%
    filter(!(`Alloc Prob Rand XY` == "c(0.333333333333333, 0.333333333333333, 0.333333333333333)" & `Alloc
      Prob Rand Z` == "c(0.25, 0.25, 0.25, 0.25)") &
             !(`Blocksize XY` == 2 & `Alloc Prob Rand XY` == "c(0.333333333333333, 0.333333333333333,
      0.333333333333333)") &
             !(`Blocksize XY` == 3 & `Alloc Prob Rand XY` == "c(0.5, 0.5)") &
             !(`Blocksize Z` == 4 & `Alloc Prob Rand Z` == "c(0.333333333333333, 0.333333333333333,
      0.333333333333333)") &
             !(`Blocksize Z` == 3 & `Alloc Prob Rand Z` == "c(0.25, 0.25, 0.25, 0.25)"))


#  only keep the combinations of the three control means that are of interest
```

```
  grid1_mat <- grid1_mat %>%
    dplyr::filter((`Mean Control X` == 0 & `Mean Control Y` == 0 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.1 & `Mean Control Y` == 0.1 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.2 & `Mean Control Y` == 0.2 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.3 & `Mean Control Y` == 0.3 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.4 & `Mean Control Y` == 0.4 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.5 & `Mean Control Y` == 0.5 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.6 & `Mean Control Y` == 0.6 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.7 & `Mean Control Y` == 0.7 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.8 & `Mean Control Y` == 0.8 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.9 & `Mean Control Y` == 0.9 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -1 & `Mean Control Y` == 1 & `Mean Control Z` == 0) |
                  (`Mean Control X` == -0.1 & `Mean Control Y` == 0 & `Mean Control Z` == 0.1) |
                  (`Mean Control X` == -0.2 & `Mean Control Y` == 0 & `Mean Control Z` == 0.2) |
                  (`Mean Control X` == -0.3 & `Mean Control Y` == 0 & `Mean Control Z` == 0.3) |
                  (`Mean Control X` == -0.4 & `Mean Control Y` == 0 & `Mean Control Z` == 0.4) |
                  (`Mean Control X` == -0.5 & `Mean Control Y` == 0 & `Mean Control Z` == 0.5) |
                  (`Mean Control X` == -0.6 & `Mean Control Y` == 0 & `Mean Control Z` == 0.6) |
                  (`Mean Control X` == -0.7 & `Mean Control Y` == 0 & `Mean Control Z` == 0.7) |
                  (`Mean Control X` == -0.8 & `Mean Control Y` == 0 & `Mean Control Z` == 0.8) |
                  (`Mean Control X` == -0.9 & `Mean Control Y` == 0 & `Mean Control Z` == 0.9) |
                  (`Mean Control X` == -1 & `Mean Control Y` == 0 & `Mean Control Z` == 1))


  # iterate over the grid and call the function for all combinations save the output in a list
  results_list <- list()
  for(i in 1:nrow(grid1_mat)) {
    results_list[[i]] <- callSimulation(n = grid1_mat$`Sample Size`[i],
                                        allocProb_vec = grid1_mat$`Prevalence`[[i]],
                                        armsNumb_vec = 1:3,
                                        t = grid1_mat$`Time periods`[i],
                                        deltaT1 = grid1_mat$`Treatment effect T1`[i],
                                        deltaT2 = grid1_mat$`Treatment effect T2`[i],
                                        meanCX = grid1_mat$`Mean Control X`[i],
                                        meanCY = grid1_mat$`Mean Control Y`[i],
                                        meanCZ = grid1_mat$`Mean Control Z`[i],
                                        sdCX = grid1_mat$`Std Control X`[i],
                                        sdCY = grid1_mat$`Std Control Y`[i],
                                        sdCZ = grid1_mat$`Std Control Z`[i],
                                        sdT1X = grid1_mat$`Std T1 X`[i],
                                        sdT1Z = grid1_mat$`Std T1 Z`[i],
                                        sdT2Y = grid1_mat$`Std T2 Y`[i],
                                        sdT2Z = grid1_mat$`Std T2 Z`[i],
                                        allocProbRandXY_vec = grid1_mat$`Alloc Prob Rand XY`[[i]],
                                        allocProbRandZ_vec = grid1_mat$`Alloc Prob Rand Z`[[i]],
                                        blocksizeXY = grid1_mat$`Blocksize XY`[i],
                                        blocksizeZ = grid1_mat$`Blocksize Z`[i],
                                        complete = grid1_mat$`Complete Randomization`[i],
                                        random = grid1_mat$`Random Allocation`[i],
                                        full = grid1_mat$`Full Output`[i],
                                        iterations = grid1_mat$`Number of Iterations`[i],
                                        alpha = grid1_mat$`Signficance level`[i])

  }

  # turn the list into a data frame
  data <- do.call(rbind,
                  results_list)

  # save data frame as csv file with the date of the respective day
  today <- format(Sys.Date(), "%Y_%m_%d")
  write.csv2(data,
             paste0("data_", today, ".csv"),
             row.names = TRUE)

}
```

Universität Bremen

Offizielle Erklärungen von

Nachname: _____ Vorname: _____

Matrikelnr.: _____

A) Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.
Alle Teile meiner Arbeit, die wortwörtlich oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Gleiches gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet.
Die Arbeit wurde in gleicher oder ähnlicher Form noch nicht als Prüfungsleistung eingereicht.
Die elektronische Fassung der Arbeit stimmt mit der gedruckten Version überein.
Mir ist bewusst, dass wahrheitswidrige Angaben als Täuschung behandelt werden.

B) Erklärung zur Veröffentlichung von Bachelor- oder Masterarbeiten

Die Abschlussarbeit wird zwei Jahre nach Studienabschluss dem Archiv der Universität Bremen zur dauerhaften Archivierung angeboten. Archiviert werden:

1) Masterarbeiten mit lokalem oder regionalem Bezug sowie pro Studienfach und Studienjahr 10 % aller Abschlussarbeiten
2) Bachelorarbeiten des jeweils ersten und letzten Bachelorabschlusses pro Studienfach u. Jahr.

☐ Ich bin damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

☐ Ich bin damit einverstanden, dass meine Abschlussarbeit nach 30 Jahren (gem. §7 Abs. 2 BremArchivG) im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

☐ Ich bin nicht damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

C) Einverständniserklärung über die Bereitstellung und Nutzung der Bachelorarbeit / Masterarbeit / Hausarbeit in elektronischer Form zur Überprüfung durch Plagiatssoftware

Eingereichte Arbeiten können mit der Software *Plagscan* auf einen hauseigenen Server auf Übereinstimmung mit externen Quellen und der institutionseigenen Datenbank untersucht werden. Zum Zweck des Abgleichs mit zukünftig zu überprüfenden Studien- und Prüfungsarbeiten kann die Arbeit dauerhaft in der institutionseigenen Datenbank der Universität Bremen gespeichert werden.

☐ Ich bin damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum Zweck der Überprüfung auf Plagiate auf den *Plagscan*-Server der Universität Bremen hochgeladen wird.

☐ Ich bin ebenfalls damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum o.g. Zweck auf dem *Plagscan*-Server der Universität Bremen hochgeladen u. dauerhaft auf dem *Plagscan*-Server gespeichert wird.

☐ Ich bin nicht damit einverstanden, dass die von mir vorgelegte u. verfasste Arbeit zum o.g. Zweck auf dem *Plagscan*-Server der Universität Bremen hochgeladen u. dauerhaft gespeichert wird.

Mit meiner Unterschrift versichere ich, dass ich die oben stehenden Erklärungen gelesen und verstanden habe. Mit meiner Unterschrift bestätige ich die Richtigkeit der oben gemachten Angaben.

Datum, Ort _____ Unterschrift _____