Adaptive group sequential designs with control of the population-wise error rate

Eine dem
Fachbereich 03: Mathematik/Informatik
der Universität Bremen
eingereichte Dissertation zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat)

von
Charlie Hillner
geb. am 16.09.1992 in Delmenhorst

Einreichung am: 15.08.2021
Kolloquium am: 15.09.2021

Erstgutachter: Prof. Dr. Werner Brannath
Zweitgutachter: Prof. Dr. Frank Bretz

Acknowledgements

I would like to thank my family and friends for motivating and emotionally supporting me during my work at my PhD-thesis.

I would also like to thank Prof. Dr. Werner Brannath, who always took his time to answer any type of question in great detail and who kept motivating me to not give up and present my results to a larger audience, especially in times where I doubted myself.

Abstract

The aim of individualized medicine is to provide each patient with a therapy tailored to his or her genetic profile. This is particularly important in diseases where the efficacy of a treatment depends on various individual-specific factors. Especially in rarer diseases or in highly stratified patient populations, proof of superiority of a new therapy may now prove difficult to achieve, as the necessary test power cannot be reached due to too small sample sizes. A good example is the field of pediatric oncology, where individualization of therapies plays an increasingly important role, but the underlying study populations are so limited that proof of superiority of therapy and stratification strategies is hardly possible under classical statistical principles. The aim of this work is to combine the flexibility of adaptive designs for clinical trials with the new requirements and dynamic development in individualized medicine. For this purpose, situations are considered in which the superiority of potentially different treatments is to be investigated in different, not necessarily disjoint subgroups of an overall population. In particular, these subgroups may thus be overlapping or nested. Since a multiplicity problem arises from testing several hypotheses on partly the same data material, but the family-wise error rate (FWER) often used here is too conservative, a new, less conservative multiple type I error criterion tailored to the particular subgroup structures is used in this work. This error criterion, termed the population-wise error rate (PWER), will be used as the basis for developing new multiple, sequential, and adaptive trial designs for testing individualized therapies. Specifically, single-stage test designs with PWER control were first developed and compared with corresponding FWER-controlling designs using various special cases. Next, group sequential designs controlling for PWER were constructed, here adapting various methods from the classical theory of group sequential designs. Last, adaptive designs with PWER-control were conceived and tested in numerical examples and simulations.

Kurzfassung

Die individualisierte Medizin hat zum Ziel, jedem Patienten die auf sein genetisches Profil zugeschnittene Therapie zu geben. Dies ist insbesondere bei Krankheiten von Bedeutung, bei welchen die Wirksamkeit einer Behandlung von verschiedenen individuenspezifischen Faktoren abhängen. Gerade bei selteneren Krankheiten oder in stark stratifizierten Patientenpopulationen kann sich nun ein Überlegenheitsnachweis einer neuen Therapie als schwer möglich gestalten, da die nötige Teststärke aufgrund zu geringer Stichprobengrößen nicht erreicht werden kann. Ein gutes Beispiel ist der Bereich der pädiatrischen Onkologie, in dem die Individualisierung von Therapien eine immer stärkere Rolle spielt, die zugrunde liegenden Studienpopulationen aber so beschränkt sind, dass ein Überlegenheitsnachweis von Therapie- und Stratifizierungsstrategien unter klassischen statistischen Prinzipien kaum noch möglich ist. Ziel dieser Arbeit ist es, die Flexibilität adaptiver Designs für klinische Studien mit den neuen Anforderungen und dynamischen Entwicklung in der individualisierten Medizin zu kombinieren. Hierzu werden Situationen betrachtet, in denen die Überlegenheit potentiell verschiedener Behandlungen in verschiedenen nicht notwendigerweise disjunkter Subgruppen einer Gesamtpopulation untersucht werden soll. Insbesondere können diese Subgruppen also überlappen oder ineinander genestet sein. Da durch das Testen mehrerer Hypothesen am teilweise gleichen Datenmaterial ein Multiplizitätsproblem entsteht, die hierbei oftmals herangezogene family-wise error rate (FWER) aber zu konservativ ist, wird in dieser Arbeit ein neues, weniger konservatives multiples Typ-I-Fehlerkriterium, welches auf die besonderen Subgruppenstrukturen zugeschnitten ist, herangezogen. Dieses Fehlerkriterium, das als population-wise error rate (PWER) bezeichnet wird, soll als Basis für die Entwicklung neuer multipler, sequentieller und adaptiver Studiendesigns zur Prüfung individualisierter Therapien benutzt werden. Konkret wurden zunächst einstufige Testdesigns mit PWER-Kontrolle entwickelt und anhand verschiedener Spezialfälle mit entsprechenden FWER-kontrollierenden Designs vergleichen. Anschließend wurden gruppensequentielle Designs konstruiert, welche die PWER kontrollieren, wobei hier verschiedene Methoden aus der klassischen Theorie der gruppensequentiellen Designs adaptiert wurden. Zuletzt wurden adaptive Designs mit PWER-Kontrolle konzipiert und in numerischen Beispielen und Simulationen getestet.

Contents

List of Figures

List of Tables

Introduction

Clinical trials assess and verify the effectiveness of drugs or treatment strategies with the aid of statistical methods. Often times the pursued goal is to prove the effectiveness in a very broad patient population, so that on average as many future patients as possible can benefit from this new treatment method. Inevitably, some portion of the overall population will then benefit from the new treatment and some will not. So an effective and safe treatment for the, so to say, 'average patient', is the main focus. While this is generally a valid approach for diseases with large patient populations whose progression is mostly unaffected by a patient's individual characteristics like their genome, their age or their environment, it might be unfeasible for diseases that are vastly dependent on these very individual characteristics. An initiative called *personalized medicine* (also referred to as *precision medicine* or *individualized medicine*) is based on the fact that the efficacy of these treatments can of course still differ on an individual patient basis, e.g. due to the genetic variability of the individuals within the overall population, and that each patient should therefore ideally be given the treatment tailored to his or her individual patient profile. Especially in the field of oncology, this variability is a key problem. Nowadays cancer is understood as a complex genetic disease caused by gene mutations. Now, a treatment's efficacy for a patient who suffers from a certain cancer type (say, colon or breast cancer) may differ from that for another patient with the same cancer type because of a different mutation causing the cancer. Therefore, stratified therapeutic approaches are necessary to find an effective and safe treatment for each individual patient. So, in general, there is a need for trials that investigate one or more treatments in multiple different sub-populations. Different new types of trials that all belong to the so-called 'Master protocols' (see e.g. Woodcock and LaVange, 2017, [57]) emerged in the past decade in order to reach this goal. Examples for such trials are umbrella and basket trials, where in an umbrella trial patients with the same cancer type but different molecular alterations are enrolled and the treatments are tailored to the specific target sub-populations, while in a basket trial patients with different cancer types but one common molecular alteration are enrolled with the aim to study one targeted treatment [57, 50].

One issue that inevitably comes to mind is the need for multiplicity adjustments when multiple treatments are investigated in multiple subgroups. The term multiplicity generally describes the phenomenon that the probability of making any wrong decision regarding the rejection of the null hypotheses increases, the more hypotheses are being tested on the same dataset. Now, in some cases, the overall patient population may be structured in a way that only certain parts (subgroups) of the population can be affected by multiple wrong decisions. In the FOCUS4 study (Kaplan et al., 2013, [27]), for instance, biomarker tests were conducted to define subgroups based on the mutations present in the patients' tumour DNA. Some

patients belonged to more than one subgroup and thus the subgroups were made disjoint by means of a hierarchical ordering structure defined for the different mutations and therefore, there is no further need to adjust for multiplicity anymore. In this thesis, however, we explicitly want to allow biomarker-defined sub-populations to have an overlap implying that patients belonging to such an overlap become eligible for multiple targeted therapies. Consequently, this means that future patients of the overlap may be exposed to more than a single inefficient treatment by the trial results. Also, one has to think of a way to appropriately allocate patients belonging to an overlap of sub-populations. Issues of eligibility for multiple target therapies have been addressed e.g. in Malik et al. (2014, [37]), Collignon et al. (2020, [18]) and Kesselmeier et al. (2020, [28]).

Typically, in conformatory trials with tests of multiple hypotheses, the family-wise error rate (FWER), i.e. the probability of rejecting at least one true null hypothesis, is used to control the multiple type I error rate at a reasonably small level. When dealing with studies or sequences of studies with small and/or highly stratified therapy groups, classical multiple testing methods with strong control of the FWER can be too restrictive, too conservative and too inflexible for a consistently successful application. Especially in the field of paediatric oncology, where small therapy groups are very common, a more liberal approach for controlling the multiple type I error rate is therefore desirable (e.g. Sposto et al., 1999, [48] and Fletcher et al., 2018, [21]). To meet these growing demands in precision medicine a little better, we consider a new multiple type I error rate called *population-wise error rate (PWER)* which has already been proposed in [12]. To motivate this error rate, consider the following example.

Let us assume, for illustration, a study demonstrating the efficacy of a specific therapy in two different patient populations, defined by two different biomarkers. Assume that the efficacy of that therapy is tested in each of the two populations by means of a hypothesis test. If the populations are disjoint no multiplicity adjustments are needed since different observations are used for different hypotheses tests (e.g. Collignon et al., 2020, [18]). If the populations are intersecting, though, i.e. if there are patients that, for example, belong to both biomarker-groups, then those patients can potentially be affected by more than one falsely rejected null hypothesis and therefore multiplicity adjustments have to be made for those patients [18]. The remaining patients in the complements can only be affected by at most one erroneous rejection. Regarding error rate control, the FWER does not account for the possibility that only a certain portion of the whole population can be affected by multiplicity and would therefore be too conservative in this case. To control the PWER, on the other hand, we would basically compute an individual FWER-expression for each of the three disjoint subgroups (the intersection and the two complements), weigh them with their respective population size relative to the full population and would then eventually form a weighted mean of family-wise error rates that only concern the respective subgroups to account for the whole population structure. A rejection of the null hypothesis in the first population would only concern patients that are either exclusively in that population or in the intersection. Those who are only in the second population (and not in the intersection) should not be affected by the type I error. The FWER would not account for that whereas the PWER would.

In this thesis, we want to formally describe the PWER and propose single-stage, group-sequential and adaptive designs that control the PWER. In various examples, we will compare PWER-control and FWER-control with respect to different mea-

sures like the sample size, power and type I error rate. In detail, in Chapter 1 we will give a short overview of the mathematical background needed to understand the rest of the thesis, namely, multiple testing theory and adaptive group sequential designs both on their own and also combined. In Chapter 2 the PWER will be introduced formally, some examples for some potentially practically relevant population structures are given and a power concept based on the PWER is proposed. Chapter 3 then deals with single-step single-stage designs with PWER-control which are illustrated by means of the special case of two intersecting populations. Some parts of this chapter have also already been covered by [12] in a more specific statistical setting. Afterwards, group sequential designs with PWER control, which allow for testing certain pre-defined sets of hypotheses at each stage, are presented. Here, already existing concepts from group sequential theory like the use of the Wang & Tsiatis power family or the error spending approach are adjusted to our PWER-concept. At last, in Chapter 5, we will present an adoption of the CRP-principle by Müller & Schäfer (2004,[41]) that leads to PWER-control. This concept will be illustrated by examples and simulations.
All methods and simulations are written down in R-script files which can be found in the CD attached at the back of the thesis or in the github-repository at the URL https://github.com/chillner/RCode-PhD-Thesis.

1. Mathematical background

In this first chapter we intend to go through the key concepts needed to understand the content of this thesis. Naturally, since this thesis is about the application of a new multiple type I error rate concept in adaptive group sequential designs, this chapter will cover basics in multiple testing theory and adaptive designs. While the multiple testing part is mainly based on [20] and [19], the information about adaptive designs is taken from [55].

1.1 Multiple testing theory

First, we will give a brief introduction to some of the most important concepts of multiple testing theory. The key problem a multiple testing procedure tries to solve is type I error rate inflation. The type I error rate originally stems from the problem of trying to control the probability of falsely rejecting one single hypothesis. Say it is planned, for instance, to compare the mean efficacy of a treatment θ_T to that of a placebo θ_C by means of testing the one -sided hypotheses pair

$$H : \theta_T - \theta_C \leq \delta, \quad K : \theta_T - \theta_C > \delta$$

where $\delta \in \mathbb{R}_{\geq 0}$ is some pre-defined constant. Assuming a continuous endpoint a standard two-sample t-test would control the type-I-error rate at a level $\alpha \in (0, 1)$ by simply choosing the $(1 - \alpha)$-quantile of the given t-distribution under the null hypothesis. Now, say if we compared $m > 1$ different doses of said treatment with the same control by means of tests of

$$H_i : \theta_{T,i} - \theta_C \leq \delta, \quad K : \theta_{T,i} - \theta_C > \delta$$

with $\theta_{T,i}$ now being the true mean efficacy of the ith dose level of treatment T, $i = 1, \ldots, m$, this collection of hypotheses is considered a *family* of hypotheses. That is, if a type-I-error for at least one of the hypotheses in $\{H_i \mid i = 1, \ldots, m\}$ is made, it is seen as an incorrect decision. This motivates the so-called *family-wise error rate*, which is the probability of rejecting at least one true null hypothesis from a family of hypotheses $\mathcal{H} = \{H_1, \ldots, H_m\}$. In general, when dealing with the simultaneous test of multiple hypotheses, the control of the family-wise error rate

	No. of hypotheses m				
α	1	2	4	6	10
0.025	0.025	0.049	0.096	0.141	0.224
0.05	0.05	0.098	0.185	0.265	0.401
0.1	0.1	0.19	0.344	0.469	0.651

Table 1.1: Inflation of the type I error rate depending on the number of simultaneously tested hypotheses m and the significance levels used for each test. The test statistics are assumed to be independent on each other.

is aimed for. Controlling this error rate generally means that each hypothesis H_i, $i = 1, \ldots, m$, cannot be tested with a level α test anymore if family-wise error rate control at level α is desired. This can easily be seen by a simple example: Suppose each of the m hypotheses is tested with a level α test and that the test statistics in use are all stochastically independent. Under the assumption that all null hypotheses are true, the probability of rejecting at least one null hypothesis is then given by $1 - (1 - \alpha)^m$. Table 1.1 shows that even for smaller numbers of hypotheses like $m = 2$ or $m = 4$ the probability of making *any* type I error is significantly inflated. Using $\alpha = 5\%$ and $m = 10$ hypotheses it even goes up to around 40%. Clearly, this is not a desired way of testing multiple hypotheses at once. Significance levels for each hypothesis test will have to be stricter (smaller than α) to obtain family-wise error control.

To describe the matter more formally, we first need a handful definitions. Let $(\Omega', \mathcal{A}', \mathbb{P})$ be a probability space with some non-empty set Ω', a σ-algebra $\mathcal{A}'$ on Ω' and a probability measure $\mathbb{P} : \mathcal{A}' \to [0, 1]$. Now, define a random variable $X : (\Omega', \mathcal{A}', \mathbb{P}) \to (\Omega, \mathcal{A})$, where $\Omega := X(\Omega')$ contains all possible realizations of X and $\mathcal{A} \subseteq 2^\Omega$ denotes a σ-algebra on Ω such that $(\Omega, \mathcal{A})$ constitutes a measurable space. In statistical inference, we now search for the distribution of X, which we denote as $\mathbb{P}^X = \mathbb{P} \circ X^{-1} : \mathcal{A} \to [0, 1]$. We assume a parametric distribution for X, so we assume that $\mathbb{P}^X$ lies in some set $\mathcal{M}_\Theta := \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^m$ with $m \in \mathbb{N}$. That is, we assume $\mathbb{P}^X$ to be equal to some distribution depending on an (to us) unknown parameter $\theta \in \Theta$. In order to create a framework allowing us to make decisions on the true shape of $\mathbb{P}^X$, we need the definition of a *statistical model*.

Definition 1.1.1 (Statistical model)**.** *A triple $(\Omega, \mathcal{A}, \mathcal{M}_\Theta)$ containing of a non-empty set Ω, a σ-Algebra $\mathcal{A} \subseteq 2^\Omega$ on Ω and a (here parametrized) family of distributions $\mathcal{M}_\Theta$ is called* (parametrized) statistical model *with parameter space Θ.*

Based on a statistical model, it is possible to make statistical inference by means of data we collected in a statistical experiment. Let us denote this data as $x \in \Omega$. Since x is in Ω it is a realization of the random variable X. A decision is now made using a test decision function.

Definition 1.1.2 (Statistical test)**.** *A measurable function $\varphi : \Omega \to \{0, 1\}$ is called a (non-randomized) statistical test.*

Traditionally, if $\varphi(x) = 1$, then the corresponding null hypothesis is rejected, otherwise if $\varphi(x) = 0$, it is said to be accepted/not rejected. That is, a type-I-error

occurs, whenever $\varphi(x) = 1$ and the null hypothesis is actually true. A type-II-error happens, if $\varphi(x) = 0$ and the null hypothesis is actually false.
As already touched on in the introductory paragraph, the use of the type-I-error rate for testing one hypothesis is inappropriate when simultaneously testing more than one hypothesis because the family-wise type I error rate is generally inflated. Other reasons as to why handling multiple testing procedures is a non-trivial endeavor are that the involved statistics for conducting the simultaneous tests are typically stochastically dependent on each other and that their joint distribution function is often hard or impossible to compute analytically. Therefore, a theory for multiple tests is needed, from which we would like to present an excerpt that is relevant to the subsequent sections of this thesis here. We start by defining what a multuple testing problem is.

Definition 1.1.3. *(Multiple testing problem)*

(i) A multiple testing problem is a four-tuple $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$, where $\mathcal{H} = \{H_i \mid i \in I\}$ is the set of null hypotheses $H_i \subseteq \Theta$ with index set $I = \{1, \dots, m\}$ for $m \in \mathbb{N}$. For a hypothesis H_i we say it to be true whenever $\theta \in H_i$ and denote the index set of all true null hypotheses as $I(\theta) = \{i \in I \mid \theta \in H_i\}$.

(ii) The intersection hypothesis $H_0 := \bigcap_{i \in I} H_i$ is called global null hypothesis.

(iii) We denote a non-randomized test for $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$ as $\boldsymbol{\varphi} = (\varphi_i)_{i \in I} : \Omega \to \{0,1\}^m$ and call it a multiple test. *So $\varphi_i(x) = 1$ if and only if H_i is rejected.*

The main problem the multiple testing theory is concerned with is how to control different types of type-I-errors. Two of the most important multiple type I error rates are given in the following definition.

Definition 1.1.4 (FWER & FDR)**.** *Let $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$ be a multiple test problem and $\boldsymbol{\varphi}$ a multiple test for $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$. Further, let $V \equiv V(\theta)$ denote the random variable that counts the number of false rejections made when testing all H_i, $i \in I$, and $R \equiv R(\theta)$ the random variable counting the total amount of rejections made (regardless of them being correct or incorrect).*

(i) The family-wise error rate (FWER) *is defined as the probability of rejecting at least one true null hypothesis from $\mathcal{H}$:*

$$FWER_\theta(\boldsymbol{\varphi}) := \mathbb{P}_\theta\left(V > 0\right) = \mathbb{P}_\theta\left(\bigcup_{i \in I(\theta)} \{\varphi_i = 1\}\right) \tag{1.1}$$

(ii) The false discovery rate *is the expected proportion of the falsely rejected hypotheses in relation to the number of all rejected hypotheses:*

$$FDR_\theta(\boldsymbol{\varphi}) := \mathbb{E}_\theta\left(\frac{V}{\max(R, 1)}\right) \tag{1.2}$$

One can easily show that the FDR is always more liberal than the FWER.
As already mentioned in the introduction, we will focus on comparisons with the FWER in this thesis, but will mention some potential uses of the FDR in the Conclusion as well.
Typically, there are considered two different ways of controlling the FWER.

Definition 1.1.5 (Weak and strong FWER-control)**.** *A test $\boldsymbol{\varphi}$ for a multiple test problem $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$* weakly *controls the FWER at a level $\alpha \in (0,1)$ if*

$$\forall \theta \in H_0 : FWER_\theta(\boldsymbol{\varphi}) \leq \alpha. \tag{1.3}$$

$\boldsymbol{\varphi}$ strongly *controls the FWER at level α if*

$$\sup_{\theta \in \Theta} FWER_\theta(\boldsymbol{\varphi}) \leq \alpha \tag{1.4}$$

Obviously, (1.4) is equivalent to $\text{FWER}_\theta(\boldsymbol{\varphi})$ being bounded by α for all $\theta \in \Theta$ and thus strong control always implies weak control.

Now, for the remainder of this section, we will list some FWER-controlling multiple test procedures. Dickhaus (2014) [19] lists mainly three classes of multiple test procedures,

- margin-based multiple test procedures,
- multivariate multiple test procedures and
- closed test procedures.

Margin-based multiple test procedures encompass procedures where the significance levels of the marginal tests φ_i of $\boldsymbol{\varphi}$ can individually be chosen such that some multiple type I error rate is controlled. These tests can basically be subdivided into three further categories, namely, single-step procedures, step-wise rejective procedures and data-adaptive procedures, of which only the first type will be of interest in this thesis. Multivariate multiple test procedures use the dependency structure of the data, i.e. the joint distribution of all test statistics, to optimize the procedure's power. Here, methods based on resampling, central limit theorems or copulae are to be named. Closed test procedures are a class of multiple tests that are coherent by construction. For the first and third class of tests, we now give some of the most prominent examples.

In a single step procedure, a universal local significance level $\alpha^* = \alpha^*(\alpha)$ is chosen such that the FWER is bounded by α. Here, each marginal test φ_i, $i \in I$, is then conducted as a level α^* test. Two of the most popular methods are the so-called Bonferroni-correction (1936, [9]) and the Šidàk-correction (1967, [58]).

Example 1.1.1 (Bonferroni- & Šidàk-correction)**.** *The* Bonferroni-correction *uses $\alpha^* = \alpha/m$ as the local siginifance level for each φ_i, $i \in I = \{1, \ldots, m\}$, that is, for each φ_i and each $\theta \in \Theta$ it holds $\mathbb{P}_\theta(\{\varphi_i = 1\}) \leq \alpha/m$. This procedure then yields strong FWER-control which can easily be seen by applying the Bonferroni-inequality:*

$$FWER_\theta(\boldsymbol{\varphi}) = \mathbb{P}_\theta\left(\bigcup_{i \in I(\theta)} \{\varphi_i = 1\}\right) \leq \sum_{i \in I(\theta)} \underbrace{\mathbb{P}_\theta(\{\varphi_i = 1\})}_{\leq \alpha/m} \leq \alpha$$

The Šidàk-correction *chooses $\alpha^* = 1 - (1-\alpha)^{1/m}$ for each marginal test. This also yields strong FWER-control if the test statistics used for each marginal test are stochastically independent. There are also cases where strong error control is achieved if the test statistics are dependent (cf. [19], Ch.4).*

One of the most influential classes of FWER-controlling tests are closed testing procedures which were introduced by Marcus et al. (1976, [38]). The following theorem is first needed before these testing procedures can be introduced. It basically states that as long as the hypotheses are in an intersection closed system any coherent multiple test for a given multiple testing problem at a local level α will lead to strong FWER-control at level α.

Theorem 1.1.1 (Dickhaus 2014 [19]). *Let $\mathcal{H} = \{H_i \mid i \in I\}$ be a $\cap$-closed system of hypotheses and $\boldsymbol{\varphi} = (\varphi_i)_{i \in I}$ be a coherent multiple test for $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$ at a local level α. Then, $\boldsymbol{\varphi}$ is a strongly FWER-controlling multiple test at FWER level α for $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$.*

Proof. See [19]. □

Now, since coherence is a needed property for the above theorem to hold, a general construction principle for coherent multiple test procedures was found [38, 47].

Theorem 1.1.2 (Closure principle). *Let $\mathcal{H} = \{H_i \mid i \in I\}$ denote a $\cap$-closed system of hypotheses and $\boldsymbol{\varphi} = (\varphi_i)_{i \in I}$ an (arbitrary) multiple test for $(\Omega, \mathcal{A}, \mathcal{M}_\Theta, \mathcal{H})$ at local level α. Then, we define the* closed multiple test procedure (closed test) *$\tilde{\boldsymbol{\varphi}} = (\tilde{\varphi}_i)_{i \in I}$ based on $\boldsymbol{\varphi}$ by*

$$\forall i \in I : \tilde{\varphi}_i(x) = \min_{j : H_j \subseteq H_i} \varphi_j(x). \tag{1.5}$$

It holds: The closed test $\tilde{\boldsymbol{\varphi}}$ is coherent and strongly controls the FWER at level α.

Proof. Immediately follows by construction of $\tilde{\boldsymbol{\varphi}}$ and Theorem 1.1.1. □

The closure principle implies that such a test $\tilde{\boldsymbol{\varphi}}$ rejects some hypothesis $H_i \in \mathcal{H}$ if and only if all hypotheses $H_j \in \mathcal{H}$ that are a superset of H_i are rejected by $\boldsymbol{\varphi}$. One 'problem' of the closure principle is that it requires an intersection closed set of hypotheses. If only a set $\mathcal{H}' := \{H_i \mid i \in I\}$ without this property is available, one can extend this family to an intersection closed one by simply taking all $2^{|I|} - 1$ combinations of intersections of the elements in $\mathcal{H}$. Obviously, the number of involved hypotheses increases drastically with increasing size of I. In many special cases it is unnecessary to test all $2^{|I|} - 1$ intersection hypotheses, though. There are also works among the multiple testing literature (e.g. Brannath & Bretz [11]) that found ways to drastically reduce this number when certain additional conditions hold.

1.2 Adaptive group sequential designs

Another type of multiplicity that has yet been untouched arises when a hypothesis is tested sequentially. Sequential analysis can be dated back to Abraham Wald (1945) [53] and was concerned with testing a hypothesis with a sample size that has not been fixed in advance. A hypothesis has more so been tested as soon as the data has been collected and certain pre-defined stopping rules have been imposed to define a point at which further sampling should stop. So a sequential test tests a hypothesis several times after more and more data comes in. In 1969

(and also later in 1971 and 1975) Peter Armitage first made use of this concept in the area of clinical trials [3, 4, 40] by introducing a recursive integration formula which constitutes the building block for repeated significance tests and consequently also classical group sequential clinical trial designs. Repeated significance testing was involved with finding suitable decision regions regions ensuring that the overall type I error probability does not exceed a pre-specified α, which is possible if the number of times new data is accumulated is fixed in advance. Using the same rejection boundary every time the hypothesis is retested leads to an inflation of the type I error rate. Group sequential designs collect groups of new data at certain pre-defined schedules that are also called stages. The total number of stages is generally fixed in advance which contrasts classical sequential plans with a continuous sampling scheme. Later, key contributions from Pocock and O'Brien and Fleming further boosted the implementation of group sequential testing in medical research and many other extensions and refinements followed and are still following nowadays. The major advantage of a group sequential design over a fixed sample size design is that the total sample size actually needed until rejection is much lower meaning they are ethically and financially more justifiable. But, since group sequential designs require a fixed collection of hypotheses, stages and also sample size in a sense, there has been a major influx of theory about so-called adaptive designs starting from the 1990s. These designs allow mid-trial adjustments like a deviation from the originally planned sample sizes, an introduction or deletion of a hypothesis or a change of the whole trial design in general without harming the overall validity of the trial, that is, without inflating the overall type I error rate. To achieve this, several methods have been proposed. Among those, two of the most influential ideas have been the use of *Combination functions* [5], which suitably combine stage-wise p-values such that the overall type I error is still bounded by α, and *Conditional error functions* [43], which define new significance levels conditional on all data collected before a mid-trial adaptation has been conducted. Conditional error functions in particular are also a key component of the so-called *Conditional rejection probability principle* by Müller & Schäfer [41] who proposed a principle allowing for mid-trial changes at any time during a trial.
In this section we will present some of the most important concepts needed to understand adaptive group sequential designs. The general layout of the presented information is mainly based on Jennison & Turnbull [26], Wassmer & Brannath [55] and other literature that will be cited if needed.

1.2.1 Group sequential designs

First, we will introduce the concept of classical group sequential designs. To this end, suppose we are about to test a null hypothesis H_0 against its alternative H_1. For simplicity, assume that our observations are realizations of independent normally distributed random variables $X_1, \ldots, X_n$ with unknown expectation θ and known variance σ^2 and that a treatment is tested against a control. The (continuous) true efficacy difference is given by a parameter $\theta \in \Theta \subseteq \mathbb{R}$. We are interested in testing a hypothesis pair out of the following three,

$$\begin{aligned} H_0 : \theta = \theta_0 \quad &\text{vs.} \quad H_1 : \theta \neq \theta_0 \\ H_0 : \theta \leq \theta_0 \quad &\text{vs.} \quad H_1 : \theta > \theta_0 \\ H_0 : \theta \geq \theta_0 \quad &\text{vs.} \quad H_1 : \theta < \theta_0 \end{aligned}$$

with $\theta_0 \in \mathbb{R}$ being some pre-specified constant. Typically, one would reject H_0 at a level α if and only if the absolute value of the observed value of the test statistic

$$Z = \frac{\bar{X} - \theta_0}{\sigma}\sqrt{n} \overset{H_0}{\sim} \mathrm{N}(0,1)$$

is larger than the critical value $c = \Phi(1-\alpha/2)$, with Φ denoting the cdf of the standard normal distribution. In group sequential designs, however, data is evaluated after a certain group of observations is available. These so-called *interim analyses* are done at $K > 1$ stages and let n_k denote the number of observations collected at stage $k = 1, \ldots, K$. Also, let $N = \sum_{k=1}^{K} n_k$ denote the maximum sample size of the test. So in this group sequential setting we have the normally distributed, independent observations $X_{k,i}$ where $k = 1, \ldots, K$ denotes the stage and $i = 1, \ldots, n_k$ the i-th observation among the observations from stage k. To test H_0 against H_1 at stage k we use all data collected up to stage k and accumulate it into the following *accrued test statistic*,

$$Z^{(k)} := \frac{\bar{X}^{(k)} - \theta_0}{\sigma}\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}, \tag{1.6}$$

where $\bar{X}^{(k)}$ denotes the cumulative mean of the observations up to stage k and is given by

$$\bar{X}^{(k)} = \frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \sum_{\tilde{k}=1}^{k} \sum_{i=1}^{n_{\tilde{k}}} X_{k,i}.$$

We can also express $Z^{(k)}$ in terms of the *stage-wise test statistics* $\tilde{Z}^{(k)}$ by writing

$$\tilde{Z}^{(k)} = \sum_{\tilde{k}=1}^{k} \sqrt{\frac{n_{\tilde{k}}}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}}\tilde{Z}^{\tilde{k}}. \tag{1.7}$$

Note that while the statistics $Z^{(1)}, \ldots, Z^{(K)}$ are not independent due to consisting of data from all previous stages, the stage-wise statistics $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(K)}$ are all independent from each other. This simplifies the derivation of the covariance matrix of $\mathbf{Z} = (Z^{(1)}, \ldots, Z^{(K)})$ since one can just apply standard rules for computing covariances of linear combinations of random variables. For $k < k'$ one can easily see that

$$\mathrm{Cov}(Z^{(k)}, Z^{(k')}) = \sqrt{\frac{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}'}}} \tag{1.8}$$

which equals the correlation between $Z^{(k)}$ and $Z^{(k')}$ because the test statistics are standardized. The expectation of $Z^{(k)}$ is given by

$$\nu^{(k)} := \mathbb{E}(Z^{(k)}) = \delta\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \tag{1.9}$$

with $\delta = \frac{\theta - \theta_0}{\sigma}$ being the standardized effect size. Thus, the vector $\mathbf{Z}$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\nu} = (\nu^{(k)})_{k=1}^{K} \quad \text{and} \quad \boldsymbol{\Sigma} = (\mathrm{Cov}(Z^{(k)}, Z^{(k')}))_{k,k'=1}^{K}.$$

Now, how is the test procedure actually conducted? Generally, one defines so-called continuation regions $\mathcal{C}^{(k)}$ and rejection regions $\mathcal{R}^{(k)}$ for each stage $k = 1, \ldots, K$. For $k < K$, if the accumulated test statistic $Z^{(k)}$ lies in $\mathcal{C}^{(k)}$, then the test proceeds to its next stage without rejection of H_0, where, if it lies in $\mathcal{R}^{(k)}$, then H_0 can be rejected and the procedure stops. For $k = K$ it is $\bar{\mathcal{C}}^{(k)} = \mathcal{R}^{(k)}$ where $\bar{\mathcal{C}}$ denotes the complement of $\mathcal{C}$. The type I error quantity we want to control is the probability of falsely rejecting H_0 which equals the probability of falsely rejecting H_0 at stage 1 to K. Using the continuation and rejection regions and assuming that no stops for futility are planned we then get the requirement

$$\begin{aligned}
\mathbb{P}_{\theta_0}\left(\bigcup_{k=1}^{K}\{\text{rej. } H_0 \text{ at stage } k\}\right) &= \mathbb{P}_{\theta_0}\left(\bigcup_{k=1}^{K}\{Z^{(k)} \in \bar{\mathcal{C}}^{(k)}\}\right) && (1.10)\\
&= 1 - \mathbb{P}_{\theta_0}\left(\bigcap_{k=1}^{K}\{Z^{(k)} \in \mathcal{C}^{(k)}\}\right) && (1.11)\\
&= 1 - \int_{\mathcal{C}^{(K)}} \cdots \int_{\mathcal{C}^{(1)}} \phi(z_1, \ldots, z_K)\mathrm{d}z_1 \ldots \mathrm{d}z_K && (1.12)\\
&\overset{!}{=} \alpha. && (1.13)
\end{aligned}$$

Here ϕ denotes the probability density function (pdf) of the K-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}$ as in (1.8). The above multiple integral can, e.g., be computed by means of the recursive integration formula by [4].
If we say H_0 is rejected at stage k if the accumulated stage k test statistic exceeds (falls short of) a critical value $c^{(k)}$ for $k = 1, \ldots, K$, then the continuation regions and rejection regions can be given as

$$\begin{aligned}
\mathcal{C}^{(k)} = (-c^{(k)}, c^{(k)}) \quad &\text{and} \quad \mathcal{R}^{(k)} = \bar{\mathcal{C}}^{(k)} = (-\infty, -c^{(k)}] \cup [c^{(k)}, \infty) && (1.14)\\
\mathcal{C}^{(k)} = (-\infty, c^{(k)}) \quad &\text{and} \quad \mathcal{R}^{(k)} = \bar{\mathcal{C}}^{(k)} = [c^{(k)}, \infty) && (1.15)\\
\mathcal{C}^{(k)} = (-c^{(k)}, \infty) \quad &\text{and} \quad \mathcal{R}^{(k)} = \bar{\mathcal{C}}^{(k)} = (-\infty, -c^{(k)}] && (1.16)
\end{aligned}$$

for two-sided, upper and lower hypotheses tests, repsectively.

Wang and Tsiatis power family: The type I error rate can then be controlled at level α by finding a set of critical values $(c^{(k)})_{k=1}^{K}$. Since the requirement in (1.10) is one equation with K unknowns this solution is not unique in general. Although numerical search methods can be used to find possible values of $(c^{(k)})_{k=1}^{K}$ that solve (1.10), a different and easier to handle solution was found by parametrizing the space the critical values lie in. Wang and Tsiatis [54] suggested a class of critical values that only depends on one parameter $\xi \in \mathbb{R}$. In general, the continuation regions are given as in (1.14) where

$$c^{(k)} = c_{WT}(K, \alpha, \xi)\left(\frac{\tau_k}{\tau_1}\right)^{\xi - 0.5}, \qquad (1.17)$$

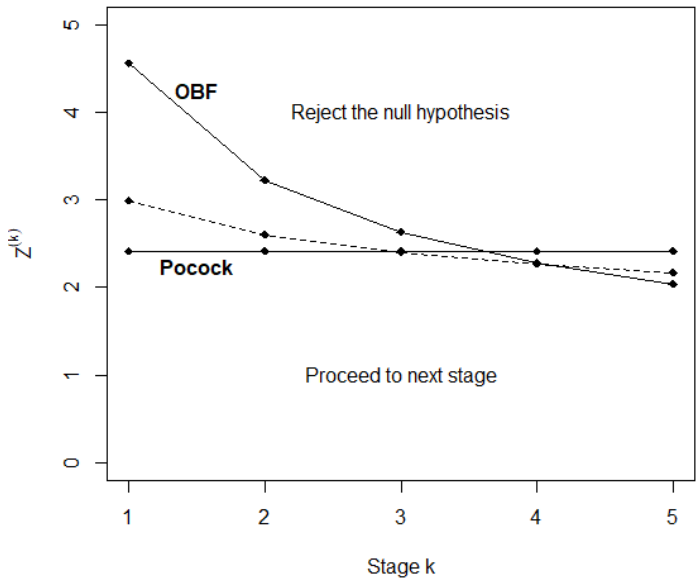

Figure 1.1: Decision regions of the upper one-sided Wang and Tsiatis test (WT) for $\xi = 0.3$ (dashed line) as compared to O'Brien and Fleming's (OBF) and Pocock's (P) design (solid lines). Dots indicate critical values under $K = 5$ and $\alpha = 0.025$ (one-sided). Picture based on [55, p. 38]

.

where $\tau_k = \sum_{\tilde{k}=1}^{k} n_{\tilde{k}}/N$ are the so-called *information rates* or *information times*, which summarize how much information relative to N is obtained up to stage k. The value $c_{WT}(K, \alpha, \xi)$ is just a constant depending on K, α and ξ and is equal to $c^{(1)}$. For equally spaced stages, i.e. if $n_1 = \cdots = n_K$, formula (1.17) simply reduces to $c^{(k)} = c_{WT}(K, \alpha, \xi)k^{\xi-0.5}$. Nonetheless, (1.17) only depends on the unknown parameter ξ and thus the type I error in (1.10) can be controlled at a level α by choosing the right value for ξ, that is, it can be controlled by solving a univariate root-finding problem in ξ. R's `uniroot`-function (`stats` package) for example uses algorithms described in [15] to find a root of a one-dimensional function.
Two of the most prominent special cases of the Wang & Tsiatis power family are the *Pocock* design, where $\xi = 0.5$ is chosen, and the *O'Brien & Fleming* design with $\xi = 0$. We see that for a K-stage Pocock design it holds $c^{(1)} = \cdots = c^{(K)}$, whereas for the O'Brien & Fleming design it is $c^{(k)} = c_{WT}\sqrt{\tau_k/\tau_1}$. So, while in a Pocock design the rejection regions stay the same for each stage, an O'Brien & Fleming design starts with a larger critical value (than that of a Pocock design) and the critical values of the subsequent stages monotonically decrease with increasing stage number. An O'Brien & Fleming design is therefore said to be stricter at the beginning stages but it becomes easier to reject the null hypothesis as more data is accrued. Designs with $\xi \in (0, 0.5)$ describe somewhat of a compromise between the Pocock and O'Brien & Fleming design as their rejection boundaries lie between those of these two designs. Figure 1.1 illustrates this behaviour for $\xi = 0.3$.

Error spending approach: One problem with the group sequential designs presented so far is that the sample sizes for each group have to be specified in advance

because the full covariance matrix of the distribution of $\mathbf{Z}$ has to be known to calculate the set of critical values. A more flexible approach that also allows for sample sizes that develop throught the course of the trial is given by the *error-spending approach* by Lan and DeMets [31] in 1983 (and also in 1987 [30]). In particular, their approach allows for interim analyses to happen at certain calendar times instead of being placed after a certain number of observations has been reached. To carry out this approach, a so-called *error-spending function* (or also alpha-spending function) $\alpha^* : [0,1] \to [0,\alpha]$ is needed which describes the cumulative type I error rate that has been spent at a certain information time $\tau \in [0,1]$ during the trial (0 being the start and 1 the end of the trial). Like in [32, 33] and other works, we will assume that the point where we are in the trial is defined by the information rates $\tau_k = \sum_{\tilde{k}=1}^{k} n_{\tilde{k}}/N$ for each $k = 1, \dots, K$. The alpha-spending function α^* can be any non-decreasing function mapping from $[0,1]$ to $[0,\alpha]$ as long as $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$. This function and the maximum sample size N have to be specified in advance. Finding critical values is done iteratively at each stage:
At stage 1 we have an information time of τ_1 and compute the first stage critical value by solving

$$P_1 := \mathbb{P}_{\theta_0}(Z^{(1)} \in \mathcal{R}^{(1)}) = \alpha^*(\tau_1). \tag{1.18}$$

Having found $c^{(1)}$ we continue at stage 2 (when reaching it) to compute $c^{(2)}$ by solving

$$P_2 := \mathbb{P}_{\theta_0}(Z^{(1)} \in \mathcal{C}^{(1)}, Z^{(2)} \in \mathcal{R}^{(2)}) = \alpha^*(\tau_2) - \alpha^*(\tau_1). \tag{1.19}$$

At stage $k > 1$ where we observe the information rate τ_k we use all previously computed critical values $c^{(1)}, \dots, c^{(k-1)}$ and find $c^{(k)}$ by solving

$$P_k := \mathbb{P}_{\theta_0}\left(\bigcap_{\tilde{k}=1}^{k-1} \left\{ Z^{(\tilde{k})} \in \mathcal{C}^{(\tilde{k})}, Z^{(k)} \in \mathcal{R}^{(k)} \right\}\right) = \alpha^*(\tau_k) - \alpha^*(\tau_{k-1}) \tag{1.20}$$

Note that this controls the overall type I error because

$$\mathbb{P}_{\theta_0}(\text{rej. } H_0 \text{ at any stage}) = \sum_{k=1}^{K} P_k = \sum_{k=1}^{K} (\alpha^*(\tau_k) - \alpha^*(\tau_{k-1})) = \alpha.$$

Example 1.2.1. *Some examples for families of error spending functions depending on a parameter ξ are:*

(i) Kim and DeMets [30] family with $\xi \in \mathbb{R}_{>0}$:

$$\alpha^*(\xi, \tau) = \alpha\tau^{\xi} \tag{1.21}$$

(ii) Hwang et al. [25] family with $\xi \in \mathbb{R}$:

$$\alpha^*(\xi, \tau) = \begin{cases} \alpha\frac{1-e^{-\xi\tau}}{1-e^{-\xi}}, & \xi \neq 0 \\ \alpha\tau, & \xi = 0 \end{cases} \tag{1.22}$$

Power and average sample size: The correct choice of the sample size in a group sequential design is a bit more complicated than with a fixed sample (1-stage) design, since we have to account for the fact that stages might be unequally spaced. In general, the power of a group sequential design under a parameter $\theta \in H_1$ is given by (again, without having any futility stops incorporated)

$$\text{Pow}_\theta = \mathbb{P}_\theta\,(\text{rej. } H_0 \text{ at any stage}) = 1 - \mathbb{P}_{\theta_0}\left(\bigcap_{k=1}^{K}\{Z^{(k)} \in \mathcal{C}^{(k)} - \nu^{(k)}\}\right) \tag{1.23}$$

with $\nu^{(k)} = \delta\sqrt{\sum_{k'=1}^{k} n^{(k)}} = \sqrt{N\tau_k}$ and $\mathcal{C}^{(k)} - \nu^{(k)} = \{c - \nu^{(k)} \mid c \in \mathcal{C}^{(k)}\}$. Here $\mathbf{Z} = (Z^{(k)})_{k=1}^{K}$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{\Sigma} = \text{Cov}(Z^{(k)}, Z^{(l)})_{k,l=1,\ldots,K}$ with correlations as in (1.8). For a given α, K, β and information rates τ_k, $k = 1, \ldots, K$, the equation $\text{Pow}_\theta = 1-\beta$ can be solved by defining the shift value ν^* such that the power with $\nu^{(k)} = \nu^*\sqrt{\tau_k/\tau_1}$, $k = 1, \ldots, K$, equals $1 - \beta$. The maximum sample size N is then equal to $N = (\nu^*/\delta)^2/\tau_1$ and the accumulated sample sizes are found via $n^{(k)} = N\tau_k$ for all k.
Since a group sequential design provides the possibility to stop early at a stage $k < K$, it is also interesting to look at the *average sample size/number* denoted as ASN_θ for a $\theta \in H_1$. If we denote N^* as a random variable with values in $[n^{(1)}, N]$, then the ASN is given by the expectation of this random variable and indicates how much sample size on average is expected to be used in this group sequential design:

$$\text{ASN}_\theta = \mathbb{E}_\theta(N^*) = n^{(1)} + \sum_{k=2}^{K} n^{(k)}\mathbb{P}_{\theta_0}\left(\bigcap_{\tilde{k}=1}^{k-1}\{Z^{(k)} \in \mathcal{C}^{(k)} - \nu^{(k)}\}\right) \tag{1.24}$$

Note that $\text{ASN}_\theta \geq n^{(1)}$ since H_0 is always tested at stage 1.

1.2.2 Adaptive designs

As already touched on in the introduction, adaptive designs allow for changes in the trial design without inflating the overall type I error rate. A very common mid-trial change is the adjustment of the sample size needed to achieve a certain power based on interim data. Normally the sample size is determined beforehand by means of the treatment effect that study should be powered for. But in many instances, prior information about this treatment effect is not compelling enough to be totally sure about its right choice. An adaptive design would allow for the use of interim data (and also other new external information) to adjust the sample size for the remaining course of the trial such that, for instance, the desired power can be reached. We have already mentioned the two main approaches in adaptive designs: combination functions and conditional error functions. Many developed principles are based on either of the two, and even though, they are conducted in a different manner, they share the same base assumptions, labelled as the *conditional invariance principle*. In the following, before presenting some of these principles for conducting adaptive designs, we need to explain the conditional invariance principle. This explanation will be mainly based on a rather heuristic description by [13]. Other mathematically deeper explanations can for example be found in [10] or [24].

Conditional invariance principle: Let us suppose a two-stage design where a null

hypothesis H_0 is to be tested, non-superiority of a treatment versus a control, for example. At stage 1, H_0 is tested as usual by means of some test statistic T_1 and the stage 2 design characteristics are planned to be based on the results of the interim analysis. If the trial proceeds to stage 2, assume that data from a group of patients independent from stage 1 has been drawn and let T_2 be the test statistic summarizing this stage 2 data. Generally, the null distribution of T_2 will depend on the interim data because the design of the second stage was chosen based on the first stage. The conditional invariance principle now states that T_2 can be transformed in a way that its null distribution conditioned on the interim data and the second stage design is equal to a fixed pre-specified null distribution, hence being invariant with respect to the interim data and any mid-trial changes [55]. Such transformations are often just transformations of T_2 to a p-value p_2 or, in case of T_2 being normally distributed, a standardization to a test statistic Z_2. Due to the invariance of the conditional null distribution of the transformed T_2 (so of p_2 or Z_2), the stage two statistic is stochastically independent from the interim data and because the null distribution of the interim data is commonly known beforehand, the joint distribution of the interim data and p_2 (or Z_2) is known as well and also invariant to the adaptation rule applied at stage 2. This yields level α rejection regions depending on the interim data and the invariant stage two test statistic, which in turn gives us a type I error controlling test procedure independent from the mid-trial adaptation rule. In general, if p_1 is the p-value for stage 1 and p_2 the stage two p-value after the adaptation has been made, Brannath et al. [14] formulated a condition these two p-values have to satisfy such that type I error control holds:

$$\mathbb{P}_{\theta_0}(p_1 \leq u) \leq u \quad \text{and} \quad \mathbb{P}_{\theta_0}(p_2 \leq u | p_1 = v) \leq u, \quad \forall u, v \in [0,1]. \tag{1.25}$$

So whenever the distribution of the p-value p_1 is stochastically larger than the uniform distribution and the conditional distribution of p_2 given p_1 is stochastically larger than the uniform as well, the above described two-stage adaptive design controls the type I error at level α. This condition is called *p-clud* (**c**onditionally **l**arger than the **u**niform **d**istribution) and generally holds if, for instance, the stage-wise observations are independent and conservative tests are used for each stage-wise p-value.

We now want to present the principles of the combination function and the conditional error function approach, respectively. Based on the latter we will then also discuss the CRP-principle by Müller and Schäfer. At last, we will briefly look into testing multiple hypotheses in an adaptive design. The following explanations will be done for two-stage adaptive designs, but general principles for more stages can be applied as well.

Combination function approach: Combination tests basically combine stage-wise test statistics (p-values) into one new quantity that can be used for a test decision after an adaptation has been conducted. We denote such a combination function as $C : [0,1]^2 \to [0,1]$ which takes the two stage-wise p-values p_1 and p_2 as its argument. The function C is chosen such that the overall type I error is still under control, even after the mid-trial adaptation. If we conduct a test design where at stage 1 H_0 is rejected if $p_1 \leq \alpha_1 < \alpha$ (where the α_1 might come from a two-stage group sequential trial, for example) and otherwise it continues to stage 2, then instead of simply comparing p_2 to some critical boundary, we take the design change at interim into

account by combining p_1 and p_2 by means of the combination function and compare that to some critical value instead. So at stage 2, one would compare $C(p_1, p_2)$ to some value c which in turn is found such that the overall type I error probability is equal to α,

$$\alpha_1 + \int_{\alpha_1}^{1} \int_{0}^{1} \mathbf{1}_{\{C(p_1,p_2)\leq c\}} \mathrm{d}p_2 \mathrm{d}p_1 = \alpha. \tag{1.26}$$

Two of the most prominent examples are the Fisher combination function [5, 6] and the inverse normal combination function [34].

Example 1.2.2. *Some examples for combination functions used for two-stage adaptive designs are:*

(i) Fisher combination function:

$$C(p_1, p_2) = p_1 p_2 \tag{1.27}$$

(ii) Inverse normal combination function:

$$C(p_1, p_2) = 1 - \Phi(w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)) \tag{1.28}$$

with weights w_1, w_2 such that $w_1^2 + w_2^2 = 1$.

One quite pleasant property of the inverse normal combination test is that boundaries from classic group sequential designs can be used. The test statistics $\tilde{Z}_i = \Phi^{-1}(1-p_i)$, $i = 1, 2$, are independent and standard normal under the assumption that p_1 and p_2 are independent and uniformly distributed. By defining $Z_2 := w_1\tilde{Z}_1 + w_2\tilde{Z}_2$, one can see that the bivariate normal distribution of $(\tilde{Z}_1, Z_2)$ is the same as that of a two-stage group sequential design with information rates $\tau_1 = \sqrt{w_1}$ and $\tau_2 = 1$. This is why critical boundaries of group sequential designs like a Pocock or O'Brien and Fleming design can simply be used for α_1 and c. Another thing to keep in mind, though, is that if w_1 and w_2 are chosen as information rates, i.e. dependant on the sample size, the weights cannot be altered even if a previously unplanned sample size recalculation for stage 2 has been done. So the weights remain fixed, but Lehmacher and Wassmer [34] have found that the implied power loss of the adaptive design is not that large (around 3%).

Conditional error function approach: The second approach was proposed by Proschan and Hunsberger in 1995 [43] and involves choosing a new conditional type I error boundary given the interim data for testing at stage 2. Suppose the same two-stage design as described in the combination function paragraph. So if $p_1 \leq \alpha_1$, the null hypothesis is rejected at stage 1 and the design stops, otherwise it proceeds to stage 2. Now, at stage 2 we use a non-decreasing function A with values in $[0, 1]$, which is called the *conditional error function* of the design. If and only if $p_2 \leq A(p_1)$, we reject H_0 at stage 2. As with the combination function, the conditional error function is also chosen such that the overall type I error is bounded by α:

$$\alpha_1 + \int_{\alpha_1}^{1} \int_{0}^{1} A(p_1) \mathrm{d}p_2 \mathrm{d}p_1 = \alpha. \tag{1.29}$$

Two error functions that are equivalent to the Fisher and inverse normal combination functions, respectively, are as follows.

Example 1.2.3. *(i) Equivalent to Fisher combination function:*

$$A(p_1) = \begin{cases} 1, & p_1 \leq \alpha_1 \\ c/p_1, & p_1 > \alpha_1 \end{cases} \tag{1.30}$$

(ii) Equivalent to inverse normal combination function:

$$A(p_1) = \begin{cases} 1, & p_1 \leq \alpha_1 \\ 1 - \Phi\left(\frac{u_{1-c} - w_1 \Phi^{-1}(1-p_1)}{w_2}\right), & p_1 > \alpha_1 \end{cases} \tag{1.31}$$

with $u_{1-c} = \Phi^{-1}(1-c)$.

Now, Müller and Schäfer proposed a general extension to this principle in 2004 [41], the so-called CRP-principle. Heuristically, the CRP-principle allows us to start with any non-adaptive design (like a classical group sequential design) and if an unforseen mid-trial change is about to be made, one can change to the test procedure of a different design that encompasses the change we want to make. In general, let φ be the test decision function of the initial non-adaptive test for H_0 which rejects H_0 if $\varphi = 1$ and accepts it if $\varphi = 0$. Now, in an interim analysis after a part of our planned-for sample size has been collected, we find out from the interim data and possibly other new external information that we want to change the design characteristics of our initial procedure and follow a different design plan at stage 2. Let us denote this new test decision function for H_0 as $\tilde{\varphi}$ and let x_1 denote the interim data from stage 1 (this could either be the stage 1 p-value p_1 or a stage 1 test statistic observation z_1, for example). Given x_1 the conditional rejection probability of the initial design φ is given by

$$A(x_1) = \mathbb{E}_{\theta_0}(\varphi | X_1 = x_1) \tag{1.32}$$

and is a known quantity. The CRP-principle now requires the conditional rejection probability $\tilde{A}(x_1)$ of the new test $\tilde{\varphi}$ has to be bounded by $A(x_1)$, i.e.

$$\tilde{A}(x_1) = \mathbb{E}_{\theta_0}(\tilde{\varphi} | X_1 = x_1) \overset{!}{\leq} \mathbb{E}_{\theta_0}(\varphi | X_1 = x_1) = A(x_1). \tag{1.33}$$

Obviously $A(x_1)$ itself satisfies condition (1.33) and thus, even if we decide to not deviate from the initially planned design φ we can just follow this old design again which by construction controls the type I error at level α. One can now show that the CRP-principle controls the overall type I error rate at level α even if a mid-trial adaptation is carried out.

Theorem 1.2.1. *Let φ and $\tilde{\varphi}$ be test decision functions of the initial and new design, respectively, and let X_1 denote the random variable with values in the same space the stage 1 interim data lies in. Let A and $\tilde{A}$ be defined as $A(X_1) = \mathbb{E}_{\theta_0}(\varphi \mid X_1)$ and $\tilde{A} = \mathbb{E}_{\theta_0}(\tilde{\varphi} \mid X_1)$ similar to (1.33). Then it holds that*

$$\mathbb{E}_{\theta_0}(\tilde{A}(X_1)) \leq \alpha. \tag{1.34}$$

Proof. Using the monotonicity of the expectation and the law of total expectation (*) we immediately get

$$\mathbb{E}_{\theta_0}(\tilde{A}(X_1)) \leq \mathbb{E}_{\theta_0}(A(X_1)) = \mathbb{E}_{\theta_0}(\mathbb{E}_{\theta_0}(\varphi | X_1)) \overset{(*)}{=} \mathbb{E}_{\theta_0}(\varphi) \leq \alpha.$$

□

The proof for this statement in a mathematically more rigorous setting can be found in [10]. One main advantage of the CRP-principle is that it can be applied to any design be it a group sequential or an adaptive one. A rather concerning disadvantage, however, is that the conditional rejection probability under the initial design has to be known and depending on the complexity of φ, this can become complicated (see [42, 23], for example).

Multiple testing in adaptive designs: So far, we have presented the concepts of adaptive group sequential designs for the special case of one null hypothesis of interest. But since, as already described in Section 1.2.1, there are many clinical studies in which multiple hypotheses are to be tested at once and one typically still wishes to keep the advantages of sequential testing and adaptive designs, there is need for an extension of the previous theory. When combining sequential testing and multiple testing theory, we have to deal with multiple types of multiplicity at once – the multiplicity that comes with testing multiple hypotheses simultaneously and the multiplicity that comes with testing these hypotheses repeatedly in an adaptive (and) group sequential manner (cf. [20] pp. 46). The difficulty now is to embed the adaptive group sequential designs into a suitable multiple testing procedure that controls some multiple type I error rate. Thanks to Theorem 1.1.2 we know that there is a way to construct coherent multiple testing procedures with strong FWER-control. Several authors like Bauer and Kieser (1999, [8]) , Bauer et al. (1999, [29]), Lehmacher et al. (2000, [35]) and Hommel (2001, [24]) worked out an implementation of a so-called 'adaptive closed test procedure' that basically works as follows: Say we are interested in testing hypotheses $H_1, \ldots, H_m$ in a two-stage design with one interim analysis. Let $\mathcal{H}$ be the $\cap$-closed set of hypotheses $\mathcal{H} = \{H_J \mid J \subseteq \{1, \ldots, m\}\}$. Then the hypothesis H_i is rejected if all H_J with $i \in J$ are rejected by the combination test or the CRP-principle at a local level α (depending on which of the two adaptive test procedures is chosen). This controls the FWER at a level α.

As an ending note for this chapter, we want to point out that the closed test procedure has only been included here for the sake of completeness. We will not explicitly apply it in this thesis because, as we will see in a subsequent chapter, there is no closed test principle for the multiple type I error measure we will investigate.

2. The population-wise error rate

This chapter is dedicated to the mathematical conceptualization, definition and interpretation of the already mentioned population-wise error rate. Several examples for practically relevant settings are given and discussed to further deepen the understanding. Moreover, population-wise and family-wise error rate are compared with each other in each respective example. Furthermore, a conceivable way to define a power measure analogue to the population-wise error rate is presented.

2.1 Motivation

Suppose we have a $m \geq 1$ different populations $\mathcal{P}_1, \ldots, \mathcal{P}_m$ with some of them overlapping with each other and let $\mathcal{P} = \bigcup_{j=1}^{m} \mathcal{P}_j$ be the full population. The subpopulations $\mathcal{P}_j$, $j \in I := \{1, \ldots, m\}$, could, for instance, have been defined by certain patient characteristics which in turn are defined by different biomarkers. To define the population-wise error rate we need to partition the full population $\mathcal{P}$ into disjoint subpopulations, that is, let

$$\left\{ \mathcal{P}_J = \bigcap_{j \in J} \mathcal{P}_j \setminus \bigcup_{k \in I \setminus J} \mathcal{P}_k : J \subseteq I \right\} \tag{2.1}$$

be a partition of $\mathcal{P}$, that is $\mathcal{P} = \uplus_{J \subseteq I} \mathcal{P}_J$. Here $A \uplus B$ denotes the union of two disjoint sets A and B. An example for $m = 4$ intersecting populations can be seen in Figure 2.1. Furthermore, for each $J \subseteq I$ let π_J be the *relative population size (or prevalence) of* $\mathcal{P}_J$, where $\sum_{J \subseteq I} \pi_J = 1$. For convenience, we will write the collection of all π_J, $J \subseteq I$, as a vector $\boldsymbol{\pi} = (\pi_J)_{J \subseteq I}$. For the case in Figure 2.1, a value of, say, $\pi_{\{1,2,3\}} = 0.2$ would mean that $\mathcal{P}_{\{1,2,3\}} = (\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3) \setminus \mathcal{P}_4$ makes up 20% of the whole population $\mathcal{P}$. Note that these population sizes are relative to $\mathcal{P}$ and are typically unknown quantities in practice for which an estimator has to be used. For now, we will treat them as known, however, and will discuss this matter in a later section.

For the sake of simplicity, let us first assume that a test φ_j of a hypothesis $H_j : \theta_j \leq 0$ is conducted in $\mathcal{P}_j$ for each $j = 1, \ldots, m$, respectively, where $\theta_j \in \mathbb{R}$ denotes the

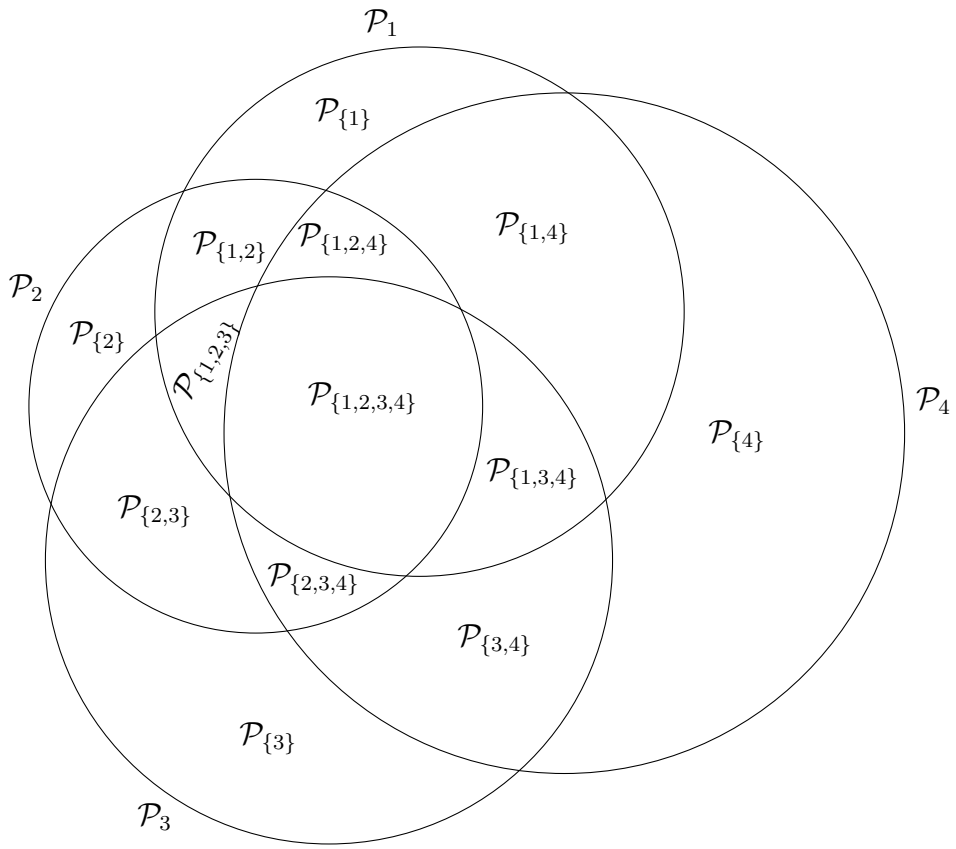

Figure 2.1: Partition $\{\mathcal{P}_J \mid J \subseteq I\}$ of the union of four overlapping populations $\mathcal{P}_j$, $j \in I = \{1, \ldots, 4\}$. Since all populations have a non-empty intersection with each other, there are $2^4 - 1 = 15$ disjoint subgroups $\mathcal{P}_J$ the union $\bigcup_{j=1}^{4} \mathcal{P}_j$ can be decomposed into.

parameter of interest for $\mathcal{P}_j$ (e.g. difference between a treatment and a control). As already outlined in the introduction, heuristically, we want the population-wise error rate to be defined as a weighted sum over different family-wise error rate expressions that concern an erroneous rejection made in $\mathcal{P}_J$, i.e. we aim for an expression resembling

$$\text{PWER} = \sum_{J \subseteq I} \pi_J \text{FWER}_{\mathcal{P}_J}. \tag{2.2}$$

Here $\text{FWER}_{\mathcal{P}_J}$ denotes the probability of erroneously rejecting at least one null hypothesis H_j that concerns patients from $\mathcal{P}_J$ – a family-wise error rate for patients in $\mathcal{P}_J$, so to say. Now what exactly do we mean by saying 'an erroneous rejection of a null hypothesis H_j that concerns patients in $\mathcal{P}_J$'? Let us assume that a hypothesis $H_j : \theta_j \leq 0$ is tested in $\mathcal{P}_j$, where θ_j might describe the true mean efficacy of a new treatment (vs. a control). If a decision is made for $\mathcal{P}_j$, be it rejection or acceptance of H_j, this decision is made for every (future) patient in $\mathcal{P}_j$ including those patients belonging to a subset of $\mathcal{P}_j$. For example, if H_j is erroneously rejected, this error will also affect patients in $\mathcal{P}_{\{j\}} \subset \mathcal{P}_j$ which are then assigned to an inefficient treatment. This does obviously not mean that we can translate a type I error made in $\mathcal{P}_j$ to a type I error in a subgroup – the true parameter of interest of said subgroup might even be in the alternative, after all – but rather that even if the alternative holds in the subgroup, we still treat it like the null hypothesis actually holds (when H_j is true). This ensures that a false rejection in a larger population can be translated to a smaller population and leads to a rather conservative treatment of type I errors in this case. On the other hand, if the alternative holds in $\mathcal{P}_j$ and the null hypothesis is true in $\mathcal{P}_{\{j\}}$, then patients in $\mathcal{P}_{\{j\}}$ would be assigned to a treatment being inefficient for them describing a more anti-conservative approach. Also have in mind that this does not mean that the PWER describes the risk of an *individual* patient of getting an inefficient treatment, which is beyond any error control, but moreso the risk that a *population* is assigned to an inefficient treatment.
All in all, with the above reasoning in mind each $\text{FWER}_{\mathcal{P}_J}$-expression can (informally) be written as

$$\text{FWER}_{\mathcal{P}_J} = \mathbb{P}(\text{at least one type I error in } \mathcal{P}_J), \tag{2.3}$$

where a type I error in $\mathcal{P}_J$ is made whenever a hypothesis is erroneously rejected for a superset of $\mathcal{P}_J$. In the next section we will describe how to formally write out this expression.

2.2 Mathematical setting and definition

Formally, we need to properly define the parametric family of probability measures and their respective measure space for our population-wise setting. In general, we might not only be interested in testing hypotheses in any of the $\mathcal{P}_j$, $j = 1, \ldots, m$ or $\mathcal{P}_J$, $J \subseteq I$, but more so in any union of disjoint subgroups $\mathcal{P}_J$. To this end, we define the set

$$\mathcal{C}_{\mathcal{P}} := \{J \subseteq I : \mathcal{P}_J \neq \emptyset\} \subseteq 2^I \setminus \{\emptyset\} \tag{2.4}$$

containing all non-empty partitions $\mathcal{P}_J$. The $\mathcal{P}$ is added to make clear what overall population this model is based on. Note that $\mathcal{C}_{\mathcal{P}} = 2^I \setminus \{\emptyset\}$ if there is no set $\mathcal{P}_j$ that

does not overlap with any of the other populations. Now an element $U \subseteq \mathcal{C}_{\mathcal{P}}$ defines a union of disjoint sub-populations,

$$\mathcal{P}^U := \bigcup_{J \in U} \mathcal{P}_J. \tag{2.5}$$

Note that a union of sub-populations is indicated by an upper-case index U, whereas the letter U is conveniently chosen to address a '**u**nion'. Also, by this definition it is $\mathcal{P}^{\{\{J\}\}} = \mathcal{P}_J$ and $\mathcal{P}^{\{J|j\in J\}} = \mathcal{P}_j$ for each $j = 1, \ldots, m$. So the upper-case index is a set of sets.
In general we are now interested in testing the efficacy of a treatment T^U in $\mathcal{P}^U$ by testing a corresponding null hypothesis H^U against some alternative K^U for each $U \in \mathfrak{U} \subseteq \mathcal{C}_{\mathcal{P}}$ with $\mathfrak{U}$ denoting some pre-defined subset containing all unions of disjoint subgroups we are actually interested in testing in. For instance, in the example of $m = 4$ overlapping populations (as in Figure 2.1), if we were interested in testing the efficacy of a treatment vs. a control in each $\mathcal{P}_j$, $j = 1, \ldots, 4$, the set $\mathfrak{U}$ would be equal to $\mathfrak{U} = \{U_1, \ldots, U_4\}$ with $U_j := \{J \subseteq I \mid j \in J\}$. For each $U \in \mathfrak{U}$ we now intend to test

$$H^U : \theta(\mathcal{P}^U, T^U) \leq 0 \quad \text{vs.} \quad K^U : \theta(\mathcal{P}^U, T^U) > 0, \tag{2.6}$$

where $\theta^U := \theta(\mathcal{P}^U, T^U)$ denotes the true mean efficacy of treatment T^U in $\mathcal{P}^U$ (in comparison to a control). Since $\mathcal{P}^U$ is a union of disjoint sub-populations $\mathcal{P}_J$, we define each parameter $\boldsymbol{\theta}^U$ as a weighted sum of all parameters θ_J with $J \in U$,

$$\theta^U := \sum_{J \in U} \pi_J^U \theta_J, \quad \forall U \in \mathfrak{U}, \tag{2.7}$$

with $\pi_J^U = \pi_J / \pi^U$ for $J \in U$ and $\pi^U = \sum_{J \in U} \pi_J$ for $U \in \mathfrak{U}$. For these parameters, we further define an L-dimensional parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^L$, where $L \geq |\mathfrak{U}|$. In general, we will work with cases where $L = |\mathfrak{U}|$ and thus

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^L \mid \boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}} \text{ with } \theta^U = \sum_{J \in U} \pi_J^U \theta_J, \quad \forall U \in \mathfrak{U} \subseteq \mathcal{C}_{\mathcal{P}} \right\}, \tag{2.8}$$

but L can generally be larger since our probability distribution can depend on further unknown parameters (like nuisance parameters) we do not test for [1]. In those cases we would need to consider a parameter space $\tilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta} \times \boldsymbol{\Upsilon}$ with $\boldsymbol{\Upsilon}$ being $L - |\mathfrak{U}|$-dimensional. If not stated otherwise, however, we will assume that $\boldsymbol{\Upsilon} = \emptyset$.
Furthermore, note that the above definition of $\mathfrak{U}$ does not account for the possibility of investigating more than one treatment in some union $\mathcal{P}^U$ as we will mostly deal with the case of one treatment per population. This more general case is addressed in Section 3.2 for single-stage and in Section 4.3.2 for group sequential designs.
The definition in (2.8) has some implications on $\boldsymbol{\Theta}$. If some U and U' are overlapping, i.e. if $U \cap U' \neq \emptyset$, then the parameters θ^U and $\theta^{U'}$ are dependent on each other because their weighted sums share all θ_J with $J \in U \cap U'$. Therefore, $\boldsymbol{\Theta}$ is a proper subset of $\mathbb{R}^L$. For later considerations we write $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$, where

$$\boldsymbol{\Theta}_0 := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid \boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}}, \theta^U \in H^U \forall U \in \mathfrak{U}\} \tag{2.9}$$

[1] Think of a normal distribution where both mean μ and standard deviation σ are unknown.

and $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. So $\boldsymbol{\Theta}_0$ is the set of all parameter constellations, where each component belongs to their respective null hypothesis and $\boldsymbol{\Theta}_1$ is its complementary set.
Simultaneously testing all H^U requires a $|\mathfrak{U}|$-dimensional random vector of test statistics $\mathbf{Z}$ for each hypothesis and therefore we have a probability distribution $\mathbb{P}^{\mathbf{Z}}$ that we want to find. In a parametric statistical model it is $\mathbb{P}^{\mathbf{Z}} \in \{\mathbb{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Finally, denoting Ω as the sample space of $\mathbf{Z}$ and $\mathcal{A}$ as a σ-Algebra on Ω, we end up with the parametric statistical model

$$\mathcal{M}_{\boldsymbol{\Theta}} := (\Omega, \mathcal{A}, \{\mathbb{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}). \tag{2.10}$$

From now on we denote such a multiple testing problem as *population-wise testing problem* and write it as

$$(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H}). \tag{2.11}$$

The different elements of this four-tuple have the following definitions.

- $\mathcal{C}_{\mathcal{P}}$, containing the (index) set of all non-empty disjoint sub-populations of a partition of $\mathcal{P}$
- $\boldsymbol{\pi} = (\pi_J)_{\{J \in I\}}$ the vector of relative population sizes of all disjoint subgroups
- $\mathcal{M}_{\boldsymbol{\Theta}}$ the statistical model for the multiple testing problem
- $\mathcal{H}$ the set of all hypotheses that are to be tested

This 4-tuple basically contains every information needed to fully define a population-wise multiple testing problem. The aforementioned index set $\mathfrak{U}$ of all populations we plan to test is implicitly defined by $\mathcal{H}$.

Now, since we are interested in formalizing $\text{FWER}_{\mathcal{P}_J}$ for each $J \subseteq I$, we need to formalize which hypotheses actually affect $\mathcal{P}_J$. For $J \subseteq I$ let $A_J := \{U \in \mathfrak{U} \mid J \in U\}$ be the index set of all hypotheses H^U whose erroneous rejection affects $\mathcal{P}_J$ and for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ let $I(\boldsymbol{\theta}) := \{U \in \mathfrak{U} \mid \theta^U \in H^U\}$ the index set of all true null hypotheses. Then $I_J(\boldsymbol{\theta}) := I(\boldsymbol{\theta}) \cap A_J$ is the set of all $U \in \mathfrak{U}$ where the erroneous rejection of the corresponding H^U affects $\mathcal{P}_J$. For each $U \in \mathfrak{U}$ let

$$\varphi^U = \begin{cases} 1, & H^U \text{ rejected} \\ 0, & \text{otherwise} \end{cases} \tag{2.12}$$

be a test decision function for testing H^U. We can then formalize $\text{FWER}_{\mathcal{P}_J}$ as

$$\text{FWER}_{\mathcal{P}_J} = \mathbb{P}_{\boldsymbol{\theta}}\left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \{\varphi^U = 1\}\right) \tag{2.13}$$

and the FWER of the full design equals

$$\text{FWER} = \mathbb{P}_{\boldsymbol{\theta}}\left(\bigcup_{U \in I(\boldsymbol{\theta})} \{\varphi^U = 1\}\right). \tag{2.14}$$

We can now give a formal definition of the population-wise error rate (PWER).

Definition 2.2.1 (Population-wise error rate)**.** *Let $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_\boldsymbol{\Theta}, \mathcal{H})$ be a population-wise testing problem. For each $\mathcal{P}_J$, $J \in \mathcal{C}_\mathcal{P}$, and each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ let $I_J(\boldsymbol{\theta})$ be the index set of all true null hypotheses H^U, $U \in \mathfrak{U}$, whose rejection concern $\mathcal{P}_J$. Further, let each H^U be tested via a decision function φ^U as defined in (2.13). Then the* population-wise error rate (PWER) *is defined as*

$$PWER_{\boldsymbol{\theta}} = \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J FWER_{\mathcal{P}_J} = \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \{\varphi^U = 1\} \right), \tag{2.15}$$

where $\bigcup_{U \in I_J(\boldsymbol{\theta})} \{\varphi^U = 1\} := \emptyset$ whenever $I_J(\boldsymbol{\theta}) = \emptyset$.

So for each partition $\mathcal{P}_J$ of the overall population we need to know the probability of rejecting at least one true null hypothesis that affects this partition. Moreover, note that compared to the family-wise error rate, which controls the maximum risk for future patients to be assigned to an inefficient treatment strategy, the PWER is an average risk measure as it is a weighted mean of family-wise error rates.
A direct consequence of Definition 2.2.1 is now given by the following theorem.

Theorem 2.2.1. *Let $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_\boldsymbol{\Theta}, \mathcal{H})$ be a population-wise multiple testing problem. For all parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ it then holds $PWER_{\boldsymbol{\theta}} \leq FWER_{\boldsymbol{\theta}}$.*

Proof. The FWER, as the probability of rejecting at least one true null, is given by

$$FWER_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I(\boldsymbol{\theta})} \{\varphi^U = 1\} \right).$$

The assertion now immediately follows since for all $J \subseteq I$ we have

$$\mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \{\varphi^U = 1\} \right) \leq FWER_{\boldsymbol{\theta}}$$

and it is $\sum_{J \subseteq I} \pi_J = 1$. □

We conclude this section with a couple of remarks.

Remark. *(i) For $\mathcal{P}_1 = \cdots = \mathcal{P}_m = \mathcal{P}$ it holds $PWER_{\boldsymbol{\theta}} = FWER_{\boldsymbol{\theta}}$, since in this case there is only one partition, namely $J = I$.*

(ii) Suppose we intended to test exactly one hypothesis H_j in each respective population $\mathcal{P}_j$ by means of a level α_j test φ_j, $j = 1, \ldots, m$. Then $\mathfrak{U} = \{U_j\}_{j=1}^m$ with $U_j = \{J \subseteq I \mid j \in J\}$. With $I(\boldsymbol{\theta}) = \{j \in I \mid \theta_j \in H_j\}$, $A_J = \{j \in I \mid j \in J\}$ for all $J \subseteq I$ and $I_J(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) \cap A_J$, the FWER is given by

$$FWER_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{j \in I(\boldsymbol{\theta})} \{\varphi_j = 1\} \right) \tag{2.16}$$

and the PWER simplifies to

$$PWER_{\boldsymbol{\theta}} = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{j \in J \cap I(\boldsymbol{\theta})} \{\varphi_j = 1\} \right). \tag{2.17}$$

Here, the addend for $J = I$ is equal to $\pi_I FWER_{\boldsymbol{\theta}}$, so for the extreme case of $\pi_I = 1$, PWER and FWER are exactly the same. If all subpopulations $\mathcal{P}_1, \ldots, \mathcal{P}_m$ are disjoint, then the PWER is bounded by $\alpha_{max} := \max_{i \in I} \alpha_i$ because of

$$PWER_{\boldsymbol{\theta}} = \sum_{j \in I} \pi_{\{j\}} \underbrace{\mathbb{P}_{\boldsymbol{\theta}}(\{\varphi_j = 1\})}_{\leq \alpha_{max}} \leq \alpha_{max} \tag{2.18}$$

Hence, in the case of disjoint subpopulations, no multiplicity adjustments are necessary.

In later chapters we will use another way of expressing the PWER, given in the follwing remark.

Remark. *(PWER as expectation) Assuming the same setup as in Definition 2.2.1, let $\varphi_J^{\boldsymbol{\theta}}$ be a function mapping to $\{0, 1\}$ given by*

$$\varphi_J^{\boldsymbol{\theta}} := \begin{cases} 1, & \text{if } \sum_{U \in I_J(\boldsymbol{\theta})} \varphi^U \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.19}$$

returning whether any type I error for $\mathcal{P}_J$, $J \in \mathcal{C}_{\mathcal{P}}$, has bee made. Further, let $\varphi^{\boldsymbol{\theta}} = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \varphi_J^{\boldsymbol{\theta}}$ be the percentage of the full population $\mathcal{P}$ that has been affected by a type I error. Then due to the linearity of the expectation the PWER under some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can be written as

$$PWER_{\boldsymbol{\theta}}(\boldsymbol{\varphi}) = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\varphi_J^{\boldsymbol{\theta}} = 1 \right) = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{E}_{\boldsymbol{\theta}} \left(\varphi_J^{\boldsymbol{\theta}} \right) = \mathbb{E}_{\boldsymbol{\theta}}(\varphi^{\boldsymbol{\theta}}) \tag{2.20}$$

This property will be central for the considerations in Chapter 6.

2.3 Controlling the PWER

As already stated, each test decision function φ^U for H^U can be expressed in terms of test statistics Z^U and a critical value c^U via $\mathbf{1}_{\{Z^U \geq c^U\}}$. Before we go on with some properties of the FWER, we want to make two other assumptions on the vector of these test statistics $\mathbf{Z} = (Z^U)_{U \in \mathfrak{U}}$. Since the PWER involves probabilities of sub-vectors $\mathbf{Z}_{I_J(\boldsymbol{\theta})} = \{Z^U : U \in I_J(\boldsymbol{\theta})\}$ for $J \subseteq \mathcal{C}_{\mathcal{P}}$ in each addend of the weighted sum, we need to make sure that the joint distribution of these sub-vectors does not depend on whether hypotheses which are tested with any test statistic not in $\mathbf{Z}_{I_J(\boldsymbol{\theta})}$ are true or false. This assumption is basically the subset-pivotaliy assumption from Westfall & Young (1993, [56]). Also, we want to assume stochastic monotonicity of the test statistics in all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$:

(A1) **Monotonicity:** For component Z^V of $\mathbf{Z}$, $V \in \mathfrak{U}$, it holds

$$\mathbb{P}_{\boldsymbol{\theta}_1}\left(\{Z^V \geq c^V\}\right) \leq \mathbb{P}_{\boldsymbol{\theta}_2}\left(\{Z^V \geq c^V\}\right) \tag{2.21}$$

for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_0$ with $\theta_1^U = \theta_2^U$ for all $U \neq V$ and $\theta_1^V \leq \theta_2^V$.

(A2) **Subset pivotality:** For $J \in \mathcal{C}_{\mathcal{P}}$ the joint distribution of each sub-vector $\mathbf{Z}_{I_J(\boldsymbol{\theta})}$ does not depend on θ^U for $U \notin I_J(\boldsymbol{\theta})$.

Analogously to error control concepts for the FWER, we define *strong* and *weak* control of the PWER as follows.

Definition 2.3.1 (Strong and weak PWER-control)**.** *Let $\boldsymbol{\varphi}$ be a test decision function for a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ and $\alpha \in (0,1)$ a predefined significance level. Then $\boldsymbol{\varphi}$ controls the PWER in the* weak *sense if*

$$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 : PWER_{\boldsymbol{\theta}}(\boldsymbol{\varphi}) \leq \alpha \tag{2.22}$$

and in the strong *sense if*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} PWER_{\boldsymbol{\theta}}(\boldsymbol{\varphi}) \leq \alpha. \tag{2.23}$$

Obviously, strong PWER-control implies weak PWER-control.
Now, to obtain strong PWER-control we can try to find the least favorable configuration $\boldsymbol{\theta}^*$ which maximizes the PWER on $\boldsymbol{\Theta}$. The PWER is maximized under $\boldsymbol{\theta}^*$ if all sets $I_J(\boldsymbol{\theta}^*)$, $J \in \mathcal{C}_{\mathcal{P}}$, are maximal, i.e. if $I_J(\boldsymbol{\theta}^*) = A_J = \{U \in \mathfrak{U} \mid J \in U\}$. This is, of course, the case if all components of $\boldsymbol{\theta}^*$ are less or equal to 0., or in general, if $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$. Because the PWER is not constant on $\boldsymbol{\Theta}_0$, we have to find a value $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$ maximizing the PWER on $\boldsymbol{\Theta}_0$. The assumptions (A1) and (A2) now simplify the search for the least favorable configuration as the next theorem states.

Theorem 2.3.1. *Let $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ be a population-wise testing problem and $\boldsymbol{\varphi} = (\varphi^U)_{U \in \mathfrak{U}}$ be a corresponding decision function with $\varphi^U = \mathbf{1}(Z^U \geq c^U)$, where Z^U are test statistics fulfilling (A1) and (A2) and $c^U \in \mathbb{R}$ a critical value for testing H^U. Then it holds*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} PWER_{\boldsymbol{\theta}}(\boldsymbol{\varphi}) = PWER_{\boldsymbol{\theta}^*}(\boldsymbol{\varphi}), \tag{2.24}$$

where $\boldsymbol{\theta}^* = \boldsymbol{0} \in \mathbb{R}^L$.

Proof. Assume $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_0$, then $I(\boldsymbol{\theta}^*) = \mathfrak{U}$ and $I_J(\boldsymbol{\theta}^*) = A_J$ for all $J \in \mathcal{C}_{\mathcal{P}}$ and all components $\theta^{*U} \leq 0$. Due to the monotonicity assumption (A1), it follows that each probability term

$$p_{J,\boldsymbol{\theta}^*} := \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \{\varphi^U = 1\}\right) = \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \{Z^U > c^U\}\right),$$

$J \in \mathcal{C}_{\mathcal{P}}$, is maximized if $\boldsymbol{\theta}^* = \mathbf{0}$. So $\mathbf{0}$ maximizes the PWER in $\boldsymbol{\Theta}_0$.
If $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_1$, there are some $U \in \mathfrak{U}$ with $U \notin I(\boldsymbol{\theta}^*)$ (i.e. with $\theta^{*U} > 0$) and consequently $I_J(\boldsymbol{\theta}^*) \subset A_J$ for all $J \in U$ (here $\subset$ describes the proper subset relation). For these J it then holds $p_{J,\boldsymbol{\theta}^*} < p_{J,\mathbf{0}}$ because it is a probability over a union of less rejection regions. Thus, $\boldsymbol{\theta}^* = \mathbf{0}$ must be the least favorable configuration. □

In this thesis we will mostly assume a multivariate normal distribution for the vector of test statistics $\mathbf{Z}$ which satisfies assumption (A1) because the univariate normal distribution satisfies it and (A2) since each sub-vector of dimension $l < L$ is again multivariate normal with corresponding sub-expectation-vector and sub-correlation-matrix. The assumption of a multivariate normal distribution is often justified by the multidimensional central limit theorem [52]. Thus, if the components of test statistic vector are of a different distribution (say binomial) and can be written as a scaled sum, asymptotic PWER-control is still possible.

2.4 Some special cases

To further illustrate the concept of the PWER we want to go through some special cases.

2.4.1 Two overlapping populations

First, consider two populations $\mathcal{P}_1$ and $\mathcal{P}_2$ and two treatments T_1 and T_2 to be tested by means of the hypotheses $H_1 : \theta(\mathcal{P}_1, T_1) \leq 0$ and $H_2 : \theta(\mathcal{P}_2, T_2) \leq 0$. As before, the parameters $\theta_j := \theta(\mathcal{P}_j, T_j)$ describe the mean efficiency of treatment T_j that gets administered to patients in $\mathcal{P}_j$, $j = 1, 2$. The overall population $\mathcal{P}$ can now be partitioned into three disjoint sub-populations, $\mathcal{P}_{\{1\}} := \mathcal{P}_1 \setminus \mathcal{P}_2$, $\mathcal{P}_{\{2\}} := \mathcal{P}_2 \setminus \mathcal{P}_1$ and $\mathcal{P}_{\{1,2\}} := \mathcal{P}_1 \cap \mathcal{P}_2$. This setup corresponds to the population-wise testing problem $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_\mathbf{\Theta}, \mathcal{H})$ with $\mathcal{P} = \mathcal{P}_{\{1\}} \cup \mathcal{P}_{\{2\}} \cup \mathcal{P}_{\{1,2\}}$ such that $\mathcal{C}_\mathcal{P} = \{\mathcal{P}_{\{1\}}, \mathcal{P}_{\{2\}}, \mathcal{P}_{\{1,2\}}\}$, $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$, $\mathbf{\Theta} = \mathbb{R}^2$ and $\mathcal{H} = \{H_1, H_2\}$. Also, as already mentioned, here $\mathfrak{U} = \{U_1, U_2\}$ with $U_j = \{\{j\}, \{1, 2\}\}$, for $j = 1, 2$.
Due to $\mathcal{P}_{\{i\}} \subseteq \mathcal{P}_i$ and $\mathcal{P}_{\{1,2\}} \subseteq \mathcal{P}_1, \mathcal{P}_2$, a type I error for $\mathcal{P}_{\{i\}}$ is made whenever H_i is falsely rejected, $i = 1, 2$, and a type I error for $\mathcal{P}_{\{1,2\}}$ is made whenever H_1 or H_2 are erroneously rejected. Therefore, in case of H_1 and H_2 being true, that is for $\boldsymbol{\theta} \in \mathbf{\Theta}_0$, we have:

$$\text{PWER}_{\boldsymbol{\theta}} = \pi_{\{1\}}\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1) + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_2) + \pi_{\{1,2\}}\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1 \text{ or } H_2)$$

If H_1 is true and H_2 is false, then:

$$\text{PWER}_{\boldsymbol{\theta}} = \pi_{\{1\}}\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1) + \pi_{\{1,2\}}\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1) = (\pi_{\{1\}} + \pi_{\{1,2\}})\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1)$$

For H_1 being false and H_2 being true, i.e. for some $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbf{\Theta}_1$ with $\theta_1 > 0$ and $\theta_2 \leq 0$, we obtain:

$$\text{PWER}_{\boldsymbol{\theta}} = (\pi_{\{2\}} + \pi_{\{1,2\}})\mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_2)$$

That is, whenever only one null hypothesis H_i is true, the PWER equals the probability of rejecting H_i times the relative population size of $\mathcal{P}_i$. As a little comparison: the FWER in this case is equal to $P_{\boldsymbol{\theta}}(\text{reject } H_i)$, so the PWER is $(1-(\pi_{\{2\}}+\pi_{\{1,2\}}))\cdot 100\%$ smaller than the FWER.

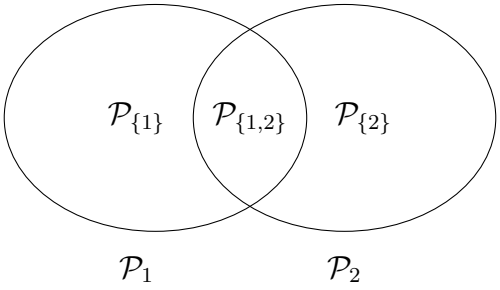

Figure 2.2: $m = 2$ intersecting populations

2.4.2 Nested populations

The PWER is also well-suited for handling the practically relevant case of nested populations. Suppose we want to test hypotheses $H_j : \theta_j := \theta(\mathcal{P}_j, T_j) \leq 0$ in populations $\mathcal{P}_j$, respectively, which are structured as a descending sequence, i.e. $\mathcal{P}_1 \supset \mathcal{P}_2 \supset \cdots \supset \mathcal{P}_m$. Define $\mathcal{P}_{\{1,\ldots,j\}} := \mathcal{P}_j \setminus \mathcal{P}_{j+1}$ for $j < m$ and $\mathcal{P}_{\{1,\ldots,m\}} = \mathcal{P}_m$ and let $\pi_{\{1,\ldots,j\}}$ be the respective relative population sizes. We commit to a type I error in $\mathcal{P}_{\{1,\ldots,j\}}$ if any correct H_j is rejected for $i \leq j$. For the PWER we then obtain

$$\text{PWER}_{\boldsymbol{\theta}} = \sum_{j=1}^{m} \pi_{\{1,\ldots,j\}} \mathbb{P}_{\boldsymbol{\theta}}(\text{reject at least one true } H_i \text{ for } i \leq j).$$

Especially, if $\mathcal{P}_j$ is defined by a continuous biomarker X (with some probability distribution $\mathbb{P}^X$ defined on some other measure space $(\Omega', \mathcal{A}',)$), i.e. $\mathcal{P}_j = \{X > t_j\}$, for cut-off points x_j, $j = 1, \ldots, m+1$ (with $x_{m+1} := \infty$), the PWER equals

$$\text{PWER}_{\boldsymbol{\theta}} = \sum_{j=1}^{m} \mathbb{P}'(x_j < X \leq x_{j+1}) \mathbb{P}_{\boldsymbol{\theta}}(\text{reject at least one true } H_i \text{ for } i \leq j).$$

Again, to fit this into the general setting, it is $\mathcal{P} = \mathcal{P}_1 = \bigcup_{j=1}^{m} \mathcal{P}_{\{1,\ldots,j\}}$, $\boldsymbol{\pi} = (\pi_{\{1,\ldots,j\}})_{j=1}^{m}$ such that $\mathcal{C}_{\mathcal{P}} = \{\mathcal{P}_{\{1\}}, \mathcal{P}_{\{1,2\}}, \ldots, \mathcal{P}_{\{1,\ldots,m\}}\}$, $\boldsymbol{\Theta} = \mathbb{R}^m$. Also $\mathcal{H} = \{H_1, \ldots, H_m\}$ and $\mathfrak{U} = \{U_1, \ldots, U_m\}$ with $U_j := \{\{1, \ldots, j\}, \ldots, \{1, \ldots, m\}\}$ for $j = 1, \ldots, m$.

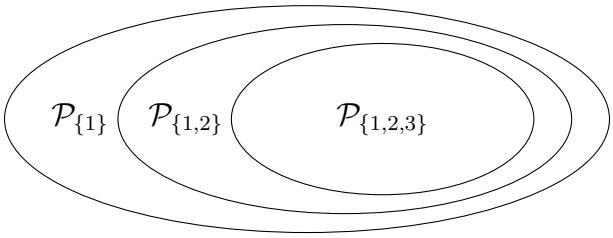

Figure 2.3: Nested population structure for $m = 3$.

2.4.3 Three populations with two intersections

At last, we want to give an example where the FWER is strictly conservative even for a control of the maximum (instead of the average) type I error rate. Consider

three populations $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ with $\mathcal{P}_1 \cap \mathcal{P}_2 \neq \emptyset$, $\mathcal{P}_2 \cap \mathcal{P}_3 \neq \emptyset$ and $\mathcal{P}_1 \cap \mathcal{P}_3 = \emptyset$, as in Figure (2.4). That is, $\mathcal{P}_{\{1,2,3\}} = \mathcal{P}_{\{1,3\}} = \emptyset$ and thus $\pi_{\{1,2,3\}} = \mathcal{P}_{\{1,3\}} = 0$. Again, hypotheses of the form $H_j : \theta(\mathcal{P}_j, T_j) \leq 0$ are to be tested in each population, respectively. Under the global null hypothesis, where all null hypotheses H_j are true, the PWER is then given by

$$PWER = \sum_{i=j}^{3} \pi_{\{j\}} \mathbb{P}(\text{reject } H_j) + \sum_{j=1}^{2} \pi_{\{j,j+1\}} \mathbb{P}(\text{reject } H_j \text{ or } H_{j+1}).$$

The FWER under the global null, however, equals

$$FWER = \mathbb{P}(\text{reject } H_1 \text{ or } H_2 \text{ or } H_3).$$

But since it is not possible for a patient to be in $\mathcal{P}_1$ and $\mathcal{P}_3$ simultaneously, the FWER corrects for multiplicity that no patient is actually affected by. Now, one may think of a different approach. Since no patient is part of $\mathcal{P}_{\{1,3\}}$, one could simply try to control the following quantity (assuming the worst case scenario is the global null):

$$\max\{\mathbb{P}(\text{reject } H_1 \text{ or reject } H_2), \mathbb{P}(\text{reject } H_2 \text{ or reject } H_3)\} \tag{2.25}$$

If we controlled (2.25) at a level α, the PWER would still be well below α since $\mathbb{P}(\text{reject } H_i) < \alpha$. Apart from the conservativeness of this approach, it is also very discontinuous in the sense that the maximum of two values is not necessarily a continuous function.

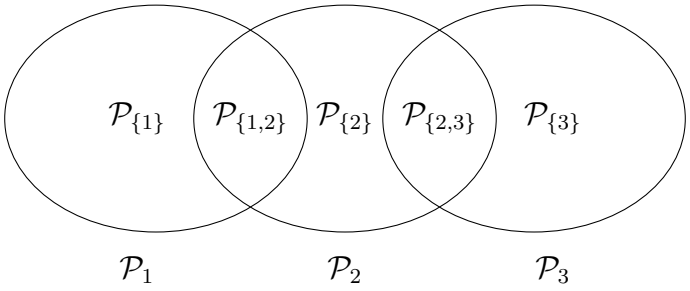

Figure 2.4: $m = 3$ populations and their disjoint subpopulations

2.5 A note on the interpretation of the PWER

To better understand how to interpret the control of the PWER, we want to briefly restate what is already written in the discussion of [12]. As briefly touched on in Section 2.1, with the PWER we do not aim to make claims on the individual patient level but rather on a population level. That is, we make claims on treatment strategies that consist of a treatment and a population the treatment is investigated in. This is also reflected by the notation of the estimand $\theta(\mathcal{P}_j, T_j)$ we have used throughout this section – it depends on a pair $(\mathcal{P}_j, T_j)$. The same rationale applies when aiming for FWER-control as individual efficacy claims are not anticipated here, either. A population-wise claim can be viewed as an approximation for an individual claim in the target population, however. Test results from more than a

single population may be used for a more informed individual decision. With PWER control, we consider the worst case scenario, where an efficacy claim for a treatment strategy will always lead to an application of the treatment to all patients in the target population. Also, note that a potential off-label use, i.e. the application of a treatment to patients that are not part of its target population, is not accounted for.

2.6 Population-wise power

To conclude this chapter, we want to propose a concept on how to obtain type II error control for a PWER-controlling test procedure. Now the definition of *power* is quite ambiguous depending on the given mathematical setting and the test procedures we are dealing with. If only a single hypothesis H_0 was to be tested, we would interpret the power of the testing procedure as *probability of rejecting H_0 under the condition that H_0 is false*. For a multiple testing procedure, however, like there are multiple ways to define a type I error quantity, there are multiple ways to define power. Since we primarily aim for PWER-controlling methods, one quite intuitive power measure would be something like a *population-wise power*. Ultimately, though, just because one uses the PWER as a type I error measure, one might still be interested in controlling a different power measure. We will list some possible choices in the end of this subsection.
To motivate the population-wise power, which we will abbreviate with PWP, let us, for simplicity, consider a single stage design where we test a hypothesis $H_i : \theta_i \leq 0$ in $\mathcal{P}_i$ for $i = 1, 2$, respectively. Of course, we assume $\mathcal{P}_1 \cap \mathcal{P}_2 \neq \emptyset$. If both hypotheses are true ($\theta_1, \theta_2 \leq 0$), the PWER is given by

$$PWER_{\boldsymbol{\theta}} = \sum_{i=1}^{2} \pi_{\{i\}} \mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_i) + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}}(\text{reject } H_1 \text{ or } H_2). \tag{2.26}$$

Now assume that both null hypotheses are false, so $\theta_i > 0$ for $i = 1, 2$. One way to define the population-wise power would simply be to define it as in (2.26) under a different parameter configuration. That is the PWP would be the exactly the PWER but instead of using probabilities of making at least one type I error in each respective $\mathcal{P}_J$ one would consider probabilities of making at least one correct rejection in each respective $\mathcal{P}_J$. Now if $\theta_1 > 0$ and $\theta_2 \leq 0$, i.e. H_1 is false and H_2 is true, the PWP will be equal to

$$PWP_{\boldsymbol{\theta}} = \underbrace{(\pi_{\{1\}} + \pi_{\{1,2\}})}_{=:\pi_1} \mathbb{P}_{\theta_1}(\text{reject } H_1) = \pi_1 \mathbb{P}_{\theta_1}(\text{reject } H_1), \tag{2.27}$$

since we can only reject H_1 correctly. This expression is clearly bounded by π_1 which can be any number between 0 and 1, so standard power values such as 80% or 90% might be impossible to reach and therefore this is not a reasonable power measure. A straightforward solution is to divide this expression by the size of all subpopulations the alternative hypothesis is holds in. Since the alternative holds in $\mathcal{P}_1$ (H_1 is false), we divide this expression by π_1 and thus obtain $PWP_{\boldsymbol{\theta}} = \mathbb{P}_{\theta_1}(\text{reject } H_1)$.
The following definition generalizes the idea described above.

Definition 2.6.1 (Population-wise Power)**.** *Let $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ be a population-wise testing problem and $\boldsymbol{\varphi} = (\varphi^U)_{U \in \mathfrak{U}}$ a corresponding test function. Further, for $\boldsymbol{\theta}_1 \in$*

$\boldsymbol{\Theta}_1$ *we define index sets* $I_1(\boldsymbol{\theta}_1) := \{J \subseteq I \mid \exists U \in \mathfrak{U} : J \in U \wedge \theta_1^U \in K^U\}$ *and* $I_{1,J}(\boldsymbol{\theta}_1) := \{U \in \mathfrak{U} \mid J \in U \wedge \theta_1^U \in K^U\}$ *for each* $J \in I_1(\boldsymbol{\theta}_1)$*. Then, for a* $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$ *the* Population-wise power (PWP) *is defined as*

$$PWP_{\boldsymbol{\theta}_1}(\boldsymbol{\varphi}) = \frac{\sum_{J \in I_1(\boldsymbol{\theta}_1)} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_{1,J}(\boldsymbol{\theta}_1)} \{\varphi^U = 1\} \right)}{\sum_{J \in I_1(\boldsymbol{\theta}_1)} \pi_J}. \tag{2.28}$$

This definition requires a further explanation. The definitions of the two index sets $I_1(\boldsymbol{\theta}_1)$ and $I_{1,J}(\boldsymbol{\theta}_1)$ are based on certain assumptions about how to transfer the validity of a hypothesis concerning some union $\mathcal{P}^U$ to a subpopulation, $\mathcal{P}_J$ in this case. Let us say that θ^U is the true effect in some $\mathcal{P}^U$ and θ_J the true effect in $\mathcal{P}_J \subseteq \mathcal{P}^U$. If θ^U is positive (negative), there is no real way to infer whether θ_J is positive (negative). For the PWER we made the assumption that a type I error is made in a partition $\mathcal{P}_J$ whenever at least one hypothesis is erroneously rejected that is a superset of $\mathcal{P}_J$, since even if there was a relevant effect in $\mathcal{P}_J$ (i.e. $\theta_J > 0$), pretending that there is indeed none ($\theta_J \leq 0$) would only make the procedure more conservative. Likewise, for the PWP we assume that the alternative hypothesis holds in a partition $\mathcal{P}_J$ whenever at least one effect θ^U with $J \in U$ is greater than zero.

Of course, as previously commented on, any other power measure can still be used depending on the practical setting. Examples for other measures are (see Maurer and Mellein, 1988, [39])

1. The probability of rejecting at least one false null hypothesis

$$\text{Pow}_{\boldsymbol{\theta}_1,1} = \mathbb{P}_{\boldsymbol{\theta}_1} \left(\bigcup_{J \subseteq I} \bigcup_{U \in I_{1,J}(\boldsymbol{\theta}_1)} \{\varphi^U = 1\} \right) \tag{2.29}$$

2. The "total power", i.e. the probability of rejecting all false null hypotheses

$$\text{Pow}_{\boldsymbol{\theta}_1,\text{all}} = \mathbb{P}_{\boldsymbol{\theta}_1} \left(\bigcap_{J \subseteq I} \bigcap_{U \in I_{1,J}(\boldsymbol{\theta}_1)} \{\varphi^U = 1\} \right) \tag{2.30}$$

3. The expected number of correctly rejected null hypotheses $S(\boldsymbol{\theta}_1)$, i.e. $\mathbb{E}(S(\boldsymbol{\theta}_1))$.

In the remainder of this thesis, we will mainly use $\text{Pow}_{\boldsymbol{\theta}_1,1}$ and the PWP as power measures, if not stated otherwise.

3. Single-stage designs for studies with overlapping populations

In today's practice of clinical studies, especially in oncology, trials are typically conducted sequentially in multiple stages. This encompasses classical group sequential designs and adaptive designs. Before tackling these two design types, we want to focus on single stage designs first. There are generally a couple of reasons why a single stage design might be preferable over a multi-stage one. Even though sequential designs are typically preferred due to economical and ethical reasons such as savings in time and money, as already described in Section 1, they can also pose some problems in terms of statistical methodology, interpretation and logistics. For example, in the final analysis of a sequential design, quantities such as p-values and confidence intervals are still defined but need to be interpreted differently [45]. Also according to Bauer et al. (2004, [7]) in an adaptive design the test statistics used after an adaptation are different from the sufficient test statistics in the accrued sample, which might be an unacceptable drawback to some practitioners. Moreover, with a smaller sample size, an early rejection of a null hypothesis can lead to an overestimation of the true effect size [44]. Also from a purely didactic standpoint it makes sense to start off with an easier set of designs such that the understanding of the more general design concepts can be grasped more easily.
This chapter is outlined in the following way. First off, it is described how the PWER is controlled in a single stage design, by means finding a suitable set of critical values. Also the existance of PWER-adjusted p-values and confidence intervals is described. By now, we have also only dealt with one treatment per population. The investigation of different treatments for the different populations $\mathcal{P}_1, \ldots, \mathcal{P}_m$ can pose some problems regarding how the randomization in each stratum $\mathcal{P}_J$ has to be conducted. We will propose a general way to deal with multiple treatments in a single stage design. In a subsequent section this method will be applied to the special case of two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$ for which savings in sample size compared to a similar FWER-controlling method will be discussed. Furthermore, based on this example the effect of estimating the population sizes $\boldsymbol{\pi}$ on the PWER is investigated via simulations. Lastly, we apply the PWER-concept to a multiple testing approach for umbrella trials suggested in Sun et al. (2016, [51]).

3.1 PWER-control and statistical inference

This section is dedicated to demonstrating how to achieve control of the population-wise error rate in a single stage design at a prespecified level α. For the following mathematical setup we will show how PWER-controlling critical values are obtained and how p-values and confidence intervals can be constructed.
In general, assume again that a testing problem $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ is given, where an overall population $\mathcal{P}$ consists of m possibly intersecting subpopulations $\mathcal{P}_i$, $i \in I = \{1, \ldots, m\}$. The partition is again given by $\mathcal{C}_\mathcal{P} := \{J \subseteq I : \mathcal{P}_J \neq \emptyset\}$, whereas $\mathcal{P}_J$ is defined as in (2.1). For a set $\mathfrak{U} \subseteq \mathcal{C}_\mathcal{P}$ of unions of populations $\mathcal{P}^U$, $U \in \mathfrak{U}$, we intend to test hypothesis pairs of the form

$$H^U : \theta(\mathcal{P}^U, T^U) \leq 0 \quad \text{vs.} \quad K^U : \theta(\mathcal{P}^U, T^U) > 0$$

with T^U being a treatment strategy to be investigated in $\mathcal{P}^U$ and $\theta^U := \theta(\mathcal{P}^U, T^U)$ being the mean efficacy difference between T^U and a control. Therefore, $\mathcal{H} = \{H^U \mid U \in \mathfrak{U}\}$ and $\boldsymbol{\Theta} = \mathbb{R}^{|\mathfrak{U}|}$.

3.1.1 PWER-control with critical values

Let us suppose that each H^U can be tested with a test statistic Z^U and assume that the joint distribution of $\mathbf{Z} = \{Z^U\}_{U \in \mathfrak{U}}$ is known at least approximately for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Further assume that the test statistics Z^U are constructed such that H^U is rejected if and only if Z^U exceeds a critical value $c^* \in \mathbb{R}$ (think of a Wald-type test statistic). In order to control the PWER at a prespecified significance level $\alpha \in (0, 1)$, we need to find $c^* = c^*(\alpha, \boldsymbol{\pi})$ such that $PWER_{\boldsymbol{\theta}^*} \leq \alpha$ under the least favorable parameter configuration $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, that is, under the configuration that maximizes the PWER. Under $\boldsymbol{\theta}^*$ the PWER is equal to the sum over all probabilities of making at least one type I error for the partition $\mathcal{P}_J$, $J \subseteq I$, weighted by the population size π_J of $\mathcal{P}_J$:

$$PWER_{\boldsymbol{\theta}^*} = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*} \left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \{Z^U > c^*\} \right) \leq \alpha. \tag{3.1}$$

As established in Theorem 2.3.1, under certain assumptions on the test statistics, the maximal PWER is obtained under the global null hypothesis, i.e. under $\boldsymbol{\theta}^* = \mathbf{0} = (0, \ldots, 0)$. If the (asymptotic) correlation between the test statistics only depends on the population prevalences π_J, $J \subseteq I$, which is often the case, then the PWER-level can be fully exhausted. Bear in mind, however, that this is only possible if the π_J are either known or can be estimated.
Since (3.1) only depends on one unknown parameter, the critical value c^*, this equation can be solved uniquely by applying a univariate root finding method. The root is unique since the PWER as a function of $c^* \in \mathbb{R}$ lies in $[0, 1]$ (when allowing $c^* \in \{-\infty, \infty\}$) and is strictly monotonously decreasing if the joint distribution of $\mathbf{Z}$ is continuous [1]. Since the PWER is always bounded by the FWER, this critical value is smaller than the one for FWER-control. Therefore the PWER leads to a higher power and a lower sample size to achieve a certain power. Note that the

[1] Or if the measure $\mathbb{P}$ has no atoms, i.e. there are no sets $A \subseteq \Omega$ with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) < \mathbb{P}(A)$ for all measurable $B \subset A$.

above approach may use population-specific critical values c^{*U}. Uniqueness of the solution could then be obtained by using weights for the individual critical values, for example by setting $c^{*U} = w^U c^{*U}$, where w^U denotes the population-specific weight for $U \in \mathfrak{U}$. In the cases presented in this thesis, however, we will only focus on the case of $w^U = 1$ for all U.
Another point to take note of is that the definition of the PWER as a weighted mean suggests that if we wanted to control the PWER at a level α, a small sub-population $\mathcal{P}_J$ would lead to higher probabilities of making a type I error in that sub-population. More precisely, the smaller π_J, the larger is the corresponding probability $\mathbb{P}_{\boldsymbol{\theta}^*}(\bigcup_{i\in J}\{Z_i \geq c^*\})$ meaning that the procedure can be seen as more liberal in $\mathcal{P}_J$. We will illustrate and critically discuss this with an example at the end of Section 3.3.2.

3.1.2 PWER-adjusted p-values and confidence intervals

Other very common ways to report statistical results of multiple testing procedures are simultaneous p-values and confidence intervals. In this section we will shortly describe how to extend these concepts in our PWER-setting. For a less general case, where treatments are investigated specifically in $\mathcal{P}_1, \ldots, \mathcal{P}_m$ instead of in unions $\mathcal{P}^U$, this has already been covered in [12].

PWER-adjusted p-values:
When testing a single (one-sided) hypothesis H_0 at a level α using a z-test statistic Z and a critical value c^* one definition is *the smallest significance level at which the test procedure rejects the null* which immediately leads to the p-value p being equal to $p = \mathbb{P}_{H_0}(Z \geq z)$ with z being the realization of Z. Here, H_0 is rejected if $p \leq \alpha$ and hence

$$p \leq \alpha \Leftrightarrow z \geq c^*.$$

In a multiple testing procedure, so called adjusted p-values are commonly used as they can be interpreted similarly as in the one hypothesis case: *the adjusted p-value p_j for a hypothesis H_j describes the smallest significance value at which the multiple test procedure rejects H_j.* To find an adjusted p-value for our PWER-controlling procedure, let us assume that each H^U, $U \in \mathfrak{U}$, is tested by means of some test statistic Z^U again and that H^U is rejected if and only if $z^U \geq c^{U*}$ with z^U being the observed value of Z^U and c^{U*} being critical values ensuring that the PWER is bounded by α. Let $\boldsymbol{\theta}^*$ be the least favorable configuration under which the c^{U*} have been determined and let p^U_{PWER} denote our supposed *PWER-adjusted p-value* for testing H^U. With

$$p^U_{PWER} = \sum_{J\subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U\in I_J(\boldsymbol{\theta})} \{Z^U \geq z^U\}\right) \tag{3.2}$$

we find that the equivalence

$$\forall U \in \mathfrak{U}: \quad p^U_{PWER} \leq \alpha \Leftrightarrow z^U \geq c^{U*} \tag{3.3}$$

holds for all $\alpha \in (0,1)$. That is because if $p^U_{PWER} \leq \alpha$ it is $z_j \geq c^*_j$ since c^*_j is the smallest critical value such that $PWER \leq \alpha$ and the converse is immediately clear. Hence, the test PWER-controlling test procedure for each H^U can also be conducted

by using the respective adjusted p-values p^U_{PWER}. Note that p^U_{PWER} also yields the smallest PWER-level the hypothesis H^U can be rejected with. This is because the smallest critical value that leads to a rejection of H^U is simply z^U. Since the equivalence between $p^U_{PWER} \leq \alpha$ and $z^U \geq c^{U*}$ holds for all significance levels α (and corresponding c^{U*}), we see that $P(p^U_{PWER} \leq \alpha) \leq \alpha$ and so p^U_{PWER} are conservative p-values in the classical univariate sense as well.

Simultaneous confidence intervals:
In this subsection we will show how to extend the PWER-controlling multiple test procedure to simultaneous confidence intervals for the parameter $\theta^U = \theta(\mathcal{P}^U, T^U)$, $U \in \mathfrak{U}$. We do this by using the duality between multiple hypothesis tests and simultaneous confidence intervals.
To this end, let $\boldsymbol{\vartheta} = (\vartheta^U)_{U\in\mathfrak{U}}$ be a vector of possible values for $\boldsymbol{\theta} = (\theta^U)_{U\in\mathfrak{U}} \in \boldsymbol{\Theta}$. Furthermore, consider the null hypotheses

$$H^{\theta^U} : \theta^U = \vartheta^U, \quad U \in \mathfrak{U}.$$

and assume that Z^{ϑ^U}, $U \in \mathfrak{U}$ are (asymptotically) pivotal test statistics for H^{ϑ^U} meaning that the (asymptotic) joint distribution of $(Z^{\vartheta^U})_{U\in\mathfrak{U}}$ under $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ is identical for all $\boldsymbol{\vartheta}$. One can build one-sided lower confidence intervals $\mathcal{C}^U = (-\infty, \tilde{\theta}^U]$ for the case of Z^{ϑ^U} decreasing in ϑ^U for the given data by using the lower boundary

$$\tilde{\theta}^U := \min\{\vartheta^U : Z^{\vartheta^U} \leq c^*\}. \tag{3.4}$$

c^* here denotes the critical value defined in (3.1) for $\boldsymbol{\theta}^* = \boldsymbol{\vartheta}$. This critical value is independent from $\boldsymbol{\vartheta}$ due to the pivotality of (Z^{ϑ^U}). In general, many (one sided) tests use a test statistic T^{ϑ^U} that is monotonous in ϑ^U like tests that use Wald-type test statistics $Z^{\vartheta^U} = (\hat{\theta}^U - \vartheta^U)/\mathrm{S}^U$ where $\hat{\theta}^U$ is an estimate of θ^U (e.g. the MLE) with a standard error S^U that is independent of the parameter value ϑ^U. In the case of a lower confidence interval we now get $\tilde{\theta}^U = \hat{\theta}^U - c^*\mathrm{S}^U$. Upper confidence bounds can be found analogously, where we get $(-\infty, \tilde{\theta}^U]$ with $\tilde{\theta}^U = \hat{\theta}^U + c^*\mathrm{S}^U$. Two-sided confidence intervals can then be obtained through the intersection lower and upper one-sided confidence intervals.
To show the coverage properties of the lower confidence intervals (the other two are analogous) consider that we have randomly drawn a patient P from $\mathcal{P}$. We define the random set $I_P = \{U \in \mathfrak{U} : P \in \mathcal{P}^U\}$ containing all indices of U the subpopulations $\mathcal{P}^U$ the patient P belongs to. This set is random because P is drawn from $\mathcal{P}$ at random. In particular, this set also gives us all population-specific efficacy parameters θ^U, $U \in I_P$, relevant for patient P. For the true (unknown) parameter of interest θ^U we observe that $\tilde{\theta}^U > \theta^U$ if and only if $Z^{\theta^U} > c^*$ by the definition of $\tilde{\theta}^U$ in (3.4). Since the dual tests for H^{θ^U}, $U \in \mathfrak{U}$, control the PWER, the (simultaneous) probability that any of the lower confidence bounds $\tilde{\theta}^U$, $U \in I_P$, exceed the true value θ^U is bounded by α leading to the coverage property

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\tilde{\theta}^U \leq \theta^U, \quad \forall U \in I_P\right) \geq 1 - \alpha. \tag{3.5}$$

So for a randomly chosen future patient the lower confidence intervals the lower confidence intervals $[\tilde{\theta}^U, \infty)$, $U \in \mathfrak{U}$, cover all true $\theta^U = \theta(\mathcal{P}^U, T^U)$ relevant to this patient with a probability of at least $1 - \alpha$. Because it is $I_P = \{U \in \mathfrak{U} \mid J \in U\}$ if and only if $P \in \mathcal{P}_J$ the coverage probability above can be expressed as

$$\sum_{J\subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}}\left(\tilde{\theta}^U \leq \theta^U \text{ for all } U \in \mathfrak{U} \text{ with } J \in U\right).$$

implying that (3.5) in a way controls an average of simultaneous coverage probability where we focus in each stratum on the relevant confidence statements and average the strata-wise coverage probability over the entire population $\mathcal{P}$.
The upper confidence bounds and two-sided confidence intervals control the same type of average simultaneous coverage probability. As for the classical confidence intervals, the two-sided interval have a twice as large non-coverage probability as the one-sided intervals.
Although the concepts described above are all available to us in the single-step single-stage case they require a different approach in (adaptive) group sequential designs (cf. [26], for example).

3.2 Investigating several treatments in each population

To apply the ideas from the previous section, let us assume that we intend to investigate more than one treatment in each of the unions of sub-populations $\mathcal{P}^U$, $U \in \mathfrak{U}$, against one common control C. For each U let

$$\mathcal{T}^U = \{T_1^U, \ldots, T_{m^U}^U\}$$

be a set of $m^U \geq 1$ treatments that are to be investigated in $\mathcal{P}^U$. Further, let $\mu_{T_l^U}^U \in \mathbb{R}$ be the true mean efficacy of treatment T_l^U in $\mathcal{P}^U$ and $\mu_C^U \in \mathbb{R}$ be the true mean efficacy of the control C in $\mathcal{P}^U$, respectively. We again make the assumption that each μ_C^U and each $\mu_{T_l^U}^U$ is a weighted sum of their involved stratum-wise effects. So, if $\theta_{T_l^U}^J$ and θ_C^J denote the true mean effect of T_l^U and C in $\mathcal{P}_J$ for $U \in \mathfrak{U}$ and $J \in U$, respectively, we say that the respective overall effects for $\mathcal{P}^U$ are of the form

$$\mu_{T_l^U}^U = \sum_{J \in U} \pi_J^U \theta_{T_l^U}^J \quad \text{and} \quad \mu_C^U = \sum_{J \in U} \pi_J^U \theta_C^J, \tag{3.6}$$

where $\pi_J^U = \pi_J / \pi^U$. We intend to test the hypotheses

$$H_l^U : \theta(\mathcal{P}^U, T_l^U) \leq 0 \quad \text{vs.} \quad K_l^U : \theta(\mathcal{P}^U, T_l^U) > 0, \quad \text{for} \quad l = 1, \ldots, m^U, \tag{3.7}$$

where the parameter $\theta(\mathcal{P}^U, T_l^U) = \mu_{T_l^U}^U - \mu_C^U$ here describes the efficacy of treatment T_l^U vs. the control C in $\mathcal{P}^U$. We allow for the sets $\mathcal{T}^U$ to be intersecting, i.e. we allow for a treatment to be tested in more than one population. So we are dealing with a population-wise test problem $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ with

$$\mathcal{H} = \{H_l^U \mid U \in \mathfrak{U}; \quad l = 1, \ldots, m^U\}$$

and

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\theta(\mathcal{P}^U, T_l^U))_{U,l} \mid U \in \mathfrak{U}; \quad l = 1, \ldots, m^U\}.$$

3.2.1 General case of multiple unequal treatments

One issue we face is the way patients from the disjoint subgroups $\mathcal{P}_J$, $J \subseteq I$, are supposed to be randomized to each treatment. Let N be the total number of patients recruited and $n_J = N\pi_J$ the stratum-wise sample size. Because each patient from $\mathcal{P}_J$ is by definition part of exactly $|A_J|$ different unions of sub-populations, patients from $\mathcal{P}_J$ are allocated to one of $t_J := |\mathcal{T}^J|$ treatments or the control,

where $\mathcal{T}^J := \bigcup_{U \in A_J} \mathcal{T}^U$. So if n_J denotes the number of observations drawn from $\mathcal{P}_J$, then, assuming equal allocation ratios, $n_J/(t_J+1)$ patients are allocated to each treatment from $\bigcup_{U \in A_J} \mathcal{T}^U$ and the control, respectively. In general, one can introduce randomization weights $r_{U,l} \in (0,1)$ with $\sum_{l=1}^{m_U} r_{U,l} = 1$ and then replace $n_J/(t_J+1)$ by $n_J r_{U,l}$ in all calculations below.
Now, to construct a test statistic Z_l^U for testing each H_l^U, we need to consider observations from the different strata $\mathcal{P}_J$. To this end, for any $J \subseteq I$ and $U \in A_J$ let $X^J_{T_l^U,i} \sim \mathrm{N}(\theta^J_{T_l^U}, (\sigma^J_{T_l^U})^2)$ and $X^J_C \sim \mathrm{N}(\theta^J_C, (\sigma^J_C)^2)$ be the i-th observation from treatment group T_l^U and C, respectively, for $i = 1, \ldots, n_J/(t_J+1)$. Technically, it is possible for $n_J/(t_J+1)$ to be a fractional number (not an integer) in which case it has to be rounded up to the next greatest integer, but for simplicity we will ignore this issue. Variances are assumed to be known and all observations are assumed to be independent from each other for the sake of simplicity as well. We define estimators for $\theta^J_{T_l^U}$ and θ^J_C as means given by

$$\hat{\theta}^J_{T_l^U} = \frac{1}{n_J/(t_J+1)} \sum_{i=1}^{n_J/(t_J+1)} X^J_{T_l^U,i} \quad \text{and} \quad \hat{\theta}^J_C = \frac{1}{n_J/(t_J+1)} \sum_{i=1}^{n_J/(t_J+1)} X^J_{C,i} \tag{3.8}$$

and estimators for the overall effects are

$$\hat{\mu}^U_{T_l^U} = \sum_{J \in U} \pi^U_J \hat{\theta}^J_{T_l^U} \quad \text{and} \quad \hat{\mu}^U_C = \sum_{J \in U} \pi^U_J \hat{\theta}^J_C. \tag{3.9}$$

Obviously, both estimators are unbiased because the means in (3.8) are unbiased and because of (3.6) and $\sum_{J \in U} (\pi_J/\pi^U) = 1$. Thus, we can construct Z_l^U for testing H_l^U as Wald-type test statistic by defining

$$Z_l^U = \frac{\hat{\mu}^U_{T_l^U} - \hat{\mu}^U_C}{\sqrt{\mathrm{Var}(\hat{\mu}^U_{T_l^U} - \hat{\mu}^U_C)}}. \tag{3.10}$$

The variance in the denominator can easily be computed by using the independence of the observations and basic properties of the variance,

$$\begin{aligned} V_l^U &:= \mathrm{Var}(\hat{\mu}^U_{T_l^U} - \hat{\mu}^U_C) \\ &= \sum_{J \in U} \left(\frac{\pi_J}{\pi^U}\right)^2 \mathrm{Var}(\hat{\theta}^J_{T_l^U} - \hat{\theta}^J_C) \\ &= \sum_{J \in U} \left(\frac{\pi_J}{\pi^U}\right)^2 \left(\mathrm{Var}(\hat{\theta}^J_{T_l^U}) + \mathrm{Var}(\hat{\theta}^J_C)\right) \\ &= \sum_{J \in U} \left(\frac{\pi_J}{\pi^U}\right)^2 \left(\frac{(\sigma^J_{T_l^U})^2}{n_J/(t_J+1)} + \frac{(\sigma^J_C)^2}{n_J/(t_J+1)}\right). \end{aligned} \tag{3.11}$$

We assume (as approximation) a multivariate normal distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ for the random test statistic vector[2] $\mathbf{Z} = (Z_l^U)_{U,l}$ with expectation vector $\boldsymbol{\nu}$ and correlation matrix $\boldsymbol{\Sigma}$. The expectation vector is simply given by $\boldsymbol{\nu} = (\nu_l^U)_{U,l}$ with

$$\nu_l^U = \theta(\mathcal{P}^U, T_l^U)/\sqrt{V_l^U}, \quad j = 1, \ldots, m \quad \text{and} \quad l = 1, \ldots, m_j. \tag{3.12}$$

[2]We note here that $\mathbf{Z}$ is indeed meant to be a vector although the notation $\mathbf{Z} = (\mathbf{Z}_l^U)_{U,l}$ suggests that $\mathbf{Z}$ is a matrix with entries Z_l^U.

The correlation matrix of $\mathbf{Z}$ is found by computing $\text{Cov}(Z_l^U, Z_{l'}^{U'})$ for each possible combination of $U, U' \in \mathfrak{U}$, $l = 1, \ldots, m^U$ and $l' = 1, \ldots, m^{U'}$. We can first notice that due to independence $\text{Cov}(Z_l^U, Z_{l'}^{U'}) = 0$ for all U, U' such that $U \cap U' = \emptyset$. We first see that

$$\begin{aligned}\text{Cov}(Z_l^U, Z_{l'}^{U'}) &= \frac{\text{Cov}(\hat{\mu}_{T_l^U}^U - \hat{\mu}_C^U, \hat{\mu}_{T_{l'}^{U'}}^{U'} - \hat{\mu}_C^{U'})}{\sqrt{V_l^U V_{l'}^{U'}}} \\ &= \sum_{J \in U} \sum_{J' \in U'} \left(\frac{\pi_J}{\pi^U}\right)\left(\frac{\pi_{J'}}{\pi^{U'}}\right) \frac{\text{Cov}(\hat{\theta}_{T_l^U}^J - \hat{\theta}_C^J, \hat{\theta}_{T_{l'}^{U'}}^{J'} - \hat{\theta}_C^{J'})}{\sqrt{V_l^U V_{l'}^{U'}}}\end{aligned}$$

Now, if $J \neq J'$ the covariance term will be 0 because of the independence assumption. Hence, the covariance reduces to

$$\begin{aligned}\text{Cov}(Z_l^U, Z_{l'}^{U'}) &= \sum_{J \in U \cap U'} \left(\frac{\pi_J^2}{\pi^U \pi^{U'}}\right) \frac{\text{Cov}(\hat{\theta}_{T_l^U}^J - \hat{\theta}_C^J, \hat{\theta}_{T_{l'}^{U'}}^J - \hat{\theta}_C^J)}{\sqrt{V_l^U V_{l'}^{U'}}} \\ &= \sum_{J \in U \cap U'} \left(\frac{\pi_J^2}{\pi^U \pi^{U'}}\right) \frac{\text{Cov}(\hat{\theta}_{T_l^U}^J, \hat{\theta}_{T_{l'}^{U'}}^J) + \text{Cov}(\hat{\theta}_C^J, \hat{\theta}_C^J)}{\sqrt{V_l^U V_{l'}^{U'}}},\end{aligned}$$

where the second equality holds due to $\text{Cov}(\hat{\theta}_{T_l^U}^J, \hat{\theta}_C^J) = 0$ (different treatment groups). We now also have $\text{Cov}(\hat{\theta}_C^J, \hat{\theta}_C^J) = \text{Var}(\hat{\theta}_C^J) = (\sigma_C^J)^2/(n_J/(t_J+1))$. For the other covariance expression we find that

$$\text{Cov}(\hat{\theta}_{T_l^U}^J, \hat{\theta}_{T_{l'}^{U'}}^J) = \begin{cases} 0, & \text{if} \quad T_l^U \neq T_{l'}^{U'} \\ \dfrac{(\sigma_{T_l^U}^J)^2}{n_J/(t_J+1)}, & \text{if} \quad T_l^U = T_{l'}^{U'} \end{cases}$$

because only if the same treatment is tested in both $\mathcal{P}^U$ and $\mathcal{P}^{U'}$ that treatment will be part of every $\mathcal{P}_J$ with $J \in U \cap U'$. To sum up we can write

$$\begin{aligned}\rho_{l,l'}^{U,U'} &:= \text{Cov}(Z_l^U, Z_{l'}^{U'}) \\ &= \frac{1}{\sqrt{V_l^U V_{l'}^{U'}}} \sum_{J \in U \cap U'} \left(\frac{\pi_J^2}{\pi^U \pi^{U'}} \frac{(\sigma_C^J)^2 + (\sigma_{T_l^U}^J)^2 \mathbf{1}(T_l^U = T_{l'}^{U'})}{n_J/(t_J+1)}\right)\end{aligned} \tag{3.13}$$

By now knowing the full distribution structure of $\mathbf{Z}$ the PWER for the parameter $\boldsymbol{\theta}^* = (\theta_l^{U*})_{U,l} = \mathbf{0}$ can be expressed in terms of the multivariate normal cdf:

$$\begin{aligned}PWER_{\boldsymbol{\theta}^*} &= \sum_{J \subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U \in A_J} \bigcup_{l=1}^{m^U} \{Z_l^U \geq c_l^U\}\right) \\ &= \sum_{J \subseteq I} \pi_J (1 - \Phi_{\boldsymbol{\Sigma}_J}(\mathbf{c}_J))\end{aligned}$$

Here, $\boldsymbol{\Sigma}_J$ denotes the correlation matrix of all Z_l^U with $U \in A_J$ and $l = 1, \ldots, m^U$ and $\mathbf{c}_J$ the corresponding critical values for rejecting the corresponding hypotheses H_l^U. Assuming equal critical values $c_l^U = c$ or some type of weighting approach like $c_l^U = w_l^U c$, with pre-specified weights w_l^U and $c \in \mathbb{R}$, for all U and l, the equation $PWER_{\boldsymbol{\theta}^*} = \alpha$ is uniquely solvable for some $\alpha \in (0, 1)$.

Example 3.2.1 (Two intersecting populations)**.** *We return to the case of $m = 2$ populations $\mathcal{P}_1, \mathcal{P}_2$ with $\mathfrak{U} = \{U_1, U_2\}$ and $U_j = \{\{j\}, \{1,2\}\}$ for $j = 1, 2$. Say, we are interested in testing whether treatment T_1 is efficient in $\mathcal{P}_1$ and T_2 is efficient in $\mathcal{P}_2$, respectively. So we have $\mathcal{T}^{U_j} = \{T_j\}$, $m^{U_j} = 1$ for $j = 1, 2$. For simplicity, let us assume that all $\sigma_G^J = \sigma$ for all treatment groups $G \in \{T_1, T_2, C\}$ and all $J \subseteq I$. To test $H^{U_j} : \theta(\mathcal{P}_j, T_j) \leq 0$ we use test statistics $Z_j := Z^{U_j} = \left(\hat{\theta}_{T_j}^{U_j} - \hat{\theta}_C^{U_j}\right)/\sqrt{V^{U_j}}$ for $j = 1, 2$.*
Now, we have $\mathcal{T}^{U_j} = \{T_j\}$ for $j = 1, 2$, $T_1 \neq T_2$, $t_{\{1\}} = t_{\{2\}} = 1$, $t_{\{1,2\}} = 2$ and the variance V_j for $j = 1, 2$ is computed as

$$V_j = \frac{2\sigma^2}{N}\left[\left(\frac{\pi_{\{j\}}}{\pi_j}\right)^2 \frac{2}{\pi_{\{j\}}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)^2 \frac{3}{\pi_{\{1,2\}}}\right] = \frac{2\sigma^2}{N\pi_j^2}(2\pi_{\{j\}} + 3\pi_{\{1,2\}})$$

and the covariance between Z_1 and Z_2 is given by

$$\begin{aligned}\rho_{1,2} &= \frac{N\pi_1\pi_2}{2\sigma^2\sqrt{(2\pi_{\{1\}} + 3\pi_{\{1,2\}})(2\pi_{\{2\}} + 3\pi_{\{1,2\}})}} \frac{3\sigma^2\pi_{\{1,2\}}^2}{N\pi_{\{1,2\}}\pi_1\pi_2} \\ &= \frac{3\pi_{\{1,2\}}}{2\sqrt{(2\pi_{\{1\}} + 3\pi_{\{1,2\}})(2\pi_{\{2\}} + 3\pi_{\{1,2\}})}}\end{aligned}$$

For equal population sizes, i.e. for $\pi_{\{1\}} = \pi_{\{2\}} = (1 - \pi_{\{1,2\}})/2$, this reduces to

$$\rho_{1,2} = \frac{3\pi_{\{1,2\}}}{2(1 + 2\pi_{\{1,2\}})}.$$

In both cases using the covariance (correlation) matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{1,2} \\ \rho_{1,2} & 1 \end{bmatrix}$ the PWER under the global null $\boldsymbol{\theta}^* = (0, 0)$ is equal to

$$PWER_{\boldsymbol{\theta}^*} = \sum_{j=1}^{2} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}(Z_j \geq c_j) + \pi_{\{1,2\}}\mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{j=1}^{2}\{Z_j \geq c_j\}\right) \tag{3.14}$$

$$= \sum_{j=1}^{2} \pi_{\{j\}}(1 - \Phi(c_j)) + \pi_{\{1,2\}}(1 - \Phi_{\boldsymbol{\Sigma}}(c_1, c_2)). \tag{3.15}$$

3.2.2 Investigating one single treatment in all populations

Now, imagine we want to test the efficacy of one single treatment T in each $\mathcal{P}^U$, $U \in \mathfrak{U}$, against a common control C. Thus, the set $\mathcal{T}^U$ just contains the treatment T for each $U \in \mathfrak{U}$ and our hypothesis pairs reduce to

$$H^U : \theta(\mathcal{P}^U, T) \leq 0 \quad \text{vs.} \quad K^U : \theta(\mathcal{P}^U, T) > 0, \quad U \in \mathfrak{U}, \tag{3.16}$$

where $\theta(\mathcal{P}^U, T) = \theta_T^U - \theta_C^U$ again describes the efficacy of T vs. C in $\mathcal{P}^U$.
Since the same treatments are tested in each population, there is no need to construct our test statistic via a weighted sum of the scores of the respective disjoint subpopulations. If N is the total sample size in the overall population, let $n^U = N\pi^U$ be the number of observations from $\mathcal{P}^U$. Then in each $\mathcal{P}^U$, we randomize $n^U/2$ patients to the treatment T and the other ones to the control C. For two intersecting

populations (Example 2.4.1), if $N = 100$ and $\pi_{\{1\}} = \pi_{\{2\}} = 0.4$ and $\pi_{\{1,2\}} = 0.2$, then $n_{\{1\}} = n_{\{2\}} = 40$ and $n_{\{1,2\}} = 20$. When randomizing 20 patients to T and 20 patients to control in $\mathcal{P}_{\{1\}}$ and 10 patients to T and C, respectively, in $\mathcal{P}_{\{1,2\}}$, then $20/30 = 2/3$ of the patients sampled from $\mathcal{P}_1$ assigned to T belong to $\mathcal{P}_{\{1\}}$ which exactly matches the proportion $\pi_{\{1\}}/(\pi_{\{1\}}+\pi_{\{1,2\}}) = 0.4/0.6 = 2/3$. Thus, there is no need to weigh the respective stratum-wise estimators with their relative population sizes. For each $J \in \mathcal{C}_\mathcal{P}$ let $X^J_{G,i} \sim \mathrm{N}(\theta^J_G, \left(\sigma^J_G\right)^2)$ with $G \in \{T, U\}$ and $i = 1, \ldots, n_J/2$ denote the observations in $\mathcal{P}^J$. Now, let the $\hat{\theta}^J_G$ be the mean estimator given by

$$\hat{\theta}^J_G = \frac{2}{n_J} \sum_{i=1}^{n_J/2} X^J_{G,i} \tag{3.17}$$

which is an unbiased estimator for θ^J_G. For each union $\mathcal{P}^U$ with $U \subseteq \mathcal{C}_\mathcal{P}$ we consider the sample size $n^U := \sum_{J\in U} n_J$ and define the estimator

$$\hat{\mu}^U_G = \frac{1}{n^U} \sum_{J\in U} n_J \hat{\theta}^J_G \tag{3.18}$$

which equals the mean over all observations drawn from $\mathcal{P}^U$. With $n_J/n^U = \pi_J/\pi^U$ we again get that this is an unbiased estimator for $\theta^U = \sum_{J\in U} \pi_J/\pi^U \theta^J_G$. For $U \in \mathfrak{U}$, let Z^U now be the test statistic

$$Z^U = \frac{\hat{\mu}^U_T - \hat{\mu}^U_C}{\sqrt{\mathrm{Var}(\hat{\mu}^U_T - \hat{\mu}^U_C)}} \tag{3.19}$$

for testing H^U, where the variance $V^U := \mathrm{Var}(\hat{\mu}^U_T - \hat{\mu}^U_C)$ is found by

$$V^U = \mathrm{Var}(\hat{\mu}^U_T) + \mathrm{Var}(\hat{\mu}^U_C) = \frac{2}{(n^U)^2} \sum_{J\in U} n_J \left(\left(\sigma^J_T\right)^2 + \left(\sigma^J_C\right)^2\right).$$

To calculate the correlation $\rho^{U,U'} = \mathrm{Cov}(Z^U, Z^{U'})$ for $U, U' \in \mathfrak{U}$ between two test statistics, we first note that for $U \cap U' = \emptyset$, it is $\rho^{U,U'} = 0$ due to the independence assumption of the observations. For $U \cap U' \neq \emptyset$ we find

$$\begin{aligned}\rho^{U,U'} &= \frac{\mathrm{Cov}(\hat{\mu}^U_T - \hat{\mu}^U_C, \hat{\mu}^{U'}_T - \hat{\mu}^{U'}_C)}{\sqrt{V^U V^{U'}}} = \frac{\mathrm{Cov}(\hat{\mu}^U_T, \hat{\mu}^{U'}_T) + \mathrm{Cov}(\hat{\mu}^U_C, \hat{\mu}^{U'}_C)}{\sqrt{V^U V^{U'}}} \\ &= \frac{\sum_{J\in U} \sum_{J'\in U'} n_J n_{J'} (\mathrm{Cov}(\hat{\theta}^J_T, \hat{\theta}^{J'}_T) + \mathrm{Cov}(\hat{\theta}^J_C, \hat{\theta}^{J'}_C))}{n^U n^{U'} \sqrt{V^U V^{U'}}}.\end{aligned}$$

For all $J \neq J'$ both covariance expressions are 0 due to independence. Otherwise, that is for $J \in U \cap U'$, it is $\mathrm{Cov}(\hat{\theta}^J_G, \hat{\theta}^J_G) = \mathrm{Var}(\hat{\theta}^J_G) = 2\left(\sigma^J_G\right)^2/n_J$ for $G \in \{T, C\}$. Thus, the correlation reduces to

$$\begin{aligned}\rho^{U,U'} &= \sum_{J\in U\cap U'} \frac{2n_J \left(\left(\sigma^J_T\right)^2 + \left(\sigma^J_C\right)^2\right)}{n^U n^{U'} \sqrt{V^U V^{U'}}} \\ &= \sum_{J\in U\cap U'} \frac{n_J \left(\left(\sigma^J_T\right)^2 + \left(\sigma^J_C\right)^2\right)}{\sqrt{\sum_{\tilde{J}\in U} n_{\tilde{J}} \left(\left(\sigma^{\tilde{J}}_T\right)^2 + \left(\sigma^{\tilde{J}}_C\right)^2\right)} \sqrt{\sum_{\tilde{J}'\in U'} n_{\tilde{J}'} \left(\left(\sigma^{\tilde{J}'}_T\right)^2 + \left(\sigma^{\tilde{J}'}_C\right)^2\right)}}.\end{aligned} \tag{3.20}$$

Assuming equal variances $\sigma_G^J = \sigma$ for all G and J, this formula further reduces to

$$\rho^{U,U'} = \frac{\sum_{J \in U \cap U'} n_J}{\sqrt{n^U n^{U'}}} = \frac{n^{U \cap U'}}{\sqrt{n^U n^{U'}}} = \frac{\pi^{U \cap U'}}{\sqrt{\pi^U \pi^{U'}}} \in [0,1], \tag{3.21}$$

where, again, $N\pi^U = n^U$, $U \subseteq \mathcal{C}_\mathcal{P}$, was used. Thus, for $\mathbf{Z} = (Z^U)_{U \in \mathfrak{U}}$ we find the correlation matrix $\mathbf{\Sigma} = \left(\mathrm{Cov}\left(Z^U, Z^{U'}\right)\right)_{U,U' \in \mathfrak{U}}$ by means of formulas (3.20) or (3.21) and the PWER under $\boldsymbol{\theta}^* = \mathbf{0}$ can be found via the multivariate normal distribution with correlation matrix $\mathbf{\Sigma}$ and expectation vector $\mathbf{0}$,

$$\mathrm{PWER}_{\boldsymbol{\theta}^*} = \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \left\{Z^U \geq c^U\right\}\right) \tag{3.22}$$

$$= \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J (1 - \Phi_{\mathbf{\Sigma}_J}(\mathbf{c}_J)). \tag{3.23}$$

Again, $\mathbf{\Sigma}_J$ denotes the correlation matrix of the sub-vector $\mathbf{Z}_J = \{Z^U \mid J \in U\}$ and $\mathbf{c}_J$ the vector of all critical values c^U with $J \in U$.

Example 3.2.2. *With the same setup as in Example 3.2.1, let us assume that only one treatment T is to be tested in both $\mathcal{P}_j = \mathcal{P}^{U_j}$, with $U_j = \{\{j\}, \{1,2\}\}$ for $j = 1, 2$. We further assume that the variances for each observation are equal to σ^2. With test statistics given by*

$$Z^{U_j} = \frac{\hat{\theta}_T^U - \hat{\theta}_C^U}{\sqrt{Var(\hat{\theta}_T^U - \hat{\theta}_C^U)}} = \frac{\hat{\theta}_T^U - \hat{\theta}_C^U}{2\sigma/\sqrt{n^U}}$$

we find the correlation matrix of $\boldsymbol{Z} = \left(Z^{U_1}, Z^{U_2}\right)$ by means of formula (3.21):

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \frac{\pi_{\{1,2\}}}{\sqrt{\pi_1 \pi_2}} \\ \frac{\pi_{\{1,2\}}}{\sqrt{\pi_1 \pi_2}} & 1 \end{bmatrix}$$

This matrix is then used to compute the PWER under $\boldsymbol{\theta}^* = (0,0)$:

$$PWER_{\boldsymbol{\theta}^*} = \sum_{j=1}^{2} \pi_{\{j\}} \left(1 - \Phi_{\mathbf{\Sigma}_{\{j\}}}\left(c^{U_j}\right)\right) + \pi_{\{1,2\}} \left(1 - \Phi_{\mathbf{\Sigma}}\left(c^{U_1}, c^{U_2}\right)\right) \tag{3.24}$$

Note that $\pi_{\{1,2\}}/\sqrt{\pi_1 \pi_2} = 2\pi_{\{1,2\}}(1 + \pi_{\{1,2\}})$ if $\pi_{\{1\}} = \pi_{\{2\}}$ is assumed, which is larger than the correlation in the unequal treatment case in Example 3.2.1, which in turn is to be expected because a larger number of patients in the intersection is part of the control group.

3.3 Comparing PWER- and FWER-control

In Theorem 2.2.1 we established that PWER-controlling test procedures are always more or at worst equally liberal than respective FWER-controlling procedures. Naturally, more liberality will lead to a larger power and thus to savings in sample size needed to reach a certain power. For the special case of two intersecting populations we will compare FWER- and PWER-controlling procedures with respect to power, sample size and type I error in different settings. We will do this by using the PWER-control methods described in Sections 3.1 and 3.2. In all of these settings we will deal with a population-wise testing problem of the form $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_\mathbf{\Theta}, \mathcal{H})$ where

- $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$ with complements $\mathcal{P}_{\{1\}} = \mathcal{P}_1 \setminus \mathcal{P}_2$, $\mathcal{P}_{\{2\}} = \mathcal{P}_2 \setminus \mathcal{P}_1$ and $\mathcal{P}_{\{1,2\}} = \mathcal{P}_1 \cap \mathcal{P}_2$,
- $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}}) \in [0,1]^3$ with $\sum_{J \subseteq \{1,2\}} \pi_J = 1$,
- $\boldsymbol{\Theta} = \mathbb{R}^2$,
- $\mathcal{H} = \{H_1, H_2\}$, with $H_j : \theta_j = \theta(\mathcal{P}_j, T_j) \leq 0$ for $j = 1, 2$,
- $\mathfrak{U} = \{U_1, U_2\}$ with $U_j = \{\{j\}, \{1,2\}\}$ for $j = 1, 2$.

First we compare both control methods in (i) the case of dealing with a combination of two independent studies whose patient populations intersect, (ii) a study where the treatments for $\mathcal{P}_1$ and $\mathcal{P}_2$ are identical ($T_1 = T_2$) and (iii) where they are different ($T_1 \neq T_2$). For all cases we will mainly use the results from Example 3.2.1 and Example 3.2.2 from the previous two sections. Finally, we will apply our PWER-concept to the multiple testing approach for umbrella trials examined in Sun *et al.* (2016) [51] and compare it to the originally used FWER-control. The results in this section can also be found in [12].

3.3.1 Combination of independent samples

We first want to start with a rather hypothetical situation. Let us assume two different intersecting sub-populations $\mathcal{P}_1$ and $\mathcal{P}_2$, defined by two different biomarkers, in both of which a treatment $T = T_1 = T_2$ is planned to be investigated. Moreover, suppose that it has been decided that the effect of the treatment T for the two biomarker positive groups is to tested in a clinical trial with two different parallel samples. Since the analysis of the two samples is submitted as a package to regulatory authorities, a multiple testing approach can be seen as a reasonable option to control the multiple type I error. Due to the overlap of both populations, to bound the overall probability for a future patient to be affected by an inefficient treatment, we aim to control the PWER, which can be seen as a compromise between FWER-control and testing the efficacy of T in each population one after another.
Following the setup of Example 3.2.2, we have two two treatment strategies $(\mathcal{P}_j, T)$, $j = 1, 2$, which are investigated in two independent samples. Thus, the respective test statistics $Z_j \sim \mathrm{N}(\theta_j/\sigma_j, 1)$, $j = 1, 2$, as in Example 3.2.2, are stochastically independent, so the correlation matrix $\boldsymbol{\Sigma}$ is simply the two-dimensional unit matrix. In the following we want to quantify the power gain when using PWER-control instead of FWER-control. To this end, for a prespecified $\alpha \in (0,1)$ let $c^*_{PWER} = c^*_{PWER}(\alpha)$ and $c^*_{FWER} = c^*_{FWER}(\alpha)$ be critical values that guarantee a level α PWER- and FWER-control, respectively.
The FWER under the global null $\boldsymbol{\theta}^* = (\theta^*_1, \theta^*_2) = (0,0)$ is equal to

$$\mathrm{FWER}_{\boldsymbol{\theta}^*} = 1 - \Phi(c^*_{FWER})^2$$

and is controlled at α by the Šidák's critical value $c^*_{FWER} = \Phi^{-1}(\sqrt{1-\alpha})$. From Example 3.2.2 we know that the PWER under $\boldsymbol{\theta}^*$ equals

$$\mathrm{PWER}_{\boldsymbol{\theta}^*} = (1 - \pi_{\{1,2\}})\,(1 - \Phi(c^*_{PWER})) + \pi_{\{1,2\}}\left(1 - \Phi(c^*_{PWER})^2\right),$$

which is only depending on the relative size of the overlap $\pi_{\{1,2\}}$. So the amount of multiplicity adjustment needed to control the PWER is solely determined by the

size of $\pi_{\{1,2\}}$. Finding c^*_{PWER} requires us to solve $PWER_{\boldsymbol{\theta}^*} = \alpha$, which is simply a quadratic equation with one unknown (namely, $\Phi(c^*_{\text{PWER}})$). Using standard solving techniques gives us the roots

$$c^*_{PWER} = \Phi^{-1}\left(\frac{-(1-\pi_{\{1,2\}}) + \sqrt{(1-\pi_{\{1,2\}})^2 + 4\pi_{\{1,2\}}(1-\alpha)}}{2\pi_{\{1,2\}}}\right). \tag{3.25}$$

One can see that for $\pi_{\{1,2\}}$ tending to 0 this critical value decreases to $\Phi^{-1}(1-\alpha)$ which is simply the critical value for the unadjusted case, whereas for $\pi_{\{1,2\}}$ tending to 1 one can see that c^*_{PWER} monotonically increases towards c^*_{FWER}. Also, because this critical value is only dependent on $\pi_{\{1,2\}}$ and not on $\pi_{\{1\}}$ and $\pi_{\{2\}}$, for a fixed value of $\pi_{\{1,2\}} = p_{1,2} \in [0,1]$ this critical value is the same for all population sizes $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, p_{1,2}) \in [0,1]^3$ with $\pi_{\{1\}} + \pi_{\{2\}} + p_{1,2} = 1$.

We can now quantify the gain in power associated with using the PWER as a type I error control measure. We do this by considering the ratio of the sample size needed for PWER- or FWER-control and the sample size that would be needed for the case where no multiplicity adjustments are made. For $c \in \{c^*_{PWER}, c^*_{FWER}\}$ and non-centrality parameter $\delta_j = \theta_j/\sigma_j$, the sample size for testing H_j in $\mathcal{P}_j$ must satisfy the inequality $n_c \geq (\Phi^{-1}(1-\beta) + c)^2/\delta_j^2$, when we opt for a marginal power of at least $1-\beta$. Now, we consider the ratios

$$q_\alpha(c) := \frac{n_c}{n_{\Phi^{-1}(1-\alpha)}} = \left(\frac{\Phi^{-1}(1-\beta) + c}{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha)}\right)^2 \quad \text{for } c \in \{c^*_{PWER}, c^*_{FWER}\}, \tag{3.26}$$

which quantify how much more sample size one needs for a marginal power of $1-\beta$ when using either of the methods to account for multiplicity (numerator) compared to the unadjusted method (denominator). We calculated different values of $q_\alpha(c)$ for $c \in \{c^*_{PWER}, c^*_{FWER}\}$ by varying $\pi_{\{1,2\}} \in [0,1]$. Figure 3.1 shows a plot of the calculations for $\alpha = 0.025$ and $1-\beta = 0.8$. We see that $q_\alpha(c^*_{FWER}) = 21\%$ for all $\pi_{\{1,2\}}$, where $c^*_{FWER} = \Phi^{-1}(\sqrt{0.975}) \approx 2.24$. However, depending on the size of $\pi_{\{1,2\}}$, PWER-control needs substantially less sample size, which makes sense because only a fraction ($\pi_{\{1,2\}}$) of the overall population can be potentially affected by more than one false rejection. But the larger the intersection, the closer the sample size increase gets to the value for FWER-control because, as mentioned above, the critical value increases with increasing $\pi_{\{1,2\}}$. To give a concrete example: The sample size increase for PWER-control is only at around 10% for a relative intersection size of $\pi_{\{1,2\}} = 40\%$ which is not even half of what is necessary if FWER-control is chosen. At last, note that for $\pi_{\{1,2\}} = 1$ the sample sizes are the same for PWER- and FWER-control, simply due to the fact that PWER and FWER coincide in this case.

3.3.2 Testing population specific effects in one study

To extend on Examples 3.2.1 and 3.2.2, let us consider a single study with two overlapping populations $\mathcal{P}_1$ and $\mathcal{P}_2$. In each $\mathcal{P}_j$, $j = 1, 2$, we compare a treatment T_j to the control C. Just as in the two examples 3.2.1 and 3.2.2 we will again consider the two scenarios of (i) unequal treatments and (ii) equal treatments. Assumptions on the distribution of the data is also the same as in the examples, i.e. that observations from each disjoint sub-population are i.i.d. normal with mean treatment

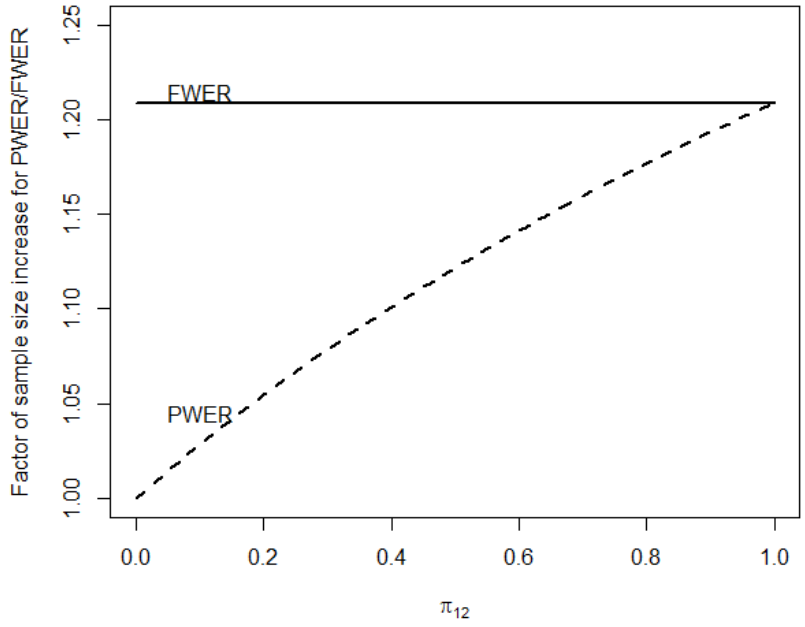

Figure 3.1: Factor of sample size increase compared to the unadjusted case to achieve a marginal power of $1-\beta = 80\%$ with PWER- and FWER-control at $\alpha = 0.025$ in a combination of two independent studies with different but overlapping populations.

difference $\theta_j = \theta(\mathcal{P}_j, T_j)$ (that is, treatment minus control) and a common known variance σ^2 across treatments and subgroups. We will also z-tests to test the hypotheses $H_j : \theta_j \leq 0$ at a level α. For the sample sizes in each disjoint subgroup $\mathcal{P}_J$, $J \subseteq I = \{1,2\}$, let $n_J = N \cdot \pi_J$ be the sample size in $\mathcal{P}_J$ and let $N = \sum_{J\subseteq\{1,2\}} n_J$ the overall total sample size. Test statistics and their correlation matrices are derived by means of the theory provided in Section 3.2.

(i) Unequal treatments: In case of $T_1 \neq T_2$ we have two different treatments, i.e. $\mathcal{T}_j = \{T_j\}$ and test for parameters $\theta_j = \theta(\mathcal{P}_j, T_j) = \mu_{T_j,j} - \mu_{C,j}$, for $j = 1, 2$, where

$$\mu_{T_j,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\theta_{T_j}^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\theta_{T_j}^{\{1,2\}} \quad \text{and} \quad \mu_{C,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\theta_{C}^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\theta_{C}^{\{1,2\}}$$

are defined over the strata-wise treatment effects. For each $J \subseteq I$, $j \in J$ and $G_j \in \{T_j, C\}$ we assume to have drawn realizations $X^J_{G_j,1}, \ldots, X^J_{G_j,n_J} \sim \mathrm{N}(\theta^J_G, \sigma^2)$ with a known value $\sigma > 0$. We estimate the above parameters by

$$\hat{\mu}_{T_j,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\hat{\theta}_{T_j}^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\hat{\theta}_{T_j}^{\{1,2\}} \quad \text{and} \quad \hat{\theta}_{C,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\hat{\theta}_{C}^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\theta_{C}^{\{1,2\}}$$

with $\hat{\theta}^J_{G_j}$, $G_j \in \{T_j, C\}$, given as the means in (3.8) with $t_J = 1$ for $J = \{j\}$ and $t_J = 2$ for $J = \{1,2\}$. Test statistics

$$Z_j = \frac{\hat{\theta}_{T_j,j} - \hat{\theta}_{C,j}}{\sqrt{\frac{2\sigma^2}{N\pi_j^2}(2\pi_{\{j\}} + 3\pi_{\{1,2\}})}} \sim \mathrm{N}(\delta_j, 1) \tag{3.27}$$

are used to test $H_j : \theta_j \leq 0$. Here $\delta_j := \theta_j/\sigma$ denotes the non-centrality parameter of the test. Since we have two different treatments, $n_{\{1,2\}}/3$ patients are randomized to the treatments T_1, T_2 and the control C, respectively, whereas in $\mathcal{P}_{\{j\}}$, $j = 1, 2$, there are $n_{\{j\}}/2$ patients randomized to T_j and the control, respectively. Assuming $\pi_{\{1\}} = \pi_{\{2\}}$, we see from Example 3.2.1 that $\rho_{1,2} = \frac{3\pi_{\{1,2\}}}{2(1+2\pi_{\{1,2\}})}$.

(ii) Equal treatments: In case of having $T_1 = T_2 = T$ the same treatment is investigated in both populations and by the considerations of Example 3.2.2 we a 1:1 randomization to every stratum $\mathcal{P}_J$, $J \subseteq I$, is conducted. From Example 3.2.2, we find that it is $\mathcal{T}_1 = \mathcal{T}_2 = \{T\}$ (one treatment for both populations) and $\theta(\mathcal{P}_j, T) = \mu_{T,j} - \mu_{C,j}$ for $j = 1, 2$, where

$$\mu_{T,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\theta_T^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\theta_T^{\{1,2\}} \quad \text{and} \quad \mu_{C,j} = \left(\frac{\pi_{\{j\}}}{\pi_j}\right)\theta_C^{\{j\}} + \left(\frac{\pi_{\{1,2\}}}{\pi_j}\right)\theta_C^{\{1,2\}}.$$

For $J \subseteq I$ and $G \in \{T, C\}$ we assume to have drawn realizations $X_{G,1}^J, \ldots, X_{G,n_J}^J \sim \mathrm{N}(\theta_G^J, \sigma^2)$ with a known value $\sigma > 0$ and $n_J = N\pi_J$. Test statistics

$$Z_j = \frac{\hat{\theta}_{T,j} - \hat{\theta}_{C,j}}{2\sigma/\sqrt{n_j}} \sim \mathrm{N}(\delta_j, 1) \tag{3.28}$$

are used to test $H_j : \theta_j \leq 0$. Here $\delta_j := \theta(\mathcal{P}_j, T)/\sigma$ denotes the non-centrality parameter of the test. As seen in 3.2.2, for $\pi_{\{1\}} = \pi_{\{2\}}$ the correlation between Z_1 and Z_2 equals $\rho_{1,2} = 2\pi_{\{1,2\}}/\big(1 + \pi_{\{1,2\}}\big)$ which is larger than the correlation in scenario (i) for all $\pi_{\{1,2\}} \in [0, 1]$.

Error control for both scenarios: Now, to obtain PWER- and FWER-control, respectively, we follow Examples 3.2.1 and 3.2.2, from where we see that the PWER under the global null is given by

$$PWER_{\boldsymbol{\theta}^*} = (1 - \pi_{\{1,2\}})\,(1 - \Phi(c^*_{PWER})) + \pi_{\{1,2\}}\,(1 - \Phi_{\boldsymbol{\Sigma}}(c^*_{PWER}, c^*_{\mathrm{PWER}})) \tag{3.29}$$

with correlation matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{1,2} \\ \rho_{1,2} & 1 \end{bmatrix}$ and the FWER is given by $\mathrm{FWER}_{\boldsymbol{\theta}^*} = 1 - \Phi_{\boldsymbol{\Sigma}}(c^*_{\mathrm{FWER}}, c^*_{\mathrm{FWER}})$. Setting both expressions equal to some $\alpha \in (0, 1)$, respectively, yields equations with only one unknown, which can easily be solved with a univariate root-finder. For instance, for scenario (i) ($T_1 \neq T_2$), let us assume we know that $\pi_{\{1\}} = \pi_{\{2\}} = 0.4$, $\pi_{\{1,2\}} = 0.2$. Furthermore, we set $\beta = 0.2$ and $\alpha = 0.025$. For the correlation, we then obtain $\rho_{1,2} = \mathrm{Corr}(Z_1, Z_2) \approx 0.01$, which we then use to construct the covariance (correlation) matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.01 \\ 0.01 & 1 \end{bmatrix}$. Solving $\mathrm{FWER}_{\boldsymbol{\theta}^*} = \alpha$ gives us $c^*_{\mathrm{FWER}} \approx 2.23$ and solving $\mathrm{PWER}_{\boldsymbol{\theta}^*} = \alpha$ yields $c^*_{\mathrm{PWER}} \approx 2.03$. Plugging these values into (3.26), we observe a sample size increase of approximately 20% for the FWER and only of 5% for the PWER.
In Figure 3.2 plots of sample size increases for (i) $T_1 \neq T_2$ and (ii) $T_1 = T_2$ are shown for varying values of $\pi_{\{1,2\}}$ for both PWER- and FWER-control. Again, in case of disjoint populations, i.e. if $\pi_{\{1,2\}} = 0$ we observe no sample size increase for the PWER, but a sample size increase of more than 20% when using FWER-control. The larger $\pi_{\{1,2\}}$ gets the smaller the difference between the sample size increases for

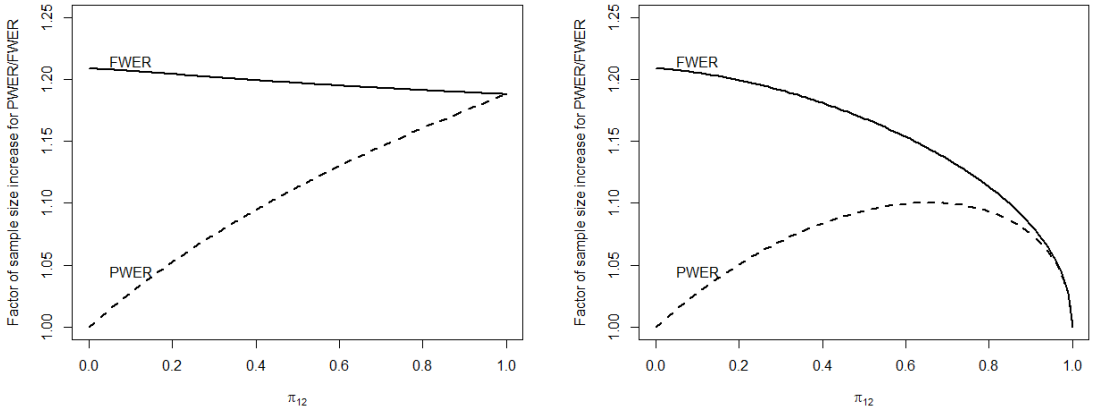

Figure 3.2: Factor of sample size increase compared to the unadjusted case for FWER- and PWER-control at $\alpha = 0.025$ in a single study with two overlapping populations depending on the size of the intersection $\pi_{\{1,2\}}$. The left panel is for scenario (i) with different experimental treatments and a common control; the right panel is for scenario (ii) with the equal experimental treatments. The power is $1 - \beta = 80\%$ in both scenarios.

PWER and FWER control become. At $\pi_{\{1,2\}} = 1$, where PWER = FWER holds, the two values coincide again. The differences regarding the curve shapes can be explained by the different correlations between Z_1 and Z_2. In both scenarios the correlation equals 0 when the populations are disjoint. In scenario (i), at $\pi_{\{1,2\}} = 1$ we have $\rho_{1,2} = 0.5$, whereas in scenario (ii) it is $\rho_{1,2} = 1$. Especially for scenario (ii) we see that PWER-control at $\pi_{\{1,2\}} = 1$ leads to the same value as if no multiplicity adjustments are made because this describes a situation where one treatment is tested in one population ($\mathcal{P}_1 \cap \mathcal{P}_2 = \mathcal{P}$). Also, since the correlation between the test statistics is higher in scenario (ii), the difference between the two curves is smaller for each value of $\pi_{\{1,2\}}$ than in scenario (i).
Further analyzing the PWER-curve of scenario (ii), we can conclude that the correlation of the test statistics and the degree of multiplicity adjustment that is needed in a way 'work against each other'. At $\pi_{\{1,2\}} = 0$ it is $\text{Cov}(Z_1, Z_2) = 0$, but there is also no need for multiplicity adjustments when using PWER-control because the populations are disjoint, while at $\pi_{\{1,2\}} = 1$ we only have one population $\mathcal{P}_1 = \mathcal{P}_2 = \mathcal{P}$, implying a correlation of 1, so no multiplicity adjustments are necessary, even though formally two hypotheses are tested for everyone. Therefore, there is no sample size increase here, either. This means that for $\pi_{\{1,2\}} \in (0, 1)$ there is a maximum $\pi^*_{\{1,2\}} \in (0, 1)$ yielding a maximal sample size increase for the PWER. For values $\pi_{\{1,2\}} < \pi^*_{\{1,2\}}$ the influence of $\text{Cov}(Z_1, Z_2)$ is smaller than the need for multiplicity adjustment. For values $\pi_{\{1,2\}} > \pi^*_{\{1,2\}}$ however the influence of the correlation dominates leading to a decrease in needed sample size increase. This can also be seen mathematically be expressing the PWER like this:

$$1 - \Phi(c^*_{\text{PWER}}) + \pi_{\{1,2\}}\{\Phi(c^*_{\text{PWER}}) - \Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}})\}$$

If $\pi_{\{1,2\}}$ is small, then the term $1 - \Phi(c^*_{\text{PWER}})$ dominates, which is independent of $\pi_{\{1,2\}}$, which leads to a critical value c^*_{PWER} being close to $\Phi^{-1}(1 - \alpha)$ – the critical value of the unadjusted case. For increasing $\pi_{\{1,2\}}$ the stronger the influence of

$\pi_{\{1,2\}}\{\Phi(c^*_{\text{PWER}}) - \Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}})\}$ becomes while $\Phi(c^*_{\text{PWER}}) - \Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}})$ decreases at a much slower rate. For larger $\pi_{\{1,2\}}$ this difference also vanishes, since

$$\Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}}) \to \Phi(c^*_{\text{PWER}})$$

for $\pi_{\{1,2\}} \to 1$. Note that the FWER always becomes maximal for $\pi_{\{1,2\}} = 0$, the case where the multiplicity adjustment is most questionable and the PWER equals the unadjusted level.

At last, we want to show a certain property of the PWER, namely, that the multiple type I error rate implicitly applied to the individual strata is increasing with decreasing strata-prevalence. This increase also exceeds the pre-defined significance level α in general. We illustrate this with scenario (ii). Aiming for a PWER-control at level $\alpha = 0.025$, the critical value c^*_{PWER} in (3.29) depends on $\pi_{\{1,2\}}$. Table 3.1 shows the multiple type I error $\mathbb{P}_{\boldsymbol{\theta}^*}(\bigcup_{j=1}^{2}\{Z_j \geq c^*\}) = 1 - \Phi_{\mathbf{\Sigma}}(c^*, c^*)$ for $\mathcal{P}_{\{1,2\}}$ with decreasing value of $\pi_{\{1,2\}}$ and respective critical value c^*_{PWER}. The values for $\mathbb{P}_{\boldsymbol{\theta}^*}(Z_j \geq c^*)$ are not listed because only one hypothesis is relevant for $\mathcal{P}_{\{j\}}$, $j = 1, 2$, and by looking at the critical values it is clear that these probabilities are less than α. We can basically see that for reasonably sized intersections the error probability does not exceed α by a large amount. The smaller the intersection gets, however, the closer the probability gets to $1 - (1 - \alpha)^2 \approx 0.0497$. So patients from the intersection would be exposed to a multiple error probability of almost 5%, which might still be seen as reasonably small. Assuming an analogous setup with 3 populations, however, would then mean that patients in a small intersection would be exposed to an error probability of around $1 - (1 - \alpha)^3 \approx 0.0731$, so α being almost tripled. The more populations are involved, the worse this problem will get, but one can also argue that really small intersections should probably just be excluded from the study anyway. Another measure one could take could be to define an upper boundary for these multiple type I error probabilities if it is ethically not justifiable. On the other hand, however, we find that the behaviour of the strata-wise type I errors is quite reasonable, since it improves power where required, namely for small strata and small sub-populations.
We want to conclude this subsection with another remark that has already been done in Brannath et al. (2021, [12]). One could ask whether the single step designs in scenario (i) and (ii) could be uniformly improved by a step-down procedure. The answer to this question is *no*. Let us assume that both H_1 and H_2 are true. One could now ask whether H_2 can be tested with a smaller critical value $c < c_{\text{PWER}}$ if H_1 has already been rejected with critical value c_{PWER}. Surely the type I error probabilities for $\mathcal{P}_{\{2\}}$ and $\mathcal{P}_{\{1,2\}}$ would increase because $\{Z_2 \geq c_{\text{PWER}}\}$ would be replaced by $\{Z_2 \geq c_{\text{PWER}}\} \cup \{Z_2 \geq c, Z_1 \geq c_{\text{PWER}}\}$, but since c_{PWER} is the smallest critical value such that the PWER is equal to α, we do not control the PWER at a level α for any $c < c_{\text{PWER}}$. One could conduct a test with a larger c_{PWER} in order to use a smaller c but this would not uniformly improve the single-step procedure.

3.3.3 Estimation of population prevalences

So far we have always assumed that the relative prevalences or population sizes $\boldsymbol{\pi} = (\pi_J)_{J \subseteq I}$ are known. In practice, this assumption is hard to justify, however. The question one now might ask is whether the use of an estimator $\hat{\boldsymbol{\pi}} = (\hat{\pi}_J)_{J \subseteq I}$ for $\boldsymbol{\pi}$ is safe to use, or in other words, whether the PWER is inflated substantially.

Table 3.1: Testing efficacy of an experimental treatment in two overlapping sub-populations with PWER-control of $\alpha = 0.025$. Critical value c^*_{PWER} and multiple type I error probability $1 - \Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}})$ for the intersection $\mathcal{P}_{\{1,2\}}$ of the two populations in dependence of its relative prevalence $\pi_{\{1,2\}}$.

	$\pi_{\{1,2\}}$				
	0.5	0.25	0.2	0.1	0.05
c^*_{PWER}	2.09	2.04	2.03	1.99	1.98
$1 - \Phi_{\mathbf{\Sigma}}(c^*_{\text{PWER}}, c^*_{\text{PWER}})$	0.031	0.038	0.04	0.044	0.047

Again defining $N = \sum_{J \subseteq I} n_J$ where n_J describes the number of observations for $\mathcal{P}_J$, $J \subseteq I$, we choose the the maximum likelihood estimator (MLE) of the multinomial distribution (see [1]) $\text{Mult}(N, \boldsymbol{\pi})$ as an estimator for $\boldsymbol{\pi}$. The probability density function of this multinomial distribution is given by

$$f((n_J)_{J \subseteq I}, \boldsymbol{\pi}) \;=\; \begin{cases} \frac{N}{\prod_{J \subseteq I} n_J!} \prod_{J \subseteq I} p_J^{n_J}, & \text{if } \sum_{J \subseteq I} n_J = N \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

The replacement of $\boldsymbol{\pi}$ by $\hat{\boldsymbol{\pi}}$ enables us to find a critical value $c^* = c^*(\hat{\boldsymbol{\pi}})$ by solving

$$\widehat{PWER} := (1 - \hat{\pi}_{\{1,2\}})\{1 - \Phi(c^*)\} + \hat{\pi}_{\{1,2\}}\{1 - \Phi_{\hat{\mathbf{\Sigma}}}(c^*, c^*)\} = \alpha, \tag{3.31}$$

where $\hat{\mathbf{\Sigma}}$ being the estimated correlation matrix of Z_1 and Z_2. This procedure implies asymptotic PWER-control because of the consistency of $\hat{\boldsymbol{\pi}}$ and the fact that the joint distribution of (Z_1, Z_2) used in the calculation c^* is conditional on $(n_J)_{J \subseteq I}$. We examine the behaviour of the PWER in the examples given by scenarios (i) and (ii) of Section 3.3.2. For a vector of true prevalences $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$, we generate sample size vectors $(\hat{n}_J)_{J \subseteq I}$ from the multinomial distribution with success probabilities $\boldsymbol{\pi}$ and a given value of N. Then we compute the MLE $\hat{\boldsymbol{\pi}}$ with components given by $\hat{\pi}_J = \hat{n}_J / N$ for $J \subseteq I$. In R this can be done with the `stats`-function `rmultinom`. To assess the degree of PWER-inflation we calculate the probabilities for a type I error for each sub-population $\mathcal{P}_J$ by using our "estimated" critical value $\hat{c}^*$ and the conditional correlation structure of the test statistics. We find the true PWER with the with the "estimated" critical value plugged in by weighting each of these probabilities by their respective true population prevalence π_J:

$$(1 - \pi_{\{1,2\}})\{1 - \Phi(\hat{c}^*)\} + \pi_{\{1,2\}}\{1 - \Phi_\rho(\hat{c}^*, \hat{c}^*)\} \tag{3.32}$$

For each constellation of $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$ this procedure was repeated 10.000 times. The true prevalence values for $(\pi_{\{1\}}, \pi_{\{2\}})$ were taken from a grid

$$\{0, 0.05, 0.1, \ldots, 1\}^2$$

(note that $\pi_{\{1,2\}} = 1 - \pi_{\{1\}} - \pi_{\{2\}}$). The mean of each of the above true PWER-expressions was taken as approximation of the actual overall PWER. This approximation was done for scenarios (i) and (ii) and $N = 50$ and $N = 100$, respectively. We found that the target PWER of 0.025 is only missed very slightly, even when using a small sample size like $N = 50$. More precisely, among all these four cases,

the mean true PWER values range from around $2.49 \cdot 10^{-2}$ to at worst $2.51 \cdot 10^{-2}$ where the largest standard error we observed is at around $1.2 \cdot 10^{-5}$, which all in all shows a fairly negligible deviation 0.025. To be fair, however, this also implies that the standard deviation of a single true PWER amounts to $1.2 \cdot 10^{-3}$, meaning that values of around 0.026 are for sure possible. The attached CD-Rom contains contour plots showing these results in more detail for each grid point.
So in a way, $\boldsymbol{\pi}$ can be treated as a nuisance parameter when aiming for PWER-control in a single-stage, single-step design. Therefore, for this thesis we will from now on assume known prevalences. In case of a small relative population size π_J it might be that, by chance, no patient is recruited in this group. This would mean that we would not account for all the multiplicity for these patients completely. If such intersection cannot be excluded theoretically from the inclusion and exclusion criteria or due to medical arguments, we could introduce a small minimal number π_{min} for all π_J in order to be more conservative. Also different approaches like shrinkage methods or Bayesian estimation of the π_J are conceivable options in future research.

3.3.4 Multiple testing approaches for umbrella trials

To assess the gain in power when using PWER-control instead of FWER-control in a single-stage design, we want to apply our method to a multiple testing approach for umbrella trials suggested in Sun *et al.* (2016, [51]). This application of our PWER-based procedure can also be found in [12]. Like in Sun *et al.* (2016), we assume m disjoint population strata $\mathcal{S}_1, \ldots, \mathcal{S}_m$. A certain experimental treatment E_i is supposed to be compared to a control C_i in each stratum. To keep it simple, we make the assumption that each population has the relative population size/prevalence $\pi_i = n_i/N$ with n_i being the number of individuals in $\mathcal{S}_i$ and $N = \sum_{i=1}^{m} n_i$ being the total number of individuals. By Section 3.3.3 we know that this assumption for π_i holds in practice at least approximately.
Because it is difficult or outright impossible to establish a treatment effect in the individual strata with a sufficiently high power in case small values of the n_i, study designs have been proposed that compare the global treatment strategy E which assigns treatment E_i and control C_i to the population strata $\mathcal{S}_i$. An overall comparison of the strategy E with C like this only requires the total sample size N and will not need any multiplicity adjustments, i.e. no multiple testing is required. If we are interested in the effect for a sub-population, however, then this method does not allow such claims if the effect of E is heterogeneous, and therefore Sun *et al.* (2016) propose to test all sub-strategies E^S, $S \subseteq \{1, \ldots, m\}$, that consider only the union $\mathcal{P}^S = \cup_{i \in S} \mathcal{S}_i$ with treatment assignments as in E, against the control in $\mathcal{P}^S$. This test procedure however requires adjustments for multiplicity for which Sun *et al.* (2016) provide a single-step procedure with FWER-control.
To formally describe of the the procedure of Sun *et al.* and our PWER-based procedure, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \boldsymbol{\Theta} = \mathbb{R}^m$ be the vector of unknown treatment effects (mean differences) in the populations $\mathcal{S}_i$, and for each $S \subseteq \{1, \ldots, m\}$ let the average treatment effect in $\mathcal{P}^S$ be given by

$$\theta^S = \sum_{i \in S} (\pi_i / \pi^S) \theta_i$$

where $\pi^S = \sum_{i \in S} \pi_i$ denotes the relative prevalence of $\mathcal{P}^S$. With X_{ij} denoting the treatment indicator for patient j in group i, which equals 1 if assigned to the

experimental treatment E_i and otherwise 0, and θ_i describing the treatment effect of E_i in population $\mathcal{S}_i$, Sun *et al.* assume the linear model

$$Y_{ij} = \mu_i + \theta_i X_{ij} + \varepsilon_{ij}. \tag{3.33}$$

Here the error terms ε_{ij} are all assumed to be i.i.d. normally distributed with expectation 0 and (homogeneous) variance σ^2. Sun *et al.* tested the hypotheses

$$H^S : \theta^S \leq 0 \quad \text{vs.} \quad K^S : \theta^S > 0 \quad \text{for all } S \subseteq L = \{1, \ldots, m\}. \tag{3.34}$$

Note that the $\mathcal{P}^S$ and H^S, $S \subseteq L$, correspond to the $\mathcal{P}^U$ and H^U from Section 2.2 in the following way: We have a population-wise testing problem $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ with an overall population $\mathcal{P} = \bigcup_{j=1}^m \mathcal{P}_j$ and a (here trivial) partition $\mathcal{C}_\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_m\}$, prevalences $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$, hypotheses $\mathcal{H} = \{H^S \mid S \subseteq L\}$ and parameter space $\boldsymbol{\Theta} = \{(\theta^S)_{S \subseteq L} \mid \theta^S \in \mathbb{R}\}$. The index set $\mathfrak{U}$ of $\mathcal{H}$ is therefore the whole power set (minus the empty set) of L.
One can obtain one-sided t-test statistics T^S for testing H^S for each $S \subseteq L$ based on the least squares estimates of the linear model. Sun *et al.* (2016) proposed a single-step procedure that compares the observed value of each T^S to the upper α-quantile c^*_{FWER} of the distribution of $\max\{T^S \mid S \subseteq L\}$ under the global null hypothesis, which corresponds to the case that there is no treatment E_i that is superior to the control C_i. The authors then eventually select one subset $S^*_{\text{FWER}} \subseteq L$ for which a positive treatment effect is claimed and define it as the set S for which the test statistic T^S is the largest:

$$S^*_{\text{FWER}} = \begin{cases} \arg\max_{S \subseteq I} T^S, & \text{if } \max\{T^S \mid S \subseteq L\} > c^*_{\text{FWER}} \\ \emptyset, & \text{else.} \end{cases} \tag{3.35}$$

To achieve PWER-control at the same level α, we determine the critical value c^*_{PWER} such that $\text{PWER} = \alpha$ holds under the global null hypothesis $\boldsymbol{\theta}^* = \mathbf{0}$. Note that while the populations $\mathcal{S}_1, \ldots, \mathcal{S}_m$ are disjoint, some of their unions $\mathcal{P}^S$ have a non-empty intersection and some do not. Because not all $\mathcal{P}^S$ overlap, the FWER aims to correct the multiple type I error rate for cases that cannot occur and can hence be seen as unnecessarily conservative (similarly to the example in Section 2.4.3).

The PWER under the global null hypothesis is given by

$$\text{PWER}_\mathbf{0} = \sum_{i \in L} \pi_i \mathbb{P}_\mathbf{0} \left(\bigcup_{S \ni i} \left\{ T^S \geq c^*_{\text{PWER}} \right\} \right), \tag{3.36}$$

with "$S \ni i$" denoting all $S \subseteq L$ that contain the index i because the population $\mathcal{S}_i$ is affected by at least one type I error if a hypothesis H^S corresponding to a population $\mathcal{P}^S$ for which $i \in S$ (and thus $\mathcal{S}_i \subseteq \mathcal{P}^S$) is erroneously rejected.

Now, a homogenous residual variance σ^2 was assumed and there are $2m$ mean parameters (treatment and control) in the linear model (3.33), the test statistics $\{T^S\}_{S \subseteq L}$ follow a joint t-distribution with $N - 2m$ degrees of freedom. In R one can use the distribution function of the multivariate t-distribution from the `mvtnorm`-package (see Genz *et al.*, 2017, [22]) by using function `pmvt`. This function requires the degrees of freedom and the correlation matrix of the test statistics as input (see e.g. Bretz *et al.*, 2016, [16]), where the latter can be found by using the contrast matrix

and the design matrix of the linear model. Probabilities in each addend of (3.36) can then be found by choosing the corresponding sub-matrices of the correlation matrix enabling us to numerically find the critical value c^*_{PWER} for known values of π_i, $i \in L$, and m to ensure PWER-control at level α.
Finally, note that because of $c^*_{\text{FWER}} > c^*_{\text{PWER}}$ we can already say that if the FWER-based approach selects a non-empty set S^*_{FWER}, this same set will also be selected by our PWER-approach. Conversely, though, the case of $S^*_{\text{FWER}} = \emptyset$ while $S^*_{\text{PWER}} \neq \emptyset$ can still occur, so PWER-control will always lead to higher power.

Performance measures: Sun *et al.* (2016) examined a handful of performance measures to evaluate the selected subset S^*. For example, they considered the average effect in the overall population when applying treatment strategy E^{S^*} in $\mathcal{P}^{S^*}$ and the control in the rest of the population. We will consider the relative quantity $\text{RAE} = 100\,\mathbb{E}\left(\sum_{i\in S^*} \pi_i\theta_i\right)/\theta_{\text{overall}}$ (relative average effect) where $\mathbb{E}$ is the expectation with respect to the sample distribution and θ_{overall} is the weighted average of the positive treatment effects,

$$\theta_{\text{overall}} = \sum_{i\in L_+} \pi_i\theta_i / \sum_{i\in L_+} \pi_i \quad \text{for} \quad L_+ = \{i = 1, \ldots, m : \theta_i > 0\},$$

that describes how efficient the experimental treatment strategy E is for the union of sub-populations that benefit from E. Note that the argument of the expectation in the RAE is random because S^* is random. Since the PWER-procedure chooses a non-empty S^* more often as the FWER-procedure, this quantity will always be larger for the PWER-approach.

In addition to this measure we will investigate the average size of the 'correctly' chosen subgroups within the selected ones, i.e. the average of $\pi^{S^*_+}/\pi^{S^*}$ where $S^*_+ = \{i \in S^* | \theta_i > 0\}$ and $\pi^{S^*_+} = \sum_{i\in S^*_+} \pi_i$. This gives the fraction of the patient cohort that benefits from the experimental treatment strategy within the one that is exposed to E^{S^*} by the results of the study. Analogously, we are interested in the average of the relative size of the 'falsely' chosen subgroups within the chosen ones: $\pi^{S^*_0}/\pi^{S^*}$ with $S^*_0 = \{i \in S^* | \theta_i \leq 0\}$. Lastly, like the authors we consider the probability of rejecting at least one false null hypothesis,

$$\text{Pow}_1 = \mathbb{P}(\text{ reject any } H^S \text{ with } \theta^S > 0,\, S \subseteq L\,),$$

as a way to measure the power of the procedures (see Section 2.6).

Design of the simulation: To make our results comparable to those of Sun *et al.* (2016), we conducted simulations with roughly the same parameters. That is, for the cases of $m = 2, 4, 6$ sub-populations and a significance level $\alpha = 0.025$, we chose a total sample size of $N = 1056$ and assume that all group-specific intercepts μ_i are equal to 0. Also, for simplicity, each group is assumed to be of equal size, i.e. $\pi_1 = \cdots = \pi_m$.
Like Sun *et al.* (2016), we assume non-negative effects $\theta_i \geq 0$ and choose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ depending on the number of sub-populations m and three other quantities. The first one is given by $q = m_0/m$ where m_0 denotes the size of $L_0 = \{i = 1, \ldots, m : \theta_i = 0\}$ which in turn describes the percentage of true null hypotheses. The second one characterizes the treatment effect heterogeneity and is defined as

$$\tau = (\theta_{\max} - \theta_{\min})/(\theta_{\max} + \theta_{\min})$$

where $\theta_{\max} = \max_{i \in L_+} \theta_i$ and $\theta_{\min} = \min_{i \in L_+} \theta_i$. This τ is equal to the relative half-range of the positive θ_i, that is, half their range divided by the average of their extremes. This means that a large τ implies a strong heterogeneity between the positive θ_i. The third one is the weighted average θ_{overall} as shown above.
One can find a grid of m equidistant points such that the three characteristics are satisfied given the values for q, m, τ and θ_{overall} and one can easily see that this grid is uniquely determined by these for quantities. We chose q such that $q \cdot m$ is always an integer. Note that if $q \geq (m-1)/m$, then there is at most one $\theta_i \neq 0$ implying that $\tau = 0$ (no heterogeneity) is the only possible value for τ.

Results: The simulation results for $m = 2$ and $m = 4$ are given in Table 3.2 and for $m = 6$ and $m = 8$ in Appendix A. From all tables we observe that PWER-control leads to a considerably larger power, average proportion of 'correctly' chosen subgroups and a larger average effect, when compared to FWER-control. On the other hand, though, PWER-control also leads to an increase of the proportion of 'falsely' chosen subgroups because a subgroup is selected more frequently with PWER-control.
While the proportion of 'falsely' chosen subgroups is increased by at most 2.2% (percentage points), which is fairly low, and remains below 5% (one-sided), the proportion of 'correctly' chosen subgroups (among the selected ones) and the power are increased by up to 10% and often by more than 5%. Also, the expected effect RAE is always larger with PWER-control.
Under the global null hypothesis ($q = 1$) the average proportion of 'falsely' selected populations is in theory equal to the one-sided FWER. With PWER-control at level 2.5% the FWER was found to be between 3.6% and 4.5% for $m = 2, 4, 6, 8$. Note that the average proportion of 'falsely' selected populations exceeds the level of 2.5% also with FWER-control when there is an effect in some but not all population strata.

All in all, we can summarize that PWER-control greatly increases the chance for delivering treatments that are actually efficient for patients in the selected subgroup. For the risk of receiving an inefficient treatment as well as the risk of patients not benefiting from the decisions regarding the treatment we see a moderate increase and can overall say that this risk remains comparable to the FWER-based procedure.

Table 3.2: Simulation results for $m = 2$ and $m = 4$. Results for power (%), the percentage of correctly and falsely chosen sub-populations and the relative average effect (RAE) for PWER- and FWER-control under parameter configurations $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ that depend on the fraction of true null hypotheses q and the relative half-range τ of the positive θ_i's.

		Pow	true	false	RAE	Pow	true	false	RAE
	$m = 2$	$q = 0$							
$\tau = 0$	PWER	36.4	36.4	0	2.9				
	FWER	31.0	31.0	0	2.5				
$\tau = 0.4$	PWER	40.4	40.4	0	3.3				
	FWER	34.6	34.6	0	2.8				
$\tau = 0.8$	PWER	51.2	51.2	0	4.7				
	FWER	45.2	45.2	0	4.2				
	$m = 2$	$q = 1/2$				$q = 1$			
$\tau = 0$	PWER	57.7	52.9	4.8	5.8	0	0	3.6	0
	FWER	52.0	47.8	4.2	5.2	0	0	2.4	0
	$m = 4$	$q = 0$				$q = 1/4$			
$\tau = 0$	PWER	36.2	36.2	0	2.3	42.2	38.8	3.5	3.0
	FWER	27.4	27.4	0	1.7	32.7	30.1	2.6	2.4
$\tau = 0.4$	PWER	37.9	37.9	0	2.5	44.8	41.3	3.6	3.3
	FWER	29.1	29.1	0	1.9	35.5	32.7	2.7	2.6
$\tau = 0.8$	PWER	43.0	43.0	0	3.0	52.7	48.9	3.7	4.2
	FWER	33.7	33.7	0	2.4	43.2	40.2	3.0	3.5
	$m = 4$	$q = 2/4$							
$\tau = 0$	PWER	53.2	45.7	7.6	4.6				
	FWER	43.8	37.8	6.1	3.8				
$\tau = 0.4$	PWER	58.8	51.1	7.8	5.0				
	FWER	49.5	43.1	6.4	4.2				
$\tau = 0.8$	PWER	73.9	65.9	8.0	6.8				
	FWER	65.3	58.5	6.8	6.0				
	$m = 4$	$q = 3/4$				$q = 1$			
$\tau = 0$	PWER	81.5	70.1	11.5	8.1	0	0	4.2	0
	FWER	75.1	64.9	10.2	7.5	0	0	2.4	0

4. General K-stage group sequential designs with PWER-control

As already discussed in the Introduction and Chapter 3, the use of single step designs has limitations not only due to financial and ethical reasons but also in terms of their flexibility. The ability to stop a trial early, which saves time and costs, is of even more importance when dealing with cancer patients. Particularly in our case of several overlapping and/or highly stratified patient populations, being able to stop the trial early, either for efficacy or futility, can save valuable resources. For this reason, and also in preparation for adaptive designs, this chapter will deal with group sequential clinical trial designs (GSDs) for intersection populations which ensure control of the PWER. In general, such a group sequential design will be defined by a number of predefined stages for each of which there is a set of hypotheses that is to be tested at each of these stages. A 'set' of hypotheses might sound rather vague and unspecific. The reason for this choice of words is that in a two-stage GSD with, say, two populations there are a plethora of options that one could follow. Say we have two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$ and at stage 1 we intend to simultaneously test a hypothesis H_1 with patients from $\mathcal{P}_1$ and H_2 with patients from $\mathcal{P}_2$. Assume for simplicity that both patient populations get the same treatment T. At stage two we could now follow different paths depending on the overall aim of the study. If we wanted to know whether a treatment performs significantly better than a control in each subgroup $\mathcal{P}_i$, we might want to keep testing H_i if it has not been rejected at stage 1 so far. If we wanted to demonstrate an effect in the intersection $\mathcal{P}_{\{1,2\}}$, we might want to test in the intersection at stage 2. Other possible options will be covered in the following chapter.

The outline of this chapter is as follows. At first, we present a general K-stage group sequential design allowing us to test pre-defined sets of hypotheses at each stage for arbitrary population structures while also controlling the PWER. In particular, this class will encompass the designs mentioned in the previous sections. Afterwards, some ways to control the PWER are proposed, in particular, adoptions of the Wang & Tsiatis power family and the error spending approach to ensure PWER-control. Theoretical foundation for both the one and the multiple treatment case are given,

similarly to Section 3.2. Then, we will discuss some special cases and will apply the above concepts to them.

4.1 Setup of the design

In the single-stage case from Section 3, we assumed a test $\boldsymbol{\varphi}$ for a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$, where we intended to test hypotheses of the form

$$H^U : \theta^U \leq 0 \quad \text{vs.} \quad K^U : \theta^U > 0, \quad \forall U \in \mathfrak{U},$$

for the superiority of the mean treatment effects $\theta^U = \theta(T^U, \mathcal{P}^U)$ in each union of partitioned sub-populations $\mathcal{P}^U$ in the case of one treatment per population $\mathcal{P}^U$. The test $\boldsymbol{\varphi}$ was given by a vector of individual decision functions φ^U for each hypothesis H^U, $U \in \mathfrak{U}$. If we now intend to test each hypothesis in a K-stage group sequential design, we have to redefine our test decision functions.
To this end, for each H^U, $U \in \mathfrak{U}$, let $\boldsymbol{\phi}^U = (\phi^{U,(k)})_{k \in I_K^U}$ be a vector of decision functions for testing H^U at stages $k \in I_K^U$, with $I_K^U \subseteq \{1, \ldots, K\}$ being the set of all stages H^U is planned to be tested at. It is $\phi^{U,(k)} = 1$ if and only if H^U has been rejected at stage k and 0 otherwise. We are thus allowing each H^U to be tested at some arbitrary, predefined number of stages. If, for example, H^U is planned to be tested at every stage, then $I_K^U = \{1, \ldots, K\}$, if it is only planned to be tested at stage K, then $I_K^U = \{K\}$ etc. For $k = 1, \ldots, K$, let us further define $\mathcal{H}_k := \{H^U \in \mathcal{H} : k \in I_K^U\}$ as the set of all hypotheses that are to be tested at stage k and let $\mathfrak{U}_k \subseteq \mathfrak{U}$ be the index set of $\mathcal{H}_k$. For testing each H^U at stage k we use some (Wald-type) accrued test statistic $Z^{U,(k)}$ depending on an estimator $\hat{\theta}^{U,(k)}$ for θ^U and some information $\mathcal{I}^{U,(k)}$. If $Z^{U,(k)}$ exceeds some critical value $c^{U,(k)}$, then H^U is said to be rejected at stage k. Thus, we define the test decision functions $\phi^{U,(k)}$ as $\phi^{U,(k)} = \mathbf{1}(Z^{U,(k)} \geq c^{U,(k)})$. The specifics of the distribution of $\mathbf{Z} = (Z^{U,(k)})$ are given in the subsequent sections where concrete examples are discussed. Now, as in classical GSDs, as soon as a hypothesis is rejected, it is not tested again at the subsequent stages it has been planned for, and a type I error is made if the hypothesis is rejected at any of those stages. Thus, for testing each H^U we use the test decision function $\boldsymbol{\varphi}$ with components

$$\varphi^U = \begin{cases} 1, & \text{if } \sum_{k \in I_K^U} \phi^{U,(k)} \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{for} \quad U \in \mathfrak{U}. \tag{4.1}$$

The PWER for this design is now simply given by

$$\begin{aligned} \mathrm{PWER}_{\boldsymbol{\theta}}(\boldsymbol{\varphi}) &= \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \{\varphi^U = 1\} \right) && (4.2) \\ &= \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \bigcup_{k \in I_K^U} \{\phi^{U,(k)} = 1\} \right) && (4.3) \\ &= \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \bigcup_{k \in I_K^U} \{Z^{U,(k)} \geq c^{U,(k)}\} \right). && (4.4) \end{aligned}$$

If we assume, say, a multivariate normal distribution for $\mathbf{Z}$ with a covariance matrix independent of all $\boldsymbol{\theta}$, then by Theorem 2.3.1 we know that $\boldsymbol{\theta}^* = \mathbf{0}$ is the least favorable configuration and control under it yields strong control of the PWER.

4.2 Methods for finding critical values

An issue classical GSD theory already has to face is the search for suitable critical values. As summarized in Chapter 1 several methods have been proposed to ensure type I error control. A straightforward way to find critical values to guarantee PWER-control is to adopt these methods. In principle, they encompass parameter families for the critical values themselves or the use of error spending functions to iteratively find critical values for each stage. Below, we will make use of both approaches for our K-stage GSDs. The former method can directly be applied to PWER-controlling methods by means of a Wang and Tsiatis parameterization of the critical value space. For the error spending method more work has to be done. The error spending approach described in Chapter 1 only works because the probability of falsely rejecting the null hypothesis at any stage can be decomposed into sum of probabilities that only concern the false rejection up to a certain stage. In the following, we show that the PWER can be decomposed into stage-wise PWER-expressions which only deal with false rejections up to a certain stage. This will allow us compute critical values analogously to the classical error spending approach.

4.2.1 The Wang & Tsiatis family

As described in Section 1.2.1, the critical value space is parametrized using a parameter $\xi \in \mathbb{R}$, such that the resulting equation $PWER_{\mathbf{0}}(\xi) = \alpha$ has a unique solution. However, we need a sensible definition of the term *information rate* with regards to the intersecting population structures we are working with. Given a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ with a K-stage group sequential test $\boldsymbol{\varphi}$ we consider two ways of defining information rates:

(i) Let $\tilde{n}^{(k)}$ be the (stage-wise) overall number of stage-k observations among all disjoint sub-populations and $\mathcal{J}_k := \bigcup_{U \in \mathfrak{U}_k} U$ be the set of all $\mathcal{P}_J$, $J \in \mathcal{C}_{\mathcal{P}}$, that are involved in tests at stage k. Then $\tilde{n}^{(k)}$ is given by

$$\tilde{n}^{(k)} = \sum_{J \in \mathfrak{J}_k} \tilde{n}_J^{(k)}, \quad \forall k = 1, \ldots, K. \tag{4.5}$$

Furthermore, let $\sum_{k=1}^{K} \tilde{n}^{(k)}$ be the maximal sample size of the whole design. Then we define the information rate at stage k as

$$\tau^{(k)} = \frac{\sum_{l=1}^{k} \tilde{n}^{(l)}}{\sum_{k=1}^{K} \tilde{n}^{(k)}}, \quad \forall k = 1, \ldots, K. \tag{4.6}$$

This definition mostly resembles that of the information rates used in classical GSDs. Note that by summing up all stage-wise sample sizes as in (4.5) the information rates $\tau^{(k)}$ in a way lose their dependency on each individual sub-population $\mathcal{P}_J$ since every possible combination of $(\tilde{n}_J^{(k)})_{J \in \mathcal{C}_{\mathcal{P}}}$ that results in the same sum will yield the exact same $\tau^{(k)}$.

(ii) A more population-oriented approach could be to define vectors of information rates at each stage, i.e. let $\boldsymbol{\tau}^{(k)} = (\tau^{U,(k)})_{U \in \mathfrak{U}_k}$ be a vector of stage k information

rates for each $\mathcal{P}^U$ with $H^U \in \mathcal{H}_k$. With $\tilde{n}^{U,(k)} = \sum_{J \in U} \tilde{n}_J^{(k)}$ each population-wise stage k information rate $\tau^{U,(k)}$ can be defined as

$$\tau^{U,(k)} = \frac{\sum_{l=1}^{k} \tilde{n}^{U,(l)}}{\sum_{l=1}^{K} \tilde{n}^{U,(l)}}. \tag{4.7}$$

That is, for each $H^U \in \mathcal{H}_k$ the rate $\tau^{U,(k)}$ is the ratio of all observations from $\mathcal{P}^U$ up to stage k divided by the maximal number of observations in $\mathcal{P}^U$. While this approach puts a stronger focus on each individual population, it is probably more unconventional in the sense that the term *information rate* usually quantifies the progress of the entire trial rather than the progress of testing in each population $\mathcal{P}^U$ individually. Therefore, this approach could also be seen as way of "extending" the notion of an information rate.

Now, let $\xi \in \mathbb{R}$ be a known (prespecified) parameter and $c \in \mathbb{R}$ be some unknown constant that implicitly depends on α, $\boldsymbol{\pi}$ and the information rates. Depending on the definition of our information rates, we parametrize critical values as follows.

(i) Similar to Wang & Tsiats we parametrize critical values $c^{(k)}$ for testing any hypotheses in $\mathcal{H}_k$ as

$$c^{(k)} = c \left(\frac{\tau^{(k)}}{\tau^{(1)}} \right)^{\xi - 0.5}, \quad \forall k = 1, \ldots, K.$$

(ii) For each $U \in \mathfrak{U}_k$ we test $H^U \in \mathcal{H}_k$ by using the population-specific critical value $c^{U,(k)}$ parameterized as

$$c^{U,(k)} = c \left(\frac{\tau^{U,(k)}}{\tau^{U,(1)}} \right)^{\xi - 0.5}, \quad \forall k = 1, \ldots, K. \tag{4.8}$$

Note that this parameterization implies that $c = c^{U,(1)}$ for all $U \in \mathfrak{U}_1$.

Regardless of choosing type (i) or (ii) each set of critical values is easily obtainable by means of a univariate root-finding algorithm (e.g. `uniroot` in R), due to c being the only unknown variable. In the following chapters, we will mainly use option (i), but option (ii) is also implemented as R-functions in the attached script files. The comparison of the two approaches with respect to power is a topic for future research.

4.2.2 Error-spending approach

As already described in Chapter 1, the ability to schedule interim analyses at fixed calendar times, rather than at times where a certain number observations have made, is of high importance in practice. To this end, we propose an adaptation of the classical error-spending approach that ensures PWER-control for the general K-stage GSD proposed in the previous section.

The classical error-spending approach uses an error-spending function $\alpha^* : [0, 1] \to [0, 1]$ with $\alpha^*(\tau) = 0$ if $\tau = 0$ and $\alpha^*(1) = 1$ if $\tau = 1$ which describes how much of

the prespecified type I error rate α is spent at each stage of the trial. $\tau^{(k)}$ denotes the information time of a K-stage trial at stages $k = 1, \dots, K$, which can either be fixed in advance or is observed through the course of the trial. If a hypothesis $H_0 : \theta \leq 0$ is to be tested by means of a z-test, for instance, the type I error is controlled by appropriately bounding the stage-wise rejection probabilities (under H_0) by the error-spending functions, i.e.

$$\mathbb{P}_{H_0}\left(\bigcap_{l=1}^{k-1}\{Z^{(l)} < c^{(l)}\} \cap \{Z^{(k)} \geq c^{(k)}\}\right) = \alpha^*(\tau^{(k)}) - \alpha^*(\tau^{(k-1)}), \quad k = 1, \dots, K, \tag{4.9}$$

with $\bigcap_{l=1}^{0}\{Z^{(l)} < c^{(l)}\} := \emptyset$ and $\tau^{(0)} := 0$. The critical values are then found iteratively by solving (4.9) using all the critical values of the previous stages.
We mimic this idea by decomposing the PWER into K 'stage-wise PWER' expressions. Heuristically, for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ this is done as follows:

$$\begin{aligned}
\text{PWER}_{\boldsymbol{\theta}} &= \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}}(\text{at least one type I error in } \mathcal{P}_J) \\
&= \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \sum_{k=1}^{K} \mathbb{P}_{\boldsymbol{\theta}}(\text{at least one type I error in } \mathcal{P}_J \text{ at stage } k) \\
&= \sum_{k=1}^{K} \underbrace{\sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}}(\text{at least one type I error in } \mathcal{P}_J \text{ at stage } k)}_{=:\text{PWER}_{\boldsymbol{\theta}}^{(k)}} \\
&= \sum_{k=1}^{K} \text{PWER}_{\boldsymbol{\theta}}^{(k)}
\end{aligned}$$

This decomposition would then enable us to find critical values in the same fashion as in the classical case, namely by iteratively solving the following equations under the LFC $\boldsymbol{\theta}^* = \mathbf{0}$:

$$\begin{aligned}
\text{PWER}_{\boldsymbol{\theta}^*}^{(1)}(c^{(1)}) &= \alpha^*(\tau^{(1)}) \\
\text{PWER}_{\boldsymbol{\theta}^*}^{(k)}(c^{(1)}, \dots, c^{(k)}) &= \alpha^*(\tau^{(k)}) - \alpha^*(\tau^{(k-1)}), \quad k = 2, \dots, K
\end{aligned}$$

Now, *at least one type I error in $\mathcal{P}_J$ at stage k* actually includes that we have not done any type I error relevant for $\mathcal{P}_J$ at the previous stages yet. This would translate to an intersection of rejection regions concerning stages $l < k$ intersected with a union of rejection regions concerning stage k. Computing the probability of such a set is not immediate and requires us to further rewrite these sets such that we are dealing with probabilities of either only intersections or unions of the individual rejection regions. The following considerations show that we can rewrite the PWER of our general K-stage GSD in exactly such a fashion. For the sake of simplicity, we assume that at each stage only one critical value is used (so we drop the dependence on the population $\mathcal{P}^U$), i.e. there are K critical values $c^{(1)}, \dots, c^{(K)}$.

Similar to the definition in Section 2.2, let $I_J^{(k)}(\boldsymbol{\theta}) = \{U \in \mathfrak{U}_k \mid \theta^U \in H^U \wedge J \in U\}$ for each $k = 1, \ldots, K$. Then the stage 1 critical value $c^{(1)}$ is found by solving

$$\mathrm{PWER}_{\boldsymbol{\theta}^*}^{(1)}(c^{(1)}) = \sum_{J \subseteq \mathcal{C}_\mathcal{P}} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U \in I_J^{(1)}(\boldsymbol{\theta}^*)} \{Z^{U,(1)} \geq c^{(1)}\} \right) = \alpha^*(\tau^{(1)}) \tag{4.10}$$

Using this critical value, we can find $c^{(2)}, \ldots, c^{(K)}$ by iteratively solving

$$\mathrm{PWER}_{\boldsymbol{\theta}^*}^{(k)}(c^{(k)}) = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcap_{l=1}^{k-1} \bigcap_{U \in I_J^{(l)}(\boldsymbol{\theta}^*)} \{Z^{U,(l)} < c^{(l)}\} \cap \bigcup_{U \in I_J^{(k)}(\boldsymbol{\theta}^*)} \{Z^{U,(k)} \geq c^{(k)}\} \right) \tag{4.11}$$

Set $M_J^k := |I_J^{(k)}(\boldsymbol{\theta}^*)|$. Since $M_J^k < \infty$ we can number the set $I_J^{(k)}(\boldsymbol{\theta}^*)$ as $I_J^{(k)}(\boldsymbol{\theta}^*) = \{U_1, \ldots, U_{M_J^k}\}$ and thus define sets $A_i^k := \{Z^{U_i,(k)} \geq c^{(k)}\}$ for $i = 1, \ldots, M_J^k$. We then get

$$\bigcup_{U \in M_J^k} \{Z^{U,(k)} \geq c^{(k)}\} = \bigcup_{i=1}^{M_J^k} A_i^k.$$

Now by defining $B_i^k := A_i^k \setminus \bigcup_{j=1}^{i-1} A_j^k = A_i^k \cap \bigcap_{j=1}^{i-1} (A_j^k)^c$ (with $B_1^k := A_1^k$) the above union can be further rewritten as disjoint union,

$$\bigcup_{i=1}^{M_J^k} A_i^k = \uplus_{i=1}^{M_J^k} B_i^k$$

and thus by the additivity of $\mathbb{P}_{\boldsymbol{\theta}^*}$ it follows

$$\mathrm{PWER}_{\boldsymbol{\theta}^*}^{(k)}(c^{(k)}) = \sum_{J \subseteq I} \pi_J \sum_{i=1}^{M_J^k} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcap_{l=1}^{k-1} \bigcap_{U \in M_J^l} \{Z^{U,(l)} < c^{(l)}\} \cap B_i^k \right),$$

which is an expression consisting of sums of probabilities over an intersection of sets. Therefore, commonly used cdfs like that of the multivariate normal distribution can directly be applied here. Note that since this error-spending method is based on decomposing the PWER into further addends, the result from Theorem 2.3.1 still holds and the LFC is just $\boldsymbol{\theta}^* = \mathbf{0}$. Regarding the choice of the error-spending function α^*, one can use any of the listed functions in Section 1.2.1. However, be mindful of the fact that error spending functions approximating the Pocock and O'Brien & Fleming boundaries in the classical setting (cf. [26, 55]) lose this approximative property in the PWER-setting. To sum up, the algorithm for finding critical values by means of our error-spending approach is as follows:

0. Prespecify the prevalences $\boldsymbol{\pi}$, the significance level α, the error spending function α^*, the number of stages K and the sets of hypotheses $\mathcal{H}_1, \ldots, \mathcal{H}_K$.
1. Find $c^{(1)}$ by solving $\mathrm{PWER}_{\mathbf{0}}^{(1)}(c^{(1)}) = \alpha^*(\tau^{(1)})$.

⋮ ⋮

k. At stage $k = 2, \ldots, K$ use $c^{(1)}, \ldots, c^{(k-1)}$ to find $c^{(k)}$ by solving $\text{PWER}_{\mathbf{0}}^{(k)}(c^{(1)}, \ldots, c^{(k-1)}) = \alpha^*(\tau^{(k)}) - \alpha^*(\tau^{(k-1)})$.

Finally, let us address a couple of practical problems that might be faced when using this approach. If we conduct a K-stage GSD where the hypotheses $H_1, \ldots, H_m$ are tested in $\mathcal{P}_1, \ldots, \mathcal{P}_m$, respectively (and no other ones), how are the prevalences $\boldsymbol{\pi}$ estimated throughout the trial? Surely, one could simply use stage 1 data to find an estimate $\hat{\boldsymbol{\pi}}^{(1)}$ (like the MLE of the multinomial distribution with success probabilities $\boldsymbol{\pi}$) and use this estimation for the whole trial, but sample sizes might not be high enough at an early stage to guarantee a reasonably good estimate. One other possibility could be to use accumulated sample sizes up to stage k to estimate the $\hat{\boldsymbol{\pi}}^{(k)}$ that is used to find the stage k critical values via $\text{PWER}_{\mathbf{0}}^{(k)}$. This would still ensure asymptotic control of the PWER even though the use of different estimates for $\boldsymbol{\pi}$ at each stage is inconsistent. Now, if only some hypotheses are rejected at stage 1 and some are not, then one would proceed to stage 2 only with those hypotheses that have not been rejected yet and thus would not get data from all populations necessary to estimate all $(\pi_J)_{J \subseteq I}$ (although one could technically ignore rejection at previous stages in a design without any futility bounds). One could not keep the latest estimation for a π_J that could have still been done up until stage k and find the re-estimate the remaining ones. This issue is certainly something that has to be researched more in future works.

4.3 GSDs for evaluating treatments in multiple populations

In Section 3.2 we have already discussed how to conduct single-stage designs with PWER-control where possibly multiple treatments were to be tested in each population. We constructed test statistics Z^U for H^U, $U \in \mathfrak{U}$ by writing them as a weighted sum of means of test statistics for each $J \in U$. We made a distinction between the cases of testing only one single treatment in all populations of interest and that of testing multiple treatments in those populations. In both cases, randomization in each of the strata (intersections) was done and estimations from those strata were summarized into an estimator for the parameter in each overall population of interest. With only one and the same treatment for each population, the test statistics for testing H^U in $\mathcal{P}^U$, $U \in \mathfrak{U}$, basically consisted of the mean over all observations from all the strata, whereas in the multiple treatment case, however, estimators had to be constructed by weighting each stratum-wise mean with the respective relative population sizes in order to avoid an incorrespondence of the relative sample sizes and the respective population sizes.

In this section, we will first further generalize the procedures from Section 3.2 to the group sequential case. In the single treatment case, this is straightforward as we consider a design where a treatment T is tested against a common control C as long as the associated null hypothesis can be rejected (or the maximal number of stages K is reached). Procedures from the previous sections can easily be applied if the correlation matrix of the involved test statistics is known. For the general case of testing multiple treatments, we consider a GSD that allows us to exclude certain treatments from the trial early if a superior effect to the control could have already been established at early stages, which then in turn allows us to randomize more patients to each individual treatment arm. Error rate control is not as

straightforward because this procedure will in general depend on the decisions made at previous stages since the allocation ratio can change for subsequent stages if a treatment has been dropped for efficacy at an earlier stage. We will see that the error spending approach can still be used to control the PWER at a pre-specified level, though. First, we will derive the correlation matrix and expectation vector of the involved test statistics, separately for the one treatment and multiple treatment case. Afterwards, PWER-controlling methods for both procedures are discussed. For the one treatment case, we will see that analogously to what we have learned in Section 3.2 this dependence on the decisions on the previous stages will disappear, whereas for the general multiple treatment case it won't. For the latter we propose a workaround by means of the error spending approach with the disclaimer that we have neither applied it to any example yet nor further examined it in any other way. So it should be seen as a future research topic.

4.3.1 Testing one treatment in multiple populations

We start with the case of investigating one single treatment $T^U = T$ in each population of interest $\mathcal{P}^U$, $U \in \mathfrak{U}$. As before, this treatment is compared to a common control C by means of testing the hypothesis pairs

$$H^U : \theta(\mathcal{P}^U, T) \leq 0 \quad \text{vs.} \quad \theta(\mathcal{P}^U, T) > 0 \tag{4.12}$$

with $\theta(\mathcal{P}^U, T) = \theta_T^U - \theta_C^U$ as in Section 3.2.2, which in turn are, again, given by the weighted sums of the respective stratum-wise effects as in (2.8). For the investigation of the efficacy of T vs. C in populations $\mathcal{P}^U$, for all $U \in \mathfrak{U}$, we consider a 1:1 randomization to treatment T and control C in each disjoint sub-population $\mathcal{P}_J$ that are subset of $\mathcal{P}^U$. Let $\tilde{n}_J^{(k)}$ be the stage-wise number of patients drawn from $\mathcal{P}_J$ at stage $k = 1, \ldots, K$, then $\tilde{n}_J^{(k)}/2$ patients are randomized to the treatment group and the other half to the control group. As in Section 3.2.2, for each treatment group $G \in \{T, C\}$ and each $J \in \mathcal{C}_\mathcal{P}$ we now assume independent stage k observations $X_{G,i}^{J,(k)} \sim \mathrm{N}(\theta_G^J, \left(\sigma_G^J\right)^2)$. As a stage-wise estimator for the stratum effects θ_G^J at stage k we consider the mean

$$\tilde{\hat{\theta}}_G^{J,(k)} = \frac{2}{\tilde{n}_J^{(k)}} \sum_{i=1}^{\tilde{n}_J^{(k)}/2} X_{G,i}^{J,(k)}, \quad G \in \{T, C\} \tag{4.13}$$

A stage-wise mean estimator for the $\tilde{n}^{U,(k)} = \sum_{J \in U} \tilde{n}_J^{(k)}$ observations in each $U \subseteq \mathcal{C}_\mathcal{P}$ is then given by

$$\tilde{\hat{\theta}}_G^{U,(k)} = \frac{1}{\tilde{n}^{U,(k)}} \sum_{J \in U} \tilde{n}_J^{(k)} \tilde{\hat{\theta}}_G^{J,(k)} \tag{4.14}$$

Now, to construct an accumulated test statistic $Z^{U,(k)}$ for testing H^U at stage k, we need an accumulated mean estimator for the observations up to stage k drawn from $\mathcal{P}^U$. Similarly to above, with $n^{U,(k)} = \sum_{\tilde{k}=1}^{k} \tilde{n}^{U,(\tilde{k})}$, we consider

$$\hat{\theta}_G^{U,(k)} = \frac{1}{n^{U,(k)}} \sum_{\tilde{k}=1}^{k} \tilde{n}^{U,(\tilde{k})} \tilde{\hat{\theta}}_G^{U,(\tilde{k})}. \tag{4.15}$$

Using the variance $V^{U,(k)} = \mathrm{Var}(\hat{\theta}_T^{U,(k)} - \hat{\theta}_C^{U,(k)})$ given by

$$V^{U,(k)} = \frac{2}{\left(n^{U,(k)}\right)^2} \sum_{\tilde{k}=1}^{k} \sum_{J \in U} \tilde{n}_J^{(\tilde{k})} \left(\left(\sigma_T^J\right)^2 + \left(\sigma_C^J\right)^2 \right)$$

we can define the accrued stage k test statistic for testing H^U as

$$Z^{U,(k)} = \frac{\hat{\theta}_T^{U,(k)} - \hat{\theta}_C^{U,(k)}}{\sqrt{V^{U,(k)}}}. \tag{4.16}$$

As done in Section 6, we will express the accrued test statistics as a weighted sum of stage-wise test statistics to simplify further computations. Thus, at each stage k, we consider the z-score $\tilde{Z}^{U,(k)}$ defined as

$$\tilde{Z}^{U,(k)} = \frac{\tilde{\hat{\theta}}_T^{U,(k)} - \tilde{\hat{\theta}}_C^{U,(k)}}{\sqrt{\mathrm{Var}\left(\tilde{\hat{\theta}}_T^{U,(k)} - \tilde{\hat{\theta}}_C^{U,(k)}\right)}}. \tag{4.17}$$

The variance $\tilde{V}^{U,(k)} := \mathrm{Var}(\tilde{\hat{\theta}}_T^{U,(k)} - \tilde{\hat{\theta}}_C^{U,(k)})$ in the denominator equals

$$\tilde{V}^{U,(k)} = \mathrm{Var}\left(\tilde{\hat{\theta}}_T^{U}\right) + \mathrm{Var}\left(\tilde{\hat{\theta}}_C^{U}\right) = \frac{2}{\left(\tilde{n}^{U,(k)}\right)^2} \sum_{J \in U} \tilde{n}_J^{(k)} \left(\left(\sigma_T^J\right)^2 + \left(\sigma_C^J\right)^2 \right)$$

as already seen in Section 3.2.2. As in Chapter 1, we now express the accrued test statistics for H^U as

$$Z^{U,(k)} = \sum_{\tilde{k}=1}^{k} w_k^{U,(\tilde{k})} \tilde{Z}^{U,(\tilde{k})} \tag{4.18}$$

with weights for $\tilde{k} = 1, \ldots, k$ given by

$$w_k^{U,(\tilde{k})} = \sqrt{\frac{\sum_{J \in U} \tilde{n}_J^{(k)} \left(\left(\sigma_T^J\right)^2 + \left(\sigma_C^J\right)^2 \right)}{\sum_{\tilde{k}=1}^{k} \sum_{J \in U} \tilde{n}_J^{(\tilde{k})} \left(\left(\sigma_T^J\right)^2 + \left(\sigma_C^J\right)^2 \right)}}. \tag{4.19}$$

Note that for equal variances $\sigma_T^J = \sigma_C^J = \sigma$ for all $J \in \mathcal{C}_{\mathcal{P}}$, the above weights simply reduce to $w_k^{U,(\tilde{k})} = \sqrt{\tilde{n}^{U,(\tilde{k})} / \sum_{\tilde{k}=1}^{k} \tilde{n}^{U,(\tilde{k})}}$, which are the square roots of the information rates for testing H^U, so to say.
Computation of the correlation structure of $\mathbf{Z} = \left(Z^{U,(k)}\right)$ is now quite straightforward and only depends on the stage-wise correlations between each component of $\mathbf{Z}$ due to our assumption of data being independently drawn at each stage and each population. For $1 \leq s \leq t \leq K$ and $U, U' \in \mathfrak{U}$ with $U \cap U' \neq \emptyset$, the covariance (which is equal to the correlation due to standardization) of $Z_U^{(s)}$ and $Z_{U'}^{(t)}$ is given by

$$\mathrm{Cov}(Z^{U,(s)}, Z^{U',(t)}) = \mathrm{Cov}\left(\sum_{l=1}^{s} w_s^{U,(l)} \tilde{Z}^{U,(l)}, \sum_{j=1}^{t} w_t^{U',(j)} \tilde{Z}^{U',(j)} \right)$$

$$= \sum_{l=1}^{s} \sum_{j=1}^{t} w_s^{U,(l)} w_t^{U',(j)} \mathrm{Cov}(\tilde{Z}^{U,(l)}, \tilde{Z}^{U',(j)})$$
$$= \sum_{l=1}^{s} \sum_{j=1}^{s} w_s^{U,(l)} w_t^{U',(j)} \mathrm{Cov}(\tilde{Z}^{U,(l)}, \tilde{Z}^{U',(j)}),$$

with $\rho_l^{U,U'} := \mathrm{Cov}(\tilde{Z}^{U,(l)}, \tilde{Z}^{U',(l)})$. Note that the second sum in the third equation above only goes to s since the stage-wise test statistics are independent for different stages. Thus, the pairwise covariances $Cov(\tilde{Z}_U^{(l)}, \tilde{Z}_{U'}^{(j)})$ are equal to 0 for $l \neq j$ and we only have to deal with the case of $l = j$. Ultimately the covariances of the accrued test statistics are given by

$$\mathrm{Cov}(Z^{U,(s)}, Z^{U',(t)}) = \begin{cases} \sum_{l=1}^{s} w_s^{U,(l)} w_t^{U',(l)} \rho_l^{U,U'}, & U \cap U' \neq \emptyset \\ 0, & \text{else} \end{cases}, \quad \text{for } s \leq t. \tag{4.20}$$

Especially for $U = U'$ and $s = t$ the above expression is equal to 1 since the sqaured weights sum up to 1. The computation of the correlation between two stage-wise test statistics is straightforward again because we can exploit the independence of the observations. As in the single-stage case only those $\mathcal{P}_J$ with $J \in U \cap U'$ are relevant for the correlation. Altogether, we obtain the same formula as in the single-stage case (3.20)

$$\rho_l^{U,U'} := \sum_{J \in U \cap U'} \frac{\tilde{n}_J^{(l)} \left(\left(\sigma_T^J\right)^2 + \left(\sigma_C^J\right)^2 \right)}{\sqrt{\sum_{\tilde{J} \in U} \tilde{n}_{\tilde{J}}^{(l)} \left(\left(\sigma_T^{\tilde{J}}\right)^2 + \left(\sigma_C^{\tilde{J}}\right)^2 \right)} \sqrt{\sum_{\tilde{J}' \in U'} \tilde{n}_{\tilde{J}'}^{(l)} \left(\left(\sigma_T^{\tilde{J}'}\right)^2 + \left(\sigma_C^{\tilde{J}'}\right)^2 \right)}} \tag{4.21}$$

which, again, simplifies to $\rho_l^{U,U'} = \pi^{U \cap U'} / \sqrt{\pi^U \pi^{U'}}$, when using $\tilde{n}_J^{(l)} = N^{(l)} \pi_J$, if $N^{(l)}$ denotes the number of observations from $\mathcal{P}$ at stage k.
The expectation of a $Z^{U,(k)}$ is simply given by

$$\nu^{U,(k)} := \mathbb{E}(Z^{U,(k)}) = \theta(\mathcal{P}^U, T) / \sqrt{V^{U,(k)}}. \tag{4.22}$$

With equal variances, this reduces to

$$\nu^{U,(k)} = \frac{\theta(\mathcal{P}^U, T)}{2\sigma} \sqrt{n^{U,(k)}}. \tag{4.23}$$

Now, PWER-control at level $\alpha \in (0,1)$ can be obtained by means of the multivariate normal distribution of $\mathbf{Z} = \left(Z^{U,(k)}\right)$ under the least favorable configuration $\boldsymbol{\theta}^* = \mathbf{0}$ with correlation matrix $\boldsymbol{\Sigma}$ with entries given by (4.20). From (4.2) we know the PWER equals

$$\mathrm{PWER}_{\boldsymbol{\theta}^*} = \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*} \left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \bigcup_{k=1}^{K} \left\{ Z^{U,(k)} \geq c^{U,(k)} \right\} \right) \tag{4.24}$$
$$= \sum_{J \in \mathcal{C}_\mathcal{P}} \pi_J (1 - \Phi_{\boldsymbol{\Sigma}_J}(\mathbf{c}_J)), \tag{4.25}$$

where $\boldsymbol{\Sigma}_J$ constitutes the correlation matrix of the sub-vector $\mathbf{Z}_J = \left\{ Z^{U,(k)} \mid J \in U \right\}$ and $\mathbf{c}_J$ being the vector of all critical values $c^{U,(k)}$ for $k = 1, \ldots, K$ and $U \in \mathfrak{U}$ such that $J \in U$.

4.3.2 Testing multiple treatments in multiple populations

We now want to generalize the single-stage procedure from Section 3.2, where multiple treatments per population are investigated, to a group sequential procedure with PWER-control. So similar to Section 3.2, we have a set of $m^U \geq 1$ to-be-tested treatments

$$\mathcal{T}^U := \{T_1^U, \ldots, T_{m^U}^U\}$$

for each $U \in \mathfrak{U}$. For each treatment $T^{U,l}$ we test a hypothesis pair

$$H_l^U : \theta_l^U = \theta(\mathcal{P}^U, T_l^U) \leq 0 \quad \text{vs.} \quad K_l^U : \theta_l^U = \theta(\mathcal{P}^U, T_l^U) > 0,$$

where $\theta_l^U = \theta_{T_l^U}^U - \theta_C^U$ and both $\theta_{T_l^U}^U$ and θ_C^U are given by

$$\theta_{T_l^U}^U = \sum_{J \in U} \pi_J^U \theta_{T_l^U}^J \quad \text{and} \quad \theta_C^U = \sum_{J \in U} \pi_J^U \theta_C^J,$$

a weighted sum of stratum-wise effects.

We want to conduct the following K-stage GSD. At stage 1, each treatment T_l^U in $\mathcal{T}^U$ is investigated in each $\mathcal{P}^U$, $U \in \mathfrak{U}$, by means of a hypothesis test for H_l^U by comparing T_l^U with the control C. Once H_l^U is rejected, the treatment is dropped from the study and the remaining treatments are tested again at stage 2 with a larger sample size. This procedure is repeated until stage K is reached or all hypotheses have been rejected. Excluding treatments from the trial leaves more room to expend resources for the remaining treatments so more patients can be tested with them.

Say that at stage $k = 1, \ldots, K$ we test each of these hypotheses H_l^U with a test decision function $\phi_l^{U,(k)}$ with $U \in \mathfrak{U}$ and $l = 1, \ldots, m^U$. To have $\phi^{U,(k)}$ be defined for each stage $k = 1, \ldots, K$ we set $\phi_l^{U,(k')} = 1$ for all $k > k'$ if a rejection of H_l^U at stage k occured. Similar to the considerations in Section 3.2, at stage 1, we need to randomize patients from each $\mathcal{P}_J$ to one of the treatments in $\mathcal{T}_J^{(1)} := \bigcup_{U \in A_J} \mathcal{T}^U$. Then we have $t_J^{(1)} := |\mathcal{T}_J^{(1)}| + 1$ treatment groups in each $\mathcal{P}_J$ with $J \in U$ and $U \in \mathfrak{U}$. For each $J \in \mathcal{C}_\mathcal{P}$ and stages $k = 1, \ldots, K$ now let $\tilde{n}_J^{(k)}$ be the number of patients newly recruited in $\mathcal{P}_J$ at stage k and for $k \geq 2$ let

$$\mathcal{T}_J^{(k)} = \left\{ T \in \mathcal{T}_J^{(1)} \mid T = T_l^U : \phi_l^{U,(k-1)} = 0 \quad \forall k' < k, U \in A_J \right\} \tag{4.26}$$

be the set of all treatments T_l^U with $J \in U$ that have yet to be tested at stage k. Obviously, this set is dependent on the data from previous stages because of its dependence on the test decision function. We have seen that in the one treatment case this will not pose any problems regarding error control but in the general case, it will.

With $t_J^{(k)} := |T_J^{(k)}| + 1$ treatment groups left in $\mathcal{P}_J$ at stage k, in each $\mathcal{P}_J$ we equally randomize $\tilde{n}_J^{(k)}/(t_J^{(k)} + 1)$ patients to either of the treatments in $\mathcal{T}_J^{(k)}$ or the control C.

To construct a test statistic for testing H_l^U at stage k, let $X_{T_l^U,i}^{J,(k)} \sim \mathrm{N}\left(\theta_{T_l^U}^J, \left(\sigma_{T_l^U}^J\right)^2\right)$ be the i-th stage k observation in $\mathcal{P}_J$ for treatment $T_l^U \in \mathcal{T}_J^{(k)}$, where $J \in \mathcal{C}_\mathcal{P}$ and $i = 1, \ldots, \tilde{n}_J/(t_J^{(k)} + 1)$. Consider the stage-wise and stratum-wise estimators for data in $\mathcal{P}_J$ at stage k as

$$\tilde{\hat{\theta}}_G^{J,(k)} = \frac{t_J + 1}{\tilde{n}_J^{(k)}} \sum_{i=1}^{\tilde{n}_J^{(k)}/(t_J+1)} X_{G,i}^{J,(k)}, \quad G \in \{T_l^U, C\} \tag{4.27}$$

only consisting of the data that newly emerged at stage k (note the tilde-sign above the hat). To summarize all data up to stage k into a cumulative mean, we write the stratum-wise estimators into a weighted sum of stage-wise stratum-wise estimators,

$$\hat{\theta}_G^{J,(k)} = \sum_{l=1}^{k} \left(w_{J,k}^{(l)}\right)^2 \tilde{\hat{\theta}}_G^{J,(l)}, \quad \text{with} \quad w_{J,k}^{(l)} := \sqrt{\frac{\tilde{n}_J^{(l)} / \left(t_J^{(l)} + 1\right)}{\sum_{\tilde{k}=1}^{k} \tilde{n}_J^{(\tilde{k})} / \left(t_J^{(\tilde{k})} + 1\right)}}. \tag{4.28}$$

and the accrued mean estimator for data up to stage k in a $\mathcal{P}^U$ is then given by

$$\hat{\theta}_G^{U,(k)} = \sum_{J \in U} \pi_J^U \hat{\theta}_G^{J,(k)}. \tag{4.29}$$

Since $\hat{\theta}_G^{J,(k)}$ is unbiased for θ_G^J, the estimator $\hat{\theta}_G^{U,(k)}$ is unbiased for $\theta_G^U = \sum_{J \in U} \pi_J^U \theta_G^J$ with $G \in \{T_l^U, C\}$. If $Z_l^{U,(k)}$ denotes the accrued test statistic consisting of all data from $\mathcal{P}^U$ up to stage k, then we consider the Wald-statistic (z-statistic)

$$Z_l^{U,(k)} = \frac{\hat{\theta}_{T_l^U}^{(k)} - \hat{\theta}_C^{(k)}}{\sqrt{V_l^U}}, \tag{4.30}$$

with variance $V_l^U := \mathrm{Var}(\hat{\theta}_{T_l^U}^{(k)} - \hat{\theta}_C^{(k)})$. Due to the independence assumption and basic properties of the variance, we find that the variance of $\hat{\theta}_G^{U,(k)}$ equals

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_G^{U,(k)}) &= \sum_{J \in U} (\pi_J^U)^2 \mathrm{Var}\left(\hat{\theta}_G^{J,(k)}\right) = \sum_{J \in U} (\pi_J^U)^2 \sum_{l=1}^{k} \left(w_{J,k}^{(l)}\right)^4 \mathrm{Var}\left(\tilde{\hat{\theta}}_G^{J,(l)}\right) \\
&= \sum_{J \in U} (\pi_J^U)^2 \sum_{l=1}^{k} \left(w_{J,k}^{(l)}\right)^4 \frac{(\sigma_G^J)^2}{\tilde{n}_J^{(l)} / (t_J^{(k)} + 1)} \\
&= \sum_{J \in U} (\pi_J^U)^2 (\sigma_G^J)^2 \sum_{\tilde{k}=1}^{k} \frac{\tilde{n}_J^{(\tilde{k})} / (t_J^{(\tilde{k})} + 1)}{\left(\sum_{l=1}^{k} \tilde{n}_J^{(l)} / (t_J^{(l)} + 1)\right)^2} \\
&= \sum_{J \in U} (\pi_J^U)^2 (\sigma_G^J)^2 / \left(\sum_{l=1}^{k} \frac{\tilde{n}_J^{(l)}}{(t_J^{(l)} + 1)}\right),
\end{aligned}$$

for $G \in \{T_l^U, C\}$, $k = 1, \ldots, K$ and $U \in \mathfrak{U}$. Therefore, $V_l^{U,(k)}$ is equal to

$$V_l^{U,(k)} = \sum_{J \in U} (\pi_J^U)^2 \left(\left(\sigma_{T_l^U}^J\right)^2 + \left(\sigma_C^J\right)^2\right) / \left(\sum_{l=1}^{k} \frac{\tilde{n}_J^{(l)}}{(t_J^{(l)} + 1)}\right) \tag{4.31}$$

To simplify computations of the covariance matrix of $\mathbf{Z} = (Z_l^{U,(k)})$ we rewrite $Z_l^{U,(k)}$ as a weighted sum of stage-wise test statistics by first considering

$$Z_l^{U,(k)} = \frac{\sum_{J \in U} \pi_J^U \left(\hat{\theta}_{T_l^U}^{J,(k)} - \hat{\theta}_C^{J,(k)}\right)}{\sqrt{V_l^{U,(k)}}} = \sum_{k'=1}^{k} \sum_{J \in U} \pi_J^U \left(w_{J,k}^{(k')}\right)^2 \frac{\left(\tilde{\hat{\theta}}_{T_l^U}^{J,(k')} - \tilde{\hat{\theta}}_C^{J,(k')}\right)}{\sqrt{V_l^{U,(k)}}}.$$

For an arbitrary $U \in \mathfrak{U}$ we define

$$\lambda_{J,k,l}^{U,(\tilde{k})} := \sqrt{\frac{v_{J,l}^{U,(\tilde{k})}}{V_l^{U,(k)}}} \tag{4.32}$$

with

$$v_{J,l}^{U,(\tilde{k})} := \left((\sigma_{T_l^U}^J)^2 + (\sigma_C^J)^2\right) / \left(\tilde{n}_J^{(\tilde{k})} / \left(t_J^{(\tilde{k})} + 1\right)\right) \tag{4.33}$$

being the stage $\tilde{k}$ variances of the treatment effect differences for $J \in U$, $k \in \{1, \ldots, K\}$ and $\tilde{k} \in \{1, \ldots, k\}$. Using these constants we can write the accumulated test statistics as follows,

$$Z_l^{U,(k)} = \sum_{\tilde{k}=1}^{k} \sum_{J \in U} \pi_J^U \left(w_{J,k}^{(\tilde{k})}\right)^2 \lambda_{J,k,l}^{U,(\tilde{k})} \underbrace{\frac{\left(\tilde{\hat{\theta}}_{T_l^U}^{J,(\tilde{k})} - \tilde{\hat{\theta}}_C^{J,(\tilde{k})}\right)}{\sqrt{v_{J,l}^{U,(\tilde{k})}}}}_{=:\tilde{Z}_{J,l}^{U,(\tilde{k})}}, \tag{4.34}$$

where $\tilde{Z}_{J,l}^{U,(\tilde{k})}$ is the standardized stage-wise mean difference between treatment T_l^U and control group C in $\mathcal{P}_J$ at stage $\tilde{k}$. Thus, the covariance of two accumulated test statistics $Z_l^{U,(s)}$ and $Z_{l'}^{U',(s')}$ for $U, U' \in \mathfrak{U}$, $l = 1, \ldots, m^U$, $l' = 1, \ldots, m^{U'}$ and $1 \leq s < s' \leq K$ is found by computing

$$\sum_{\tilde{s}=1}^{s} \sum_{\tilde{s}'=1}^{s'} \sum_{J \in U} \sum_{J' \in U'} \pi_J^U \pi_{J'}^{U'} \left(w_{J,s}^{(\tilde{s})}\right)^2 \left(w_{J',s'}^{(\tilde{s}')}\right)^2 \lambda_{J,s,l}^{U,(\tilde{s})} \lambda_{J',s',l'}^{U',(\tilde{s}')} \mathrm{Cov}(\tilde{Z}_{J,l}^{U,(\tilde{s})}, \tilde{Z}_{J',l'}^{U',(\tilde{s}')}).$$

Now due to the independence assumption of the observations, the covariance term is equal to zero if (i) $U \cap U' = \emptyset$, (ii) $\tilde{s} \neq \tilde{s}'$ or if (iii) $J \neq J'$. So the above covariance expression greatly reduces to a more manageable term:

$$\sum_{\tilde{s}=1}^{s} \sum_{J \in U \cap U'} \pi_J^U \pi_J^{U'} \lambda_{J,s,l}^{U,(\tilde{s})} \lambda_{J,s',l'}^{U',(\tilde{s})} \left(w_{J,s}^{(\tilde{s})}\right)^2 \left(w_{J,s'}^{(\tilde{s})}\right)^2 \mathrm{Cov}(\tilde{Z}_{J,l}^{U,(\tilde{s})}, \tilde{Z}_{J,l'}^{U',(\tilde{s})})$$

Using the independence assumption we now see that $\rho_{J,\tilde{s},l,l'}^{U,U'} := \mathrm{Cov}(\tilde{Z}_{J,l}^{U,(\tilde{s})}, \tilde{Z}_{J,l'}^{U',(\tilde{s})})$ is equal to

$$\rho_{J,\tilde{s},l,l'}^{U,U'} = \frac{\mathrm{Cov}(\tilde{\hat{\theta}}_{T_l^U}^{J,(\tilde{s})}, \tilde{\hat{\theta}}_{T_{l'}^{U'}}^{J,(\tilde{s})}) + \mathrm{Cov}(\tilde{\hat{\theta}}_C^{J,(\tilde{s})}, \tilde{\hat{\theta}}_C^{J,(\tilde{s})})}{\sqrt{v_{J,l}^{U,(\tilde{s})} v_{J,l'}^{U',(\tilde{s})}}}$$

because $\mathrm{Cov}\left(\tilde{\hat{\theta}}_{T_l^U}^{J,(\tilde{s})}, \tilde{\hat{\theta}}_C^{J,(\tilde{s})}\right) = \mathrm{Cov}\left(\tilde{\hat{\theta}}_{T_{l'}^{U'}}^{J,(\tilde{s})}, \tilde{\hat{\theta}}_C^{J,(\tilde{s})}\right) = 0$ (different treatment groups). With $\mathrm{Cov}(\tilde{\hat{\theta}}_C^{J,(\tilde{s})}, \tilde{\hat{\theta}}_C^{J,(\tilde{s})}) = \mathrm{Var}(\tilde{\hat{\theta}}_C^{J,(\tilde{s})}) = \left(\sigma_C^J\right)^2 / \left(\tilde{n}_J^{(\tilde{s})} / \left(t_J^{(\tilde{s})} + 1\right)\right)$ and

$$\mathrm{Cov}(\tilde{\hat{\theta}}_{T_l^U}^{J,(\tilde{s})}, \tilde{\hat{\theta}}_{T_{l'}^{U'}}^{J,(\tilde{s})}) = \begin{cases} 0, & \text{if} \quad T_l^U \neq T_{l'}^{U'} \\ \dfrac{\left(\sigma_T^J\right)^2}{\tilde{n}_J^{(\tilde{s})} / (t_J^{(\tilde{s})} + 1)}, & \text{if} \quad T_l^U = T_{l'}^{U'} = T \end{cases}$$

we then ultimately get

$$\rho^{U,U'}_{J,\tilde{s},l,l'} = \frac{1}{\sqrt{v^{U,(\tilde{s})}_{J,l} v^{U',(\tilde{s})}_{J,l'}}} \frac{\mathbf{1}_{\{T^U_l = T^{U'}_{l'}\}} \left(\sigma^J_T\right)^2 + \left(\sigma^J_C\right)^2}{\tilde{n}^{(\tilde{s})}_J / (t^{(\tilde{s})}_J + 1)}$$

for the stage-wise covariances/correlations and can write the correlation between two accrued test statistics as

$$\begin{aligned} \mathrm{Cov}(Z^{U,(s)}_l, Z^{U,(s')}_{l'}) &= \sum_{\tilde{s}=1}^{s} \sum_{J \in U \cap U'} \pi^U_J \pi^{U'}_J \lambda^{U,(\tilde{s})}_{J,s,l} \lambda^{U',(\tilde{s})}_{J,s',l'} \left(w^{(\tilde{s})}_{J,s}\right)^2 \left(w^{(\tilde{s})}_{J,s'}\right)^2 \rho^{U,U'}_{J,\tilde{s},l,l'} \qquad (4.35) \\ &= \sum_{\tilde{s}=1}^{s} \sum_{J \in U \cap U'} \frac{\pi^U_J \pi^{U'}_J \left(w^{(\tilde{s})}_{J,s} w^{(\tilde{s})}_{J,s'}\right)^2 \left(\mathbf{1}_{\{T^U_l = T^{U'}_{l'}\}} \left(\sigma^J_T\right)^2 + \left(\sigma^J_C\right)^2\right)}{\tilde{n}^{(\tilde{s})}_J / (t^{(\tilde{s})}_J + 1) \sqrt{V^{U,(s)}_l V^{U',(s')}_{l'}}} \qquad (4.36) \end{aligned}$$

The expectation of such a $Z^{U,(k)}_l$ is easy to find,

$$\begin{aligned} \mathbb{E}(Z^{U,(k)}) &= \sum_{\tilde{k}=1}^{k} \sum_{J \in U} \pi^U_J \left(w^{(\tilde{k})}_{J,k}\right)^2 \lambda^{U,(\tilde{k})}_{J,k,l} \mathbb{E}(\tilde{Z}^{U,(\tilde{k})}_{J.l}) = \sum_{\tilde{k}=1}^{k} \sum_{J \in U} \pi^U_J \left(w^{(\tilde{k})}_{J,k}\right)^2 \lambda^{U,(\tilde{k})}_{J,k,l} \frac{\theta^U_l}{\sqrt{v^{U,\tilde{k}}_{J,l}}} \\ &= \sum_{\tilde{k}=1}^{k} \sum_{J \in U} \pi^U_J \left(w^{(\tilde{k})}_{J,k}\right)^2 \frac{\theta^U_l}{\sqrt{V^{U,\tilde{k}}_l}} \qquad (4.37) \end{aligned}$$

Finally, the PWER under the global null hypothesis $\boldsymbol{\theta}^* = \mathbf{0}$ of this design can now be written as

$$\mathrm{PWER}_{\boldsymbol{\theta}^*} = \sum_{J \subseteq \mathcal{C}_\mathcal{P}} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*} \left(\bigcup_{U \in I_J(\boldsymbol{\theta}^*)} \bigcup_{k=1}^{K} \bigcup_{l: T^U_l \in \mathcal{T}^{(k)}_J} \left\{ Z^{U,(k)}_l \geq c^{(k)} \right\} \right). \qquad (4.38)$$

As already touched on, for $k \geq 2$ the set $\mathcal{T}^{(k)}_J$ of yet to be tested treatments and therefore also its size $t^{(k)}_J$, which the test statistics eventually depend on, are random as they depend on the decisions made in the previous stages. The value of a $t^{(k)}_J$ depends on the test decision made at stage $k-1$ for the treatments T^U_l being investigated in $J \in U$. Consequently, the correlation matrix $\boldsymbol{\Sigma}$ is not fully fixed in advance as in the single treatment case as the allocation ratios and thus also the correlations of the test statistics of subsequent stages are not known beforehand. It is important to note, however, that this dependency on the test decisions does not mean that this design has to be seen as an adaptive design in the sense that test decisions of subsequent stages explicitly depend on interim data – only the decision of whether a certain hypothesis has been rejected is of importance.
Let $\boldsymbol{\Psi}^U_k = \left\{ \phi^{U,(\tilde{k})}_l \mid \tilde{k} < k, l = 1, \ldots, m^U \right\}$ be a random vector (the set notation is used for convenience) containing the test decision functions for all H^U_l at all stages up to stage $k-1$ and let $\boldsymbol{\psi}^U_k$ be a realization of it. Then $\boldsymbol{\psi}^U_k \in \{0,1\}^{m^U(k-1)}$. Further, let $Z^{U,(k)}_l(\boldsymbol{\psi}^U_k)$ be the test statistic for testing H^U_l at stage k depending on the previous

decisions $\boldsymbol{\psi}_k^U$. Also, let $\boldsymbol{\psi}_k = (\boldsymbol{\psi}_k^U)_{U\in\mathfrak{U}}$ for each $k = 1, \ldots, K$. What one now could do is to specify a correlation matrix $\boldsymbol{\Sigma}_{\text{all}}$ of a test statistic vector $\mathbf{Z}_{\text{all}} = (Z_l^{U,(k)}(\boldsymbol{\psi}_k^U))$ containing test statistics accounting for all possible combinations of all $\boldsymbol{\psi}_k^U$ for every k, U and l. This vector contains test statistics for every possible outcome the trial can take and $\boldsymbol{\Sigma}_{\text{all}}$ all possible pairwise correlations that can be present. Applying the error spending approach, one could now solve the first stage PWER for the critical value $c^{(1)}$,

$$PWER_{\boldsymbol{\theta}_0}^{(1)} = \sum_{J\subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{U\in A_J}\bigcup_{l:T_l^U\in\mathcal{T}_J^{(1)}} \{Z_l^{U,(1)} \geq c^{(1)}\}\right) = \alpha^*(\tau^{(1)}), \tag{4.39}$$

where α^* is some pre-specified alpha spending function and $\tau^{(1)}$ the information rate for stage 1. At stage 2 a certain value $\boldsymbol{\psi}_2$ is realized and the corresponding sub-correlation matrix $\boldsymbol{\Sigma}_{\textbf{all}}^{(2)}(\boldsymbol{\psi}_2)$ of the sub-vector $\mathbf{Z}^{(2)}(\boldsymbol{\psi}_2) = \{Z_l^{U,(s)}(\boldsymbol{\psi}_s^U) \mid s < 2, U \in \mathfrak{U}\}$ can be used to find the second critical value etc. By Section 4.2.2 this would ensure PWER-control at a significance level α.
Another way to gain error control with the error spending approach is to incorporate all possible outcomes of the trial into the stage-wise PWER-expressions directly by means of conditional probabilities and the law of total expectation. First, one would again use (4.39) to get $c^{(1)}$. Then, the k-th ($k \geq 2$) stage PWER, can be found via

$$\text{PWER}_{\boldsymbol{\theta}^*}^{(k)} = \sum_{\boldsymbol{\psi}\in\{0,1\}^{m^U(k-1)}} \text{PWER}_{\boldsymbol{\theta}^*}^{(k)}(\boldsymbol{\Psi}_k = \boldsymbol{\psi})\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{\Psi}_k = \boldsymbol{\psi}) \tag{4.40}$$

with $\text{PWER}_{\boldsymbol{\theta}^*}^{(k)}(\boldsymbol{\Psi}_k = \boldsymbol{\psi})$ being the k-stage PWER from (4.11) conditioned on the event $\{\boldsymbol{\Psi}_k = \boldsymbol{\psi}\}$ The test statistics conditioned on $\boldsymbol{\Psi}_k = \boldsymbol{\psi}$ have to be known to calculate this quantity. Note that, as seen in Section 2, the PWER can be written as an expectation and so can the stage-wise PWER. The probabilities $\mathbb{P}_{\boldsymbol{\theta}^*}(\boldsymbol{\Psi}_k = \boldsymbol{\psi})$ under the global null follow some discrete distribution depending on the probabilities for rejecting the null hypotheses for each stage $\tilde{k} < k$ and thus the stage-wise PWERs can be found and the error spending approach can be applied.

4.3.3 A small note on basket trials in studies with intersecting populations

A basket trial is a clinical trial where one treatment in tested in several patient populations defined by different tumor histologies or indications. Typically the treatment is tested on patients whose cancer was caused by one specific mutation and are grouped based on the different histologies the mutation is present in. To our knowledge, these populations have always been disjoint or at least artificially made disjoint in clinical practice. But what if a patient suffers from a cancer mutation in more than one histology due to metastasis (i.e. the cancer spread to different body parts)? If the mutation of such a patient is present in, for instance, two different histologies, that patient may just be allocated to that histology population that is believed to be more important. This way, there would be no intersections between the different populations making an additional multiple testing correction unnecessary. But strictly speaking, populations are overlapping in this case and patients from the intersections are exposed to potentially more than one false rejection. In this case, the PWER constitutes a good compromise between not correcting for multiplicity at all and using the FWER.

4.4 Power, ASN and optimality of critical values

An important question that has yet to be addressed is the appropriate choice of the sample sizes. By equation (1.8) we know that in classical GSDs the critical values depend on the sample size ratios between the different stages through the correlation matrix of the test statistics, whereas the power of the procedure also depends on the concrete sample sizes. Typically, the sample size is chosen such that that a certain power quantitiy does not fall short of a prespecified value, typically 80% or 90%. Now, since a GSD can stop before reaching its last stage K, we also need to think of the required sample size we realistically have to *expect* leading us to the average sample size/number (ASN). In this section we will show how to control different power measures in our K-stage GSDs. Based on that we will show how the ASN can be computed. At last, we will briefly touch on how to find critical values that satisfy a certain optimality criterion, e.g. critical values that minimize the ASN or maximal sample size while still guaranteeing a certain power-value.

4.4.1 Power control

Let us first consider a special case from the design in Section 4.3.1 with two stages $K = 2$ and $m = 2$ populations $\mathcal{P}_1$ and $\mathcal{P}_2$, where hypotheses $H_j : \theta_j = \theta(\mathcal{P}_j, T) \leq 0$, $j = 1, 2$, are to be tested. So $\mathfrak{U} = \{U_1, U_2\}$ with $U_j = \{\{j\}, \{1, 2\}\}$ and $\mathcal{P}_j = \mathcal{P}^{U_j}$ for $j = 1, 2$. We also assume equal variances σ^2 in all disjoint sub-populations $\mathcal{P}_J$ with $J \subseteq \{1, 2\}$. The test statistics $\mathbf{Z} = (Z_1^{(1)}, Z_2^{(1)}, Z_1^{(2)}, Z_2^{(2)})$ follow a multivariate normal distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ with correlation matrix $\boldsymbol{\Sigma}$ (the explicit form is found by formula (4.20) in this case but is not important here) and expectation

$$\boldsymbol{\nu} = (\nu_1^{(1)}, \nu_2^{(1)}, \nu_1^{(2)}, \nu_2^{(2)}) = \left(\delta_1\sqrt{\tilde{n}_1^{(1)}}, \delta_2\sqrt{\tilde{n}_2^{(1)}}, \delta_1\sqrt{\tilde{n}_1^{(1)} + \tilde{n}_1^{(2)}}, \delta_2\sqrt{\tilde{n}_2^{(1)} + \tilde{n}_2^{(2)}}\right),$$

where $\delta_j = \theta_j/(2\sigma)$ and the stage-wise sample sizes are given by $\tilde{n}_j^{(k)} = \sum_{k'=1}^{k}(\tilde{n}_{\{j\}}^{(k')} + \tilde{n}_{\{1,2\}}^{(k')})$ for $j = 1, 2$. Therefore, we have

$$(Z_1^{(1)} - \nu_1^{(1)}, Z_2^{(1)} - \nu_2^{(1)}, Z_1^{(2)} - \nu_1^{(2)}, Z_2^{(2)} - \nu_2^{(2)}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

and computing any power expression will be similar to the computation of type I error probabilities with a shift of the rejection regions. Regardless of the power measure, we need to set $\mathbf{n} = \left(\tilde{n}_J^{(k)}\right)_{\substack{J\subseteq I \\ k=1,2}}$ such that the power expression is at least $1 - \beta$ with $\beta \in (0, 1)$. Without any other constraints this is not a uniquely solvable problem. In reasonably low dimensions a grid search for different possible values of $\mathbf{n}$ might still be computationally feasible, but in higher dimensions some constraints are needed. An easy constraint would be to set all components in $\mathbf{n}$ equal. Then the expectation $\boldsymbol{\nu}$ simplifies to $\boldsymbol{\nu} = \sqrt{n}(\delta_1, \delta_2, \delta_1\sqrt{2}, \delta_2\sqrt{2})$ yielding a uniquely solvable problem for $n \in \mathbb{N}$. However, having the same number of observations in each population and at each stage is quite unrealistic in many practical settings, especially if we already have an at least vague understanding about the prevalence in some or maybe even all sub-populations. A more sophisticated way could be to prespecify factors $\gamma_J > 0$ for each population $\mathcal{P}_J$, $J \subseteq I$, such that

$$\tilde{n}_J^{(2)} = \gamma_J \tilde{n}_J^{(1)} \tag{4.41}$$

that is, a factor that describes by how much the stage 2 sample size differs from the stage 1 sample size. By that definition every $\tilde{n}_J^{(2)}$ is only dependent on $\tilde{n}_J^{(1)}$, which in turn can be written as $\tilde{n}_J^{(1)} = N \cdot \pi_J$ with $N = \sum_{J \subseteq I} \tilde{n}_J^{(1)}$ being the overall (stage 1) sample size in $\mathcal{P}$. Thus, each stage 1 and 2 sample size for $\mathcal{P}_J$ can be written as

$$\tilde{n}_j^{(1)} = N\pi_J \quad \text{and} \quad \tilde{n}_J^{(2)} = N\pi_J\gamma_J, \tag{4.42}$$

which only depend on one unknown parameter N. Here we, again, have be mindful of the fact that the π_J are in fact unknown in most practical cases and therefore an estimate for each π_J has to be found beforehand. Nonetheless, assuming each π_J is either known or available through a guesstimate, we can control any power measure $Pow_{\boldsymbol{\theta}_A}$ under an alternative $\boldsymbol{\theta}_A$ by finding the smallest N such that

$$Pow_{\boldsymbol{\theta}_A}(N) \geq 1 - \beta.$$

This N can then be plugged into (4.42) to compute each $\tilde{n}_J^{(k)}$ for $k \geq 1$ and consequently also $\mathbf{n}$. In practice, since the sample sizes in (4.42) will be non-natural numbers due to their dependence on the real numbers π_J and γ_J one will have to use $\lceil \tilde{n}_J^{(k)} \rceil$ instead, which will lead to a slightly higher power than $1 - \beta$.

Generally, for the K-stage GSD we test hypotheses $H^U \in \mathcal{H}_k$ of the form $H^U : \theta(\mathcal{P}^U, T^U) \leq 0$, $U \subseteq \mathcal{U}$ using test statistics $\mathbf{Z} = (Z^{U,(k)})$ which we assume to follow a multivariate normal distribution $\mathrm{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ with correlation matrix $\boldsymbol{\Sigma}$ given by (4.20) and expectation vector $\boldsymbol{\nu} = (\nu^{U,(k)})$ equal to (4.22). Thus, $\mathbf{Z} - \boldsymbol{\nu}$ follows a mean zero multivariate normal distribution with the same correlation matrix and we again only have to consider probabilities of a shift of the involved rejection regions. We define prespecified population-wise factors $\gamma_J^{(k)}$ for each $J \subseteq \mathcal{C}_\mathcal{P}$ with $\mathcal{H}^U \in \mathcal{H}_k$ for any $1 \leq k \leq K - 1$ such that

$$\tilde{n}_J^{(k+1)} = \gamma_J^{U,(k)} \tilde{n}_J^{(k)}. \tag{4.43}$$

This recursion allows us to write each stage $k+1$ sample size $n_J^{(k)}$ as

$$\tilde{n}_J^{(k+1)} = N\pi_J \tilde{\gamma}_J^{(k)}, \quad 1 \leq k \leq K - 1, \tag{4.44}$$

with $\tilde{\gamma}_J^{(k)} = \prod_{l=1}^{k} \gamma_J^{(l)}$. As in the example above we simply need to find N such that $Pow_{\boldsymbol{\theta}_A}(N) \geq 1 - \beta$ by means of a univariate root finding algorithm.

4.4.2 Average sample size

A group sequential test is not only assessed by its power but also by the number of observations needed to reach a test decision. Since GSDs can stop at any stage $k = 1, \ldots, K$ this sample size is a random number, which we for now denote as $\tilde{N}$. The average sample size/number (ASN) is the expected value of $\tilde{N}$ under a certain parameter configuration $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, i.e.

$$ASN_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}} \left(\tilde{N} \right). \tag{4.45}$$

For each stage k, respectively, $ASN_{\boldsymbol{\theta}_A}$ consists of weighted sums over all $J \subseteq I$ of each stage k sample size $\tilde{n}_J^{(k)}$ for $\mathcal{P}_J$ times the respective probability of that $\tilde{n}_J^{(k)}$ is

realized. For the K-stage GSD from Section 4.1 at each stage $1 < k \leq K$ the stage-wise sample size $\tilde{n}_J^{(k)}$ is used for testing a H^U with $J \in U$ and $k \in I_K^U$, whenever any of these hypotheses have not been rejected at stage $k-1$ and H^U is considered for stage k. That is, the ASN under some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is given by

$$\begin{aligned} ASN_{\boldsymbol{\theta}} &= N + \sum_{k=2}^{K} \sum_{J \in \mathcal{C}_{\mathcal{P}}} \tilde{n}_J^{(k)} \mathbb{P}_{\boldsymbol{\theta}} \left(\text{any } H^U \text{ with } J \in U \text{ not rej. until stage } k-1 \right) \\ &= N + \sum_{k=2}^{K} \sum_{J \in \mathcal{C}_{\mathcal{P}}} \tilde{n}_J^{(k)} \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J(\boldsymbol{\theta})} \bigcap_{\tilde{k} \in I_{k-1}^U} \left\{ Z^{U,(\tilde{k})} < c^{U,(\tilde{k})} \right\} \right), \end{aligned} \tag{4.46}$$

with N being the overall stage 1 sample size. Each of the above probabilities are easily evaluated by means of the `pmvnorm()`-function if a multivariate normal distribution for the test statistics $\mathbf{Z}$ is assumed.

4.4.3 On optimal critical values

In the group sequential literature much work has been done on methods for designing an *optimal* GSD. In most cases 'optimal critical values' are referred to as critical values that guarantee type I error control while also satisfying some optimality criterion. One criterion typically aims at obtaining optimality by minimizing the ASN (or the maximal sample size) under some constraint for the power. Of course, in a statistical test procedure, and even more so in a GSD, it is of special interest to find the minimal (average/maximal) sample size necessary to guarantee a power of at least 80% or 90%. If a parameterization is used for the critical values (say Wang & Tsiatis) the task of finding optimal critical values reduces to the task of finding the parameter value that satisfies our optimality criterion.
Suppose we plan a study that can be described by our K-stage GSD and we aim for PWER-control at a level α and power-control at a level $1-\beta$. The power measure itself can be any of those listed in Section 2.6, for instance. That is we have a set of hypotheses $\mathcal{H}$ and assume the accrued test statistics for testing any $H^U \in \mathcal{H}$ in a sequential manner to jointly follow a multivariate normal distribution. Relative population prevalences $\boldsymbol{\pi}$ are assumed to be known, for example from similar studies conducted in the past. For our optimality criterion we choose to keep the average sample size as small as possible while still ensuring that the power is not below $1-\beta$ under a prespecified $\boldsymbol{\theta}_A \in \boldsymbol{\Theta}_1$. Following the strategy depicted in Section 4.3.1, each population-wise sample size can uniquely be expressed by some yet unknown N through a product of sample size factors and the prevalence of the respective population. The so found sample sizes are then used to compute the ASN. By parameterizing our critical values using a parameter ξ, be it through the Wang & Tsiatis family or an error spending function depending on a parameter, we can now find N and therefore also the ASN in dependence on ξ. Thus, we can find ξ^* such that

$$\xi^* = \underset{\xi \in \mathbb{R}}{\arg\min}\, ASN_{\boldsymbol{\theta}_A}(\xi). \tag{4.47}$$

Obviously, we can replace $ASN_{\boldsymbol{\theta}_A}(\xi)$ above with any other function of ξ we want to optimize. For example we could be interested in a minimal overall sample size $N_{max} = \sum_{k=1}^{K} N^{(k)}$, some mixture of N_{max} and ASN or a mean of N_{max} or ASN

values under different values of $\boldsymbol{\theta}_A$.
The use of a parameter family is a quite simple approach to this problem and many other approaches based on simulations [2] or backwards induction [17] could potentially be adopted for PWER-controlling procedures.

5. Numerical examples

We want to apply the GSD from Section 4.1 to several practically relevant examples. First we will revisit the population structure of two intersecting populations, but this time we are also interested in the treatment effects in the intersection and possibly also the complements. Different possible two-stage designs for testing one treatment are suggested and PWER- and power-control are dervied and discuss for all of them. In Section 2 we also briefly talked about PWER-control in nested populations which we want to further expand on in this section by means of a group sequential example. Lastly, Magnusson & Turnbull (2013, [36]) described a group sequential design where a sub-population that is likely to benefit the most from a treatment is chosen at the first stage and the subsequent stages are then used to confirm whether this claim about this supposed benefit is actually correct. We will compare PWER- and FWER-controlling methods for this example and describe the potential gain in power and sample size.

5.1 Group sequential designs for two intersecting populations

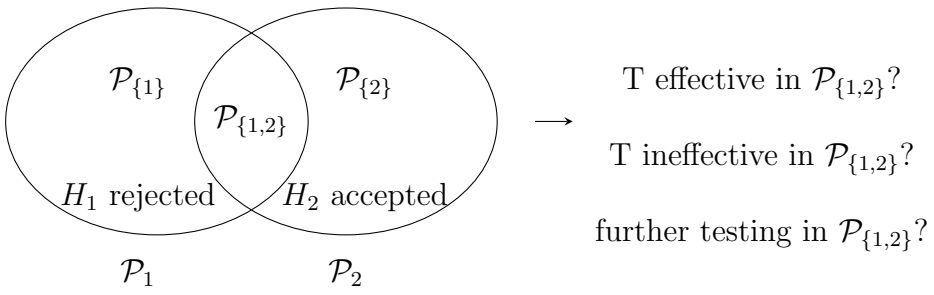

Figure 5.1: Illustration of the problem of how to proceed with a treatment T that shows significant efficacy in only one of two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$.

In Section 3 we considered a single stage test for two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$ defined by two respective biomarkers. We were mainly interested in the treatment

effects in those two sub-populations, in order to decide whether future patients of $\mathcal{P}_j$ can significantly benefit from the said treatment. However, one conceptual issue we have neglected so far is the following: in the one treatment case, what if the treatment is found to be efficacious in $\mathcal{P}_1$, but not in $\mathcal{P}_2$? That is, should future patients belonging to $\mathcal{P}_{\{1,2\}}$, and thus also to $\mathcal{P}_1$ and $\mathcal{P}_2$, get the treatment if the null can be rejected in only one of the two sub-populations? Technically, this question is impossible to answer without explicitly testing in $\mathcal{P}_{\{1,2\}}$ itself via $H_{\{1,2\}} : \theta_{\{1,2\}} \leq 0$ because each θ_i depends on the treatment effect $\theta_{\{1,2\}}$ in the intersection through

$$\theta_j = (\pi_{\{j\}}/\pi_j)\theta_{\{j\}} + (\pi_{\{1,2\}}/\pi_j)\theta_{\{1,2\}}$$

and in the same way $\hat{\theta}_i$ depends on $\theta_{\{1,2\}}$ and $\theta_{\{i\}}$. So even if a relevant treatment effect is found in $\mathcal{P}_1$ and $\mathcal{P}_2$, it still might hold that $\hat{\theta}_{\{1,2\}} \leq 0$ and rejection only happened due to patients in the complements $\mathcal{P}_{\{j\}}$ benefiting well from the treatment, i.e. due to $\hat{\theta}_{\{j\}}$ lying significantly 'far' in the alternative. Looking at it the other way around, a rejection of H_1 and/or H_2 technically does also not say anything about how beneficial the treatment is for patients in the complements $\mathcal{P}_{\{j\}}$, since a large value of $\hat{\theta}_{\{1,2\}}$ alone could lead to $\hat{\theta}_1$ or $\hat{\theta}_2$ being large enough for rejection. Of course, the larger the intersection the lower is the chance that $\hat{\theta}_{\{1,2\}}$ is negative or 0 while $\hat{\theta}_j$ is far greater than 0 (and vice versa) due to the positive correlation between $\hat{\theta}_j$ and $\hat{\theta}_{\{1,2\}}$. There are many conceivable options to solve this problem. A quite conservative approach would be to only allow the treatment for patients from the intersection if it is shown to be efficacious in both populations, whereas a more 'optimisitc' approach would be to deem the treatment efficacious in $\mathcal{P}_{\{1,2\}}$ if at least one of H_1 or H_2 are rejected. Even though the optimistic approach bears the risk of administering an in fact ineffective treatment to the intersection, the conservative variant bears the problem of withholding potentially effective treatments from those patients, especially when dealing with smaller intersection sizes. Another quite naive option is to simply plan a single stage study that solely tests in $\mathcal{P}_{\{1,2\}}$ and the two $\mathcal{P}_{\{j\}}$s in the first place, but this involves the problem of needing to screen a lot of patients to obtain a large enough sample size of patients from any disjoint sub-population, especially if some are quite small. Lastly, one could try to extend the study by a second stage (or more stages) in order to keep the option of further testing in any $\mathcal{P}_J$ or $\mathcal{P}_j$ open, if necessary. This option can be seen as a special case of our K-stage GSD with $m = K = 2$ and $\mathcal{H}_1 = \{H_1, H_2\}$ and $\mathcal{H}_2$ containing any combination of H_1, H_2, $H_{\{1\}}$, $H_{\{1,2\}}$ and $H_{\{2\}}$. In the following we want to mathematically describe, compare and critically discuss two stage designs with $\mathcal{H}_1$ and $\mathcal{H}_2$ defined as such. These types of designs will also play a role in Chapter 6 when we will deal with adaptive designs allowing us to switch between different designs mid-trial, so this section in a way serves as a preparation for this topic. In particular, we want to numerically go through the following three designs.

- **Design I:** Here we consider the case $\mathcal{H}_1 = \mathcal{H}_2 = \{H_1, H_2\}$. At stage 1, for $i = 1, 2$ test H_i in $\mathcal{P}_i$. If H_i is rejected, stop testing in $\mathcal{P}_i$ for efficacy, otherwise proceed to stage 2. After stage 2, if H_i cannot be rejected, stop the trial and accept H_i. If both H_1 and H_2 are rejected at stage 1, the trial stops early. This design is inspired by classical group sequential tests, where a hypothesis is testing in a fixed amount of stages until rejection occurs. Here, no tests in the partitions $\mathcal{P}_{\{1\}}$, $\mathcal{P}_{\{2\}}$ or $\mathcal{P}_{\{1,2\}}$ are foreseen.

- **Design II:** With $\mathcal{H}_1 = \{H_1, H_2\}$ and $\mathcal{H}_2 = \{H_{\{1,2\}}\}$ we consider the two-stage design described in the above paragraph. At stage 1, we test H_1 and H_2 in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Regardless of rejection or retention of any of H_1 or H_2, one has the option to proceed to stage 2 and test $H_{\{1,2\}}$ in $\mathcal{P}_{\{1,2\}}$.

- **Design III:** Here, the second stage options are designed to be more flexible than in Design I and II. That is, we allow options to potentially test in each subgroup after conducting first stage tests of $\mathcal{H}_1 = \{H_1, H_2\}$. For example, if H_1 is rejected and H_2 is retained, we allow the option to test in $\mathcal{P}_{\{1\}}$, $\mathcal{P}_{\{2\}}$ and $\mathcal{P}_{\{1,2\}}$ if there is uncertainty as to whether the rejection in $\mathcal{P}_1$ is enough to justify a rejection in $\mathcal{P}_{\{1\}}$. At stage 1, test H_1 and H_2 in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Depending on the test decision at stage 1, there are several options to proceed to stage 2.

  - If H_1 and H_2 are rejected, the trial can be stopped for efficacy, but the option to test at stage 2 can still be exercised.
  - If H_1 is rejected and H_2 retained, proceed to stage 2 with H_2, $H_{\{2\}}$ and $H_{\{1,2\}}$.
  - If H_2 is rejected and H_1 retained, proceed to stage 2 with H_1, $H_{\{1\}}$ and $H_{\{1,2\}}$.

For all these designs we will assume a population-wise testing problem $(\mathcal{C}_\mathcal{P}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\theta}}, \mathcal{H})$ with $\mathcal{C}_\mathcal{P} = 2^{\{1,2\}}$ and $\boldsymbol{\pi} = \left(\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}}\right)$. Also, for the overall effects $\theta_j = \theta(\mathcal{P}_j, T)$ (control included already) in $\mathcal{P}_j$, $j = 1, 2$, we assume

$$\theta_j = \frac{\pi_{\{j\}}}{\pi_j}\theta_{\{j\}} + \frac{\pi_{\{1,2\}}}{\pi_j}\theta_{\{1,2\}} \tag{5.1}$$

with $\pi_j = \pi_{\{j\}} + \pi_{\{1,2\}}$ as for example in Section 2.4.1. The sets $\boldsymbol{\Theta}$ and $\mathcal{H}$ will depend on the specific design. To keep things a bit simpler, we will assume equal variances in each population. For each of the three designs we will derive the distribution of the test statistics (expectation vector and correlation matrix), show methods on how to find critical values for PWER-control, formulate power and average sample size expressions and lastly, apply all these concepts to a numerical example. For the power measures, we choose the PWP (Definition 2.6.1) and Pow_1 (probability of rejecting at least one false null hypothesis). For Design I we will also discuss how to find optimal critical values and will refer to the respective R functions in the R-script files for the other two as the methods are completely analogous. In general, for the most part, to avoid unnecessary repetitions, the explanations for Design I will be a bit more detailed than those for Designs II and III.

5.1.1 Design I

As the first introductory example, suppose we have two populations $\mathcal{P}_1$ and $\mathcal{P}_2$ with intersection $\mathcal{P}_{\{1,2\}} \neq \emptyset$ and complements $\mathcal{P}_{\{1\}}$ and $\mathcal{P}_{\{2\}}$. Each disjoint sub-population $\mathcal{P}_J$, $J \subseteq \{1, 2\}$, has a relative population size π_J. Further suppose, again, that stage $k = 1, 2$ data from $\mathcal{P}_J$ is normally distributed with mean θ_J and variance σ^2, denoted by $X_{J,i}^{(k)} \sim \text{N}(\theta_J, \sigma^2)$ with $i = 1, \ldots, \tilde{n}_i^{(k)}$. We are interested in testing the hypotheses

$$H_j : \theta_j \leq 0 \quad \text{vs.} \quad K_j : \theta_j > 0 \quad j = 1, 2,$$

whether the efficacy $\theta_j := \theta(\mathcal{P}_j, T)$ of T administered to patients in $\mathcal{P}_j$ is superior to a control. As described above, we test H_j as long as it can be rejected or the analysis at stage 2 has been finished. This design basically corresponds to a classical group sequential test (like Pocock or O'Brien and Fleming (cf. [26])) in each population $\mathcal{P}_j$.

Distribution of test statistics: Following Section 4.3.1 using $U_j = \{\{j\}, \{1,2\}\}$ for $j = 1, 2$ the stage-wise test statistics $\tilde{Z}_j^{(k)}$ for testing H_j at stages $k = 1, 2$ are given by

$$\tilde{Z}_j^{(k)} = \frac{\sum_{J \in U_j} \sum_{i=1}^{\tilde{n}_J^{(k)}} X_{J,i}^{(k)}}{\sigma \sqrt{\tilde{n}_j^{(k)}}} \tag{5.2}$$

and are normally distributed with mean $\delta_j \sqrt{\tilde{n}_j^{(k)}}$ and variance 1, where $\delta_j = \theta_j / \sigma$ and $\tilde{n}_j^{(k)} = \tilde{n}_{\{j\}}^{(k)} + \tilde{n}_{\{1,2\}}^{(k)}$. The accumulated test statistics are therefore given by $Z_j^{(1)} = \tilde{Z}_j^{(1)}$ and $Z_j^{(2)} = w_j \tilde{Z}_j^{(1)} + \sqrt{1 - w_j^2} \tilde{Z}_j^{(2)}$, where $Z_j^{(k)}$ has mean $\delta_j \sqrt{\sum_{k'=1}^{k} \tilde{n}_j^{(k)}}$ and variance 1 and weight

$$w_j = \sqrt{\tilde{n}_j^{(1)} / \left(\tilde{n}_j^{(1)} + \tilde{n}_j^{(2)} \right)}.$$

Now, consider the random vector $\mathbf{Z} = (Z_1^{(1)}, Z_1^{(2)}, Z_2^{(1)}, Z_2^{(2)}) \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ and let $\rho = \mathrm{Cov}\left(\tilde{Z}_1^{(1)}, \tilde{Z}_2^{(1)} \right)$. The mean vector $\boldsymbol{\nu}$ and the correlation matrix $\boldsymbol{\Sigma}$ are given by (lower triangular matrix omitted since $\boldsymbol{\Sigma}$ is symmetrical)

$$\boldsymbol{\nu} = \begin{bmatrix} \delta_1 \sqrt{\tilde{n}_1^{(1)}} \\ \delta_1 \sqrt{\tilde{n}_1^{(1)} + \tilde{n}_1^{(2)}} \\ \delta_2 \sqrt{\tilde{n}_2^{(1)}} \\ \delta_2 \sqrt{\tilde{n}_2^{(1)} + \tilde{n}_2^{(2)}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & w_1 & \rho & w_2 \rho \\ & 1 & w_1 \rho & \rho(w_1 w_2 + \sqrt{(1 - w_1^2)(1 - w_2^2)}) \\ & & 1 & w_2 \\ & & & 1 \end{bmatrix} \tag{5.3}$$

From Section 4.3.1 we know that $\rho = \frac{\pi_{\{12\}}}{\sqrt{\pi_1 \pi_2}}$.

Determination of critical values: From Section 4.2 we know that we can find PWER-controlling critical values via the Wang & Tsiatis power family or the error spending approach. We use information times as

$$\tau^{(k)} = \frac{\sum_{J \in \mathcal{C}_\mathcal{P}} \sum_{k'=1}^{k} \tilde{n}_J^{(k)}}{\sum_{J \in \mathcal{C}_\mathcal{P}} \sum_{k'=1}^{2} \tilde{n}_J^{(k)}}, \quad k = 1, 2,$$

as introduced in (4.6). So using a parameter $\xi \in \mathbb{R}$ and a constant c we consider the following parameterization:

$$c^{(1)} = c, \quad c^{(2)} = c \left(\frac{\tau^{(k)}}{\tau^{(1)}} \right)^{\xi - 0.5} \tag{5.4}$$

To incorporate population-wise critical values as in (4.8), namely by considering vectors of information times $\left(\tau_1^{(k)}, \tau_2^{(k)}\right)$ for each stage k. So in this case the critical values are given by

$$c^{(1)} = c, \quad c_1^{(2)} = c\left(\frac{\tau_1^{(k)}}{\tau_1^{(1)}}\right)^{\xi-0.5}, \quad c_2^{(2)} = c\left(\frac{\tau_2^{(k)}}{\tau_2^{(1)}}\right)^{\xi-0.5} \tag{5.5}$$

with $c_j^{(2)}$ being the stage 2 critical value for population $\mathcal{P}_i$, $i = 1, 2$.
Rejection of H_j at stage k occurs if the accrued test statistic $Z_j^{(k)}$ exceeds a critical value $c^{(k)}$, $j, k = 1, 2$. These critical values can be found by solving $\text{PWER}_{\boldsymbol{\theta}^*} = \alpha$ under $\boldsymbol{\theta}^* = (0, 0)$. The PWER under $\boldsymbol{\theta}^*$ is given by

$$\text{PWER}_{\boldsymbol{\theta}^*} = \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{k=1}^{2}\{Z_j^{(k)} \geq c^{(k)}\}\right) + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{l=1}^{2}\bigcup_{k=1}^{2}\{Z_l^{(k)} \geq c^{(k)}\}\right) \tag{5.6}$$

$$= \sum_{j=1}^{2} \pi_{\{j\}}(1 - \Phi_{\boldsymbol{\Sigma}_j}(c^{(1)}, c^{(2)})) + \pi_{\{1,2\}}(1 - \Phi_{\boldsymbol{\Sigma}}(c^{(1)}, c^{(1)}, c^{(2)}, c^{(2)})) \tag{5.7}$$

with $\boldsymbol{\Sigma}_j$ being the correlation matrix of the sub-vector $\mathbf{Z}_j = \left(Z_j^{(1)}, Z_j^{(2)}\right)$. Analogously to the single step case, we commit to a type I error in $\mathcal{P}_J$ if a hypothesis concerning this population is rejected. This means that we make a type I error in $\mathcal{P}_{\{j\}}$ if we erroneously reject H_j at stage 1 or 2 and we make a type I error in $\mathcal{P}_{\{1,2\}}$ if we falsely reject H_1 or H_2 at either stage 1 or stage 2. The critical values can now be found numerically with the function `critWT` in the R-script file `DesignI`.
Another way of finding appropriate critical values is to use the error spending approach described in Section 4.2.2. So, we decompose the PWER into a sum of two stage-wise PWER-expressions $PWER^{(k)}$, $k = 1, 2$ that only depend on critical values $c^{(k)}$ up to stage k. In particular, using an error spending function α^* and information times $\tau^{(k)}$ one can then control the PWER by iteratively finding critical values $c^{(1)}$ and $c^{(2)}$ such that the equations

$$PWER^{(1)}(c^{(1)}) = \alpha^*(\tau^{(1)})$$
$$PWER^{(2)}(c^{(1)}, c^{(2)}) = \alpha - \alpha^*(\tau^{(1)})$$

are satisfied. Obviously, this leads to $PWER(c^{(1)}, c^{(2)}) = \alpha$. At stage 1, one can commit to a type I error by erroneously rejecting H_1 or H_2, which leads to

$$PWER^{(1)}(c^{(1)}) = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{j \in J}\left\{Z_j^{(1)} \geq c^{(1)}\right\}\right)$$
$$= \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\left\{Z_j^{(1)} \geq c^{(1)}\right\}\right) + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{j=1}^{2}\left\{Z_j^{(1)} \geq c^{(1)}\right\}\right)$$

This equation can be solved with respect to $c^{(1)}$ which is then plugged into $PWER^{(2)}$, given by

$$PWER^{(2)}(c^{(1)}, c^{(2)}) = \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}\left(\left\{Z_j^{(1)} < c^{(1)}, Z_j^{(2)} \geq c^{(2)}\right\}\right)$$

$$+ \pi_{\{1,2\}}\mathbb{P}\left(\underbrace{\bigcap_{j=1}^{2}\left\{Z_j^{(1)} < c^{(1)}\right\} \cap \bigcup_{j=1}^{2}\left\{Z_j^{(2)} \geq c^{(2)}\right\}}_{=:A}\right),$$

to ultimately solve $PWER^{(2)}(c^{(1)}, c^{(2)}) = \alpha - \alpha^*(\tau^{(1)})$ with respect to $c^{(2)}$. Note that the error probability in the second row above can be compute by decomposing the set into a disjoint union of sets via

$$\begin{aligned} A &= \bigcap_{j=1}^{2}\left\{Z_j^{(1)} < c^{(1)}\right\} \cap \left(\left\{Z_1^{(2)} \geq c^{(2)}\right\} \uplus \left\{Z_1^{(2)} < c^{(2)}, Z_2^{(2)} \geq c^{(2)}\right\}\right) \\ &= \{Z_1^{(1)} < c^{(1)}, Z_2^{(1)} < c^{(1)}, Z_1^{(2)} \geq c^{(2)}\} \\ &\uplus \{Z_1^{(1)} < c^{(1)}, Z_2^{(1)} < c^{(1)}, Z_1^{(2)} < c^{(2)}, Z_2^{(2)} \geq c^{(2)}\} \end{aligned}$$

Power and sample size: In the above example the choices for the sample sizes for each subgroup are quite arbitrary because neither we did account for the power of the test procedure nor did we choose sample sizes that are minimal with respect to, for instance, the maximal or the average sample size. To take those aspects into consideration we need to specify a power concept for our design. In Section 2.6 we listed (among others) two different power quantities: the probability of rejecting at least one false null (Pow_1) and the population-wise power (PWP) and in Section 4.4.1 we showed how power-control is obtained for a K-stage GSD.
For a value $\boldsymbol{\delta}_A = \left(\frac{\theta_{A,1}}{\sigma}, \frac{\theta_{A,2}}{\sigma}\right)$ the procedure is aimed to be powered for, the test statistics $\mathbf{Z} = \left(Z_1^{(1)}, Z_2^{(1)}, Z_1^{(2)}, Z_2^{(2)}\right)$ follow a multivariate normal distribution with expectation

$$\boldsymbol{\nu} = \left(\nu_1^{(1)}, \nu_2^{(1)}, \nu_1^{(2)}, \nu_2^{(2)}\right) = \left(\delta_1\sqrt{\tilde{n}_1^{(1)}}, \delta_2\sqrt{\tilde{n}_2^{(1)}}, \delta_1\sqrt{\tilde{n}_1^{(1)} + \tilde{n}_1^{(2)}}, \delta_2\sqrt{\tilde{n}_2^{(1)} + \tilde{n}_2^{(2)}}\right).$$

Therefore, $\mathbf{Z} - \boldsymbol{\nu}$ is multivariate normal with zero mean and correlation matrix $\boldsymbol{\Sigma}$. As described in Section 4.4.1, we set factors $\boldsymbol{\gamma} = (\gamma_{\{1\}}, \gamma_{\{2\}}, \gamma_{\{1,2\}})$ such that $\tilde{n}_J^{(2)} = \gamma_J \tilde{n}_J^{(1)} = \gamma_J N \pi_J$ for all $J \subseteq I$ allowing us to find the overall stage 1 sample size N such that the desired power level is reached. The weights w_j and the information times $\tau^{(k)}$ are then given by

$$\begin{aligned} w_j &= \left(1 + \frac{\tilde{n}_j^{(2)}}{\tilde{n}_j^{(1)}}\right)^{-\frac{1}{2}} = \left(1 + \frac{N(\pi_{\{j\}}\gamma_{\{j\}} + \pi_{\{1,2\}}\gamma_{\{1,2\}})}{N\pi_j}\right)^{-\frac{1}{2}} \\ &= \left(1 + \frac{\pi_{\{j\}}\gamma_{\{j\}} + \pi_{\{1,2\}}\gamma_{\{1,2\}}}{\pi_j}\right)^{-\frac{1}{2}}, \quad j = 1, 2, \end{aligned}$$

and

$$\tau^{(1)} = \frac{N}{N + \sum_{J\subseteq I}\tilde{n}_J^{(2)}} = \frac{1}{1 + \sum_{J\subseteq I}\pi_J\gamma_J}, \quad \tau^{(2)} = 1.$$

$\boldsymbol{\pi}$			$c = c(\boldsymbol{\pi}, \alpha, \xi)$		N_{PWP}		N_{Pow_1}	
$\pi_{\{1\}}$	$\pi_{\{2\}}$	$\pi_{\{1,2\}}$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$
0.3	0.3	0.4	2.954	2.290	82	90	67	74
0.35	0.35	0.3	2.927	2.271	89	97	68	75
0.4	0.4	0.2	2.892	2.246	97	106	68	75
0.4	0.2	0.4	2.953	2.290	80	88	67	73
0.4	0.3	0.3	2.927	2.271	88	97	68	75
0.6	0.2	0.2	2.891	2.246	92	101	66	73

Table 5.1: Wang & Tsiatis critical value constants $c = c(\boldsymbol{\pi}, \alpha, \xi)$ for Design I under different constellations of prevalences $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$ and Wang & Tsiatis parameters $\xi = 0, 0.5$ (OBF, Pocock) using a significance level of $\alpha = 0.025$ as well as the (stage 1) sample size needed to achieve a PWP or Pow_1 of $1 - \beta = 90\%$ under the alternative $\boldsymbol{\delta}_A = (\delta_{A,1}, \delta_{A,2}) = (0.3, 0.3)$. $\boldsymbol{\gamma} = (\gamma_{\{1\}}, \gamma_{\{2\}}, \gamma_{\{1,2\}}) = (1, 1, 1)$ was assumed.

For a $\boldsymbol{\delta}_A$ with $\delta_{A,1}, \delta_{A,2} > 0$, we can write $\text{Pow}_{1,\boldsymbol{\delta}_A}$ and $\text{PWP}_{\boldsymbol{\delta}_A}$ as

$$\text{Pow}_{1,\boldsymbol{\delta}_A} = \mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{j=1}^{2}\bigcup_{k=1}^{2}\left\{Z_j^{(k)} \geq c^{(k)}\right\}\right)$$

$$\text{PWP}_{\boldsymbol{\delta}_A} = \sum_{j=1}^{2}\pi_{\{j\}}\mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{k=1}^{2}\left\{Z_j^{(k)} \geq c^{(k)}\right\}\right) + \pi_{\{1,2\}}\text{Pow}_{1,\boldsymbol{\delta}_A}$$

Via the expectations above we can then find an N such that one of the above power measures reach $1 - \beta$.
This technically also allows us to find critical values $\mathbf{c}$ that satisfy a certain optimality criterion. We focus on the criterion

$$ASN_{\boldsymbol{\delta}_A}(\mathbf{c}) \longrightarrow \min!, \quad \text{while} \quad \text{Pow}_{\boldsymbol{\delta}_A} \geq 1 - \beta, \tag{5.8}$$

with

$$ASN_{\boldsymbol{\delta}_A} = N + \sum_{j=1}^{2}\tilde{n}_{\{j\}}^{(2)}\mathbb{P}_{\boldsymbol{\delta}_A}\left(Z_j^{(1)} < c^{(1)}\right) + \tilde{n}_{\{1,2\}}^{(2)}\mathbb{P}_{\boldsymbol{\delta}_A}\left(Z_1^{(1)} < c^{(1)} \vee Z_2^{(1)} < c^{(1)}\right) \tag{5.9}$$

and $\text{Pow}_{\boldsymbol{\delta}_A}$ being either of the power quantities above. If all γ_J are equal to some constant γ_{const}, then $w_1 = w_2 = 1/\sqrt{1 + \gamma_{\text{const}}}$ and $\tau^{(1)} = 1/2$, which are both independent on N and $\boldsymbol{\pi}$. For equally spaced stages it is $\gamma_{\text{const}} = 1$. Thus, $\boldsymbol{\Sigma}$ and the critical values – be it Wang & Tsiatis or error spending values – are only dependent on $\boldsymbol{\gamma}$. In case of unequal components of $\boldsymbol{\gamma}$ each relative population size π_J is weighted by a γ_J in both the w_j's and $\tau^{(1)}$, so the critical values will certainly be dependant on $\boldsymbol{\pi}$ as well. So the values the optimal ξ depends on are $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}_A$.

Sample calculation: Suppose we want to conduct a 2-stage group sequential design with two overlapping populations where one treatment T is administered to patients from both populations and the PWER is aimed to be controlled at a significance level of $\alpha = 0.025$. We test the hypotheses $H_j : \theta_j \leq 0$ for $j = 1, 2$. For

the sake of simplicity, we assume that we have planned for equally sized stages, i.e. with $\boldsymbol{\gamma} = (\gamma_{\{1\}}, \gamma_{\{2\}}, \gamma_{\{1,2\}}) = (1,1,1)$ yielding $\tilde{n}_J^{(1)} = \tilde{n}_J^{(2)}$ for all $J \in \mathcal{C}_{\mathcal{P}}$. Table 5.1 now contains the Wang & Tsiatis constants $c = c^{(2)}$ for the Pocock ($\xi = 0.5$) and OBF-design $\xi = 0$ for different values of $\boldsymbol{\pi}$ for this specific $\boldsymbol{\gamma}$. Thus, we have information times $\tau^{(1)} = 1/2$ and $\tau^{(2)} = 1$. Also the N needed to obtain a power of at least $1 - \beta = 90\%$ under the alternative $\boldsymbol{\delta}_A = (0.3, 0.3)$ is given for both power measures. Under the population sizes $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}}) = (0.4, 0.4, 0.2)$ we, for example, see that we would need an $N_{PWP} = 106$ for a Pocock design to satisfy $\text{PWP}_{\boldsymbol{\delta}_A} = 0.9$ which would lead to sample sizes of

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	42.4	42.4	21.2
$\tilde{n}_J^{(2)}$	42.4	42.4	21.2

and a maximal sample size of $2N = 212$ (in practice the values $\tilde{n}_J^{(k)}$ would need to be rounded up, of course). The correlation matrix $\boldsymbol{\Sigma}$ was found by using $\rho = \frac{\pi_{\{12\}}}{\sqrt{\pi_1 \pi_2}} = \frac{0.2}{\sqrt{0.6^2}} = \frac{1}{3}$ and weights are given by $w_1 = \sqrt{1/2} = w_2$, which result by the assumptions of the equally sized stage. This led to a correlation matrix of

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ & 1 & \frac{1}{3\sqrt{2}} & \frac{1}{\sqrt{2}} \\ & & 1 & \frac{1}{3} \\ & & & 1 \end{bmatrix} \tag{5.10}$$

Using the `pmvnorm()`-function from the `mvtnorm`-package [22] probabilities in (5.6) were easily evaluated by choosing the according sub-matrices of $\boldsymbol{\Sigma}$ and the required critical values can be found by means of the `uniroot`-function yielding $\mathbf{c} = \left(c^{(1)}, c^{(2)}\right) = (2.24648, 2.24648)$ as seen in Table 5.1. Note that c in a classical two-stage Pocock-design (in one population) using the same significance level is given by $c = 2.1783$ [55] only being slightly smaller than our critical value, showing that even though we have to deal with a second population (and thus a second source of multiplicity) the liberal nature of the PWER and the positive correlation between the test statistics keep the critical values considerably small. We can observe the same behaviour in an O'Brien and Fleming design ($\xi = 0$), where we get $\mathbf{c} = (2.892, 2.045)$ and c being equal to 2.7965 for a classical O'Brien and Fleming design. Also by Table 5.1 we see that for the OBF-design an N_{PWP} of 97 which is lower than for the Pocock design because the the second stage critical value is lower than the Pocock critical values and the PWP (and Pow_1) consisting of probabilities that *any* false null hypothesis is rejected.

We also want to calculate an optimal critical value that minimizes the ASN under a certain alternative $\boldsymbol{\delta}_A$. We can use the function `ASNroot` in the script file `GSDDesignI` to find the ASN under some alternative $\boldsymbol{\delta}_A$ in dependence on some Wang & Tsiatis (or error spending function) parameter ξ^*. The sample size N in the ASN-expression (5.9) is found such that the PWP is equal to 90% under $\boldsymbol{\delta}_A = (0.3, 0.3)$. On a grid for $\xi \in [0, 1]$ one can then find that $\xi^* \approx 0.4744$ yields a minimal $\text{ASN}_{\boldsymbol{\delta}_A} = 156.3324$ which is practically the Pocock design ($\xi = 0.5$ yields 156.3796 which is negligible). The critical values under $\xi^* = 0.4744$ are $\mathbf{c} = (2.2662.227)$ whereas for the Pocock

design we have $\mathbf{c} = (2.246, 2.246)$ which is more or less the same. Interestingly, choosing equal components of $\boldsymbol{\delta}_A$, like $\boldsymbol{\delta}_A = (0.2, 0.2)$ will lead to the same optimal parameter $\xi^* = 0.4744$ while optimizing under an alternative like $\boldsymbol{\delta}_A = (0.2, 0.4)$ will lead to $\xi^* = 0.3472$, which is due to the dependence of the ASN on the non-centrality-parameters $\boldsymbol{\nu}$.

5.1.2 Design II

Another design would be to only test in the intersection at stage 2, that is to define $\mathcal{H} = \{H_1, H_2, H_{\{1,2\}}\}$ with $\mathcal{H}_1 = \{H_1, H_2\}$ and $\mathcal{H}_2 = \{H_{\{1,2\}}\}$. One reasoning for using the design could be the fundamental question as to whether patients from $\mathcal{P}_{\{1,2\}}$ should get the treatment in question or not in the case where we observe a strong effect in one population ($\mathcal{P}_1$ say) but a very weak (or negative) one in the other population ($\mathcal{P}_2$). If we rejected in both $\mathcal{P}_1$ and $\mathcal{P}_2$, we could much more safely assume that there is a strong evidence for the treatment to be beneficial for patients in $\mathcal{P}_{\{j\}}$, $j = 1, 2$, as well as a 'twofold' evidence for it being efficient in $\mathcal{P}_{\{1,2\}}$. If we neither rejected in $\mathcal{P}_1$ nor $\mathcal{P}_2$, one could still proceed to the second stage and test for an effect in the intersection. If only one hypothesis was rejected, there is still uncertainty as to whether the treatment is efficient in the intersection, so proceeding to solely test $H_{\{1,2\}}$ at stage 2, while also reusing the stage 1 data drawn from $\mathcal{P}_{\{1,2\}}$, could be seen as a reasonable approach. Explicitly, we formulate this two-stage design as follows:
At stage 1, if

- H_1 and H_2 can be rejected, we can stop the trial for efficacy (but it is still possible to exercise the option to test in the intersection at stage 2).
- If at least one hypothesis cannot be rejected, proceed to stage 2 and test $H_{\{1,2\}}$.

Under the global null $\theta^* = (\theta_1, \theta_2, \theta_{\{1,2\}}) = (0, 0, 0)$ we commit to a type I error in $\mathcal{P}_{\{1,2\}}$ if and only if either H_1 or H_2 are rejected at stage 1 or if at least one hypothesis is retained at stage 1 and $H_{\{1,2\}}$ is rejected at stage 2. One can easily verify that this event is equal to $\{\text{reject } H_1 \vee \text{reject } H_2 \vee \text{reject } H_{\{1,2\}}\}$.

Distribution of test statistics: Again, we use an approximately multivariate normal z-test statistics vector $\mathbf{Z} = \left(Z_1^{(1)}, Z_2^{(1)}, Z_{\{1,2\}}^{(2)}\right)$ to test each respective hypothesis. Note that $Z_j^{(1)}$ consists of data from both $\mathcal{P}_{\{j\}}$ and $\mathcal{P}_{\{1,2\}}$, $j = 1, 2$. Formally, let $X_{J,i}^{(k)} \sim \mathrm{N}(\theta_J, \sigma^2)$ be the ith stage k observation drawn from $\mathcal{P}_J$ with $J \in \mathcal{C}_{\mathcal{P}}$. Again, with $U_j = \{\{j\}, \{1,2\}\}$ we have $Z_j^{(1)}$, $j = 1, 2$, given by

$$Z_j^{(1)} := Z^{U_j,(1)} = \frac{\sum_{J \in U_j} \sum_{i=1}^{\tilde{n}_J^{(1)}} X_{J,i}^{(1)}}{\sigma \sqrt{\tilde{n}_{\{j\}}^{(1)} + \tilde{n}_{\{1,2\}}^{(1)}}} \tag{5.11}$$

and $Z_{\{1,2\}}^{(2)}$ is equal to

$$Z_{\{1,2\}}^{(2)} = \frac{\sum_{k=1}^{2} \sum_{j=1}^{\tilde{n}_{\{1,2\}}^{(k)}} X_{\{1,2\},i}^{(k)}}{\sigma \sqrt{\tilde{n}_{\{1,2\}}^{(1)} + \tilde{n}_{\{1,2\}}^{(2)}}}. \tag{5.12}$$

Assuming equal variances across all sub-populations using the results from Section. 4.2.1 we find the expectation and correlation matrix of $\mathbf{Z}$ under an arbitrary parameter configuration $\boldsymbol{\theta} \in \mathbb{R}^3$ to be equal to

$$\boldsymbol{\nu} = \begin{bmatrix} \delta_1 \sqrt{\tilde{n}^{(1)}_{\{1\}} + \tilde{n}^{(1)}_{\{1,2\}}} \\ \delta_2 \sqrt{\tilde{n}^{(1)}_{\{2\}} + \tilde{n}^{(1)}_{\{1,2\}}} \\ \delta_{\{1,2\}} \sqrt{\tilde{n}^{(1)}_{\{1,2\}} + \tilde{n}^{(2)}_{\{1,2\}}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{\pi_{\{1,2\}}}{\sqrt{\pi_1 \pi_2}} & w_{\{1,2\}} \sqrt{\frac{\pi_{\{1,2\}}}{\pi_1}} \\ & 1 & w_{\{1,2\}} \sqrt{\frac{\pi_{\{1,2\}}}{\pi_2}} \\ & & 1 \end{bmatrix} \tag{5.13}$$

Here $\delta_l = \theta_l / \sigma$ is the standardized mean treatment effect for all $l \in \{1, 2, \{1,2\}\}$ and $w_{\{1,2\}} = \sqrt{\tilde{n}^{(1)}_{\{1,2\}} / (\tilde{n}^{(1)}_{\{1,2\}} + \tilde{n}^{(2)}_{\{1,2\}})}$ is a weight arising from testing $H_{\{1,2\}}$ at stage 2.

Determination of critical values: We consider critical values $c^{(k)}$ for each stage $k = 1, 2$ to find the PWER for this design as

$$\begin{aligned} \text{PWER}_{\boldsymbol{\theta}^*} &= \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(Z^{(1)}_j \geq c^{(1)} \right) && (5.14) \\ &\quad + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(\{ Z^{(1)}_1 \geq c^{(1)} \} \cup \{ Z^{(1)}_2 \geq c^{(1)} \} \cup \{ Z^{(2)}_{\{1,2\}} \geq c^{(2)} \} \right) && (5.15) \\ &= (1 - \pi_{\{1,2\}}) \Phi \left(-c^{(1)} \right) + \pi_{\{1,2\}} \left[1 - \Phi_{\boldsymbol{\Sigma}} \left(c^{(1)}, c^{(1)}, c^{(2)} \right) \right]. && (5.16) \end{aligned}$$

With regards to the error spending approach, this expression can be rewritten as sum of the stage-wise PWERs

$$\begin{aligned} \text{PWER}^{(1)}_{\boldsymbol{\theta}^*} &= \pi_{\{1\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(Z^{(1)}_1 \geq c^{(1)} \right) + \pi_{\{2\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(Z^{(1)}_2 \geq c^{(1)} \right) \\ &\quad + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(\{ Z^{(1)}_1 \geq c^{(1)} \} \cup \{ Z^{(1)}_2 \geq c^{(1)} \} \right) \\ &= (1 - \pi_{\{1,2\}}) \Phi \left(-c^{(1)} \right) + \pi_{\{1,2\}} \left[1 - \Phi_{\boldsymbol{\Sigma}} \left(c^{(1)}, c^{(1)} \right) \right], \\ \text{PWER}^{(2)}_{\boldsymbol{\theta}^*} &= \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*} \left(\{ Z^{(1)}_1 < c^{(1)} \} \cap \{ Z^{(1)}_2 < c^{(1)} \} \cap \{ Z^{(2)}_{\{1,2\}} \geq c^{(2)} \} \right) \\ &= \pi_{\{1,2\}} \left[\Phi_{\boldsymbol{\Sigma}}(c^{(1)}, c^{(1)}) - \Phi_{\boldsymbol{\Sigma}} \left(c^{(1)}, c^{(1)}, c^{(2)} \right) \right] \end{aligned}$$

Note that if $\pi_{\{1,2\}} < \alpha - \alpha^*(\tau^{(1)})$ then $PWER^{(2)}_{\boldsymbol{\theta}^*} = \alpha - \alpha^*(\tau^{(1)})$ is never solvable. In this case however $\pi_{\{1,2\}}$ is probably so small (around 1% for $\alpha = 0.025$) that treating $\mathcal{P}_1$ and $\mathcal{P}_2$ as disjoint is most likely the better solution. Also, addressing $PWER^{(2)}_{\boldsymbol{\theta}^*}$ as 'stage 2 population-wise error rate' might at first seem a little odd since the expressions for the other complements are absent, but since there is no way of committing to a type I error in $\mathcal{P}_{\{1\}}$ and $\mathcal{P}_{\{2\}}$ at stage 2, the corresponding stage 2 error probabilities can simply be seen as zero.
Again, the Wang & Tsiatis or the error spending approach can be used to find $\mathbf{c} = \left(c^{(1)}, c^{(2)} \right)$. For both approaches, the information rates have to be chosen with respect to how much data is drawn from $\mathcal{P}_{\{1,2\}}$ at stages 1 and 2, respectively, since $H_{\{1,2\}}$ is the only hypothesis whose test uses data from both stages. So one can choose

$$\tau = \tau^{(1)}_{\{1,2\}} = \frac{\tilde{n}^{(1)}_{\{1,2\}}}{\tilde{n}^{(1)}_{\{1,2\}} + \tilde{n}^{(2)}_{\{1,2\}}} = \frac{1}{1 + \gamma}$$

with $\gamma = \tilde{n}^{(2)}_{\{1,2\}} / \tilde{n}^{(1)}_{\{1,2\}}$ being a prespecified factor.

$\boldsymbol{\pi}$			$c = c(\boldsymbol{\pi}, \alpha, \xi)$		N_{PWP}		N_{Pow_1}	
$\pi_{\{1\}}$	$\pi_{\{2\}}$	$\pi_{\{1,2\}}$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$
0.3	0.3	0.4	2.434	2.156	201	171	105	113
0.35	0.35	0.3	2.343	2.123	209	184	116	121
0.4	0.4	0.2	2.237	2.082	218	198	125	128
0.4	0.2	0.4	2.433	2.155	197	167	104	112
0.4	0.3	0.3	2.343	2.123	208	183	116	121
0.6	0.2	0.2	2.235	2.080	207	188	121	124

Table 5.2: Wang & Tsiatis critical value constants $c = c(\boldsymbol{\pi}, \alpha, \xi)$ for Design II under different constellations of prevalences $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$ and Wang & Tsiatis parameters $\xi = 0, 0.5$ (OBF, Pocock) using a significance level of $\alpha = 0.025$ as well as the (stage 1) sample size needed to achieve a PWP or Pow_1 of $1 - \beta = 90\%$ under the alternative $\boldsymbol{\delta}_A = (\delta_{A,1}, \delta_{A,2}, \delta_{A,\{1,2\}}) = (0.3, 0.3, 0.3)$. $\gamma = 1$ was used for the sample size factor between stage 1 and 2.

Power and sample size: With regard to satisfying a certain power level we can rewrite $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ as described in 4.3.1. Let $\gamma > 0$ be a factor such that $n^{(2)}_{\{1,2\}} = \gamma \cdot n^{(1)}_{\{1,2\}}$ and N the overall stage 1 sample size such that $n^{(1)}_J = N \cdot \pi_J$ for $J \subseteq \{1, 2\}$. Then we have $w_{\{1,2\}} = \sqrt{1/(1+\gamma)}$ and

$$\boldsymbol{\nu} = \sqrt{N}\left(\delta_1\sqrt{\pi_1}, \delta_2\sqrt{\pi_2}, \delta_{\{1,2\}}\sqrt{(1+\gamma)\pi_{\{1,2\}}}\right) \tag{5.17}$$

and under a desired alternative $\boldsymbol{\delta}_A$ we can find N such that a given power criterion $\text{Pow}_{\boldsymbol{\delta}_A}(N)$ of our choice does not fall short of $1 - \beta$. For instance, under a $\boldsymbol{\delta}_A = (\delta_1, \delta_2, \delta_{\{1,2\}})$ with $\delta_1, \delta_2, \delta_{\{1,2\}} > 0$ we can write $\text{Pow}_{1,\boldsymbol{\delta}_A}$ and $\text{PWP}_{1,\boldsymbol{\delta}_A}$ as

$$\text{Pow}_{1,\boldsymbol{\delta}_A} = \mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{j=1}^{2}\left\{Z^{(1)}_j \geq c^{(1)}\right\} \cup \left\{Z^{(2)}_{\{1,2\}} \geq c^{(2)}\right\}\right)$$

$$\text{PWP}_{\boldsymbol{\delta}_A} = \sum_{j=1}^{2}\pi_{\{j\}}\mathbb{P}_{\boldsymbol{\delta}_A}\left(Z^{(1)}_j \geq c^{(1)}\right) + \pi_{\{1,2\}}\text{Pow}_{1,\boldsymbol{\delta}_A}$$

whereas with $\delta_1 > 0$ and $\delta_2, \delta_{\{1,2\}} \leq 0$ it is

$$\text{Pow}_{1,\boldsymbol{\delta}_A} = \mathbb{P}_{\boldsymbol{\delta}_A}\left(\left\{Z^{(1)}_j \geq c^{(1)}\right\}\right) = \text{PWP}_{\boldsymbol{\delta}_A}.$$

If we find N for a power criterion we decided on we can use it to compute the ASN under $\boldsymbol{\theta}$ as follows,

$$ASN_{\boldsymbol{\delta}_A} = N + n^{(2)}_{\{1,2\}}\left[1 - \mathbb{P}_{\boldsymbol{\delta}_A}\left(Z^{(1)}_1 \geq c^{(1)}_1, Z^{(1)}_2 \geq c^{(1)}_2\right)\right] \tag{5.18}$$

$$= N + N\gamma\pi_{\{1,2\}}\left[1 - \mathbb{P}_{\boldsymbol{\delta}_A}\left(Z^{(1)}_1 \geq c^{(1)}_1, Z^{(1)}_2 \geq c^{(1)}_2\right)\right], \tag{5.19}$$

since the design proceeds to stage 2 if and only if H_1 and H_2 are not simultaneously rejected at stage 1.

Sample calculation: Similar to the sample calculation for Design I, we assume $\gamma = 1$ such that $\tilde{n}^{(1)}_{\{1,2\}} = \tilde{n}^{(2)}_{\{1,2\}}$ yielding $\tau^{(1)} = 1/2$ and $w^{(1)}_{\{1,2\}} = w^{(2)}_{\{1,2\}} = \sqrt{1/2}$. Table

5.2 shows Pocock and OBF critical values for different $\boldsymbol{\pi}$ and the corresponding values N needed to ensure a power of 90% again. For a Pocock design, we have $N_{PWP} = 198$ for $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$ yielding sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	79.2	79.2	39.6
$\tilde{n}_J^{(2)}$	0	0	39.6

and a maximal sample size of $N + \tilde{n}^{(2)}_{\{1,2\}} = 198 + 39.6 = 237.6$. For the correlation matrix, we used the the quantities $\pi_{\{1,2\}}/\sqrt{\pi_1\pi_2} = 1/3$ and $\sqrt{\pi_{\{1,2\}}/\pi_j} = 1/\sqrt{3}$ for $j = 1, 2$ to find

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{\pi_{\{1,2\}}}{\sqrt{\pi_1\pi_2}} & w_{\{1,2\}}\sqrt{\frac{\pi_{\{1,2\}}}{\pi_1}} \\ & 1 & w_{\{1,2\}}\sqrt{\frac{\pi_{\{1,2\}}}{\pi_2}} \\ & & 1 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{\sqrt{6}} \\ & 1 & \frac{1}{\sqrt{6}} \\ & & 1 \end{bmatrix}.$$

to find Pocock critical values $\mathbf{c} = (2.344, 2.344)$. For comparison, an OBF-design needs critical values $\mathbf{c} = (2.237, 2.237/\sqrt{2}) = (2.237, 1.582)$ with an N_{PWP} of 218.

5.1.3 Design III

Since Design II only investigates the intersection at stage 2, let us consider an example that, in a way, combines Design I and II. At stage 1, again, H_1 and H_2 are tested in $\mathcal{P}_1$ and $\mathcal{P}_2$. The methodology is basically the same as in Design I: if either of these hypotheses H_j is rejected, we do not test in $\mathcal{P}_j$ anymore, as one might see this rejection as enough to claim a beneficial effect for every patient in drawn from it. If a H_j is accepted, though, one has the option to proceed to stage 2 not only in $\mathcal{P}_j$ again, as in Design I, but also in all disjoint subgroups $\mathcal{P}_J$ with $J \in U_j = \{\{j\}, \{1, 2\}\}$. So we have hypotheses $\mathcal{H} = \{H_1, H_2, H_{\{1\}}, H_{\{2\}}, H_{\{1,2\}}\}$ where H_1, H_2 can potentially be tested at stages 1 and 2 and $H_{\{1\}}, H_{\{2\}}, H_{\{1,2\}}$ only at stage 2.

Distribution of test statistics: At stage 1 we again test $H_j : \theta_j \leq 0$, $j = 1, 2$. At stage 2, there is the option to test any hypothesis in $\mathcal{H}$, where the hypotheses concerning patients from $\mathcal{P}_J$, $J \subseteq \{1, 2\}$, are defined as $H_J : \theta_J = \theta(\mathcal{P}_J, T) \leq 0$. Thus, we consider a vector of test statistics

$$\mathbf{Z} = \left(Z_1^{(1)}, Z_1^{(2)}, Z_2^{(1)}, Z_2^{(2)}, Z_{\{1\}}^{(2)}, Z_{\{2\}}^{(2)}, Z_{\{1,2\}}^{(2)}\right) \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$$

where $Z_j^{(k)}$ are accumulated test statistics used to test H_j, $j = 1, 2$, at stage $k = 1, 2$ and $Z_J^{(2)}$ are accumulated test statistics for testing H_J, $J \subseteq \{1, 2\}$ at stage 2. All

test statistics are, again, given by (4.18) similarly to the previous two designs. The expectation vector is equal to

$$\boldsymbol{\nu} = \begin{bmatrix} \delta_1\sqrt{\tilde{n}_1^{(1)}} \\ \delta_1\sqrt{\tilde{n}_1^{(1)}+\tilde{n}_1^{(2)}} \\ \delta_2\sqrt{\tilde{n}_2^{(1)}} \\ \delta_2\sqrt{\tilde{n}_2^{(1)}+\tilde{n}_2^{(2)}} \\ \delta_{\{1\}}\sqrt{\tilde{n}_{\{1\}}^{(1)}+\tilde{n}_{\{1\}}^{(2)}} \\ \delta_{\{2\}}\sqrt{\tilde{n}_{\{2\}}^{(1)}+\tilde{n}_{\{2\}}^{(2)}} \\ \delta_{\{1,2\}}\sqrt{\tilde{n}_{\{1,2\}}^{(1)}+\tilde{n}_{\{1,2\}}^{(2)}} \end{bmatrix} \tag{5.20}$$

while the correlation matrix $\boldsymbol{\Sigma}$ is found via formula (4.20) with weights

$$w_2^{U_j,(1)} = \sqrt{\frac{\tilde{n}_j^{(1)}}{\tilde{n}_j^{(1)}+\tilde{n}_j^{(2)}}} \quad \text{and} \quad w_2^{\{\{J\}\},(1)} = \sqrt{\frac{\tilde{n}_J^{(1)}}{\tilde{n}_J^{(1)}+\tilde{n}_J^{(2)}}}$$

and stage-wise correlations ($j = 1, 2$):

$$\begin{aligned} \mathrm{Cov}(Z_1^{(1)}, Z_2^{(1)}) &= \frac{\pi_{\{1,2\}}}{\sqrt{\pi_1\pi_2}}, \\ \mathrm{Cov}(Z_j^{(1)}, Z_{\{j\}}^{(1)}) &= \sqrt{\frac{\pi_{\{j\}}}{\pi_j}}, \\ \mathrm{Cov}(Z_{\{1,2\}}^{(1)}, Z_j^{(1)}) &= \sqrt{\frac{\pi_{\{1,2\}}}{\pi_j}}. \end{aligned}$$

Determination of critical values: The PWER under $\boldsymbol{\theta}^* = \mathbf{0} \in \mathbb{R}^5$ is given by

$$\mathrm{PWER}_{\boldsymbol{\theta}^*} = \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{k=1}^{2}\left\{Z_j^{(k)} \geq c^{(k)}\right\} \cup \left\{Z_{\{j\}}^{(2)} \geq c^{(2)}\right\}\right) \tag{5.21}$$

$$+ \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{j=1}^{2}\bigcup_{k=1}^{2}\left\{Z_j^{(k)} \geq c^{(k)}\right\} \cup \left\{Z_{\{1,2\}}^{(2)} \geq c^{(2)}\right\}\right) \tag{5.22}$$

where the probabilities can again all be computed with the multivariate normal cdf of the respective test statistic sub-vectors. For the error spending approach we can find

$$\mathrm{PWER}_{\boldsymbol{\theta}^*}^{(1)} = \sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\left\{Z_j^{(1)} \geq c^{(1)}\right\}\right) + \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{j=1}^{2}\left\{Z_j^{(1)} \geq c^{(1)}\right\}\right), \tag{5.23}$$

$$\begin{aligned} \mathrm{PWER}_{\boldsymbol{\theta}^*}^{(2)} = &\sum_{j=1}^{2} \pi_{\{j\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\left\{Z_j^{(1)} < c^{(1)}\right\} \cap \left(\left\{Z_j^{(2)} \geq c^{(2)}\right\} \cup \left\{Z_{\{j\}}^{(2)} \geq c^{(2)}\right\}\right)\right) \\ &+ \pi_{\{1,2\}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcap_{j=1}^{2}\left\{Z_j^{(1)} < c^{(1)}\right\} \cap \left(\bigcup_{j=1}^{2}\left\{Z_j^{(2)} \geq c^{(2)}\right\} \cup \left\{Z_{\{1,2\}}^{(2)} \geq c^{(2)}\right\}\right)\right). \end{aligned} \tag{5.24}$$

$\boldsymbol{\pi}$			$c = c(\boldsymbol{\pi}, \alpha, \xi)$		N_{PWP}		N_{Pow_1}	
$\pi_{\{1\}}$	$\pi_{\{2\}}$	$\pi_{\{1,2\}}$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$	$\xi = 0$	$\xi = 0.5$
0.3	0.3	0.4	3.168	2.385	88	93	68	73
0.35	0.35	0.3	3.152	2.370	95	100	68	73
0.4	0.4	0.2	3.115	2.344	103	109	69	74
0.4	0.2	0.4	3.163	2.382	86	91	67	72
0.4	0.3	0.3	3.151	2.370	95	100	68	73
0.6	0.2	0.2	3.104	2.339	97	103	67	71

Table 5.3: Wang & Tsiatis critical value constants $c = c(\boldsymbol{\pi}, \alpha, \xi)$ for Design I under different constellations of prevalences $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}})$ and Wang & Tsiatis parameters $\xi = 0, 0.5$ (OBF, Pocock) using a significance level of $\alpha = 0.025$ as well as the (stage 1) sample size needed to achieve a PWP or Pow_1 of $1-\beta = 90\%$ under the alternative $\boldsymbol{\delta}_A = (\delta_{A,1}, \delta_{A,2}, \delta_{A,\{1\}}, \delta_{A,\{2\}}, \delta_{A,\{1,2\}}) = (0.3, 0.3, 0.3, 0.3, 0.3)$.

These quantities can be computed by considering the respective sums over suitable partitions of the sets inside the probability measures as described in Section 4.2.2.

Power and sample size: With regard to satisfying a certain power level we can rewrite $\boldsymbol{\nu}$ again by using factors γ_J for each $J \subseteq \{1, 2\}$ such that $\tilde{n}_J^{(2)} = \gamma_J \tilde{n}_J^{(1)} = \gamma_J N \pi_J$ in order to be able to find a value an N guaranteeing a PWP or Pow_1 of at least $1-\beta$. The concrete values are found analogously to the other two designs. For example for

$$\boldsymbol{\delta}_A = (\delta_{A,1}, \delta_{A,2}, \delta_{A,\{1\}}, \delta_{A,\{2\}}, \delta_{A,\{1,2\}})$$

where every component is greater than 0, the $\text{PWP}_{\boldsymbol{\delta}_A}$ and $\text{Pow}_{1,\boldsymbol{\delta}_A}$ equal to

$$\text{Pow}_{1,\boldsymbol{\delta}_A} = \mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{j=1}^{2}\bigcup_{k=1}^{2}\left\{Z_j^{U,(k)} \geq c^{U,(k)}\right\} \cup \bigcup_{J\in\mathcal{C}_\mathcal{P}}\left\{Z_J^{(2)} \geq c^{(2)}\right\}\right)$$

$$\text{PWP}_{\boldsymbol{\delta}_A} = \sum_{j=1}^{2}\pi_{\{j\}}\mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{k=1}^{2}\left\{Z_j^{(k)} \geq c^{(k)}\right\} \cup \left\{Z_{\{j\}}^{(2)} \geq c^{(2)}\right\}\right) + \pi_{\{1,2\}}\text{Pow}_{1,\boldsymbol{\delta}_A}$$

If we find an N such that the either of the above power measures equals some prespecified value, one can find the ASN via

$$\text{ASN}_{\boldsymbol{\delta}_A} = N + \sum_{J\in\mathcal{C}_\mathcal{P}}\tilde{n}_J^{(2)}\mathbb{P}_{\boldsymbol{\delta}_A}\left(\bigcup_{j\in J}\left\{Z_j^{(1)} \in \mathcal{C}^{(1)}\right\}\right).$$

Sample calculation: Similar to the sample calculation for Designs I and II, we assume $\gamma = (1, 1, 1)$ such that $\tilde{n}_J^{(1)} = \tilde{n}_J^{(2)}$ yielding $\tau^{(1)} = 1/2$ and $w_J^{(1)} = w_J^{(2)} = w_j^{(1)} = w_j^{(2)} = \sqrt{1/2}$ for $j = 1, 2$ and $J \in \mathcal{C}_\mathcal{P}$. Table 5.3 shows Pocock and OBF critical values for different $\boldsymbol{\pi}$ and the corresponding values N needed to ensure a power of 90% again. For a Pocock design under $\boldsymbol{\delta}_A = (0.3, 0.3, 0.3, 0.3, 0.3)$ we have $N_{PWP} = 109$ for $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$ yielding sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	43.6	43.6	21.8
$\tilde{n}_J^{(2)}$	43.6	43.6	21.8

and a maximal sample size of $2N = 218$. Pocock critical values of around $\mathbf{c} = (2.344, 2.344)$ are then again found with the appropriate correlation matrix. For comparison, an OBF-design needs critical values $\mathbf{c} = (3.115, 3.115/\sqrt{2}) = (3.115, 2.203)$ with an N_{PWP} of 103.

5.2 An example design for nested populations

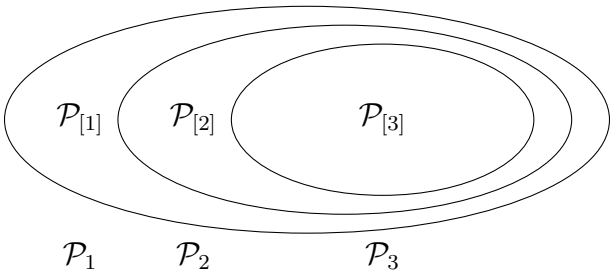

Figure 5.2: Nested population structure for $m = 3$.

To show that PWER-control also works in nested population structures, we want to consider a simple example. Suppose we intend to test the efficacy of a treatment vs. a control in a number of nested populations $\mathcal{P}_1 \supset \cdots \supset \mathcal{P}_m \neq \emptyset$ with respective population sizes $1 = \pi_1 > \cdots > \pi_m$ for $j = 1, \ldots, m$. Let us say that we aim for efficacy in the largest possible sub-population. This could, for instance, make sense if we wanted to make an efficient treatment accessible to as many people as possible. Therefore, we consider a group sequential trial design, where the efficacy of the treatment is sequentially tested in $\mathcal{P}_k$ at stage $k = 1, \ldots, K = m$ as long as rejection is reached. As in Section 2.4.2 we construct a partition of this population structure by considering disjoint subgroups $\mathcal{P}_{[j]} := \mathcal{P}_{\{1,\ldots,j\}} = \mathcal{P}_j \setminus \mathcal{P}_{j-1}$ with $\mathcal{P}_{\{0\}} = \mathcal{P}_0 := \emptyset$. Figure 5.2 is an illustration for $m = 3$ nested populations. Note that here the π_j do not sum to 1, since the populations are not disjoint, but if we define $\pi_{[j]} = \pi_j - \pi_{j+1}$ (where $\pi_{[m+1]} := 0$), then $\sum_{j=1}^{m} \pi_{[j]} = 1$. Say that observations from $\mathcal{P}_j$ are normally distributed with mean $\theta_j \in \mathbb{R}$ and variance σ_j^2 for $j = 1, \ldots, m$. Formally, we intend to test hypotheses for $\theta_j \in \mathbb{R}$, $j = 1, \ldots, m$,

$$H_j : \theta_j \leq 0 \quad \text{vs.} \quad K_j : \theta_j > 0$$

by means of test standardized statistics $\mathbf{Z} = (Z_1, \ldots, Z_m)$ which we again assume to be jointly normal distributed with expectation $\boldsymbol{\nu}$ and correlation matrix $\boldsymbol{\Sigma}$. Note that since $K = m$ the here defined Z_j equals the formerly defined $Z^{U,(j)}$ with stages $k = j$ and U such that $\mathcal{P}_j = \mathcal{P}^U$ but to provide greater clarity we will use the former notation for this subsection. At each stage k, $\pi_k N$ patients are recruited from $\mathcal{P}_k$ and H_k is tested using all patients that have been recruited from $\mathcal{P}_k$ across all previous stages. So at stage 1, N patients are used from the full population $\mathcal{P}_1$, at stage 2, the $N\pi_2$ patients from stage 1 as well as additional new collected data from $N\pi_2$ stage 2 patients are used and so on. We consider the following sequential test procedure:

- Stage 1: Test H_1 using N observations from $\mathcal{P}_1$. If H_1 is rejected, stop the trial for efficacy, otherwise proceed to stage 2.

- Stage 2: Test H_2 using $2\pi_2 N$ patients from $\mathcal{P}_2$. If H_2 is rejected, stop the trial for efficacy, otherwise proceed to stage 3.
- Stage k: Test H_k using $k\pi_k N$ patients from $\mathcal{P}_k$. If H_k is rejected, stop the trial for efficacy, otherwise proceed to stage k+1.
- Stage m: Test H_m using $m\pi_m N$ patients from $\mathcal{P}_m$. The trial is stopped either with rejection (efficacy) or acceptance (failure to show significant efficacy of treatment) of H_m.

As in previous sections, we can find that $\boldsymbol{\nu}$ is given by

$$\boldsymbol{\nu} = \left(\delta_k \sqrt{k\pi_k N}\right)_{k=1,\dots,m},$$

where $\delta_k = \theta_k/\sigma_k$. For the correlation between Z_i and Z_j, $i < j$, we refer to the expression in (4.21) with $U = U_i := \{\{1,\dots,i\},\dots,\{1,\dots,m\}\}$ and $U' = U_j := \{\{1,\dots,j\},\dots,\{1,\dots,m\}\}$ and find, due to $U_j \subset U_i$, that

$$\rho_{i,j} := \rho^{U_i,U_j} = \frac{\pi^{U_i \cap U_j}}{\sqrt{\pi^{U_i}\pi^{U_j}}} = \frac{\pi^{U_j}}{\sqrt{\pi^{U_i}\pi^{U_j}}} = \sqrt{\frac{\pi^{U_j}}{\pi^{U_i}}} = \sqrt{\frac{\pi_j}{\pi_i}}$$

and with stage k weights $w_k^{U_k,(l)} = \sqrt{\frac{N\pi_j}{kN\pi_j}} = \sqrt{\frac{1}{k}}$, for $l = 1,\dots,k$, the correlation for all $i < j$ results in

$$\operatorname{Cov}(Z_i, Z_j) = \rho_{i,j} \sum_{l=1}^{i} w_i^{U_i,(l)} w_j^{U_j,(l)} = \sqrt{\frac{\pi_j}{\pi_i}} \sum_{l=1}^{i} \sqrt{\frac{1}{ij}} = \sqrt{\frac{i\pi_j}{j\pi_i}} \tag{5.25}$$

At each stage k, the observed value of Z_k is compared to a critical value c_k, which again can be found by formulating the PWER of the design and setting it equal to α. A type I error in $\mathcal{P}_{[j]}$ is made if any hypothesis concerning a superset of $\mathcal{P}_{[j]}$, can be rejected, that is, whenever H_i, $1 \le i \le j$, is rejected. So the PWER under $\boldsymbol{\theta}^* = \mathbf{0} \in \mathbb{R}^m$ is given by

$$\begin{aligned}
\mathrm{PWER}_{\boldsymbol{\theta}^*} &= \sum_{j=1}^{m} \pi_{[j]} \mathbb{P}_{\boldsymbol{\theta}^*}(\text{rej. at least one true } H_i \text{ for } i \le j) \\
&= \sum_{j=1}^{m} \pi_{[j]} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcup_{i \le j} \{Z_i \ge c_i\}\right) \qquad (5.26) \\
&= \sum_{j=1}^{m} \pi_j \mathbb{P}_{\boldsymbol{\theta}^*}\left(\bigcap_{i<j} \{Z_i < c_i\} \cap \{Z_j \ge c_j\}\right). \qquad (5.27)
\end{aligned}$$

with $\bigcap_{i<1}\{Z_i < c_i\} := \emptyset$. The last equation holds since the union $\bigcup_{i\le j}\{Z_i \ge c_i\}$ can be written as disjoint union of the sets $\bigcap_{i<k}\{Z_i < c_i\} \cap \{Z_k \ge c_k\}$, $k \le j$, and $\pi_j = \sum_{k=j}^{m} \pi_{[k]}$. Critical values can then, again, be found by either using a parameter family as the Wang and Tsiatis family or via error spending approach in combination with the multivariate normal distribution of the sub-vectors $\mathbf{Z}_j = (Z_i)_{i\le j}$. For the error spending, the expression in (5.27) can conveniently be used to write each addend as a stage-wise PWER. Each addend in (5.27) can be interpreted as the

probability of erroneously rejecting the null in $\mathcal{P}_j$ weighted by π_j. So by simply writing

$$\text{PWER}^{(k)} = \pi_k \mathbb{P}\left(\bigcap_{i<k} \{Z_i < c_i\} \cap \{Z_k \geq c_k\} \right) \tag{5.28}$$

we can find critical values by choosing an error spending function α^* and iteratively solving $\text{PWER}^{(k)} = \alpha^*(\tau_k) - \alpha^*(\tau_{k-1})$ for each $k = 1, \ldots, m$. Of course, for $\pi_k < \alpha^*(\tau_k) - \alpha^*(\tau_{k-1})$ this equation is not solvable. A possible soultion leading to a conservative procedure could be as follows. Choose a c_k by convenience (e.g. $c_k = \Phi^{-1}(1-\alpha)$) and redefine $\alpha^*(\tau_k)$ to be equal to the resulting $\sum_{l=1}^{k} \text{PWER}^{(l)}$ such that the unused level is kept for the later stage.
The maximal sample size N_{max} of the trial is given by $N_{max} = N \sum_{j=1}^{m} \pi_j$ and the sample size $N^{(k)}$ containing all information used up to stage k is given by $N^{(k)} = N \sum_{j=1}^{k} \pi_k$. We use information rates $\tau_k = N^{(k)}/N_{max} = \sum_{j=1}^{k} \pi_k / \sum_{j=1}^{m} \pi_k$ to describe the percentage of how far through the trial one has already gone. For example, for $m = 2$ populations/stages, this would mean $\tau_1 = 1/(1+\pi_2)$ and $\tau_2 = 1$. To control a certain level of power, we can again consider the PWP under some parameter constellation $\boldsymbol{\delta}_A = (\theta_k/\sigma_k)_{k=1}^m$. Here, the PWP under the assumption that all alternative hypotheses are true (all components of $\boldsymbol{\delta}_A$ are greater than 0) is simply given by

$$\text{PWP}_{\boldsymbol{\delta}_A} = \sum_{j=1}^{m} \pi_{[j]} \mathbb{P}_{\boldsymbol{\delta}_A} \left(\bigcup_{i \leq j} \{\tilde{Z}_i \geq c_i - \nu_i\} \right), \tag{5.29}$$

with $\nu_i = \mathbb{E}(Z_i) = \theta_i \sqrt{i \pi_i N}$ and $\tilde{Z}_i \sim \text{N}(0,1)$. Setting this quantity equal to $1-\beta$ one can now simply solve for N. The same could, of course, also be done with other power measures like $\text{Pow}_{1,\boldsymbol{\theta}_A}$. The ASN under some $\boldsymbol{\delta}_A$ is given by

$$\text{ASN}_{\boldsymbol{\delta}_A} = N + N \sum_{j=2}^{m} \pi_j \mathbb{P}_{\boldsymbol{\delta}_A} \left(\bigcap_{i<j} \{Z_i < c_i\} \right). \tag{5.30}$$

As a little example, say we have $m = 3$ nested populations as in Figure 5.2 with $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3) = (1, 0.6, 0.2)$ implying that $\pi_{[1]} = \pi_{[2]} = 0.4$ and $\pi_{[3]} = 0.2$. So the correlation matrix $\boldsymbol{\Sigma}$ of $\mathbf{Z} = (Z_1, Z_2, Z_3)$ equals

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.55 & 0.26 \\ & 1 & 0.47 \\ & & 1 \end{bmatrix}$$

which is found by the formula in (5.25). For the information rates we find $\tau_1 = 5/9$, $\tau_2 = 8/9$ and $\tau_3 = 1$ and can then calculate Wang & Tsiatis (or error spending) critical values. Pocock critical values are then equal to $\mathbf{c} = (c_1, c_2, c_3) = (2.17, 2.17, 2.17)$ and OBF-values are given by $\mathbf{c} = (2.49, 1.97, 1.86)$. Under $\boldsymbol{\delta}_A = (0.2, 0.25, 0.3)$ we find an $N = 253$ needed to satisfy $\text{PWP}_{\boldsymbol{\delta}_A} = 0.9$ for the Pocock design and $N = 245$ for the OBF design, yielding maximal sample sizes of $N_{max} = N \sum_{k=1}^{3} \pi_k = 1.8N = 455.4$ and 441, respectively. The average sample sizes for these values are given by 255.2 for the Pocock and 246.8 for the OBF design under the same $\boldsymbol{\delta}_A$.

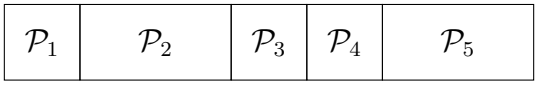

Figure 5.3: Example for $m = 5$ disjoint subpopulations.

5.3 Example from Magnusson & Turnbull (2013)

We want to exemplify the benefit of using PWER-control by applying it to the procedure presented in Magnusson & Turnbull (2013, [36]), who computed critical values by means of an error-spending approach in order to control the FWER at level α. The basic setting of their procedure, which we will denote by GSDS (DR-I in their publication), is as follows. Suppose an intervention is to be tested on a population $\mathcal{P}_0$. It is assumed that $\mathcal{P}_0$ can be partitioned into m disjoint subgroups $\mathcal{P}_1, \dots, \mathcal{P}_m$ as illustrated in in Figure 5.3 for $m = 5$. Furthermore, for $S \subseteq I = \{1, \dots, m\}$ define pooled subgroups $\mathcal{P}_S = \bigcup_{j \in S} \mathcal{P}_j$. For $j \in I$ denote prevalences $\pi_j \in [0, 1]$ of $\mathcal{P}_j$ which are not necessarily known, but can be prespecified in settings, e.g. when stratified randomization is used, and let $\pi_S = \sum_{j \in S} \pi_j$ be the prevalence in $\mathcal{P}_S$. At last, let treatment efficacy compared to control in $\mathcal{P}_j$ be denoted by θ_j and let the effect in $\mathcal{P}_S$ be given by $\theta_S = \sum_{j \in S} \pi_{S,j} \theta_j$, where $\pi_{S,j} = \pi_j / \pi_S$. For simplicity, $\theta_j \in [0, \infty)$ is assumed as in [36]. The null scenario in each population $\mathcal{P}_j$ is given by $\theta_j = 0$, similarly in pooled population $\theta_S = 0$.
Assume interim analysis times $0 < t_1 < \dots < t_K = 1$ have been planned. Define Y_{kj} and $\mathcal{I}_{kj}$ as the efficient score statistic and the observed Fisher's information for testing $H_j : \theta_j = 0$ at stage k. Given stagewise increments $X_{kj} = Y_{kj} - Y_{k-1,j}$ and $\Delta_{kj} = \mathcal{I}_{kj} - \mathcal{I}_{k-1,j}$, where $Y_{0j} = \mathcal{I}_{0j} = 0$. They assume that

$$X_{kj} \sim N(\theta_j \Delta_{kj}, \Delta_{kj}), \quad \forall k = 1, \dots, K \quad \text{and} \quad j \in I \tag{5.31}$$

For pooled populations $\mathcal{P}_S$ the above quantities are defined similarly:

$$Y_{k,S} = \sum_{j \in S} Y_{kj}, \quad X_{k,S} = Y_{k,S} - Y_{k-1,S}, \quad \mathcal{I}_{k,S} = \sum_{j \in S} \mathcal{I}_{kj}, \quad \Delta_{k,S} = \mathcal{I}_{k,S} - \mathcal{I}_{k-1,S} \tag{5.32}$$

Lastly, define a standardized test statistic $Z_{kj} = Y_{kj} / \sqrt{\mathcal{I}_{kj}}$, which will be compared to stagewise boundaries (l_k, u_k), $l_k \leq u_k$.
Their design GSDS can be summarized as follows: At stage 1 a pooled population is selected. Define $I^* = \{j \in I : Z_{1j} > l_1\}$ as the index set of all populations $\mathcal{P}_j$ which are to be pooled together, the remaining populations $\mathcal{P}_{j'}$, $j' \notin I^*$ are dropped from the trial and are not being tested in subsequent stages. At stages $k = 1, \dots, K$ the hypothesis $H_{I^*} : \theta_{I^*} = 0$ is tested by means of the test statistic $Z_{k,I^*} = Y_{k,I^*} / \sqrt{\mathcal{I}_{k,I^*}}$. If $Z_{k,I^*} \geq u_k$, then H_{I^*} is rejected and the trial is terminated for efficacy of the treatment. If $Z_{k,I^*} < l_k$, the trial is stopped for futility and if $Z_{k,I^*} \in (l_k, u_k)$ the trial proceeds to stage $k + 1$. At stage K, it is $l_K = u_K$ to force termination.
Critical boundaries $\mathbf{l} = (l_1, \dots, l_K)$ and $\mathbf{u} = (u_1, \dots, u_K)$ are computed by means of exit probabilities

$$\psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(\text{Select } S \text{ and reject } H_S \text{ exactly at stage } k), \quad S \subseteq I$$

which is equal to $\psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(I^* = S, Z_{1,S} \geq u_1)$ for $k = 1$ and for $k = 2, \dots, K$ it is $\psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) =$

$$\mathbb{P}_{\boldsymbol{\theta}}(I^* = S, Z_{1,S} < u_1, Z_{2,S} \in (l_2, u_2), \dots, Z_{k-1,S} \in (l_{k-1}, u_{k-1}), Z_{k,S} \geq u_k).$$

Effect size			Rej. prob. (P)				Rej. prob. (F)				$\mathbb{E}(I_T)$	
θ_0	θ_1	θ_2	H_0	H_1	H_2	Total	H_0	H_1	H_2	Total	P	F
0	0	0	.024	.006	.02	.05	.016	.004	.013	.033	6.6	6.74
1	1	1	.619	.051	.23	.9	.599	.046	.215	.86	6.4	6.7
.25	1	0	.101	.392	.011	.504	.078	.355	.007	.44	7.4	7.6
.5	2	0	.219	.653	.002	.873	.189	.652	.001	.843	6.6	7

Table 5.4: GSDS with PWER control with $\pi_1 = 1/4$, $\pi_2 = 3/4$, $K = m = 2$. Critical values for PWER-control are $l_1 = 0.436$, $u_1 = 2.125$, $u_2 = l_2 = 1.853$. Critical values under FWER-control (following Magnusson and Turnbull) are $l_1 = 0.436$, $u_1 = 2.278$, $u_2 = l_2 = 2.077$. Here P denotes our PWER-based approach while F denotes the FWER-based approach.

Analogously, they define the probability that the trial stops at stage k with acceptance of H_S as

$$\xi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(\text{Select } S \text{ and accept } H_S \text{ exactly at stage } k), \quad S \subseteq I$$

for which they define $\xi_{1,\emptyset} = \mathbb{P}_{\boldsymbol{\theta}}(I^* = \emptyset)$ and for nonempty $S \subseteq I$, $\xi_{1,S} = 0$. For $k = 2, \ldots, K$ it is $\xi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) =$

$$\mathbb{P}_{\boldsymbol{\theta}}(I^* = S, Z_{1,S} < u_1, Z_{2,S} \in (l_2, u_2), \ldots, Z_{k-1,S} \in (l_{k-1}, u_{k-1}), Z_{k,S} \leq l_k).$$

Denote marginal stopping probabilities as

$$\begin{aligned}
\psi_k(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}(\text{Stop exactly at stage } k \text{ with rejection of some } H_S) \\
&= \sum_{S \subseteq I} \psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}), \\
\xi_k(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}(\text{Stop exactly at stage } k \text{ with no rejection of any } H_S) \\
&= \sum_{S \subseteq I} \xi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}).
\end{aligned}$$

Then upper and lower bounds $\mathbf{u} = (u_1, \ldots, u_K)$ and $\mathbf{l} = (l_1, \ldots, l_K)$ are defined via

$$\psi_k(\mathbf{l}, \mathbf{u}; \mathbf{0}) = \alpha_U^*(t_k) - \alpha_U^*(t_{k-1}) \quad \text{and} \quad \xi_k(\mathbf{l}, \mathbf{u}; \mathbf{0}) = \alpha_L^*(t_k) - \alpha_L^*(t_{k-1})$$

with $\alpha_U^* : [0,1] \to [0, \alpha]$ ($\alpha_U^*(0) = 0, \alpha_U^*(1) = \alpha$) being a non-decreasing upper α-spending function and $\alpha_L^* : [0,1] \to [0, 1-\alpha]$ ($\alpha_L^*(0) = 0, \alpha_L^*(1) = 1-\alpha$) a non-decreasing $(1-\alpha)$-spending function. The definition of these functions ensure that $l_K = u_K$ is forced, since ψ_k and ξ_k are probabilities of complementary events. Note that the critical values are determined under $\boldsymbol{\theta} = \mathbf{0}$ which ensures that the familwise error rate,

$$FWER_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one } H_S,\, S \subseteq I^0(\boldsymbol{\theta})),$$

is bounded by α for all $\boldsymbol{\theta} \in \Theta$, where $I^0(\boldsymbol{\theta}) = \{i \in I : \theta_i = 0\}$. It has to be said, however, that the authors state that the proof for this statement is based on the fact that they do not allow the components of $\boldsymbol{\theta}$ to be negative.

Effect size			Rej. prob. (P)				Rej. prob. (F)				$\mathbb{E}(I_T)$	
θ_0	θ_1	θ_2	H_0	H_1	H_2	Total	H_0	H_1	H_2	Total	P	F
0	0	0	.025	.013	.013	.05	.017	.008	.008	.034	6.67	6.72
1	1	1	.676	.112	.112	.9	.653	.103	.103	.859	6.42	6.7
.5	1	0	.186	.488	.006	.68	.161	.449	.004	.614	7.19	7.41
1	2	0	.303	.685	0	.988	.298	.684	0	.983	5.31	5.51

Table 5.5: GSDS with PWER control with $\pi_1 = \pi_2 = 1/2$, $K = m = 2$. Critical values for PWER-control are $l_1 = 0.436$, $u_1 = 2.1557$, $u_2 = l_2 = 1.8707$. Critical values under FWER-control (following Magnusson and Turnbull) are $l_1 = 0.436$, $u_1 = 2.2976$, $u_2 = l_2 = 2.0980$. Here P denotes our PWER-based approach while F denotes the FWER-based approach.

Now, instead of FWER-control we try to ensure PWER-control under the global null by means of the PWER-error-spending approach from the previous sections. The PWER for GSDS is given by

$$\begin{aligned} PWER_{\boldsymbol{\theta}} &= \sum_{j=1}^{m} \pi_j \mathbb{P}_{\boldsymbol{\theta}}(\text{reject at least one } H_S,\, j \in S \subseteq I^0(\boldsymbol{\theta})) \\ &= \sum_{i=1}^{m} \pi_j \mathbb{P}_{\boldsymbol{\theta}}\Big(\bigcup_{j \in S \subseteq I^0(\boldsymbol{\theta})} \{I^* = S, \text{reject } H_S\} \Big) \\ &= \sum_{j=1}^{m} \pi_j \sum_{j \in S \subseteq I^0(\boldsymbol{\theta})} \sum_{k=1}^{K} \mathbb{P}_{\boldsymbol{\theta}}(I^* = S, \text{reject } H_S \text{ exactly at stage } k) \\ &= \sum_{k=1}^{K} \sum_{j=1}^{m} \pi_j \sum_{j \in S \subseteq I^0(\boldsymbol{\theta})} \psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}) = \sum_{k=1}^{K} PWER_{\boldsymbol{\theta}}^{(k)}(\mathbf{l}, \mathbf{u}) \end{aligned}$$

and can thus be written as sum of stagewise PWERs. The authors showed a numerical example for $K = m = 2$, $\pi_1 = 1/4$, $\pi_2 = 3/4$, $\alpha = .05$ and equally spaced stages where they compared their procedure to that of a design by Stallard & Facey (1996, [49]) in terms of rejection probabilities of the hypotheses H_1, H_2 and $H_{\{1,2\}}$ and the expected information at termination. They found that their procedure is advantageous when the treatment efficacy is strong in the small subgroup. We want to use the PWER-error spending approach to see by how much the results differ in terms of rejection probabilities and expected information. They computed the selection boundary l_1 via

$$\mathbb{P}(I^* = \emptyset) = \mathbb{P}(Z_{1,1} \leq l_1, Z_{1,2} \leq l_1) = \alpha_U^*(t_1) = \frac{1-\alpha}{2}$$

yielding $l_1 = \Phi^{-1}(\sqrt{(1-\alpha)/2}) = 0.4936$. u_1 is then found by solving $\psi_{1,1} + \psi_{1,2} + \psi_{1,\{1,2\}} = \alpha_U^*(t_1) = \alpha/2$ and $u_2 = l_2$ is found by solving $\psi_{2,1} + \psi_{2,2} + \psi_{2,\{1,2\}} = \alpha - \alpha_U^*(t_1) = \alpha/2$ (or the same equation with ξ, $(1-\alpha)/2$ and α_L^* instead of ψ, α and α_U^*). This results in $(l_1, u_1) = (0.4936, 2.2783)$ and $l_2 = u_2 = 2.0772$. For the PWER approach we use the same selection boundary l_1 to make the two approaches more comparable. Then we solve

$$PWER^{(1)}(l_1, u_1) = \pi_1(\psi_{1,1} + \psi_{1,\{1,2\}}) + \pi_2(\psi_{1,2} + \psi_{1,\{1,2\}}) = \alpha_U^*(t_1) = \alpha/2 \quad (5.33)$$

yielding $(l_1, u_1) = (0.4936, 2.1254)$ and solving

$$PWER^{(2)}(u_2) = \pi_1(\psi_{2,1} + \psi_{2,\{1,2\}}) + \pi_2(\psi_{2,2} + \psi_{2,\{1,2\}}) = \alpha/2 \tag{5.34}$$

yields $u_2 = 1.8549$. Both upper boundaries are lower than those from Magnusson and Turnbull indicating higher rejection probabilities and lower expected information. A 'power' of $1 - \beta = .9$ under a configuration $\boldsymbol{\theta}^*$ can be ensured by finding the maximal information $\mathcal{I}_{max}$ needed to ensure that the following quantity is equal to $1 - \beta$:

$$\mathbb{P}_{\boldsymbol{\theta}^*}(\text{reject some } H_S,\, S \subseteq I | \theta_j = \theta^*, \forall j \in I) = \sum_{S \subseteq I} \sum_{k=1}^{K} \psi_{k,S}(\mathbf{l}, \mathbf{u}; \boldsymbol{\theta}^*) \tag{5.35}$$

Note that the rejection probabilities for the hypotheses $H_S : \theta_S = 0$ are given by the addends of the first sum of the above power quantity. Of course, one could also use the PWP mentioned in previous sections. Magnusson & Turnbull computed a maximal information of $\mathcal{I}_{max} = 10.31$ under $\boldsymbol{\theta} = (1, 1)$ to achieve a power of 90%. When using the PWER-approach we get $\mathcal{I}_{max} = 8.99$ (13% relative difference). The resulting rejection probabilities and expected information values can be seen in Table 5.4. The first row consists of the corresponding addends of the PWER, respectively. Of course, by construction, the PWER is equal to $\alpha = 0.05$ under the PWER-approach. Under the GSDS approach the PWER is equal to 0.033, which was to be expected due to the critical values being larger. One can see larger rejection probabilities for each possible scenario under the PWER-approach compared to the GSDS-approach and in total the power increases by $4-6\%$ depending on the concrete parameter configuration. The expected information is also smaller for under all considered parameter configurations even though the differences are rather slight. So for this design, the PWER only give a rather small improvement.
Rejection probabilities of the individual hypotheses are directly linked to the values of u_1 and u_2. As seen in Table 5.4 the rejection probabilities under the PWER-approach are all higher than under the GSDS, since the critical values are lower ($\mathbf{u}_{PWER} = (2.1254, 1.8549)$ vs. $\mathbf{u} = (2.2783, 2.0772)$).
Now suppose that we modified the GSDS in a way that after the selection of a population $\mathcal{P}_{I^*}$, $I^* \subseteq I$, we not only test H_{I^*} but also all individual hypotheses H_i, where $i \in I^*$. That is, if $I^* = \{1, 2\}$, the hypotheses $H_{\{1,2\}}, H_{\{1\}}$ and $H_{\{2\}}$ are to be tested. Then the PWER of this modified design (for $K = m = 2$) is given by

$$\begin{aligned}
PWER_{\boldsymbol{\theta}} &= \sum_{i=1}^{K} \pi_i \mathbb{P}_{\boldsymbol{\theta}}(\text{At least one type I error in } \{i\}) \\
&= \sum_{i=1}^{K} \pi_i \mathbb{P}_{\boldsymbol{\theta}}(\text{reject any } H_S \text{ or } H_{\{i\}},\, S \subseteq I^0(\boldsymbol{\theta}),\, i \in S) \\
&= \pi_1(\mathbb{P}_{\boldsymbol{\theta}}(I^* = \{1\}, \text{rej. } H_{\{1\}}) + \mathbb{P}_{\boldsymbol{\theta}}(I^* = \{1, 2\}, \text{rej. } H_{\{1,2\}} \text{ or } H_{\{1\}})) \\
&+ \pi_2(\mathbb{P}_{\boldsymbol{\theta}}(I^* = \{2\}, \text{rej. } H_{\{2\}}) + \mathbb{P}_{\boldsymbol{\theta}}(I^* = \{1, 2\}, \text{rej. } H_{\{1,2\}} \text{ or } H_{\{2\}}))
\end{aligned}$$

Again, this expression can be transformed into a sum of K individual PWER-expressions such that the PWER-error spending approach can be applied to find critical values u_1 and u_2. As before, it is $\mathbb{P}_{\boldsymbol{\theta}}(I^* = \{i\}, \text{rej. } H_{\{i\}}) = \psi_{1i} + \psi_{2i}$, $i = 1, 2$. The other probability terms can be written as a sum of first and second stage rejection probabilities and can, for example, be simulated using the joint distributions of

the test statistics. The critical values obtained by applying PWER-error spending (with $\alpha = .05$ and the same error-spending function as before) are

$$\mathbf{u} = (2.1362, 1.8717)$$

which are still significantly smaller than those found under the GSDS-approach. Thus, even if we make the design more flexible in terms of the number of possible hypotheses being tested, the rejection probabilities of the individual hypotheses will still be larger than under the FWER-controlled procedure.

6. Adaptive designs with PWER-control

In Chapter 4 we have introduced a general K-stage group sequential design for arbitrary population structures and proposed conditions and methods that ensure a strong control of the PWER. Even though the error spending approach offers flexibility by allowing for PWER-control under randomly developing sample sizes (or information rates), group sequential designs are still quite restrictive in terms of planning and execution. In a group sequential design, sample sizes, error spending functions and hypotheses to be tested have to be planned beforehand and no deviations from this plan are allowed as they can potentially lead to an inflation of the overall type I error of the trial. Apart from those restrictions, study populations have to be pre-planned in advance as well. Especially when dealing with multiple possibly intersecting populations with multiple subgroups one often wants to pool certain subgroups if the subgroups display a notable amount of homogeneity. Or sometimes it is simply desirable to test in a subgroup that has not been specified beforehand but which seems promising in terms of treatment efficacy based on descriptive analyses or other external information. Adaptive group sequential designs enable us to incorporate all these potential mid-trial adjustments without undermining the validity of the trial, i.e. without risking an inflation of the overall type I error rate. As already described in Chapter 1 classical approaches on how to conduct such mid-trial adaptations are (i) the combination function approach, where test statistics (p-values) from each individual stage are combined by means of a combination function, whose value is then compared to some boundary, and (ii) the conditional error function approach where the conditional type I error (conditioned on data from previous stages of the trial) is computed and used as a local significance level for the remaining stages after adaptation was conducted. One way to understand an adaptive design is the CRP-principle by Müller & Schäfer [41] which incorporates the conditional error function approach. The basic notion of the CRP-principle is to start the trial with a conventional non-adaptive design at level α, e.g. a group-sequential design, and to use the conditional type I error rate as conditional error function when a data-driven design change is to be made. In the following subsections we want to adopt this CRP-principle to the PWER, so that

design adaptations can be made while still keeping the overall PWER at level α. In particular, this principle will make it possible to start with any K-GSD and switch to a different design structure after any particular interim analysis.
This chapter is structured as follows. First, the CRP-principle with PWER-control is introduced theoretically. After that, a numerical example is given to see how the resulting critical values after a mid-trial adaptation look like and to generally improve the understanding of the method. Lastly, the principle is used in a simulation study and power and (expected) sample size are compared to various other adaptive and purely group sequential designs.

6.1 The CRP-principle for PWER-control

Before describing the PWER-adjusted version of the CRP-principle, we want to revisit the CRP-principle originally stated by Müller and Schäfer and then build up on it. Consider an initial level-α-test decision function φ for some null hypothesis H_0. Suppose, an interim analysis is performed after following the initial test procedure and that we learn from this interim data (and/or possibly external information) that certain design features should be changed. Then, φ is replaced by a new test φ' which is required to control the conditional type-I-error of the initial test in order to avoid an inflation of the overall type-I-error. For now, let z_{int} denote interim data collected up to said interim analysis. Then the conditional type I error rate of the initial test is given by

$$A(z_{int}) = \mathbb{E}_{H_0}(\varphi|z_{int}) = \mathbb{P}_{H_0}(\varphi = 1|z_{int}) \tag{6.1}$$

Note that by writing $\mathbb{E}_{H_0}(\varphi|z_{int})$ we technically mean $\mathbb{E}_{H_0}(\varphi|Z_{int} = z_{int})$, where Z_{int} is a random variable mapping to the sample space of all possible data points up to the point when the interim analyses is conducted. The CRP-principle states that when an adaptation takes place, one can still control the overall type I error by constructing φ' such that its conditional error rate A' satisfies

$$A'(z_{int}) \leq A(z_{int}). \tag{6.2}$$

By Theorem 1.2.1, this then yields

$$\mathbb{E}_{H_0}(A'(z_{int})) \leq \mathbb{E}_{H_0}(\varphi) \leq \alpha \tag{6.3}$$

and therefore the overall type I error rate is still bounded by α. Thus, the conditional type I error will never exceed $A(z_{int})$ irrespective of whether there is a change in design features or not. This enables us to start with any K-stage group sequential design using any critical values that ensure overall PWER-control. Then, a mid-trial adjustment at some stage $1 < k' \leq K$ (e.g. using a sample size of $n' = 100$ instead of initially planned $n = 50$) can be conducted by defining a new test φ' which contains the desired changes and has to satisfy (6.2). From this equality, new, data-dependent critical values can be computed which are then used for the remainder of the trial.

We mimic this idea and use the CRP-principle to construct an adaptive design that controls the PWER. Let us say we have a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ with $\mathcal{H} = \{H^U : \theta^U \leq 0 \mid U \in \mathfrak{U}\}$ and a test decision function for a K-stage group sequential design denoted as $\boldsymbol{\varphi}$ as in 4.1. We first rewrite the PWER

in terms of an expectation of a random variable as done in (2.19). Remember that $I_K^U := \{k \in \{1,\dots,K\} \mid U \in \mathfrak{U}_k\}$ describes the stages a certain H^U is planned to be tested at. So by using

$$\varphi_J^{\boldsymbol{\theta}} := \begin{cases} 1, & \text{if} \quad \sum_{U \in I_J(\boldsymbol{\theta})} \sum_{k \in I_K^U} \varphi_U^{(k)} \geq 1 \\ 0, & \text{else} \end{cases}, \tag{6.4}$$

which describes an indicator function that is 1 iff any type I error relevant to $\mathcal{P}_J$ has been made, we can define the weighted mean of indicator functions

$$\varphi^{\boldsymbol{\theta}} = \sum_{J \subseteq I} \pi_J \varphi_J^{\boldsymbol{\theta}} \tag{6.5}$$

to rewrite the PWER as

$$PWER_{\boldsymbol{\theta}} = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{P}_{\boldsymbol{\theta}} \left(\varphi_J^{\boldsymbol{\theta}} = 1\right) = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J \mathbb{E}_{\boldsymbol{\theta}} \left(\varphi_J^{\boldsymbol{\theta}}\right) = \mathbb{E}_{\boldsymbol{\theta}} \left(\varphi^{\boldsymbol{\theta}}\right). \tag{6.6}$$

The expectation $\mathbb{E}_{\boldsymbol{\theta}}(\varphi^{\boldsymbol{\theta}})$ describes the expected percentage of random future patients that are affected by at least one erroneous rejection.
For simplicity, we will assume $K = 2$ for the remainder of this chapter, since the following ideas can all be generalized for an arbitrary number of stages. If at stage $K = 2$ a deviation from the initial trial design is planned, we need to compute the PWER given the interim data $\mathbf{z}_{int}$ collected up to stage $K - 1 = 1$. We denote this conditional PWER as *cp*, which takes on the role of the conditional error function A above and is given by the following definition.

Definition 6.1.1. *(Conditional PWER) Let $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$ be a population-wise testing problem with 2-stage GSD test function $\boldsymbol{\varphi} = (\varphi^{U,(k)})$ with $U \in \mathfrak{U}$ and $k = 1, 2$. Further let $\boldsymbol{Z}_{int} = \left(Z_{int}^{U,(1)}\right)_{U \in \mathfrak{U}}$ be a vector of accumulated test statistics for testing all H^U, $U \in \mathfrak{U}$, and $\boldsymbol{z}_{int}$ a realization of it. Let $\mathcal{Z}$ be the space of all possible realizations of $\boldsymbol{Z}_{int}$. For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ let $(\boldsymbol{Z}_{int})_{I_J(\boldsymbol{\theta})}$ be the sub-vector containing test statistics for all true null hypotheses whose erroneous rejection concern $\mathcal{P}_J$, $J \in \mathcal{C}_{\mathcal{P}}$, and $\varphi_J^{\boldsymbol{\theta}}$ defined as in (6.4). Further, let $A_J^{\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ be a conditional error function defined as*

$$A_J^{\boldsymbol{\theta}}(\boldsymbol{z}_{int}) := \begin{cases} 1, & \text{if } \sum_{U \in I_J(\boldsymbol{\theta})} \varphi^{U,(1)} \geq 1 \\ \mathbb{E}_{\boldsymbol{\theta}}(\varphi_J^{\boldsymbol{\theta}} | (\boldsymbol{Z}_{int})_{I_J(\boldsymbol{\theta})} = (\boldsymbol{z}_{int})_{I_J(\boldsymbol{\theta})}), & \text{otherwise} \end{cases} \tag{6.7}$$

The conditional PWER *$cp_{\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ is then defined as*

$$cp_{\boldsymbol{\theta}}(\boldsymbol{z}_{int}) = \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J A_J^{\boldsymbol{\theta}}(\boldsymbol{z}_{int}). \tag{6.8}$$

The above expectations $\mathbb{E}_{\boldsymbol{\theta}}(\varphi_J^{\boldsymbol{\theta}} | (\mathbf{Z}_{int})_{I_J(\boldsymbol{\theta})} = (\mathbf{z}_{int})_{I_J(\boldsymbol{\theta})})$ will from now on be shortened to $\mathbb{E}_{\boldsymbol{\theta}}(\varphi_J^{\boldsymbol{\theta}} | (\mathbf{z}_{int})_{I_J(\boldsymbol{\theta})})$. Assuming a multivariate normal distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ for a test statistics vector $(\mathbf{Z}^{U,(k)})_{\substack{U \in \mathfrak{U} \\ k=1,2}}$ and using $\varphi^{U,(k)} = \mathbf{1}_{\{Z^{U,(k)} \geq c^{(k)}\}}$ one can rewrite these conditional expectations as probabilities

$$\mathbb{E}_{\boldsymbol{\theta}}(\varphi_J^{\boldsymbol{\theta}} | (\mathbf{z}_{int})_{I_J(\boldsymbol{\theta})}) = \mathbb{P}_{\boldsymbol{\theta}} \left(\bigcup_{U \in I_J^{(2)}(\boldsymbol{\theta})} \{Z^{U,(2)} \geq c^{(2)}\} \mid (\mathbf{z}_{int})_{I_J^{(2)}(\boldsymbol{\theta})} \right) \tag{6.9}$$

where $I_J^{(2)}(\boldsymbol{\theta}) = \{U \in I_J(\boldsymbol{\theta}) \mid 2 \in I_2^U\}$ being the set of all H^U that are planned to be tested at stage 2. Since we condition on all stage 1 data (and therefore stage 1 test decisions), the rejection regions $\{\varphi^{U,(1)} = 1\}$ vanish.
The general plan is now to use this definition to mimic the classical CRP-principle. Heuristically, if a mid-trial adaptation is planned to be conducted at stage 2, we compute the conditional PWER cp_b under the new design $\boldsymbol{\varphi}_b$ that encompasses all the design changes to be made and the conditional PWER cp_a under the initial design $\boldsymbol{\varphi}_a$. This cp_b depends on a stage 2 crtical value c_b that is found by solving $cp_b(c_b) \leq cp_a$ under all possible parameter constellations.
More formally, let $\boldsymbol{\varphi}_a$ be the $K = 2$-stage initial GSD for a population-wise testing problem $\mathcal{D}_a = (\mathcal{C}_{\mathcal{P}_a}, \boldsymbol{\pi}_a, \mathcal{M}_{\boldsymbol{\Theta}_a}, \mathcal{H}_a)$ and let $\boldsymbol{\varphi}_b$ be the new test for the new problem $\mathcal{D}_b = (\mathcal{C}_{\mathcal{P}_b}, \boldsymbol{\pi}_b, \mathcal{M}_{\boldsymbol{\Theta}_b}, \mathcal{H}_b)$. Generally, all components of $\mathcal{D}_a$ and $\mathcal{D}_b$ can differ from each other, allowing for the introduction of new (sub-)populations, new hypotheses etc. In our examples we will assume that $\mathcal{C}_{\mathcal{P}_a} = \mathcal{C}_{\mathcal{P}_b}$ and $\boldsymbol{\pi}_a = \boldsymbol{\pi}_b$, so the population structure does not change, but in general it is possible that, say, a whole new population is introduced to the trial which would then lead to a different set of relative population sizes $\boldsymbol{\pi}_b$ as well. For the hypothesis sets of the old and new designs, $\mathcal{H}_a$ and $\mathcal{H}_b$, we again consider index sets $\mathfrak{U}_a$ and $\mathfrak{U}_b$, respectively.

Example 6.1.1. *As an example consider a GSD $\boldsymbol{\varphi}_a$ for $\mathcal{D}_a = (\mathcal{C}_{\mathcal{P}_a}, \boldsymbol{\pi}_a, \mathcal{M}_{\boldsymbol{\Theta}_a}, \mathcal{H}_a)$ with the assumptions from Design I of Section 5.1.1, i.e. with $I = \{1, 2\}$, $\mathcal{C}_{\mathcal{P}_a} = 2^I$, $\boldsymbol{\pi}_a = (\pi_{a,J})_{J \in \mathcal{C}_{\mathcal{P}_a}}$, $\boldsymbol{\Theta}_a = \mathbb{R}^2$ and $\mathcal{H}_a = \{H_1, H_2\}$ with $H_j : \theta(\mathcal{P}_{a,j}, T) \leq 0$ for $j = 1, 2$ planned to be tested at $K = 2$ stages, respectively. Assume that H_1 can be rejected at stage 1 and H_2 cannot. Say that due to a descriptive analysis of the interim data or some external information one decides against a further test of H_2 and wants to test for an effect $\theta_{\{1,2\}}$ in $\mathcal{P}_{\{1,2\}}$ instead. This mid-trial design change could then be conducted by introducing a new GSD $\boldsymbol{\varphi}_b$ for $\mathcal{D}_b = (\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}_b}, \mathcal{H}_b)$ that describes Design II of Section 5.1.2. Thus $\mathcal{C}_{\mathcal{P}_a} = \mathcal{C}_{\mathcal{P}_b}$, $\boldsymbol{\pi}_a = \boldsymbol{\pi}_b$, $\boldsymbol{\Theta}_b = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{\{1,2\}}) \mid \boldsymbol{\theta} \in \mathbb{R}^3\}$ and $\mathcal{H}_b = \{H_1, H_2, H_{\{1,2\}}\}$ with H_1, H_2 tested at stage 1 (already happened) and $H_{\{1,2\}}$ at stage 2 (still to be done).*

The above example implicitly showed that the introduction of a new population we want to test in (here the intersection $\mathcal{P}_{\{1,2\}}$) can change the dimension of the parameter space the initial design was defined on. The problem is that parameters that implicitly determine the size of other parameters can be newly introduced, just like $\theta_{\{1,2\}}$ influences both θ_1 and θ_2. Particularly, with regard to controlling the overall PWER of the adaptive design, we will have to consider a 'combined' parameter space $\boldsymbol{\Theta}_C$ containing all parameters of the initial and new design. To this end we want to restrict ourselves to the cases of

$$\mathcal{C}_{\mathcal{P}_a} = \mathcal{C}_{\mathcal{P}_b} \quad \text{and} \quad \boldsymbol{\pi}_a = \boldsymbol{\pi}_b,$$

i.e. to cases where the overall population structure (number of populations and sizes of populations) remain the same. At the end of this section we will shortly discuss the measures one would have to take if this restriction was dropped. Generally, for $\mathcal{P}_a = \mathcal{P}_b$ and $\boldsymbol{\pi}_a = \boldsymbol{\pi}_b$ the population structure remains the same and thus a $J \in \mathcal{C}_{\mathcal{P}_a}$ describes the same disjoint subgroup $\mathcal{P}_J$ that a $J \in \mathcal{C}_{\mathcal{P}_b}$ does. The set $\mathfrak{U}_C := \mathfrak{U}_a \cup \mathfrak{U}_b \cup \mathcal{C}_{\mathcal{P}}$ contains the indices of all populations that are tested by either

the old or new design $(\mathfrak{U}_a \cup \mathfrak{U}_b)$ completed by all indices for each stratum $\mathcal{P}_J$. We then consider the combined parameter space

$$\boldsymbol{\Theta}_C := \left\{ \boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}_C} \mid \theta^U = \sum_{J \in U} \pi_J^U \theta_J \right\} \subseteq \mathbb{R}^{|\mathfrak{U}_C|}. \tag{6.10}$$

In Example 6.1.1 we therefore had constellations of the form $(\theta_1, \theta_2, \theta_{\{1\}}, \theta_{\{2\}}, \theta_{\{1,2\}})$. Now, due to $\mathfrak{U}_b \neq \mathfrak{U}_a$ in general, the conditional PWER under the new design may depend on other parameters than the conditional PWER under the old design, i.e. $cp^b_{\boldsymbol{\theta}}$ is based on a probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{new}$ whereas $cp^a_{\boldsymbol{\theta}}$ is based on a measure $\mathbb{P}'_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_a$. To this end we will simply consider a family of probability measures $\mathcal{M}_{\boldsymbol{\Theta}_C}$ over $\boldsymbol{\Theta}_C$.
Now, analogously to (6.3), we require

$$cp_{b,\boldsymbol{\theta}}(\mathbf{z}_{int}) \leq cp_{a,\boldsymbol{\theta}}(\mathbf{z}_{int}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_C, \mathbf{z}_{int} \in \mathcal{Z} \tag{6.11}$$

such that the overall type-I-error of the adaptive design is still under control. Mimicking the proof strategy from Theorem 1.2.1 proving the following theorem is quite straightforward.

Theorem 6.1.1. *Let $\mathcal{D}_a$ and $\mathcal{D}_b$ be two population-wise testing problems and $\boldsymbol{\varphi}_a$ and $\boldsymbol{\varphi}_b$ the respective α-level PWER-controlling K-stage GSDs thereof. If (6.11) holds for all interim data $\boldsymbol{Z}_{int}$, then the overall type I error of the new design $\mathcal{D}_b$ is bounded by α, i.e. $\mathbb{E}_{\boldsymbol{\theta}}(cp_{b,\boldsymbol{\theta}}) \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_C$.*

Proof. Due to a basic property of the conditional expectation we know that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ it is

$$\mathbb{E}_{\boldsymbol{\theta}}(cp_{\boldsymbol{\theta}}(\mathbf{Z}_{int})) = \mathbb{E}_{\boldsymbol{\theta}}(\mathbb{E}_{\boldsymbol{\theta}}(\varphi^{\boldsymbol{\theta}} | \mathbf{Z}_{int})) = \mathbb{E}_{\boldsymbol{\theta}}(\varphi^{\boldsymbol{\theta}}) = PWER_{\boldsymbol{\theta}}.$$

Hence, by applying (6.11) we find

$$\mathbb{E}_{\boldsymbol{\theta}}(cp'_{\boldsymbol{\theta}}(\mathbf{Z}_{int})) \leq \mathbb{E}_{\boldsymbol{\theta}}(cp_{\boldsymbol{\theta}}(\mathbf{Z}_{int})) = PWER_{\boldsymbol{\theta}} \leq \alpha, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_C.$$

□

The above result theoretically allows us to find a new stage 2 critical value c_b such that the overall error rate of the new design is still bounded by α. However, requirement (6.11) is practically quite unfeasible because it must hold for all configurations $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$, where $\boldsymbol{\Theta}_C$ usually is a higher dimensional real coordinate space. It would be easier to implement our CRP-principle if we only needed to find c_b by solving a system consisting of a finite number of inequations. We will later show that we can achieve this by demanding a somewhat weaker requirement than (6.11). Namely, instead of requiring the inequality to hold for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ we only make this requirement for all $\boldsymbol{\theta}$ whose components are either equal to 0 or in the alternative. Before we formulate this requirement, we want to point out that the system (6.11) of inequalities is not always solvable for every type of design change.

Example 6.1.2. *Suppose we start a trial with Design II from Section 5.1.2. So at stage 1 we test $H_j : \theta_j \leq 0$ with $\theta_j = (\pi_{\{j\}}/\pi_j)\theta_{\{j\}} + (\pi_{\{1,2\}}/\pi_j)\theta_{\{1,2\}}$ and a corresponding critical value $c^{(1)}$ found by some PWER-controlling method (e.g. using*

the Wang & Tsiatis family). Again this corresponds to $\mathfrak{U}_a = \{U_1, U_2, U_{\{1,2\}}\}$ *with* $U_j = \{\{j\}, \{1,2\}\}$ *and* $U_{\{1,2\}} = \{\{1,2\}\}$*. Assume that we can reject* H_1*, but not* H_2*, and now, instead of only aiming for a test in the intersection, we also want to test in* $\mathcal{P}_{\{1\}}$ *– so we aim for testing in more informative sub-populations. This may be a valid approach if the fact that we can reject* H_1 *is still not compelling enough for us or if descriptive analyses revealed that there is a heterogeneous effect in* $\mathcal{P}_{\{1\}}$ *and we thus want to take a closer look into both disjoint subgroups. That is, the rejection of* H_1 *is ignored, so* $\varphi_b^{U_1,(1)}$ *is (artificially) set to 0 in the new design. Since this is an unplanned design change, we want to apply our CRP-principle. This means that for each possible constellation of* $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{\{1\}}, \theta_{\{2\}}, \theta_{\{1,2\}}) \in \boldsymbol{\Theta}_C$ *we would need to ensure that (6.11) holds. The concrete form of the conditional PWER-expressions, however, is only dependent on the components of* $\boldsymbol{\theta}$ *being* ≤ 0 *or* > 0*. With stage 1 data* $\boldsymbol{z}_{int} = \left((z_J^{(1)})_{J\subseteq I}, (z_j^{(1)})_{j=1,2}\right)$ *we can find the conditional PWER under each* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$*. For example, if* $\boldsymbol{\theta} = \boldsymbol{0}$ *we have*

$$cp_{a,\boldsymbol{0}}(\boldsymbol{z}_{int}) = \pi_{\{1\}} \cdot 1 + \pi_{\{1,2\}} \cdot 1 + \pi_{\{2\}} \cdot 0 = \pi_1$$

because rejection of H_1 *means an error has been made in* $\mathcal{P}_1$*, so in both* $\mathcal{P}_{\{1\}}$ *and* $\mathcal{P}_{\{1,2\}}$*. Since no further test is forseen in any population that could affect patients in* $\mathcal{P}_{\{2\}}$ *the conditional error probability corresponding to* $\mathcal{P}_{\{2\}}$ *equals 0. Note that the equation above holds for every* $\boldsymbol{\theta}$ *in*

$$\left\{\boldsymbol{\theta} \in \boldsymbol{\Theta}_C \mid \boldsymbol{\theta} = (\theta^U)_{U\in\mathfrak{U}_C} \text{ with } \theta^U \leq 0\right\},$$

and not only for $\boldsymbol{\theta} = \boldsymbol{0}$*. Under the new design, we ignore the rejection of* H_1*, so the errors made in* $\mathcal{P}_{\{1\}}$ *and* $\mathcal{P}_{\{1,2\}}$ *are ignored as well. So we have*

$$cp_{b,\boldsymbol{0}}(\boldsymbol{z}_{int}) = \pi_{\{1\}}\mathbb{P}\left(Z_{\{1\}}^{(2)} \geq c_b^{(2)} | z_{\{1\}}^{(1)}\right) + \pi_{\{1,2\}}\mathbb{P}\left(Z_{\{1,2\}}^{(2)} \geq c_b^{(2)} | z_{\{1,2\}}^{(1)}\right)$$

which is obviously smaller than π_1*, so the inequality holds for any* $c_b^{(2)}$ *in this case. Now, for* $\boldsymbol{\theta}$ *such that* $\theta_1 > 0, \theta_2 = 0, \theta_{\{1\}} = 0, \theta_{\{2\}} < 0, \theta_{\{1,2\}} > 0$ *we find*

$$cp_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) = \pi_{\{1\}} \cdot 0 + \pi_{\{1,2\}} \cdot 0 + \pi_{\{2\}} \cdot 0 = 0.$$

But under the new design it is

$$cp_{b,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) = \pi_{\{1\}}\mathbb{P}\left(Z_{\{1\}}^{(2)} \geq c_b^{(2)} | z_{\{1\}}^{(1)}\right)$$

meaning that the inequality never holds and the test procedure can not be conducted. A way to overcome this issue is to replace in the old design for each false hypothesis H^U *the first and second stage test statistics by synthetic random variables* $\tilde{Z}_U$ *that follow the unconditional joint distribution of original test statistics under the global null hypothesis. For each of these hypotheses we then consider the unconditional worst case probability of making a type I error in the respective populations. That is, we write* cp_a *as*

$$cp_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) = \pi_{\{1\}}\mathbb{P}_{\boldsymbol{0}}\left(\tilde{Z}_1^{(1)} \geq c_a^{(1)}\right) + \pi_{\{1,2\}}\mathbb{P}_{\boldsymbol{0}}\left(\tilde{Z}_1^{(1)} \geq c_a^{(1)} \vee \tilde{Z}_{\{1,2\}}^{(2)} \geq c_a^{(2)}\right),$$

letting (6.11) become solvable for some $c_b^{(2)}$*.*

The example showed us that our PWER-adjusted CRP-principle is not applicable for certain mid-trial adaptations because (6.11) cannot be satisfied for all parameter constellations. We therefore replace $cp_{a,\boldsymbol{\theta}}$ by a new type of conditional error function $cp^*_{a,\boldsymbol{\theta}}$ that is alwayss larger than $cp_{a,\boldsymbol{\theta}}$ and also allows us to solve the inequality system (6.11) for only a finite number of parameter constellations, while still ensuring that the overall type I error rate is bounded by α, regardless of whether a design change is made or not. The finite set of parameter constellations the system (6.11) needs to be solved for is

$$\boldsymbol{\Theta}_{C,0} := \{\boldsymbol{\theta} = (\theta^U)_{U\in\mathfrak{U}_C} \in \boldsymbol{\Theta}_C \mid \forall U \in \mathfrak{U}_C : \theta^U > 0 \vee \theta^U = 0\}. \tag{6.12}$$

The reasoning of this set will be explained later in more detail. The precise requirements for $cp^*_{a,\boldsymbol{\theta}}$ are now as follows.

Requirement 1. *Let $\mathcal{D}_a$ and $\mathcal{D}_b$ be initial and new design with test decision functions $\boldsymbol{\varphi}_a$ and $\boldsymbol{\varphi}_b$ that control the PWER at a level α, respectively, and let $\boldsymbol{z}_{int} = \left(z_{int}^{U,(1)}\right)_{U\in\mathfrak{U}_C}$ be a fixed but arbitrary vector of stage 1 interim data in the form of z-scores for each H^U, $U \in \mathfrak{U}_C$. The function $cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ replacing the conditional PWER of the initial design $cp_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ with must fulfill the following four conditions:*

(i) $\mathbb{E}_{\boldsymbol{\theta}}(cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{Z}_{int})) \leq \alpha$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{C,0}$

(ii) $cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ *only depends on those components of* $\boldsymbol{z}_{int}$ *where the null hypothesis holds, i.e. only on the sub-vector* $(\boldsymbol{z}_{int})_{I_J(\boldsymbol{\theta})}$.

(iii) $cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int})$ *is monotonically increasing in* $\boldsymbol{z}_{int}$ *in all components* $z_{int}^{U,(1)}$ *with* $\theta^U \leq 0$.

(iv) $cp_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) \leq cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_C, \boldsymbol{z}_{int} \in \mathcal{Z}$

Using $cp^*_{a,\boldsymbol{\theta}}$ we reformulate condition (6.11) to

$$cp_{b,\boldsymbol{\theta}}(\mathbf{z}_{int}) \leq cp^*_{a,\boldsymbol{\theta}}(\mathbf{z}_{int}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_C, \mathbf{z}_{int} \in \mathcal{Z} \tag{6.13}$$

with $\mathbf{z}_{int}$ being a fixed interim data vector again. Note that $cp_{b,\boldsymbol{\theta}}$ as it is defined by (6.8) also satisfies conditions (ii) and (iii).
The proof of the following theorem now shows that there is a finite set of parameter constellations $\boldsymbol{\theta}$ that is enough to satisfy condition (6.13) if $cp^*_{a,\boldsymbol{\theta}}$ fulfills Requirement 1. This ultimately leads to overall PWER-control at level α.

Theorem 6.1.2. *Assume there exists a function $cp^*_{a,\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ satisfying conditions (i) to (iv) of Requirement 1. Then there exists a finite set $\boldsymbol{\Theta}_{rel} \subset \boldsymbol{\Theta}_{C,0}$ for which the inequalities*

$$cp_{b,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) \leq cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_{rel} \tag{6.14}$$

imply overall PWER-control of the new design $\boldsymbol{\varphi}_b$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$.

Proof. We define the partition $\left\{\boldsymbol{\Theta}_{C,0}^{\tilde{\mathfrak{U}}}\right\}_{\tilde{\mathfrak{U}}\subseteq\mathfrak{U}_C}$ of $\boldsymbol{\Theta}_{C,0}$ where

$$\boldsymbol{\Theta}_{C,0}^{\tilde{\mathfrak{U}}} := \{\boldsymbol{\theta} = (\theta^U)_{U\in\mathfrak{U}_C} \in \boldsymbol{\Theta}_C \mid \forall U \in \tilde{\mathfrak{U}} : \theta^U = 0 \wedge \forall U \in \mathfrak{U}_C \setminus \tilde{\mathfrak{U}} : \theta^U > 0\}.$$

For each $\tilde{\mathfrak{U}} \subseteq \mathfrak{U}_C$ now select a fixed representative $\tilde{\boldsymbol{\theta}}^{\tilde{\mathfrak{U}}} \in \boldsymbol{\Theta}_{C,0}^{\tilde{\mathfrak{U}}}$ and let

$$\boldsymbol{\Theta}_{rel} := \left\{ \tilde{\boldsymbol{\theta}}^{\tilde{\mathfrak{U}}} \mid \tilde{\mathfrak{U}} \subseteq \mathfrak{U}_C \right\}. \tag{6.15}$$

This set is finite because there are finitely many subsets $\tilde{\mathfrak{U}}$ of $\mathfrak{U}_C$. For each $\boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}_C} \in \boldsymbol{\Theta}_C$ we can now find a representative $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{rel}$ with components $\tilde{\theta}^U = 0$ for all $U \in \mathfrak{U}_C$ where $\theta^U \leq 0$ and $\tilde{\theta}^U > 0$ where $\theta^U > 0$. Using the monotonicity condition (iii) for $cp_{b,\boldsymbol{\theta}}$ and condition (i) for $cp^*_{a,\boldsymbol{\theta}}$ we then have

$$\mathrm{PWER}_{\boldsymbol{\theta}}(\boldsymbol{\varphi}_b) = \mathbb{E}_{\boldsymbol{\theta}}(cp_{b,\boldsymbol{\theta}}(\mathbf{Z}_{int})) \leq \mathbb{E}_{\tilde{\boldsymbol{\theta}}}(cp_{b,\tilde{\boldsymbol{\theta}}}(\mathbf{Z}_{int})) \leq \mathbb{E}_{\tilde{\boldsymbol{\theta}}}(cp^*_{a,\tilde{\boldsymbol{\theta}}}(\mathbf{Z}_{int})) \leq \alpha$$

where first equality holds due to the law of total expectation, the second inequality because of condition (iii) also holding for $cp_{b,\boldsymbol{\theta}}$, the third inequality because of (6.14) and the last inequality because of (i). If no mid-trial adjustment is made, $cp_{b,\boldsymbol{\theta}}(\mathbf{z}_i nt) = cp_{a,\boldsymbol{\theta}}(\mathbf{z}_i nt)$, and then due to condition (iv) still we have $cp_{b,\boldsymbol{\theta}}(\mathbf{z}_{int}) = cp_{a,\boldsymbol{\theta}}(\mathbf{z}_{int}) \leq cp^*_{a,\boldsymbol{\theta}}(\mathbf{z}_{int})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ and all interim sample points $\mathbf{z}_{int}$. The same arguments as above can be used to show $\mathrm{PWER}_{\boldsymbol{\theta}}(\boldsymbol{\varphi}_b) \leq \alpha$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ with and without adaptations. □

With the following definition we show that such a function $cp^*_{a,\boldsymbol{\theta}}$ exists.

Definition 6.1.2 (modified conditional PWER)**.** *Let $\boldsymbol{\varphi}_a = \left(\varphi^{U,(k)}\right)_{\substack{U \in \mathfrak{U}_C \\ k=1,2}}$ be the initial α-level 2-stage GSD for a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}}, \mathcal{H})$. For each $J \in \mathcal{C}_{\mathcal{P}}$ we define the conditional error function $A_J^{*\boldsymbol{\theta}}$ as 1 if $\sum_{U \in I_J(\boldsymbol{\theta})} \varphi^{U,(1)} \geq 1$ and otherwise as*

$$A_J^{*\boldsymbol{\theta}}(\boldsymbol{z}_{int}) := \mathbb{P}_{\boldsymbol{0}}\left(\bigcup_{U \in I_J^{(2)}(\boldsymbol{\theta})} \{\varphi^{U,(2)} = 1\} \cup \bigcup_{U \in I_{1,J}(\boldsymbol{\theta})} \bigcup_{k \in I_2^U} \{\varphi^{U,(k)} = 1\} \mid (\boldsymbol{z}_{int})_{I_J(\boldsymbol{\theta})} \right) \tag{6.16}$$

The modified conditional PWER *is then defined as*

$$cp^*_{a,\boldsymbol{\theta}}(\boldsymbol{z}_{int}) := \sum_{J \in \mathcal{C}_{\mathcal{P}}} \pi_J A_J^{\boldsymbol{\theta}}(\boldsymbol{z}_{int}). \tag{6.17}$$

If any type I error concerning $\mathcal{P}_J$ has already been made at stage 1 (before the adaptation), then the conditional error function A_J^* is equal to 1. Otherwise it equals the probability of the union of two sets. The first set is the event that any true hypothesis H^U at stage 2 is rejected given the stage 1 data $z_{int}^{U,(1)}$. The modification is now incorporated in the second set and describes the rejection of any *false* H^U at any stage a test for H^U has been planned for. This second event is evaluated under the global null hypothesis, indicated by the probability measure $\mathbb{P}_{\boldsymbol{0}}$, even though these θ^U are in the alternative. This modified $cp^*_{a,\boldsymbol{\theta}}$ fulfills all conditions (i) to (iv) of Requirement 1: (i) is satisfied because the expectation of $cp^*_{a,\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{C,0}$ is bounded by the PWER under the global null which is for all $\boldsymbol{\theta}$ at most α. (ii) holds because A_J^* is equal to a probability conditioned on $(\mathbf{Z}_{int})_{I_J(\boldsymbol{\theta})} = (\mathbf{z}_{int})_{I_J(\boldsymbol{\theta})}$. Condition (iii) is fulfilled because of the multivariate normal assumption we made

on the test statistics and (iv) holds because A_J^* additionally covers rejection regions concerning those $U \in \mathfrak{U}_C$ where $\theta^U > 0$.

One problem that needs to be addressed is the set $\boldsymbol{\Theta}_{rel}$ and how to obtain it in practice. This inevitably brings us to the set $\boldsymbol{\Theta}_{C,0}$ the set $\boldsymbol{\Theta}_{rel}$ is defined over. The attentive reader might have noticed that $\boldsymbol{\Theta}_{C,0}$, which contains $\boldsymbol{\theta}^U$ for which either $\theta^U > 0$ or $\theta^U = 0$, is not fully compatible with our mathematical framework. This is because in our case it is $\theta^U = \sum_{J \in U} \pi_J^U \theta_J$ which implies that some parameter constellations in $\boldsymbol{\Theta}_{C,0}$ are impossible. Think of Example 6.1.2 with $\theta_1 > 0$, $\theta_2 = 0$, $\theta_{\{1\}} = 0$, $\theta_{\{2\}} < 0$ and $\theta_{\{1,2\}} > 0$. This $\boldsymbol{\theta}$ is not an element of $\boldsymbol{\Theta}_{C,0}$ because of $\theta_{\{2\}} < 0$. Since there are infinitely many possibilities for $\theta_{\{2\}}$ being strictly negative and $\theta_{\{1,2\}}$ being strictly positive (such that $\theta_2 = 0$), we can consider the limit of the parameter constellation $\boldsymbol{\theta}$ with $\theta_{\{2\}} \uparrow 0$ and $\theta_{\{1,2\}} \downarrow 0$ such that in the limit we get $\theta_1 > 0$, $\theta_2 = 0$, $\theta_{\{1\}} = 0$, $\theta_{\{2\}} = 0$ and $\theta_{\{1,2\}} > 0$. The resulting $\boldsymbol{\theta}$ lies in $\boldsymbol{\Theta}_{C,0}$. To find $\boldsymbol{\Theta}_{rel}$ we have to find all possible $\boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}_C}$ with components $\theta^U \leq 0$ or $\theta^U > 0$ and then consider the limit parameter constellations. In general, the number element in $\boldsymbol{\Theta}_{rel}$ is equal to the number of possible null constellations $I(\boldsymbol{\theta})$.

At the end of this section we want to briefly state the problem of new populations being introduced with the mid-trial adaptation. If an adaptation introduces a new population to $\mathcal{C}_{\mathcal{P}_a}$ yielding a $\mathcal{C}_{\mathcal{P}_b}$ with a more complex structure one would first have to recalculate all $\boldsymbol{\pi}_a$ to some $\boldsymbol{\pi}_b$ since a J in $\mathcal{C}_{\mathcal{P}_a}$ does not yield the same π_J and θ_J than a J in $\mathcal{C}_{\mathcal{P}_b}$. For instance, the intersection $\mathcal{P}_{\{1,2\}}$ in a population structure with two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$ is not the same as in a population structure with three intersecting populations $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$ where $\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3 \neq \emptyset$. Thus, one would need to redefine $\mathfrak{U}_a$ to a $\mathfrak{U}_{a,new}$ describing the sub-populations from $\mathcal{P}_a$ in terms of the new population structure $\mathcal{P}_b$. Then one could, again, consider $\mathfrak{U}_C := \mathfrak{U}_{a,new} \cup \mathfrak{U}_b \cup \mathcal{C}_{\mathcal{P}_b}$ and the resulting combined parameter space $\boldsymbol{\Theta}_C$ containing all $\boldsymbol{\theta} = (\theta^U)_{U \in \mathfrak{U}_C}$. One practical issue is the concrete recalculation of $\boldsymbol{\pi}_b$ as one would at least need to know the proportion of the new $\pi_{b,J}$ in $\mathcal{P}_J$. For instance if we initially have two intersecting populations $\mathcal{P}_1$ and $\mathcal{P}_2$ and now a third one $\mathcal{P}_3$ is introduced such that $\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{P}_3 \neq \emptyset$, then one would need to recalculate $\pi_{b,\{1,2\}}$. Adaptive designs that add new populations are a topic of future research.

6.2 Numerical examples

Before testing our method in a simulation study, we want to apply it to the setting from Example 6.1.2 and compute new critical values for some values $\mathbf{z}_{int}$ to get a better idea of the procedure.
Assume that we start a trial where we intend to test the efficacy of one single treatment T against a control C in two overlapping populations $\mathcal{P}_1$ and $\mathcal{P}_2$ by means of testing $\mathcal{H}_a = \{H_1, H_2\}$ with $H_j = \theta_j = \theta_(\mathcal{P}_j, T) \leq 0$ following Design I. So we have a two-stage test $\boldsymbol{\varphi}$ for the population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}_a}, \mathcal{H}_a)$ with $\mathcal{C}_{\mathcal{P}} = \{\{1\}, \{2\}, \{1,2\}\}$, $\boldsymbol{\pi} = (\pi_J)_{J \in \mathcal{C}_{\mathcal{P}}}$ and $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$. Now at interim we found out that we can reject H_1 and retain H_2. Say, that our focus shifts from testing in H_2, since me might not see any hope with our observed z-score for $\mathcal{P}_2$ and thus want to further investigate the treatment effects in $\mathcal{P}_{\{1\}}$ and $\mathcal{P}_{\{1,2\}}$. So we shift to a design $\boldsymbol{\varphi}_b$ for a population-wise testing problem $(\mathcal{C}_{\mathcal{P}}, \boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\Theta}_b}, \mathcal{H}_b)$ with

$$\boldsymbol{\Theta}_b = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{\{1,2\}}) \in \mathbb{R}^3 \mid \theta_j = \frac{\pi_{\{j\}}}{\pi_j} \theta_{\{j\}} + \frac{\pi_{\{1,2\}}}{\pi_j} \theta_{\{1,2\}}, j = 1, 2 \right\}$$

and $\mathcal{H}_b = \{H_1, H_2, H_{\{1\}}, H_{\{1,2\}}\}$ with H_1, H_2 being tested at stage 1 as in the initial design and $H_J : \theta_J \leq 0$, $J \in \{\{1\}, \{1,2\}\}$, at stage 2. So, we need to consider the combined parameter set $\boldsymbol{\Theta}_C$ given by

$$\boldsymbol{\Theta}_C = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{\{1\}}, \theta_{\{2\}}, \theta_{\{1,2\}}) \in \mathbb{R}^5 \mid \theta_j = \frac{\pi_{\{j\}}}{\pi_j}\theta_{\{j\}} + \frac{\pi_{\{1,2\}}}{\pi_j}\theta_{\{1,2\}}, j = 1,2 \right\}. \tag{6.18}$$

From this set, we need to find out all relevant constellations $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{rel}$ where any sub-vector $\boldsymbol{\theta}_{\tilde{\mathfrak{U}}} = \mathbf{0}$ for $\tilde{\mathfrak{U}} \subseteq \mathfrak{U}_C = \{1, 2, \{1\}, \{2\}, \{1,2\}\}$ and the remaining components are either > 0. These are given in the first 5 columns of Table 6.1. In R, the function `ThetaC` in the script-file `ADscript` will return these constellations for a given set of populations investigated under the old and new design, respectively. Given some z-scores $\mathbf{z}_{int} = ((z_j^{(1)})_{j=1}^2, (z_J^{(1)})_{J \in \mathcal{C}_\mathcal{P}})$ containing the respective interim data from $\mathcal{P}_j$, $j = 1, 2$, and $\mathcal{P}_J$, $J \in \mathcal{C}_\mathcal{P}$, obtained from the formulas in Section 4.3.1 one can then determine the conditional PWER-expression of the old and new design, respectively, under each constellation $\boldsymbol{\theta}$. For example, the constellation $\boldsymbol{\theta}$ with $\theta_{\{1\}}, \theta_{\{2\}}, \theta_1 > 0$, $\theta_{\{1,2\}} < 0$ and $\theta_2 = 0$ yields the old (modified) conditional PWER

$$\begin{aligned}
cp^*_{a,\boldsymbol{\theta}}(\mathbf{z}_{int}) &= \pi_{\{1\}}\mathbb{P}_0\left(Z_1^{(1)} \geq c_a^{(1)}\right) + \pi_{\{12\}}\mathbb{P}_{\boldsymbol{\theta}}\left(Z_1^{(1)} \geq c_a^{(1)} \vee Z_{\{12\}}^{(2)} \geq c_a^{(2)} | z_{\{12\}}^{(1)}\right) \\
&= \pi_{\{1\}}\Phi\left(-c_a^{(1)}\right) \\
&\quad + \pi_{\{1,2\}}\mathbb{P}_{\boldsymbol{\theta}}\left(\{Z_1^{(1)} \geq c_a^{(1)}\} \cup \left\{\tilde{Z}_{\{1,2\}}^{(2)} \geq \frac{c_a^{(2)} - w_{\{1,2\}}^{(1)} z_{\{1,2\}}^{(1)}}{w_{\{1,2\}}^{(2)}}\right\}\right) \\
&\overset{\theta_{\{1,2\}}\uparrow 0}{\to} \pi_{\{1\}}\Phi\left(-c_a^{(1)}\right) + \pi_{\{1,2\}}\left(1 - \Phi_{\boldsymbol{\Sigma}_{1,\{1,2\}}}\left(c_a^{(1)}, \frac{c_a^{(2)} - w_{\{1,2\}}^{(1)} z_{\{1,2\}}^{(1)}}{w_{\{1,2\}}^{(2)}}\right)\right)
\end{aligned}$$

where the stage-wise test statistic $\tilde{Z}_{\{1,2\}}^{(2)}$ has mean $\theta_{\{1,2\}}\sqrt{\tilde{n}_{\{1,2\}}^{(2)}}/\sigma_{\{1,2\}} < 0$. The first probability statement is the unconditional the probability of making a type I error relevant for $\mathcal{P}_{\{1\}}$ because of $\theta_1 > 0$ and the definition of the modified PWER. The second probability concerning $\mathcal{P}_{\{1,2\}}$ thus consists of the unconditional event that the test statistic $\tilde{Z}_1^{(1)}$ exceeds $c^{(1)}$ or that $H_{\{1,2\}}$ is rejected at stage 2 given the stage 1 z-score $z_{\{1,2\}}^{(1)}$. The correlation matrix $\boldsymbol{\Sigma}_{1,\{1,2\}}$ for the sub-vector $\left(Z_1^{(1)}, \tilde{Z}_{\{1,2\}}^{(2)}\right)$ is needed to find this probability, but it can easily be obtained using the formulas in Section 4.3.1 (in this special case it is the unit matrix due to stage 1 and 2 observations being independent from each other). Since H_2 has been retained at stage 1 and no hypothesis whose false rejection could be relevant for $\mathcal{P}_{\{2\}}$ is planned to be tested, the respective probability for $\mathcal{P}_{\{2\}}$ is zero. In the last line, we let $\theta_{\{1,2\}} \uparrow 0$ (and $\theta_1, \theta_{\{1\}}, \theta_{\{2\}} \downarrow 0$) to obtain the limit conditional PWER. For the new design, we intend to test $H_{\{1\}}$ and $H_{\{1,2\}}$ using a new critical value $c_b^{(2)} \neq c_a^{(2)}$. Since $\theta_{\{1\}} > 0$ we cannot make a type I error for $\mathcal{P}_{\{1\}}$ implying that the limit conditional PWER of the new design is given by

$$\pi_{\{1,2\}}\mathbb{P}_0\left(Z_{\{1,2\}}^{(2)} \geq c_b^{(2)} | z_{\{1,2\}}^{(1)}\right) = \pi_{\{1,2\}}\Phi\left(-\frac{c_b^{(2)} - w_{b,\{1,2\}}^{(1)} z_{\{1,2\}}^{(1)}}{w_{b,\{1,2\}}^{(2)}}\right).$$

The actual values of $cp_{max}(c_b^{(2)} \mid \mathbf{z}_{int})$ and $\alpha_C(\mathbf{z}_{int})$ are, of course, dependent on the concrete value of $\mathbf{z}_{int}$ and thus no real general formula can be given here.

Numerical example 1: As a numerical example, assume we started Design II with with a Pocock design such that a PWP of 90% is found under some alternative $\boldsymbol{\delta}_A = (\delta_{A,1}, \delta_{A,2}, \delta_{A,\{1,2\}}) = (0.3, 0.3, 0.3)$, where $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, i.e. we have critical values $\mathbf{c} = (2.082, 2.082)$ and sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	80	80	40
$\tilde{n}_J^{(2)}$	0	0	40

as found in Section 5.1.2. We now change to a design where $H_{\{1\}}$ is tested as well, which we do by recruiting further 80 patients and thus want to change to a design with sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	80	80	40
$\tilde{n}_J^{(2)}$	80	0	40

Say, we have interim values

$$\mathbf{z}_{int} = \left(z_{\{1\}}^{(1)}, z_{\{2\}}^{(1)}, z_{\{1,2\}}^{(1)}, z_1^{(1)}, z_2^{(1)}\right) = (2, -0.7, 1.5, 2.5, 0.3)$$

which indicate that H_1 can be rejected ($z_1^{(1)} \geq 2.082$) and that there is a good chance that the treatment is also beneficial for patients in $\mathcal{P}_{\{1\}}$ and $\mathcal{P}_{\{1,2\}}$ (because $z_J^{(1)}, J \in \{\{1\}, \{1,2\}\}$ are reasonably large). The function `critAD` in the script file `ADscript` now finds a suitable critical value $c_b^{(2)}$ such that (R2) is fulfilled (see Appendix B for a more detailed computation in R). It returns $\alpha_C(\mathbf{z}_{int}) \approx 0.0145$ and $c_b^{(2)} = 2.728$. The value of $\alpha_C(\mathbf{z}_{int})$ is fairly conservative (a bit larger than $\alpha/2$) indicating that our method can probably be improved in some way. Nonetheless, this new critical value can now be used to conduct a test for $H_{\{1\}}$ and $H_{\{1,2\}}$ at stage 2 without inflating the overall population-wise type I error.

Numerical example 2: Let us assume that we have started our two-stage trial with Design I and, as before, rejection of $H_1 : \theta_1 \leq 0$ and acceptance of $H_2 : \theta_2 \leq 0$ have happened at interim. Now, say, we are not interested in further testing of H_2 anymore, but want to further investigate the effect $\theta_{\{1,2\}}$ in the intersection. So we want to switch from Design I to II at stage 2. As before, the combined parameter space is simply given by the set in (6.18). The resulting inequations are given in Table 6.2. Say, Design I was started as an equally spaced OBF-design such that a PWP of 90% was guaranteed under $\boldsymbol{\delta}_A = (0.3, 0.3)$ and $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, i.e. we have critical values $\mathbf{c} = (2.892,\ 2.045)$ and sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}_J^{(1)}$	43	43	22
$\tilde{n}_J^{(2)}$	43	43	22

as found in Section 5.1.1. Instead of only using 22 observations for testing in the intersection we want to enrich this intersection by using 43+43+22 = 108 observations. This would of course mean that the weight $w^{(1)}_{b,\{1,2\}}$ equals $\sqrt{22/(108+22)} = 0.4114$ now instead of $\sqrt{1/2}$ which it would have been equal to without the enrichment. So we switch to a design with sample sizes

	$\mathcal{P}_{\{1\}}$	$\mathcal{P}_{\{2\}}$	$\mathcal{P}_{\{1,2\}}$
$\tilde{n}^{(1)}_J$	43	43	22
$\tilde{n}^{(2)}_J$	0	0	108

Say, we have interim values

$$\mathbf{z}_{int} = \left(z^{(1)}_{\{1\}}, z^{(1)}_{\{2\}}, z^{(1)}_{\{1,2\}}, z^{(1)}_1, z^{(1)}_2\right) = (1.5, -0.7, 2, 2.38, 0.3)$$

indicating that a further investigation of the treatments efficacy in the intersection could be promising. Using `critAD` again we find an $\alpha_C(\mathbf{z}_{int}) \approx 0.025$ (so basically the full PWER of the initial design) and a critical value of $c^{(1)}_b = 1.871$ which more liberal than the 0.025-quantile of the standard normal distribution. The enrichment of the intersection facilitated rejection at stage 2 and controlling the minimal limit conditional PWER in this case.

6.3 Simulation study

At last, we want to test our PWER-adjusted CRP-principle in a simulation study. We want to simulate the situation where a two-stage trial initially starts with Design I. That is, a design for which it is planned to investigate one treatment in two overlapping populations $\mathcal{P}_1$ and $\mathcal{P}_2$ by means of testing the hypotheses $H_j : \theta_j = \theta(\mathcal{P}_j, T) \leq 0$ for $j = 1, 2$ at stages 1 and 2. In the previous section we already gave a short numerical example for the case of switching from Design I to II while shifting the whole sample size planned for stage 2 into the intersection. Of course, one can now think of other designs one might want to switch to. Depending on the circumstance one might not do any mid-trial change and conduct the initial group sequential plan. In other cases, new hypotheses for populations we have not planned for initially may be tested. In the following we want to investigate different design changes that can be made at stage 2 by using our above introduced adaptive design concept.

General setting: We say that we conduct a design change whenever either H_1 or H_2 (exclusive or) are tested. If any H_j is rejected, then we would usually not intend to test further in $\mathcal{P}_j$ (e.g. to save resources). If both H_1 and H_2 are retained at stage 1, we deem a continuation of the trial as futile as we do not expect a high chance of observing any meaningful effects in any of the disjoint sub-populations. If both H_1 and H_2 are rejected, we stop the trial for efficacy. This rule holds for all adaptive designs and the initial group sequential design we will investigate. Proceeding like this could describe a trial situation in which we are mainly interested in the effects in the larger populations $\mathcal{P}_1$ and $\mathcal{P}_2$ and if a contradictory result occurs (like rejecting H_i but not H_j, $i \neq j$) then we are willing to spend more resources to further

investigate the population we have not yet established a meaningful treatment effect for. Let

$$\mathbf{z}_{int} = \left(z_{\{1\}}^{(1)}, z_{\{2\}}^{(1)}, z_{\{1,2\}}^{(1)}, z_1^{(1)}, z_2^{(1)}\right)$$

be the vector of stage 1 z-scores for each sub-population $\mathcal{P}_J$, $J \in \mathcal{C}_\mathcal{P}$, and $\mathcal{P}_j$, $j = 1, 2$. Assume that H_j has been accepted and H_i has been rejected at stage 1, $i \neq j$, and let $U_j := \{\{j\}, \{1, 2\}\}$:

(AD1) For $J \in U_j$ we consider testing all H_J where $z_J^{(1)} \geq q$ with $J \in U_j$ and q being some prespecified constant indicating that the effect in $\mathcal{P}_J$ is 'promising enough'. If both $z_J^{(1)} \geq q$, $J \in U_j$, however, we simply continue the group sequential design, so H_j is retested at stage 2 with the previously planned sample size.

(AD2) Shift to Design II (test only $H_{\{1,2\}}$) as in the second example in Section 6.2.

(AD3) Test all H_J with $J \in U_j$.

(AD4) Test all H_J with $J \in U_j$ and H_j.

(GSD) No trial mid-trial adaptation is forseen and the GSD Design I is conducted as planned (for comparison)

(S) A single stage design, where hypotheses $H_J : \theta_J \leq 0$ and $H_j : \theta_j \leq 0$ for $J \in \mathcal{C}_\mathcal{P}$ and $j = 1, 2$ are simultaneously tested.

In each of the above adaptive designs AD1 to AD4 the total stage 2 sample size planned for the initial design is reshuffled to the union of the disjoint sub-populations $\mathcal{P}_J$ according to their relative prevalences. So if the total sample size at stage 2 planned for Design I was equal to $N^{(2)}$, then AD1 would use $\pi_J/\pi_j N^{(2)}$ observations to test H_J for $J \in U_j$ if both $z_J^{(1)} \geq q$, and $N^{(2)}$ for $H_{J'}$ if only $z_{J'}^{(1)} \geq q$. AD2 would use $N^{(2)}$ patients to test $H_{\{1,2\}}$, AD3 would use $\pi_J/\pi_j N^{(2)}$ for both H_J, $J \in U_j$ and AD4 would use the same number of observations as AD3 does for the disjoint sub-populations and would use $N^{(2)}$ for testing H_j.

Performance measures: To measure the performance of each design we simulate the power for the tests of each hypothesis $H_1, H_2, H_{\{1\}}, H_{\{2\}}, H_{\{1,2\}}$ individually by counting the number of times each hypothesis is rejected by the respective design divided by the total number of simulation runs M_{sim}. Since some designs do not continue to test H_j after it has not been rejected at stage 1, it is reasonable to also include the more informative rejection of both $H_{\{j\}}$ and $H_{\{1,2\}}$. So we also count the number of times that H_j is rejected *or* both $H_{\{j\}}$ and $H_{\{1,2\}}$ is rejected and denote this 'hypothesis' as H_j^*. This quantity, of course, only makes sense for AD3 and AD4. Due to the fact that the H_J's will not be tested in each simulation run, we also want to consider for each hypothesis the number of times it has been rejected given the fact that it has actually been tested. So we consider both $\#(H \text{ rejected})/N_{sim}$ and $\#(H \text{ rejected})/\#(H \text{ tested})$ for every $H \in \{H_1, H_2, H_{\{1\}}, H_{\{2\}}, H_{\{1,2\}}, H_1^*, H_2^*\}$. We call the first one *unconditional power* and the second one *conditional power*. For H_1^* and H_2^*, we make a special consideration when it comes to the conditional power: among all cases where both the informative rejection can happen (so when

$H_{\{1,2\}}$ and some $H_{\{j\}}$ are both tested at stage 2), we count all cases where either H_j or the informative rejection happened. In case of AD4 this means that we count $\#(H_{\{j\}} \wedge H_{\{1,2\}} \vee H_j \text{ rejected})/\#(H_{\{j\}} \wedge H_{\{1,2\}} \text{tested})$. In case of AD3 it then is $\#(H_{\{j\}} \wedge H_{\{1,2\}} \text{ rejected})/\#(H_{\{j\}} \wedge H_{\{1,2\}} \text{ tested})$ because no testing of H_j is done at stage 2 and including the test of H_j at stage 1 is non-sensical because H_j is tested in *every* simulation run anyway.
Lastly, we measure the expected sample size of each design by recording the total sample size used for each simulation run and taking the average over all these values.

Simulation setup: We conducted $M_{sim} = 50.000$ simulation runs and assumed various effects $\boldsymbol{\delta}_A = (\delta_{A,\{1\}}, \delta_{A,\{2\}}, \delta_{A,\{1,2\}}, \delta_{A,1}, \delta_{A,2})$ where $\delta_{A,J}$, $J \subseteq \{1,2\}$ are all chosen such that the mean of $\delta_{A,1}$ and $\delta_{A,2}$ is equal to 0.3. For example $\delta_{A,1} = 0.4$ and $\delta_{A,2} = 0.2$ could be chosen then to display some heterogeneity between the treatment effects in $\mathcal{P}_1$ and $\mathcal{P}_2$. We examined the constellation of $\boldsymbol{\pi} = (\pi_{\{1\}}, \pi_{\{2\}}, \pi_{\{1,2\}}) = (0.4, 0.4, 0.2)$. A stage 1 overall N of around 117 was chosen for all scenarios which basically equals the number of observations a single stage design would need to achieve a power of 90% in the overall population if the effect size there is equal to 0.3. We did this to make the results between the different scenarios a bit more comparable. The significance level was chosen to be $\alpha = 0.025$ one-sided and Pocock and O'Brien & Fleming designs were used, respectively, for the initial Design I. For AD1 we set $q = \Phi^{-1}(0.9) \approx 1.28$, so a value larger than the 90% quantile of the standard normal distribution is seen as promising.

Results: The rejection probabilities (in %) for each hypothesis and the expected sample sizes are summarized in Tables 6.3 and 6.4, where Pocock designs were used as the initial design. In Appendix B, Tables B.1 to B.2 show the same respective results just with an O'Brien & Fleming design as the initial design.
First, we describe a couple of obvious characteristics that are displayed by all the tables solely due to the setup of the simulation itself. The rejection probabilities for H_1 and H_2 of the initial GSD (first line) is always larger than those of the four adaptive designs because the GSD, by design, always tests an at stage 1 retained hypothesis again, whereas only AD1 and AD4 have the possibility of retesting a retained H_j incorporated. The probabilities in the first two columns do also not differ with regard to whether the conditional or unconditional power was measured because H_1 and H_2 are tested in every simulation run. Also, since AD2 only tests in the intersection at stage 2, the probabilities for $H_{\{1\}}$ and $H_{\{2\}}$ are zero. With respect to H_1^* and H_2^*, the rejection probabilities are only really relevant for AD3 and AD4 since these are the only designs that have the possibility of testing in both disjoint subgroups of a $\mathcal{P}_j$ incorporated. For the other designs, the values are simply the same as the respective probabilities for H_1 and H_2. Lastly, with respect to the expected sample size AD1 is expected to have the lowest value because its design has the most possibilities to stop the trial early because of the incorporation of the value q. For the single stage design, the expected sample size is always equal to $2N^{(1)} \approx 234$. Looking at the unconditional probabilities, one can generally see that the rejection probabilities are in most cases lower than in the conditional tables which makes sense because not every hypothesis $H \in \{H_{\{1\}}, H_{\{2\}}, H_{\{1,2\}}, H_1, H_2\}$ is tested in every simulation run due to the specific decision rules we defined for each design. Between Pocock and O'Brien & Fleming initial design choices one can also see that the rejection probabilities for H_1 and H_2 are a bit lower in the OBF-case

due to the larger stage 1 critical value, but especially for AD4 the conditional probability of rejecting H_j^* can get increased as seen in Table B.2, for instance. Generally, the expected sample size for AD2-AD4 is very slightly lower than under the Pocock initial design, whereas for AD1 it is slightly higher.
Let us now discuss the properties of individual designs in more detail. The GSD power values are quite high, although one has to think of the fact that a value of 117 leads to a slightly overpowered study as values of around 100 are already enough to guarantee a power of 90% (see Section 5.1.1). The expected sample size is a bit larger than that of AD1 because it does not have early stopping incorporated. Also the expected sample size of the GSD can be larger than that of the other designs as seen in Tables 6.3 and B.1 because sample size from the whole stage 2 is not reshuffled into the population $\mathcal{P}_j$ for which still has to be tested. Admittedly, for the adaptive designs one would have to screen more patients in order to conduct this reshuffling. Regarding the rejection probabilities of H_1 and H_2 the adaptive designs suffer a power loss which can go up to around 20-30% but if we count in the informative rejection of both disjoint subgroups (i.e. if we look at H_1^* and H_2^*), we see that especially AD4 gets overall the closest to the power values of the GSD in basically all scenarios. This is because AD4 tests both H_j and $H_{\{j\}}$ and $H_{\{1,2\}}$ if H_j could not be rejected at stage 1 yet. With respect to the rejection probabilities of each H_J, $J \subseteq \{1,2\}$, we can generally say that *if* the rejection of an H_i and retention of H_j, $i \neq j$, happens at stage 1, that the probability of rejecting the H_Js can generally be pretty high, as indicated by the conditional probabilities in, for example, Table 6.3. Table 6.3 in general describes a case where the adaptive design is actually not what one would go for in practice, because all effects are identical (to 0.3) in each population. Going for an adaptive design, we see that one loses from around 15% (AD4) up to 30% of power (AD2 and AD3) for testing H_1 and H_2. Still, power values for testing any of the H_J are still considerably high, even though, looking at the undconditional power values, tests for these hypotheses do not happen often. Considering the case of heterogeneous effects $\boldsymbol{\delta}_A = (\delta_{\{1\}}, \delta_{\{2\}}, \delta_{\{1,2\}}, \delta_1, \delta_2) = (0.4, -0.2, 0.7, 0.5, 0.1)$ as displayed in Table 6.4 we see large power values for H_1 and H_1^*, as expected due to the effect size of $\delta_1 = 0.5$ and very low values for H_2 and H_2^*, due to $\delta_2 = 0.1$, so the adaptive designs behave correctly. In this case the conditional and unconditional power values for $H_{\{1,2\}}$ are very large because of the large effect in the intersection, so the adaptive design tests the correct null hypotheses at stage 2. Even though $\delta_{\{1\}}$ is quite large as well the fact that rejection of H_1 almost always happens leads to power values for $H_{\{1\}}$ being quite low.

Discussion: Generally, one can say that the probability of rejecting either a H_j or doing the more informative rejection of $H_{\{j\}}$ and $H_{\{1,2\}}$ is in most cases strongly increased in the conditional case and only slightly to moderately increased in the unconditional case. With a more sophisticated approach to the rules describing when to enrich a disjoint subgroup $\mathcal{P}_J$, $J \subseteq \{1,2\}$, one could probably also create cases where the informative rejection will happen more often overall. Also, keep in mind that the establishment of a relevant treatment effect in a sub-population in general is much more valuable than in the initially planned populations $\mathcal{P}_1$ and $\mathcal{P}_2$ and that better methods to quantify/assess this gain need to be imposed to really make a fair assessment to the utility of this adaptive design method and the here used adaptive design options. A utility function approach that somehow values the rejection in a subgroup higher than in a superset could be a reasonable approach. The primary

aim of this section was more so to show that the PWER-adjusted CRP-principle can be used to a wide range of pre-planned options for mid-trial adaptations.

$(\theta_{\{1\}},\theta_{\{2\}},\theta_{\{1,2\}},\theta_1,\theta_2)$	$cp_{b,\boldsymbol{\theta}}$	$cp^*_{a,\boldsymbol{\theta}}$
(0,0,0,0,0)	$p^{\boldsymbol{\theta}}_{\{1\}} + p^{\boldsymbol{\theta}}_{\{1,2\}}$	π_1
(0,>,0,0,>)	$p^{\boldsymbol{\theta}}_{\{1\}} + p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(Z_2^{(1)} \geq c^{(1)})$
(>,0,0,>,0)	$p^{\boldsymbol{\theta}}_{\{12\}}$	$\pi_{\{1\}}\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_1^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)}\|Z_{\{12\}}^{(1)})$
(<,<,>,0,0)	$p^{\boldsymbol{\theta}}_{\{1\}}$	π_1
(>,>,<,0,0)	$p^{\boldsymbol{\theta}}_{\{1\}} + p^{\boldsymbol{\theta}}_{\{1,2\}}$	π_1
(>,0,<,0,<)	$p^{\boldsymbol{\theta}}_{\{12\}}$	π_1
(<,0,>,0,>)	$p^{\boldsymbol{\theta}}_{\{1\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(Z_2^{(1)} \geq c^{(1)})$
(0,0,>,>,>)	$p^{\boldsymbol{\theta}}_{\{1\}}$	$(1-\pi_{\{1,2\}})\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_2^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)})$
(0,<,>,>,0)	$p^{\boldsymbol{\theta}}_{\{1\}}$	$\pi_{\{1\}}\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)})$
(0,>,>,>,>)	$p^{\boldsymbol{\theta}}_{\{1\}}$	$(1-\pi_{\{1,2\}})\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_2^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)})$
(0,>,<,<,0)	$p^{\boldsymbol{\theta}}_{\{1\}} + p^{\boldsymbol{\theta}}_{\{1,2\}}$	π_1
(>,<,>,>,0)	0	$\pi_{\{1\}}\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)})$
(>,>,<,>,0)	$p^{\boldsymbol{\theta}}_{\{12\}}$	$\pi_{\{1\}}\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_1^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)}\|Z_{\{12\}}^{(1)})$
(<,>,>,0,>)	$p^{\boldsymbol{\theta}}_{\{1\}}$	$\pi_1 + \mathbb{P}_0(Z_2^{(1)} \geq c^{(1)})$
(>,>,<,0,>)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \mathbb{P}_0(Z_2^{(1)} \geq c^{(1)})$
(>,>,0,>,>)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_{\{1\}}(1-\pi_{\{1,2\}})\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_1^{(1)} \geq c^{(1)} Z_2^{(2)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)}\|Z_{\{12\}}^{(1)})$
(>,0,>,>,>)	0	$(1-\pi_{\{1,2\}})\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_2^{(1)} \geq c^{(1)} \vee Z_{\{12\}}^{(2)} \geq c^{(2)})$

Table 6.1: Conditional PWER under the old and new design for all relevant different parameter configurations $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$ in the design described in Example 5.1.1, where $H_2 : \theta_2 \leq 0$ has been retained and the originally intended rejection of $H_1 : \theta_1 \leq 0$ has been ignored in order to test in $\mathcal{P}_{\{1\}}$ and $\mathcal{P}_{\{1,2\}}$ at stage 2 instead. In the second column $p^{\boldsymbol{\theta}}_J := \pi_J \mathbb{P}_{\theta_J}\left(Z_J^{(2)} \geq c_b^{(2)} | Z_J^{(1)}\right)$. Also $\pi_1 = \pi_{\{1\}} + \pi_{\{1,2\}}$.

$(\theta_{\{1\}},\theta_{\{2\}},\theta_{\{1,2\}},\theta_1,\theta_2)$	$cp_{b,\boldsymbol{\theta}}$	$cp^*_{a,\boldsymbol{\theta}}$
(0,0,0,0,0)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(0,>,0,0,>)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})$
(>,0,0,>,0)	$p^{\boldsymbol{\theta}}_{\{12\}}$	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\} \cup \{Z_2^{(2)} \geq c^{(2)}\} \mid Z_2^{(1)})$
(<,<,>0,0,0)	0	$\pi_1 + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(>,>,<,0,0)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(>,0,<,0,<)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(<,0,>,0,>)	0	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})$
(0,0,>,>,>)	0	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{j=1}^2\bigcup_{k=1}^2\{Z_j^{(k)} \geq c^{(k)}\})$
(0,<,>,>,0)	0	$\pi_{\{1\}}\mathbb{P}_0(Z_1^{(1)} \geq c^{(1)} \vee Z_1^{(2)} \geq c^{(2)})+$ $\pi_{\{2\}}\mathbb{P}_0(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(Z_1^{(1)} \geq c^{(1)} \vee Z_1^{(2)} \geq c^{(2)} \vee Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(0,>,>,>,>)	0	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{j=1}^2\bigcup_{k=1}^2\{Z_j^{(k)} \geq c^{(k)}\})$
(0,>,<,<,0)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_{\boldsymbol{\theta}}(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})$
(>,<,>,>,0)	0	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\} \cup \{Z_2^{(2)} \geq c^{(2)}\} \mid Z_2^{(1)})$
(>,>,<0,>,0)	$p^{\boldsymbol{\theta}}_{\{12\}}$	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(Z_2^{(2)} \geq c^{(2)} \mid Z_2^{(1)})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\} \cup \{Z_2^{(2)} \geq c^{(2)}\} \mid Z_2^{(1)})$
(<,>,>,0,>)	0	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})$
(>,>,<,0,>)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_1 + \pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})$
(>,>,0,>,>)	$p^{\boldsymbol{\theta}}_{\{1,2\}}$	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{j=1}^2\bigcup_{k=1}^2\{Z_j^{(k)} \geq c^{(k)}\})$
(>,0,>,>,>)	0	$\pi_{\{1\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_1^{(k)} \geq c^{(k)}\})+$ $\pi_{\{2\}}\mathbb{P}_0(\bigcup_{k=1}^2\{Z_2^{(k)} \geq c^{(k)}\})+$ $\pi_{\{12\}}\mathbb{P}_{\mathbf{0}}(\bigcup_{j=1}^2\bigcup_{k=1}^2\{Z_j^{(k)} \geq c^{(k)}\})$

Table 6.2: Conditional PWER under the old (Design I) and new design (Design II) for all relevant $\boldsymbol{\theta} \in \boldsymbol{\Theta}_C$, where $H_1 : \theta_1 \leq 0$ has been rejected and $H_2 : \theta_2 \leq 0$ has been retained. In the second column $p_J^{\boldsymbol{\theta}} := \pi_J \mathbb{P}_{\theta_J}\left(Z_J^{(2)} \geq c_b^{(2)} | Z_J^{(1)}\right)$. Also $\pi_1 = \pi_{\{1\}} + \pi_{\{1,2\}}$.

cond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD	91	91.36	-	-	-	-	-	167.5
AD1	61.68	62.36	95.17	95.45	98.31	-	-	150.2
AD2	59.80	60.76	-	-	97.91	-	-	160.3
AD3	59.80	60.76	73.75	73.43	52.51	37.39	35.49	160.3
AD4	77.06	77.19	62.30	61.68	39.77	92.21	92.49	160.3
single	90.93	91.37	75.37	76.31	44.34	-	-	233.5
uncond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD1	91	91.36	-	-	-	-	-	167.5
AD1	61.68	62.36	5.61	5.45	14.54	-	-	150.2
AD2	59.8	60.76	-	-	36.52	-	-	160.3
AD3	59.80	60.76	14.11	13.35	19.58	66.95	67.21	160.3
AD4	77.06	77.19	11.92	11.21	14.84	77.44	77.56	160.3
single	90.93	91.37	75.37	76.31	44.34	-	-	233.5

Table 6.3: Conditional and unconditional rejection probabilities for different hypotheses and the expected sample sizes for the whole trial. We used $\alpha = 0.025$ (one-sided), $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, $\boldsymbol{\delta}_A = (\delta_{\{1\}}, \delta_{\{2\}}, \delta_{\{1,2\}}, \delta_1, \delta_2) = (0.3, 0.3, 0.3, 0.3, 0.3)$, $N = 116.7491$ (Power 90%) and $\xi = 0.5$ (Pocock Design I as initial design).

cond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD	99.99	17.55	-	-	-	-	-	182.5
AD1	97.25	7.99	0	0	1	-	-	219.5
AD2	97.25	7.99	-	-	1	-	-	221
AD3	97.25	7.99	50	0	99.89	50	0	221
AD4	97.27	10.56	50	0	99.85	1	2.87	221
single	99.99	16.01	94.87	0	99.57	-	-	233.5
uncond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD	99.99	17.55	-	-	-	-	-	182.5
AD1	97.25	7.99	0	0	88.02	-	-	219.5
AD2	97.25	7.99	-	-	89.30	-	-	221
AD3	97.25	7.99	0.01	0	89.20	97.26	7.99	221
AD4	97.27	10.56	0.01	0	89.17	97.27	10.56	221
single	99.99	16.01	94.87	0	99.57	-	-	233.5

Table 6.4: Conditional and unconditional rejection probabilities for different hypotheses and the expected sample sizes for the whole trial. We used $\alpha = 0.025$ (one-sided), $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, $\boldsymbol{\delta}_A = (\delta_{\{1\}}, \delta_{\{2\}}, \delta_{\{1,2\}}, \delta_1, \delta_2) = (0.4, -0.2, 0.7, 0.5, 0.1)$, $N = 116.7491$ (Power 90%) and $\xi = 0.5$ (Pocock Design I as initial design).

7. Conclusion and Outlook

In this thesis we have used a newly proposed multiple type I error rate, called *population-wise error rate (PWER)*, to reply to the ever so growing complexity of clinical trials in multiple, highly stratified patient populations. This error rate is tailored to population structures where certain sub-populations can overlap while other can be totally disjoint from each other. The family-wise error rate (FWER), which is usually used as a multiple type I error rate, proves to be overly conservative in such trials since it does not account for the particular population structure the patients in the trial belong to, while the PWER does by averaging all (family-wise) type I error probabilities for every disjoint subgroup the overall patient population consists of. Thus, one can technically say that whenever FWER-control is possible, PWER-control is as well since the PWER is just an average of FWER-expressions. In this thesis PWER-controlling methods were proposed for single-stage (single-step) test designs, as well as group sequential and lastly also adaptive designs and the gain in power and the associated sample size needed to obtain a certain power was illustrated by means of several theoretical example designs as well as designs already proposed in the literature. In particular, in Chapter 3 we considered the two cases of investigating one single treatment and multiple treatments in many, possibly overlapping, biomarker-defined patient populations. This was then applied to the special case of two overlapping populations and the gain in sample size and marginal power could be quantified and visualized for different intersection sizes. Lastly, we applied PWER-control to a design from Sun & et al. (2016, [51]), where hypothesis tests for superior treatment effects in all unions of a number of pre-defined disjoint sub-populations were conducted. We could demonstrate gains in power up to 10% and often above 5%. In Chapter 4 we proposed group sequential designs with PWER-control. The Wang & Tsiatis method could immediately be applied to the PWER-concept and the development of an error spending approach for PWER-control could be achieved by a straightforward adaptation of the classical approach by expressing the PWER as a sum of stage-wise PWER-terms. We compared our proposed PWER-adjusted error spending method to a design by Magnusson & Turnbull (2013, [36]) who incorporated a way to select the most benefiting patient population in which a hypothesis is then tested via a group sequential design. We found a relative decrease in maximum information needed for a certain power

level of around 15% and relative decrease in expected information of up to 6%. As a preparation for adaptive designs, we considered multiple conceivable designs, named Design I, II and II, for the case of two intersecting populations, which allowed for testing of treatment effects in the disjoint subgroups. Ways to find critical values, control power and (average) sample size were derived and applied in numerical examples.
At last, we proposed a method to control the PWER in adaptive designs in arbitrarily structured populations by means of the CRP-principle by Müller & Schäfer (2004, [41]) which allows to switch from a design one initially started with to a new design without inflating the overall type I error rate. The possibility to write the PWER as the expectation of the percentage of patients affected by a type I error in the relative to overall population allowed us to apply the CRP-principle to our PWER-concept with only a few difficulties. As in the classical CRP-principle, the overall PWER is controlled whenever the conditional population-wise error rate of the new design is bounded by that of the initial design, for every possible parameter constellation. We proposed a method where this constraint only had to be satisfied for a finite number of parameter constellations in order to guarantee overall PWER-control after a possible mid-trial adaptation. Here, a challenge was the fact that parameters (treatment effects) in larger populations are weighted sums of the effects in the respective disjoint subgroups of this population and thus cannot be seen as decoupled from each other. A modification to the conditional PWER of the initial design enabled us to guarantee overall PWER-control and provided a straightforward way to determine new critical boundaries by simultaneously solving a finite number of inequalities. In our examples we used a conservative solution for these inequality systems. To test our method, we conducted simulations for the case of two intersecting populations. Two-stage designs were considered where the initial design was given by Design I from Chapter 4 and depending on certain rules a different design was tested at stage 2. We overall found that for some parameter configurations, the power loss is not that high compared to the case where Design I was conducted without mid-trial adaptations and the power of rejecting in certain smaller sub-populations is high whenever it is chosen to test for them. Also, for homogeneous effects, where an adaptive design is mostly unnecessary, the use of any adaptive design we proposed resulted in a power loss from 15% to 30% in H_1 and H_2 depending on the design adaptation. Nonetheless, rejection probabilities for any disjoint subset of $\mathcal{P}$ was still quite high even though tests for these populations did not happen frequently. For heterogeneous treatment effects, we also see a slight power decrease for H_1 and H_2 but the power in the respective disjoint subgroups that displayed a relevant treatment effect was large. Surely, the adaptation rules defined used in Section 6.3 are very basic and lack a certain degree of sophistication. A thorough examination of the utility of such adaptive designs with more thought-out adaptation rules as well as an application to an already existing trial example is a topic for future research.

In all considerations described above we made some quite simplifying assumptions. For once, we almost always assumed the relative population sizes $\boldsymbol{\pi}$ to be known. This is quite a strong assumption considering that the PWER itself and also the parameters we tested for depend on the population sizes. In Chapter 3 we considered the possibility of estimating the prevalences with the trial data and found that the estimation of $\boldsymbol{\pi}$ does not harm PWER-control in any consideriable amount.

However, this still needs to be examined by more examples, especially for group sequential and adaptive designs. Another assumption we made was that the observations originated from a normal distribution with a known (mostly common) variance σ^2. Firstly, a topic for future research should be to examine error control in such population structures where the variance is unknown and different in each population one tests in. Assuming a multivariate t-distribution for the test statistics would also only be an approximation here since such a multivariate t-distributed random variable is defined by the quotient of a multivariate normal vector divided by one (univariate) χ^2-distributed random variable. Technically, each component of that multivariate normal vector would need to be divided by an individual χ^2 random variable when different variances are assumed for each population. Secondly, we only looked at continuous endpoints. PWER-control for binomial or survival analysis should be further examined as well. For a large enough sample size a multivariate normal distribution might suffice but otherwise, direct computation methods should be derived.
For group sequential designs in general, we also only considered designs without incorporating futility stopping or other types of early rejection decisions. For example, in the two population case, one could try to implement a rule where the trial is stopped for efficacy at stage 1 if the null hypotheses are rejected in both population $\mathcal{P}_1$ and $\mathcal{P}_2$, where some prespecified 'early rejection' critical value is used. This could strengthen the claim we implicitly made with Design II that rejection in both hypotheses gives a strong evidence for a beneficial effect in the intersection as well. With regard to adaptive multiple testing problems we already mentioned that the closed test procedure can not be applied to the PWER in the same way it can to the FWER together with the combination test approach. Future research could therefore involve the invention of a closed test principle that ensures PWER-control and then combination function or conditional error function approaches could be used to ensure PWER-control in an adaptive design, as well. Also, as already mentioned in Chapter 3, the development of step-down procedures that uniformly improve the single-step test with PWER-control is another possible future research topic.
Another interesting application could be the use in platform trials (see e.g. [46]), which are adaptive multi-arm and multi-stage trials where treatments can be dropped and newly introduced during the course of the trial. These trials generally have an open master protocol, so they especially have no defined endpoint. If populations were overlapping or highly stratified in those trials, PWER-control could be something one would want to aim for. Lastly, we want to mention that other, more liberal error rates than the FWER already exist, like the false discovery rate (FDP), for instance, and that it would probably make sense to theoretically and practically compare the PWER to those error rates in terms of type I and II error control as well.

A. Further simulation results from Section 3.3.4

The tables for the cases of $m = 6$ and $m = 8$ sub-populations are given in the following tables.

$m = 6$		Pow	correct	false	RAE	Pow	correct	false	RAE
		$q = 0$				$q = 1/6$			
$\tau = 0$	PWER	34.9	34.9	0	2.0	37.4	34.9	2.5	23
	FWER	26.0	26.0	0	1.5	28.1	26.3	1.8	17
$\tau = 0.4$	PWER	36.3	36.3	0	2.1	39.1	36.4	2.7	24
	FWER	26.6	26.6	0	1.6	29.2	27.3	1.9	19
$\tau = 0.8$	PWER	39.9	39.9	0	2.5	43.6	40.9	2.8	30
	FWER	29.9	29.9	0	1.9	33.6	31.5	2.1	23
		$q = 2/6$				$q = 3/6$			
$\tau = 0$	PWER	42.0	36.6	5.4	2.8	49.8	40.9	9.0	38
	FWER	32.3	28.2	4.1	2.2	39.5	32.6	7.0	31
$\tau = 0.4$	PWER	44.1	38.5	5.6	3.1	53.5	44.1	9.4	42
	FWER	34.4	30.2	4.3	2.4	43.2	35.9	7.4	34
$\tau = 0.8$	PWER	50.8	44.8	6.1	3.8	63.1	53.1	10.0	52
	FWER	40.4	35.7	4.7	3.0	53.2	45.1	8.1	44
		$q = 4/6$				$q = 5/6$			
$\tau = 0$	PWER	64.5	51.3	13.2	5.8	91.7	78.7	13.1	92
	FWER	54.6	43.7	10.9	5.0	87.2	75.2	12.0	87
$\tau = 0.4$	PWER	71.5	58.3	13.2	62				
	FWER	61.9	50.8	11.1	53				
$\tau = 0.8$	PWER	85.8	74.1	11.8	79				
	FWER	78.9	68.5	10.4	72				

		$q = 1$			
$\tau = 0$	PWER	0	0	4.2	0
	FWER	0	0	2.2	0

$m = 8$		Pow	correct	false	RAE	Pow	correct	false	RAE
		$q = 0$				$q = 1/8$			
$\tau = 0$	PWER	34.0	34.0	0	1.8	35.6	33.5	2.2	19
	FWER	24.3	24.3	0	1.3	26.2	24.6	1.6	14
$\tau = 0.4$	PWER	34.8	34.8	0	1.9	36.9	34.7	2.2	21
	FWER	25.5	25.5	0	1.4	27.0	25.4	1.6	15
$\tau = 0.8$	PWER	37.4	37.4	0	2.1	40.0	37.7	2.3	24
	FWER	27.9	27.9	0	1.6	30.4	28.7	1.7	19
		$q = 2/8$				$q = 3/8$			
$\tau = 0$	PWER	38.0	33.5	4.5	2.2	41.3	34.2	7.1	26
	FWER	28.4	25.1	3.3	1.7	31.1	25.9	5.2	20
$\tau = 0.4$	PWER	39.6	35.0	4.6	2.4	43.6	36.4	7.3	29
	FWER	29.8	26.4	3.4	1.8	33.1	27.5	5.5	22
$\tau = 0.8$	PWER	44.0	39.0	5.0	2.9	49.3	41.4	7.9	35
	FWER	33.6	29.8	3.8	2.2	38.4	32.3	6.1	27
		$q = 4/8$				$q = 5/8$			
$\tau = 0$	PWER	46.9	36.9	10.0	3.3	56.0	42.3	13.7	45
	FWER	36.7	28.9	7.8	2.6	45.9	34.7	11.2	38
$\tau = 0.4$	PWER	49.5	39.1	10.5	3.6	59.7	45.5	14.3	48
	FWER	39.3	31.1	8.2	2.9	49.5	37.8	11.7	41
$\tau = 0.8$	PWER	57.4	46.0	11.5	4.4	70.4	55.3	15.1	59
	FWER	46.6	37.4	9.2	3.6	60.9	48.1	12.8	51
		$q = 6/8$				$q = 7/8$			
$\tau = 0$	PWER	72.9	54.6	18.2	6.8	96.3	83.8	12.5	96
	FWER	63.8	48.1	15.7	6.0	93.4	81.6	11.8	93
$\tau = 0.4$	PWER	79.0	61.5	17.5	69				
	FWER	70.4	55.1	15.3	61				
$\tau = 0.8$	PWER	91.3	77.6	13.8	84				
	FWER	86.9	74.2	12.7	79				
		$q = 1$							
$\tau = 0$	PWER	0	0	4.5	0				
	FWER	0	0	2.3	0				

B. Further simulation results from Section 6.3

cond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD1	93.26	93.62	-	-	-	-	-	196.9
AD1	41.91	41.72	93.87	96.73	98.22	-	-	150.1
AD2	35.06	35.05	-	-	98.26	-	-	159.6
AD3	35.06	35.05	74.31	72.36	54.28	38.79	35.87	159.6
AD4	51.78	51.75	59.22	56.54	39.30	92.86	92.81	159.6
single	90.93	91.37	75.37	76.31	44.34	-	-	233.5
uncond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD1	93.26	93.62	-	-	-	-	-	196.9
AD1	41.91	41.72	3.76	3.84	12.11	-	-	150.1
AD2	35.06	35.05	-	-	36.08	-	-	159.6
AD3	35.06	35.05	13.64	13.28	19.93	42.18	41.64	159.6
AD4	51.78	51.75	10.87	10.38	14.43	52.11	52.10	159.6
single	90.93	91.37	75.37	76.31	44.34	-	-	233.5

Table B.1: Conditional and unconditional rejection probabilities for different hypotheses and the expected sample sizes for the whole trial. We used $\alpha = 0.025$ (one-sided), $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, $\boldsymbol{\delta}_A = (\delta_{\{1\}}, \delta_{\{2\}}, \delta_{\{1,2\}}, \delta_1, \delta_2) = (0.3, 0.3, 0.3, 0.3, 0.3)$, $N = 116.7491$ (Power 90%) and $\xi = 0.5$ (OBF Design I as initial design).

cond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD	99.99	20.17	-	-	-	-	-	189.9
AD1	90.33	1.98	0	0	1	-	-	220
AD2	90.33	1.95	-	-	1	-	-	220
AD3	90.33	1.95	33.33	0	99.93	33.33	0	220
AD4	90.35	5.05	0	0	99.93	1	3.51	220
single	99.99	16.00	94.87	0	99.57	-	-	233.5
uncond.	H_1	H_2	$H_{\{1\}}$	$H_{\{2\}}$	$H_{\{1,2\}}$	H_1^*	H_2^*	$E(N)$
GSD	99.99	20.17	-	-	-	-	-	189.9
AD1	90.33	1.98	0	0	87.65	-	-	219.1
AD2	90.33	1.95	-	-	88.41	-	-	220
AD3	90.33	1.95	0	0	88.35	90.34	1.95	220
AD4	90.35	5.05	0	0	88.35	90.35	5.05	220
single	99.99	16	94.87	0	99.57	-	-	233.5

Table B.2: Conditional and unconditional rejection probabilities for different hypotheses and the expected sample sizes for the whole trial. We used $\alpha = 0.025$ (one-sided), $\boldsymbol{\pi} = (0.4, 0.4, 0.2)$, $\boldsymbol{\delta}_A = (\delta_{\{1\}}, \delta_{\{2\}}, \delta_{\{1,2\}}, \delta_1, \delta_2) = (0.4, -0.2, 0.7, 0.5, 0.1)$, $N = 116.7491$ (Power 90%) and $\xi = 0$ (OBF Design I as initial design).

Bibliography

[1] https://mathworld.wolfram.com/MultinomialDistribution.html. Last accessed: 2021-06-09.

[2] Qi An and Myron Chang. "Computation of optimal group sequential designs". In: *Sequential Analysis* 35.4 (2016), pp. 453–464. eprint: https://doi.org/10.1080/07474946.2016.1238256.

[3] P. Armitage. *Sequential Medical Trials*. Halsted Press book. Wiley, 1975. ISBN: 9780470033234.

[4] P. Armitage, C. K. McPherson, and B. C. Rowe. "Repeated Significance Tests on Accumulating Data". In: *Journal of the Royal Statistical Society. Series A (General)* 132.2 (1969), pp. 235–244. ISSN: 00359238.

[5] P Bauer and K Köhne. "Evaluation of experiments with adaptive interim analyses". In: *Biometrics* 50.4 (1994), 1029—1041. ISSN: 0006-341X.

[6] Peter Bauer. "Multistage testing with adaptive designs". In: *Biometrie und Informatik in Medizin und Biologie* 20 (1989), pp. 130–148.

[7] Peter Bauer and Werner Brannath. "The advantages and disadvantages of adaptive designs for clinical trials". In: *Drug Discovery Today* 9.8 (2004), pp. 351–357. ISSN: 1359-6446.

[8] Peter Bauer and Meinhard Kieser. "Combining different phases in the development of medical treatments within a single trial". In: *Statistics in Medicine* 18.14 (1999), pp. 1833–1848.

[9] C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Libreria internazionale Seeber, 1936.

[10] W. Brannath, G. Gutjahr, and P. Bauer. "Probabilistic Foundation of Confirmatory Adaptive Designs". In: *Journal of the American Statistical Association* 107.498 (2012), pp. 824–832. eprint: https://doi.org/10.1080/01621459.2012.682540.

[11] Werner Brannath and Frank Bretz. "Shortcuts for Locally Consonant Closed Test Procedures". In: *Journal of the American Statistical Association* 105.490 (2010), pp. 660–669. eprint: https://doi.org/10.1198/jasa.2010.tm08127.

[12] Werner Brannath, Charlie Hillner, and Kornelius Rohmeyer. *A liberal type I error rate for studies in precision medicine*. 2021. arXiv: 2011.04766 [stat.ME].

[13] Werner Brannath, Franz Koenig, and Peter Bauer. "Multiplicity and flexibility in clinical trials". In: *Pharmaceutical Statistics* 6.3 (2007), pp. 205–216. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.302.

[14] Werner Brannath, Martin Posch, and Peter Bauer. "Recursive Combination Tests". In: *Journal of the American Statistical Association* 97.457 (2002), pp. 236–244. eprint: https://doi.org/10.1198/016214502753479374.

[15] R.P. Brent. *Algorithms for Minimization Without Derivatives*. Dover Books on Mathematics. Dover Publications, 2013. ISBN: 9780486143682.

[16] F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. CRC Press, 2016. ISBN: 9781420010909.

[17] Myron N. Chang. "Optimal designs for group sequential clinical trials". In: *Communications in Statistics - Theory and Methods* 25.2 (1996), pp. 361–379. eprint: https://doi.org/10.1080/03610929608831700.

[18] Olivier Collignon et al. "Current Statistical Considerations and Regulatory Perspectives on the Planning of Confirmatory Basket, Umbrella, and Platform Trials". In: *Clinical Pharmacology & Therapeutics* 107.5 (2020), pp. 1059–1067.

[19] T. Dickhaus. *Simultaneous Statistical Inference: With Applications in the Life Sciences*. SpringerLink : Bücher. Springer Berlin Heidelberg, 2014.

[20] A. Dmitrienko, A.C. Tamhane, and F. Bretz. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2009. ISBN: 9781584889854.

[21] Trahair T.N. Fletcher J.I. Ziegler D.S. "Too many targets, not enough patients: rethinking neuroblastoma clinical trials". In: *Nat Rec Vancer* 18 (2018), pp. 389–400.

[22] Alan Genz et al. *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6. 2017.

[23] Georg Gutjahr, Werner Brannath, and Peter Bauer. "An Approach to the Conditional Error Rate Principle with Nuisance Parameters". In: *Biometrics* 67.3 (2011), pp. 1039–1046. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2010.01507.x.

[24] Gerhard Hommel. "Adaptive Modifications of Hypotheses After an Interim Analysis". In: *Biometrical Journal* 43.5 (2001), pp. 581–589. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1521-4036%28200109%2943%3A5%3C581%3A%3AAID-BIMJ581%3E3.0.CO%3B2-J.

[25] Irving K. Hwang, Weichung J. Shih, and John S. De Cani. "Group sequential designs using a family of type I error probability spending functions". In: *Statistics in Medicine* 9.12 (1990), pp. 1439–1445. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780091207.

[26] C. Jennison and B.W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 1999. ISBN: 9781584888581.

[27] Richard Kaplan et al. "Evaluating Many Treatments and Biomarkers in Oncology: A New Design". In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31 (2013).

[28] Miriam Kesselmeier, Norbert Benda, and André Scherag. "Effect size estimates from umbrella designs: Handling patients with a positive test result for multiple biomarkers using random or pragmatic subtrial allocation". In: *PLOS One* 15.8 (Aug. 2020), pp. 1–24.

[29] Meinhard Kieser, Peter Bauer, and Walter Lehmacher. "Inference on Multiple Endpoints in Clinical Trials with Adaptive Interim Analyses". In: *Biometrical Journal* 41.3 (1999), pp. 261–277. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-4036%28199906%2941%3A3%3C261%3A%3AAID-BIMJ261%3E3.0.CO%3B2-U.

[30] Kyungmann Kim and David L. DeMets. "Design and analysis of group sequential tests based on the type I error spending rate function". In: *Biometrika* 74.1 (Mar. 1987), pp. 149–154. ISSN: 0006-3444. eprint: https://academic.oup.com/biomet/article-pdf/74/1/149/576837/74-1-149.pdf.

[31] K. K. Gordon Lan and David L. DeMets. "Discrete Sequential Boundaries for Clinical Trials". In: *Biometrika* 70.3 (1983), pp. 659–663. ISSN: 00063444.

[32] K. K. Gordon Lan and David L. Demets. "Group sequential procedures: Calendar versus information time". In: *Statistics in Medicine* 8.10 (1989), pp. 1191–1198.

[33] K.K. Gordon Lan, David M. Reboussin, and David L. DeMets. "Information and information fractions for design and sequential monitoring of clinical trials". In: *Communications in Statistics - Theory and Methods* 23.2 (1994), pp. 403–420. eprint: https://doi.org/10.1080/03610929408831263.

[34] Walter Lehmacher and Gernot Wassmer. "Adaptive Sample Size Calculations in Group Sequential Trials". In: *Biometrics* 55.4 (1999), pp. 1286–1290. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.1999.01286.x.

[35] Kieser M. Lehmacher W. and Hothorn L. "Sequential and Multiple Testing for Dose-Response Analysis". In: *Drug Information Journal* 34 (2000), pp. 591–597.

[36] B. P. Magnusson and B. W. Turnbull. "Group sequential enrichment design incorporating subgroup selection". In: *Statistics in Medicine* 32.16 (2013), pp. 2695–2714. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5738.

[37] Shakun M. Malik et al. "Consensus Report of a Joint NCI Thoracic Malignancies Steering Committee: FDA Workshop on Strategies for Integrating Biomarkers into Clinical Development of New Therapies for Lung Cancer Leading to the Inception of "Master Protocols" in Lung Cancer". In: *Journal of Thoracic Oncology* 9.10 (2014), pp. 1443 –1448.

[38] Ruth Marcus, Eric Peritz, and K. R. Gabriel. "On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance". In: *Biometrika* 63.3 (1976), pp. 655–660.

[39] W. Maurer and B. Mellein. "On New Multiple Tests Based on Independent p-Values and the Assessment of Their Power". In: 1988.

[40] C. K. McPherson and P. Armitage. "Repeated Significance Tests on Accumulating Data When the Null Hypothesis is Not True". In: *Journal of the Royal Statistical Society: Series A (General)* 134.1 (1971), pp. 15–25. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2343971.

[41] Hans-Helge Müller and Helmut Schäfer. "A general statistical principle for changing a design any time during the course of a trial". In: *Statistics in medicine* 23 (Aug. 2004), pp. 2497–508.

[42] Martin Posch et al. “Conditional Rejection Probabilities of Student’s t-test and Design Adaptations”. In: *Biometrical Journal* 46.4 (2004), pp. 389–403. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200410042.

[43] MA Proschan and SA Hunsberger. “Designed extension of studies based on conditional power”. In: *Biometrics* 51.4 (Dec. 1995), 1315—1324. ISSN: 0006-341X.

[44] M.A. Proschan, K.K.G. Lan, and J.T. Wittes. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Statistics for Biology and Health. Springer New York, 2006. ISBN: 9780387449708.

[45] Todd S. et al. “Interim analyses and sequential designs in phase III studies”. In: *Br J Clin Pharmacol* 51.5 (2001), pp. 394–399.

[46] Benjamin R Saville and Scott M Berry. “Efficiencies of platform clinical trials: A vision of the future”. In: *Clinical Trials* 13.3 (2016). PMID: 26908536, pp. 358–366. eprint: https://doi.org/10.1177/1740774515626362.

[47] Eckart Sonnemann. “General Solutions to Multiple Testing Problems”. In: *Biometrical Journal* 50.5 (2008), pp. 641–656. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200810462.

[48] Stram D.O. Sposto R. “A strategic view of randomized trial design in low-incidence paediatric cancer”. In: *Statistics in Medicine* 18 (1999), pp. 1183–1197.

[49] Nigel Stallard and Karen M. Facey. “Comparison of the spending function method and the christmas tree correction for group sequential trials”. In: *Journal of Biopharmaceutical Statistics* 6.3 (1996), pp. 361–373.

[50] Karolina Strzebonska and Marcin Waligora. “Umbrella and basket trials in oncology: Ethical challenges”. In: *BMC Medical Ethics* 20 (2019).

[51] Hong Sun et al. “Comparing a stratified treatment strategy with the standard treatment in randomized clinical trials”. In: *Statistics in Medicine* 35.29 (2016), pp. 5325–5337. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7091.

[52] A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN: 9780521784504.

[53] A. Wald. “Sequential Tests of Statistical Hypotheses”. In: *The Annals of Mathematical Statistics* 16.2 (1945), pp. 117 –186.

[54] Samuel K. Wang and Anastasios A. Tsiatis. “Approximately Optimal One-Parameter Boundaries for Group Sequential Trials”. In: *Biometrics* 43.1 (1987), pp. 193–199. ISSN: 0006341X, 15410420.

[55] G. Wassmer and W. Brannath. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Springer International Publishing, 2016. ISBN: 9783319325606.

[56] P.H. Westfall and S.S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Statistics. Wiley, 1993. ISBN: 9780471557616.

[57] Janet Woodcock and Lisa M. LaVange. "Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both". In: *New England Journal of Medicine* 377.1 (2017). PMID: 28679092, pp. 62–70. eprint: https://doi.org/10.1056/NEJMra1510062.

[58] Zbyněk Šidák. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633. eprint: https://doi.org/10.1080/01621459.1967.10482935.