A liberal type I error rate for studies in precision medicine

Werner Brannath\textsuperscript{1}, Charlie Hillner\textsuperscript{2}, and Kornelius Rohmeyer\textsuperscript{3}

\textsuperscript{1}University of Bremen, Institute for Statistics and Competence Center for Clinical Trials, Bremen, Germany, brannath@uni-bremen.de

\textsuperscript{2}University of Bremen, Institute for Statistics and Competence Center for Clinical Trials, Bremen, Germany

\textsuperscript{3}University of Oldenburg, Institute of Mathematics, Germany

Abstract

We introduce a new multiple type I error criterion for clinical trials with multiple populations. Such trials are of interest in precision medicine where the goal is to develop treatments that are targeted to specific sub-populations defined by genetic and/or clinical biomarkers. The new criterion is based on the observation that not all type I errors are relevant to all patients in the overall population. If disjoint sub-populations are considered, no multiplicity adjustment appears necessary, since a claim in one sub-population does not affect patients in the other ones. For intersecting sub-populations we suggest to control the average multiple type error rate, i.e. the probably that a randomly selected patient will be exposed to an inefficient treatment. We call this the population-wise error rate, exemplify it by a number of examples and illustrate how to control it with an adjustment of critical boundaries or adjusted p-values. We furthermore define corresponding simultaneous confidence intervals. We finally illustrate the power gain achieved by passing from family-wise to population-wise error rate control with two simple examples and a recently suggest multiple testing approach for umbrella trials.

Keywords: Enrichment designs; Family-wise error rate; Multiple testing; Platform trials; Population-wise error rate; Umbrella trials

1 Introduction

The aim of precision medicine is to provide each patient with an optimal treatment tailored to his or her genetic and/or clinical profile. One strategy for reaching this goal is to undertake trials where one or several treatments are investigated in multiple sub-populations. Examples for such trials are umbrella and basket trials in oncology. 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. In a basket trial patients with different cancer types but one common molecular alteration are enrolled with the aim to study one targeted treatment (see e.g. Woodcock and LaVange, 2017; Strzebonska and Waligora, 2019). In many cases the target or sub-populations are disjoint by nature, but when many different biomarkers or cancer types are used, it can also occur that patients belong to more than one sub-population. For example, in the FOCUS4 study (Kaplan \textit{et al.}, 2013) 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. In this manuscript, we explicitly allow biomarker-defined sub-populations to be overlapping such that patients become eligible for multiple targeted therapies. This means that future
patients of the overlap may be exposed to more than a single inefficient treatment by the trial results. Moreover, for such studies suitable allocation procedures have to be defined. Issues of eligibility for multiple target therapies have been addressed e.g. in Malik et al. (2014) and Collignon et al. (2014).

In confirmatory clinical trials with tests of several hypotheses the multiple type I error is usually kept small by controlling the family-wise error rate (FWER). With the growing effort of detecting new and more predictive biomarkers and an increasing focus on rare diseases, it is becoming more and more difficult to undertake clinical trials that are sufficiently powered and also provide sufficient control of type I errors. Since the control of multiple type I errors amplifies this issue, more liberal alternatives to the common approach of family-wise error rate control are of strong interest. If a treatment or a treatment strategy is tested in several disjoint populations and each population is affected by only a single hypothesis test, the overall study basically consists of separate trials that merely share the same infrastructure. Therefore, no multiplicity adjustments are needed (e.g. Glimm and Di Scala, 2015; Collignon et al., 2020). However, if some sub-populations are overlapping, these intersections will contain patients that are possibly exposed to multiple erroneously rejected null hypotheses, implying that one has to adjust for multiplicity (e.g. Collignon et al., 2020). Since only patients in the intersections are concerned with this multiplicity issue, there is no need for adjustments for patients in the complements, who can only be affected by at most one false rejection of a null hypothesis. The FWER would therefore be too conservative also in this case. Especially for small and/or highly stratified populations, as for instance encountered in paediatric oncology, a more liberal approach is desirable (e.g. Fletcher et al., 2018). The purpose of this manuscript is to propose a new concept of multiple type I error control that is less conservative. With this new error rate, which we name population-wise error rate (PWERM), we aim to keep the average multiple type I error rate at a reasonable level. This provides control of the probability that a randomly chosen future patient will be exposed to an inefficient treatment policy.

The paper is outlined as follows. First, the PWER is motivated by means of a simple example, followed by the general mathematical definition for the case of possibly intersecting populations. Then, we demonstrate how to control the PWER at a pre-specified level by adjusting critical boundaries or p-values. In the subsequent section the gain in power is investigated when using PWER instead of FWER control. This will be done by means of two illustrative examples. In the first example we will investigate the case of two overlapping populations and assess the power gain when investigating (i) two different treatments in each population, respectively, and (ii) the same treatment in both populations. The second example consists of an application of our new population-based concept to a multiple testing approach for umbrella trials suggested in Sun et al. (2016). In section 5 we extend the multiple test with PWER-control to simultaneous confidence intervals and discuss their coverage properties. The paper concludes with a discussion in Section 6.

2 The population-wise error rate

In this section the aforementioned population-wise error rate is introduced both conceptually and formally. Examples for different practically relevant settings are given to further deepen the understanding.

2.1 General framework and definition

Consider an overall population \( \mathcal{P} \) consisting of \( m \geq 2 \) possibly overlapping sub-populations \( \mathcal{P}_1, \ldots, \mathcal{P}_m \subseteq \mathcal{P} \) and suppose that we want to investigate a treatment \( T_i \) in each \( \mathcal{P}_i \) by means of a hypothesis test. In the sequel we call the tuples \( (\mathcal{P}_i, T_i) \) the treatment policies. To each treatment policy \((\mathcal{P}_i, T_i)\) we assign the null hypothesis \( H_i : \theta_i \leq 0 \), where \( \theta_i = \theta(\mathcal{P}_i, T_i) \) quantifies the efficacy of treatment \( T_i \) in comparison to a control in population \( \mathcal{P}_i \). The population-wise error rate is then given by the risk for a randomly chosen patient (same probability for each patient) to be assigned to one or more inefficient treatment policies, i.e.
to belong to at least one tuple \((P_i, T_i)\) with \(\theta_i \leq 0\) for which \(H_i\) has been rejected.

![Diagram of three populations and their disjoint sub-populations](image)

**Figure 1:** \(m = 3\) populations and their disjoint sub-populations

In order to define the PWER mathematically, we need to partition the overall population into disjoint sub-populations \(P_J := \cap_{i \in J} P_i \setminus \cup_{i \in I \setminus J} P_i\) for \(J \subseteq I := \{1, \ldots, m\}\). In Figure 1, we see an example for such a partition based on three populations \(P_i\), \(i = 1, 2, 3\). Note that \(P_{\{1,2,3\}} = \emptyset\). For each non-empty subset \(P_J\) denote its prevalence by \(\pi_J\) such that \(\sum_{J \subseteq I, P_J \neq \emptyset} \pi_J = 1\). For any future patient in \(P_J, J \subseteq I\), we commit a type I error if he/she belongs to at least one \((P_i, T_i)\) with \(\theta_i \leq 0\) for which \(H_i\) has been rejected, for each \(i \in J\). The population-wise error rate (PWER) is then defined as

\[
\text{PWER} = \sum_{J \subseteq I, P_J \neq \emptyset} \pi_J \mathbb{P}(\text{falsely reject any } H_j \text{ with } j \in J).
\]

(1)

To determine the PWER, we need to know for each subset \(P_J\) the probability of rejecting at least one true null hypothesis that affects this subset.

Compared to the FWER, which controls the maximum risk for future patients to be assigned to an inefficient treatment strategy, the PWER is an average risk. It is more liberal, because

\[
\text{PWER} = \sum_{J \subseteq I, P_J \neq \emptyset} \pi_J \mathbb{P}(\text{falsely reject any } H_j \text{ with } j \in J)
\]

\[
\leq \left( \sum_{J \subseteq I, P_J \neq \emptyset} \pi_J \right) \mathbb{P}(\text{falsely reject any } H_i \text{ for } i \in I) = \text{FWER},
\]

where equality occurs only in the extreme case when \(\pi_I = 1\) and all other \(\pi_J = 0\).

### 2.2 Two intersecting populations

As an example consider a trial with two intersecting populations \(P_1\) and \(P_2\) and two treatments \(T_1\) and \(T_2\) to be tested by means of the hypotheses \(H_1 : \theta(P_1, T_1) \leq 0\) and \(H_2 : \theta(P_2, T_2) \leq 0\). Usually, the two treatments will compared to the same control, however, the basic idea given in the subsequent sections will also apply with treatment specific controls. As illustrated in the left panel of Fig. 2, the overall population can be partitioned into three disjoint sub-populations, \(P_{\{1\}} := P_1 \setminus P_2, P_{\{2\}} := P_2 \setminus P_1\) and \(P_{\{1,2\}} := P_1 \cap P_2\). Obviously, we commit a type I error for \(P_{\{i\}}\) whenever \(H_i\) is falsely rejected, \(i = 1, 2\), and for \(P_{\{1,2\}}\) whenever \(H_1\) or \(H_2\) are falsely rejected. Hence, if \(H_1\) and \(H_2\) are both true, then

\[
\text{PWER} = \pi_{\{1\}} \mathbb{P}(\text{reject } H_1) + \pi_{\{2\}} \mathbb{P}(\text{reject } H_2) + \pi_{\{1,2\}} \mathbb{P}(\text{reject } H_1 \text{ or } H_2)
\]

(2)
If $H_1$ is true and $H_2$ is false, then:

\[
P_{\text{WER}} = \pi_{\{1\}} \mathbb{P}(\text{reject } H_1) + \pi_{\{1,2\}} \mathbb{P}(\text{reject } H_1) = (\pi_{\{1\}} + \pi_{\{1,2\}}) \mathbb{P}(\text{reject } H_1)
\]

Similarly, if $H_1$ is false and $H_2$ is true, then $P_{\text{WER}} = (\pi_{\{2\}} + \pi_{\{1,2\}}) \mathbb{P}(\text{reject } H_2)$. Note, if only one null hypothesis is true, say $H_i$, the PWER reduces to the probability of rejecting $H_i$ multiplied by the size of the population $P_i$.

2.3 Nested populations

In practice, one also often faces the problem of nested populations $P_1 \supset P_2 \supset \cdots \supset P_m$, as in the right panel of Fig. 2, and in each $P_i$ the hypothesis $H_i : \theta_i(P_i, T_i) \leq 0$. Define $P_{[i]} := P_{\{1, \ldots, i\}}$ for $i \leq m$. We commit a type I error for $P_{[i]}$ whenever any true $H_j$ is rejected for $j \leq i$. With prevalences $\pi_{[i]} := \pi_{\{1, \ldots, i\}}$ of $P_{[i]}$ the PWER is given by

\[
P_{\text{WER}} = \sum_{i=1}^{m} \pi_{[i]} \mathbb{P}(\text{reject at least one true } H_j \text{ for } j \leq i).
\]

Especially, if $P_i$ is defined by a continuous biomarker $X$, i.e. $P_i = \{X > t_i\}$ for cut-off points $t_i$, $i = 1, \ldots, m+1$ (with $t_{m+1} := \infty$), the PWER can be written as

\[
P_{\text{WER}} = \sum_{i=1}^{m} \mathbb{P}(t_i < X \leq t_{i+1}) \mathbb{P}(\text{reject at least one true } H_j \text{ for } j \leq i).
\]

2.3.1 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 $P_1, P_2$ and $P_3$ with $P_1 \cap P_2 \neq \emptyset$, $P_2 \cap P_3 \neq \emptyset$ and $P_1 \cap P_3 = \emptyset$, as in Fig. 1. Again, hypotheses of the form $H_i : \theta(P_i, T_i) \leq 0$ are to be tested in each population, respectively. Under the global null hypothesis, where all null hypotheses $H_i$ are true, the PWER is given by

\[
P_{\text{WER}} = \sum_{i=1}^{3} \pi_{\{i\}} \mathbb{P}(\text{reject } H_i) + \sum_{i=1}^{2} \pi_{\{i,i+1\}} \mathbb{P}(\text{reject } H_i \text{ or } H_{i+1}).
\]

The FWER under the global null equals $FWER = \mathbb{P}(\text{reject } H_1 \text{ or } H_2 \text{ or } H_3)$. Since it is not possible for a patient to be in $P_1$ and $P_3$ simultaneously, the FWER corrects for a multiplicity that no patient is actually affected by.
3 Control of the population-wise error rate

In this section we demonstrate how to achieve control of the PWER at a pre-specified level $\alpha$ under the general framework in Section 2.4. Suppose that each $H_i$ can be tested with a test statistic $Z_i$ where larger values of $Z_i$ speak against $H_i$. We assume further that the joint distribution of $\{Z_i\}_{i=1}^m$ is known (at least approximately). In order to control the PWER at a pre-specified significance level $\alpha \in (0, 1)$, we need to find the smallest critical value $c^* \in \mathbb{R}$ such that

$$\text{PWER}_{\theta^*} = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\theta^*} \left( \bigcup_{i \in J \cap I(\theta^*)} \{ Z_i > c^* \} \right) \leq \alpha,$$  

(3)

where $\theta^* = (\theta^*_1, \ldots, \theta^*_m)$ is the parameter configuration that maximizes the PWER and $I(\theta^*) = \{ i \in I : \theta^*_i \in H_i \}$ the index set of corresponding true null hypotheses. Usually the maximal PWER is obtained under the global null hypothesis, i.e. for $\theta^* = (0, \ldots, 0)$. If the joint distribution of the $Z_i$ is continuous, then we can reject $H_i$ also if $Z_i = c^*$, i.e. the strict inequalities in (3) can be replaced by the more familiar rules $Z_i \geq c^*$.

Since the (asymptotic) correlations between the test statistics usually depend only on the population prevalences $\pi_j$, $J \subseteq I$, the PWER-level can be exhausted under $\theta^*$. When each $H_i$ is tested by means of a $p$-value $p_i$, we can reach $\text{PWER} \leq \alpha$ by choice of an adjusted significance level $\alpha^*$ that is applied to all $p_i$.

The critical value $c^*$ in (3) or adjusted significance level $\alpha^*$ can be solved by applying a univariate root finding method. Because the PWER is always bounded by the FWER, the critical value and adjusted significance level are more liberal than the one for FWER-control. Therefore the PWER leads to a higher power and a lower sample size to achieve a certain power.

Instead of determining the critical value $c^*$ we could report the $\text{PWER}$-adjusted $p$-values

$$p_j^{\text{PWER}} = \sum_{J \subseteq I} \pi_J \mathbb{P}_{\theta^*} \left( \bigcup_{i \in J \cap I(\theta^*)} \{ Z_i > z_i^{\text{obs}} \} \right), \quad j = 1, \ldots, m,$$  

(4)

where $z_i^{\text{obs}}$ is the observed value of $Z_i$. Obviously, $p_j^{\text{PWER}} \leq \alpha$ if an only if $z_i^{\text{obs}} \geq c^*$ and hence $H_j$ can alternatively be tested with the PWER-adjusted $p$-value $p_j^{\text{PWER}}$. Furthermore, $p_j^{\text{PWER}}$ gives the smallest PWER-level the hypothesis $H_j$ can be rejected with.

Note that we could control the PWER also with population-specific critical values $\alpha^*_j$ (or adjusted levels $\alpha^*_j$). Unique solutions for $c^*_i = w_i c^*$ can be obtained by setting $c^*_i = w_i c^*$ for pre-specified weights $w_i > 0$ and searching for the $c^*$ that meets the pre-specified PWER-level. Multiplicity adjusted $p$-values can also be calculated with the weights $w_i$.

The weights may, for instance, be larger for smaller populations $\mathcal{P}_i$ in order to increase the chance of finding efficient treatment policies for small sub-populations. However, due to the weighting by $\pi_J$ in definition (1) and expression (3), the multiple type I error rate $\mathbb{P}_{\theta^*} \left( \bigcup_{i \in J \cap I(\theta^*)} \{ Z_i \geq c^* \} \right)$ for $\mathcal{P}_J$ will automatically be larger for smaller $\pi_J$. This will be illustrated by a numerical example at the end of Section 4.2. We will therefore only consider equal critical values $c^*$ in our examples below.

4 Comparison with FWER-controlling procedures

Due to the PWER being more liberal than the FWER, the next naturally arising question is how much this affects quantities like power and sample size. We will at first compare for two intersecting populations the performance of PWER-control with FWER-control when the treatments investigated in each population...
are different and when one and the same treatment is investigated in each population. Secondly, we will apply our method to the multiple testing approach for umbrella trials considered in Sun et al. (2016) and compare our results to theirs.

### 4.1 Combination of independent studies

We start with a hypothetical, but statistically simple situation. Assume that a treatment \( T \) is investigated in two intersecting populations \( \mathcal{P}_i, i = 1, 2 \), that are defined by two different biomarkers. Assume further that a sponsor has decided to test the effect of \( T \) for the two biomarker positive groups in two different but parallel clinical trials with different centres. Since the two studies are submitted as a package to regulatory authorities, a multiple testing approach is required. Let us assume that PWER-control is accepted as a compromise between control of the FWER and the unadjusted testing, the latter being the case when submitting the two studies one after another. PWER-control bounds the overall probability for a future patient to be exposed to an inefficient treatment strategy.

Since the two treatment strategies (\( \mathcal{P}_i, T \), \( i = 1, 2 \), are investigated in two independent studies, the corresponding test statistics \( Z_i \) are stochastically independent. Let us further assume that both \( Z_i \) are normally distributed with variance 1. The question is now, what we gain in terms of power by switching from FWER- to PWER-control. We will assume an overlap between the two populations \( \mathcal{P}_1 \) and \( \mathcal{P}_2 \) of probability \( \pi(1,2) \) that will be varied in our investigation.

Let \( \Phi \) and \( \Phi^{-1} \) be the standard normal distribution and quantile functions, respectively. By the independence of the test statistics, the FWER = \( 1 - \Phi(c^*_F)^2 \) is controlled at \( \alpha \) by Šidák’s critical value \( c^*_F = \Phi^{-1}(\sqrt{1-\alpha}) \). Following Example 1, the PWER is given by

\[
\text{PWER} = (1 - \pi(1,2)) \{1 - \Phi(c^*_P)\} + \pi(1,2) \{1 - \Phi(c^*_F)^2\}
\]

where \( c^*_P \) is the critical value used for control of the PWER at level \( \alpha \). Note that \( \pi(1,2) \) determines how much multiplicity adjustment is needed for PWER-control. Solving \( \text{PWER} = \alpha \) yields

\[
c^*_P = \Phi^{-1}\left(\frac{(1 - \pi(1,2)) + \sqrt{(1 - \pi(1,2))^2 + 4\pi(1,2)(1-\alpha)}}{2\pi(1,2)}\right).
\]

For \( \pi(1,2) \to 0 \) this critical value monotonically decreases to \( \Phi^{-1}(1-\alpha) \) coinciding with the unadjusted case (which is appropriate with disjoint populations) and for \( \pi(1,2) \uparrow 1 \) we have \( c^*_P \uparrow c^*_F \).

To assess the power gain by using PWER-instead of FWER-control, we consider the factor of sample size increase with PWER or FWER control in comparison to the one with no multiplicity correction. Aiming for a marginal power of at least \( 1 - \beta \), the sample size in each population \( \mathcal{P}_j \) has to be at least \( n_c \geq \left( \Phi^{-1}(1-\beta) + c \right)^2 / \delta_j^2 \) with critical value \( c \) and non-centrality parameter \( \delta_j \) in \( \mathcal{P}_j \). The fractions

\[
q_\alpha(c) := \frac{n_c}{n_{\Phi^{-1}(1-\alpha)}} = \left(\frac{\Phi^{-1}(1-\beta) + c}{\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha)}\right)^2 \quad \text{for} \quad c \in \{c^*_P, c^*_F\},
\]

describe how much more sample size one would need for a marginal power of \( 1 - \beta \) when the multiplicity adjustments are performed.

Figure 3 shows \( q_{0.025}(c) \) for \( \alpha = 0.025 \) depending on the size \( \pi(1,2) \) of \( \mathcal{P}_1 \cap \mathcal{P}_2 \) when both populations are assumed to be of equal size. FWER-control requires an increase in sample size of about 21% while PWER-control requires considerably less depending on \( \pi(1,2) \). The larger the intersection, the more patients are potentially exposed to two false rejections, therefore the critical value increases and the sample size needed to achieve a certain power value increases as well. At \( \pi(1,2) = 1 \), PWER and FWER coincide and so do the factors of sample size inflation. If, for instance, the intersection makes up 40% of the union of the two populations only around 10% sample size increase is needed when using PWER-control, less than half than what is necessary with FWER-control.
we would need to compute  

\[ \hat{\mu} \]  

the population. Hence, the population-wise means must be estimated by a weighted sum of strata-wise means:

\[ \bar{\mu}_{i,G_i} = \left( \frac{\pi_i}{\pi_i + \pi_{i,2}} \right) \bar{x}_{i,1,G_i} + \left( \frac{\pi_{i,2}}{\pi_i + \pi_{i,2}} \right) \bar{x}_{i,2,G_i}, \quad G_i \in \{T_i, C\}, \]

where \( \bar{x}_{i,G_i} \) is the mean response in strata \( \mathcal{P}_i \), \( J \subseteq \{1, 2\} \), under treatment \( G_i \). In the above example, we would need to compute \( \bar{x}_{T_i} = 0.7 \cdot \bar{x}_{T_i,1} + 0.3 \cdot \bar{x}_{T_i,2} \) for treatment \( T_i \).

The z-test statistic is finally given by \( Z_i = (\bar{x}_{i,T_i} - \hat{\mu}_{i,C}) / \sqrt{\text{Var}(\bar{x}_{i,T_i} - \hat{\mu}_{i,C})} \). Since in the intersection \( \mathcal{P}_{1,2} \) the same control group is used for both test statistics, they are positively correlated. Assuming

![Figure 3: Factor of sample size increase compared to the unadjusted case to achieve a marginal power of 80% with PWER- and FWER-control in a combination of two independent studies with different but overlapping populations.](image)

### 4.2 Testing population specific effects in one study

We consider now a single study with two overlapping populations \( \mathcal{P}_i, i = 1, 2 \), for which a treatment \( T_i \) is compared to a common control \( C \). We will investigate two possible scenarios, namely (i) \( T_1 \neq T_2 \) and (ii) \( T_1 = T_2 \). For simplicity, we assume that both populations have the same size, i.e. \( \pi_{(1)} = \pi_{(2)} \).

We assume further that the data from each population are normally distributed with mean treatment difference \( \theta_i \) and a common known variance \( \sigma^2 \) (across treatments and subgroups) and z-tests are used to test \( H_i : \theta_i \leq 0 \). For \( J \subseteq \{1, 2\} \), we denote by \( n_J = N \cdot \pi_J \) the sample size in \( \mathcal{P}_J \) and by \( N = \sum_{J \subseteq \{1, 2\}} n_J \) the overall total sample size.

In scenario (i) we have to think of a way to randomize patients to either treatment or control. In the complements \( \mathcal{P}_{(1)} \) we apply 1:1 randomization to treatment \( T_i \) or control \( C \). In the intersection \( \mathcal{P}_{1,2} \) we apply 1:1:1 randomization to the three groups \( T_1, T_2 \) and \( C \). By this we can assume that in \( \mathcal{P}_{(1)} \), there are \( n_{(1)}/2 \) patients in the treatment and control group, whereas in the intersection there are \( n_{(1,2)}/3 \) patients in each group.

Obviously, this type of allocation leads to an inconsistency between the sample and the prevalences. Say \( \mathcal{P}_1 \) has a prevalence of \( \pi_1 = \pi_{(1)} + \pi_{(1,2)} = 100/170 = 0.59 \) and of 100 patients in \( \mathcal{P}_1 \), 70 belong to \( \mathcal{P}_{(1)} \) and 30 to \( \mathcal{P}_{1,2} \). However, applying the above allocation rule implies that 35/45 \( \approx 77.7\% \) of the patients sampled from \( \mathcal{P}_i \) and assigned to treatment \( T_1 \) belong to \( \mathcal{P}_{(1)} \). This means that the proportions of the strata-wise sample sizes within a treatment group do not match their corresponding proportions in the population. Hence, the population-wise means must be estimated by a weighted sum of strata-wise means:

\[ \hat{x}_{i,G_i} = \left( \frac{\pi_i}{\pi_i + \pi_{i,2}} \right) \bar{x}_{i,1,G_i} + \left( \frac{\pi_{i,2}}{\pi_i + \pi_{i,2}} \right) \bar{x}_{i,2,G_i}, \quad G_i \in \{T_i, C\}, \]

where \( \bar{x}_{i,G_i} \) is the mean response in strata \( \mathcal{P}_i \), \( J \subseteq \{1, 2\} \), under treatment \( G_i \). In the above example, we would need to compute \( \hat{x}_{T_1} = 0.7 \cdot \bar{x}_{T_1,1} + 0.3 \cdot \bar{x}_{T_1,2} \) for treatment \( T_1 \).

The z-test statistic is finally given by \( Z_i = (\hat{x}_{i,T_i} - \hat{x}_{i,C}) / \sqrt{\text{Var}(\hat{x}_{i,T_i} - \hat{x}_{i,C})} \). Since in the intersection \( \mathcal{P}_{1,2} \) the same control group is used for both test statistics, they are positively correlated. Assuming

\[ \text{Factor of sample size increase} \]

\[ \begin{array}{c|c|c|c|c|c|c|c} \hline \text{Factor} & 0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 & 1.25 \\ \hline \text{FWER} & 1.00 & 1.05 & 1.10 & 1.15 & 1.20 & 1.25 & \text{FWER} \\ \hline \text{PWER} & 1.00 & 1.05 & 1.10 & 1.15 & 1.20 & 1.25 & \text{PWER} \\ \hline \end{array} \]
\( \pi_{(1)} = \pi_{(2)} \), we obtain \( \text{Corr}(Z_1, Z_2) = (3/2)\pi_{(1,2)}/(1 + 2\pi_{(1,2)}) \). The calculation of this correlation and an expression for the variance \( \text{Var}(\hat{x}_{1,T} - \hat{x}_{1,C}) \) can be found in Appendix B.

In scenario (ii), we investigate one and the same treatment \( T_1 = T_2 = T \) in both populations and apply the 1:1 randomization to every stratum. By this we can use for \( H_i \) the test statistic \( Z_i = (\hat{x}_{i,T} - \hat{x}_{i,C}) / \left( 2\sigma / \sqrt{n_{(1,i)} + n_{(1,2)}} \right) \). Because we are using the same treatment in both populations, we expect a higher correlation between \( Z_1 \) and \( Z_2 \). Indeed, for \( \pi_{(1)} = \pi_{(2)} \) the correlation is equal to \( \text{Corr}(Z_1, Z_2) = 2\pi_{(1,2)}/(1 + \pi_{(1,2)}) \) which is greater or equal to the correlation with different treatments for all \( \pi_{(1,2)} \in [0, 1] \); see Appendix B.

For both scenarios, we intend to find critical values to control the PWER and FWER, respectively. Following Section 2.2 the PWER under the global null is given by

\[
PWER_0 = \pi_{(1)} \mathbb{P}(\{Z_1 \geq c_p^*\}) + \pi_{(2)} \mathbb{P}(\{Z_2 \geq c_p^*\}) + \pi_{(1,2)} \mathbb{P}(\{Z_1 \geq c_p^*\} \cup \{Z_2 \geq c_p^*\}) = (1 - \pi_{(1,2)}) (1 - \Phi(c_p^*)) + \pi_{(1,2)} (1 - \Phi_p(c_{p_1}^*, c_{p_2}^*)) \]  
\( (7) \)

with \( c_p^* \) being the critical value that is to be found, and \( \Phi \) is the cumulative distribution function of the bivariate normal distribution with standard normal marginals and correlation \( \rho \). A univariate root finding algorithm can now be used to solve \( \text{PWER} = \alpha \) for \( c_p^* \).

As an example, suppose we are in scenario (i) (multiple treatments) with \( \pi_{(1)} = \pi_{(2)} = 0.4, \pi_{(1,2)} = 0.2, \beta = 0.2 \) and \( \alpha = 0.025 \). Then we have \( \rho = \text{Corr}(Z_1, Z_2) \approx 0.01 \). We solve \( \text{FWER} = 1 - \Phi_p(c_{p_1}^*, c_{p_2}^*) = \alpha \) to obtain \( c_{p_2}^* \approx 2.23 \) and \( \text{PWER} = \alpha \) to obtain \( c_{p_1}^* \approx 2.03 \). Using (8), this yields a sample size increase of around 20% for the FWER and only an increase of 5% for the PWER.

Figure 4 shows graphs of sample size increases for both types of multiple error control in dependence of \( \pi_{(1,2)} \) for both scenarios. At \( \pi_{(1,2)} = 0 \) (disjoint populations), for instance, the PWER-approach yields no sample size increase, where the FWER-based method yields an increase of over 20%. With increasing intersection size the difference between sample sizes for PWER and FWER control declines until both values fall together at \( \pi_{(1,2)} = 1 \) where the PWER is equal to the FWER. Since the correlation between the test statistics is higher in the one treatment case, the difference between the two curves is smaller for each value of \( \pi_{(1,2)} \).
In practice, however, this assumption is often not justified, so it is natural to ask whether the replacement of \( \hat{\pi} \) instead of the true prevalences, we then computed the critical value by solving

\[
\frac{1}{2} \Phi^{-1}(1-\alpha) - \rho c \Phi^{-1}(1-\alpha)
\]

where required, namely for small strata and small sub-populations.

Until now we have assumed that the prevalences \( \pi \) of the two populations in dependence of its prevalence \( \pi_{(1,2)} \).

| \( \pi_{(1,2)} \) | \( c_\rho \) | \( 1 - \Phi_\rho(c_\rho^*, c_\rho^*) \) |
|----------------|--------|----------------|
| 0.5            | 2.09   | 0.031          |
| 0.25           | 2.04   | 0.038          |
| 0.2            | 2.03   | 0.04           |
| 0.1            | 1.99   | 0.044          |
| 0.05           | 1.98   | 0.047          |

Table 1: Testing efficacy of an experimental treatment in two overlapping populations with PWER-control. Critical value \( c_\rho \) and multiple type I error probability \( 1 - \Phi_\rho(c_\rho^*, c_\rho^*) \) for the intersection \( \mathcal{P}_{(1,2)} \) of the two populations in dependence of its prevalence \( \pi_{(1,2)} \).

For the PWER, this graphic also illustrates that the correlation of the test statistics and the degree of adjustments needed to correct for multiplicity behave like opposing ‘forces’. At \( \pi_{(1,2)} = 0 \) the test statistics are uncorrelated, implying a maximum of adjustment is needed, but there is no need to adjust for multiplicity such that in summary the sample size increase factor is equal to 1. At \( \pi_{(1,2)} = 1 \) there is only one population, so the correlation is 1 which implies that no multiplicity adjustments are needed, although we are formally testing two hypotheses for everyone. In summary there is again no sample size increase required. For intersection sizes between 0 and 1 we see a maximum for the factor of sample size increase. This is because for values smaller than the maximum, the need for multiplicity-adjustment is higher than the influence of the correlation between \( Z_1 \) and \( Z_2 \). For greater values, the influence of the correlation dominates and therefore the factor is decreasing.

Mathematically, this can be seen by rewriting the PWER as:

\[
PWER = 1 - \Phi(c_\rho^*) + \pi_{(1,2)} \left\{ \Phi(c_\rho^*) - \Phi_\rho(c_\rho^*, c_\rho^*) \right\}
\]

For small values of \( \pi_{(1,2)} \) the expression \( 1 - \Phi(c_\rho^*) \), which is independent of \( \pi_{(1,2)} \), dominates which implies that the critical value is quite close to that of the unadjusted case, i.e. \( \Phi^{-1}(1-\alpha) \). The larger \( \pi_{(1,2)} \) becomes, the more influential is the expression \( \pi_{(1,2)} \left\{ \Phi(c_\rho^*) - \Phi_\rho(c_\rho^*, c_\rho^*) \right\} \) while \( \Phi(c_\rho^*) - \Phi_\rho(c_\rho^*, c_\rho^*) \) decreases at a much slower rate than \( \pi_{(1,2)} \) increases. For larger \( \pi_{(1,2)} \) this difference vanishes, since

\[
\Phi_\rho(c_\rho^*, c_\rho^*) \to \Phi(c_\rho^*)
\]

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.

We finally come back to the already mentioned consequence of PWER-control that the multiple type I error rate implicitly applied to the individual population strata is increasing with decreasing strata-prevalence. We illustrate this with scenario (ii). Aiming for a PWER-control at level \( \alpha = 0.025 \), the critical value \( c_\rho^* \) in (7) depends on \( \pi_{(1,2)} \). Table 1 shows the multiple type I error \( P_\rho(0.0)(\bigcup_{i=1}^n \{ Z_i \geq c^* \}) = 1 - \Phi_\rho(c^*, c^*) \) for \( \mathcal{P}_{(1,2)} \) with decreasing value of \( \pi_{(1,2)} \) along with the respective value of \( c_\rho^* \).

We find that this 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.

### 4.3 Estimation of population prevalences

Until now we have assumed that the prevalences \( \pi_J \) for each subset \( \mathcal{P}_J, J \subseteq I \), are known. In clinical practice, however, this assumption is often not justified, so it is natural to ask whether the replacement of \( \pi_J \) by an estimation \( \hat{\pi}_J \) will inflate the PWER by a significant amount.

We examine this behaviour by means of scenarios (i) and (ii) of Section 4.2. A suitable choice for an estimator is the maximum likelihood estimator (MLE) of the multinomial distribution \( \text{MN}(\pi, N) \) with parameters \( \pi = (\pi_J)_{J \subseteq I} \) and \( N = \sum_{J \subseteq I} n_J \). The estimators \( \hat{\pi}_J \) of \( \pi_J \) are then given by the relative frequencies \( \hat{\pi}_J = n_J/N \). For each constellation of true prevalences, we generated sample size vectors \( \{n_J\}_{J \subseteq I} \) from the \( \text{MN}(\pi, N) \)-distribution and computed the MLEs \( \{\hat{\pi}_J\}_{J \subseteq I} \). Using these estimates instead of the true prevalences, we then computed the critical value by solving \( PWER = \alpha \).
Figure 5: Contour plots of the actual PWER when using ML-estimates \( \hat{\pi}_J \) for the prevalences \( \pi_J \) in the determination of the critical value \( c^*_P \) at level \( \alpha = 0.025 \). The first row corresponds to scenario (i) and the second row to scenario (ii) from Section 4.2. Because of \( \pi_1 + \pi_2 \leq 1 \), the contour plots are restricted to the lower left rectangle of the squares.

To see by how much the true PWER is inflated by the estimation, the probabilities for a type I error for each sub-population \( \mathcal{P}_J \) are computed by using the “estimated” critical value and the respective estimated correlation structure of the involved test statistics. By weighting each of these probabilities by their respective true population prevalence \( \pi_J \), we obtain a simulated PWER by which we can assess the inflation of the true PWER due to the estimation. This procedure was repeated 10,000 times and the mean of each simulated PWER was taken. The simulations were done with R. Fig. 5 shows contour plots of these simulations for scenarios (i) and (ii) and \( N = 50 \) and \( N = 100 \), respectively. The plots indicate that the target PWER of 0.025 may be missed only slightly, even for \( N = 50 \), and that there is little to no harm to using estimated population prevalences.

4.4 Multiple testing approaches for umbrella trials

We consider now a multiple testing approach for umbrella trials suggested in Sun et al. (2016) and investigate the gain in power by switching from FWER- to PWER-control. Following Sun et al. (2016),
we assume $l$ disjoint population strata, which are denoted here by $S_1, \ldots, S_l$. In each strata a specific experimental treatment $E_i$ is compared to a control $C$. A graphical illustration for $l = 5$ is given in Fig. 6. For simplicity, we assume that each population has the prevalence $\pi_i = n_i/N$, where $n_i$ is the number of patients in $S_i$ and $N$ is the total number of patients. This holds in practice at least approximately; see also Section 4.3.

With only small $n_i$, the establishment of a treatment effect in the individual strata is difficult and impossible to achieve with sufficient power. Therefore, study designs have been suggested that compare the global treatment strategy $E$ which assigns treatment $E_i$ to population strata $S_i$, as a total with the control treatment in the overall population. Such an overall comparison of the strategy $E$ with $C$ utilizes the total sample size $N$ and does also not require multiple testing. However, it does not permit a claim for a sub-population when the effect of $E$ is heterogeneous. To improve the approach, Sun et al. (2016) suggest to test all sub-strategies $E^S$, $S \subseteq \{1, \ldots, l\}$, that consider only the union $P^S = \cup_{i \in S} S_i$ with treatment assignments as in $E$, against the control in $P^S$. This permits claims also for sub-populations and thereby increases the possibility for the efficacy conclusions. Of course, such testing requires an adjustment for multiplicity. Sun et al. (2016) provide a (single-step) procedure that controls the FWER.

For the formal description of the procedures, let $\theta = (\theta_1, \ldots, \theta_l)$ be the vector of unknown treatment effects (mean differences) in the populations, and consider for each $S \subseteq \{1, \ldots, l\}$ the average treatment effect in $P^S$:

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

with $\pi^S = \sum_{i \in S} \pi_i$ the prevalence of $P^S$. Sun et al. (2016) assume the linear model

$$Y_{ij} = \mu_i + \theta_i X_{ij} + \varepsilon_{ij},$$

where $X_{ij}$ denotes the treatment indicator for patient $j$ in group $i$ which equals 1 if assigned to the experimental treatment $E_i$ and otherwise 0, and $\theta_i$ is the treatment effect of $E_i$ in population $S_i$. The error terms $\varepsilon_{ij}$ are assumed to be i.i.d. normally distributed with mean 0 and homogeneous variance $\sigma^2$.

As mentioned above, the authors suggest to test

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

Note that the $P^S$ and $H^S$, $S \subseteq L$, correspond to the $P_i$ and $H_i$, $i \in I$, in Section 2 and 3.

From the least squares estimate of the linear model, we obtain one-sided t-test statistics $T^S$ for testing $H^S$ for each $S \subseteq L$. In order to control the FWER we can conduct a single-step procedure that compares each $T^S$ with the upper $\alpha$-quantile $c^*_F$ of the distribution of $\max \{T^S \mid S \subseteq L\}$ under the global null hypothesis, i.e. the assumption that none of the treatments $E_i$ is superior to the control. We finally select the subset $S^*_F \subseteq L$ for which a positive treatment effect is claimed and that yields the largest value of $T^S$,

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

To achieve $PWER = \alpha$ under the global null hypotheses, we introduced the $PWER$ in a setting where populations are overlapping. While $S_1, \ldots, S_l$ are disjoint, their unions $P^S$ can overlap. Since some of the $P^S$ overlap and some do not, the $PWER$ corrects the multiple type I error rate for cases that cannot occur (similar to example 3) and hence may be viewed as overly conservative.

The $PWER$ under the global null hypothesis ($\theta = 0 = (0, \ldots, 0)$) is given by

$$PWER_0 = \sum_{i=1}^m \pi_i p_0 \left( \bigcup_{S \subseteq I} \left\{ T^S \geq c^*_F \right\} \right),$$

where $c^*_F$ is the upper $\alpha$-quantile of the distribution of $\max \{T^S \mid S \subseteq L\}$ under the global null hypothesis.
where \( S \ni i \) denotes all \( S \subseteq L \) that contain the index \( i \). This is because population \( S_i \) is affected by a type I error whenever a hypothesis \( H^S \) is erroneously rejected that corresponds to a population \( P^S \) for which \( i \in S \) (or \( S_i \subseteq P^S \)).

Due to the assumption of a homogeneous residual variance and the \( 2l \) mean parameter in the linear model (8), \( \{ T^S \}_{S \subseteq L} \) follows a joint t-distribution with \( N-2l \) degrees of freedom. In R, the distribution function of the multivariate t-distribution is implemented in the `mvtnorm`-package (see Genz et al., 2017) via the `pmvt()` function and needs the degrees of freedom \( df \) and the correlation matrix \( corr \) of the test statistics as arguments (see e.g. Bretz et al., 2016). The correlation matrix can be computed using the contrast matrix and the design matrix of the linear model. Probabilities in (11) are then calculated by choosing the appropriate sub-matrices of the correlation matrix. Thus, for known values of \( \pi_i, i \in L \), and \( l \), we can numerically determine the critical value \( c^*_P \) such that \( P_{WPE} = \alpha \).

We know that \( c^*_F > c^*_P \), which implies that whenever the FWER-approach selects a non-empty \( S^*_F \), the same set is selected by the PWER-approach, \( S^*_P = S^*_F \). We may, however, select the empty set with the FWER-approach, \( S^*_F = \emptyset \), while \( S^*_P \neq \emptyset \).

### 4.4.1 Performance measures

Sun et al. (2016) examined several quality and performance measures to assess how good a selected subset \( S^* \) is. For example, they considered the average effect in the overall population when applying treatment strategy \( E^{S^*} \) in \( S^* \) and the control in the rest of the population. We will consider the relative quantity \( RAE = 100 \frac{E(\sum_{i \in S^*} \pi_i \theta_i)}{\theta_{overall}} \) where \( E \) is the expectation with respect to the sample distribution. Since the PWER-procedure chooses a non-empty \( S^* \) more often than the FWER-procedure, this quantity will always be larger for the PWER-procedure.

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 \( \pi^S_{S^*} / \pi^S \) where \( S^*_+ = \{ i \in S^+ \theta_i > 0 \} \) and \( \pi^S_{S^*} = \sum_{i \in S^*_+} \pi_i \). 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. Anlogously, 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 \( S^*_0 = \{ i \in S^+ \theta_i = 0 \} \). Lastly, we consider the probability of rejecting at least one false null hypothesis,

\[
\text{Power} = \mathbb{P}(\text{reject any } H^S \text{ with } \theta^S > 0, S \subseteq L),
\]

as a way to measure the power of the test procedures.

### 4.4.2 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 \( l = 2, 4, 6 \) sub-populations and a significance level \( \alpha = 0.025 \), we chose a total sample size if \( N = 1056 \) and assume that all group-specific intercepts \( \mu_i \) are equal to 0.
Also, for simplicity, each group is assumed to be of equal size, i.e. \( \pi_1 = \cdots = \pi_l \). All simulations were done in R.

As in Sun et al. (2016), we assume non-negative effects \( \theta_i \geq 0 \) and choose \( \theta = (\theta_1, \ldots, \theta_l) \) based on the number of subgroups \( l \) and three further characteristics. The first one is the percentage of true null hypothesis: \( q = l_0/l \) with \( l_0 \) the size of \( L_0 = \{i = 1, \ldots, l : \theta_i = 0\} \). The second one is a weighted average of the positive treatment effects,

\[
\theta_{\text{overall}} = \frac{\sum_{i \in L_+} \pi_i \theta_i}{\sum_{i \in L_+} \pi_i}
\]

for \( L_+ = \{i = 1, \ldots, l : \theta_i > 0\} \), that describes how efficient the experimental treatment strategy \( E \) is for the union of sub-populations that benefit from \( E \). The third one characterizes the treatment effect heterogeneity and is defined as

\[
\tau = (\theta_{\text{max}} - \theta_{\text{min}})/(\theta_{\text{max}} + \theta_{\text{min}})
\]

where \( \theta_{\text{max}} = \max_{i \in L_+} \theta_i \) and \( \theta_{\text{min}} = \min_{i \in L_+} \theta_i \). Note that \( \tau \) equals the relative half-range of the positive \( \theta_i \)'s, i.e. half of their range divided by the average of their extremes. Obviously, a large \( \tau \) means a large heterogeneity between the positive \( \theta_i \).

Given values for \( l, q, \theta_{\text{overall}} \) and \( \tau \) one finds a grid of \( l \) equidistant points such that the three characteristics are met. One easily verifies, that this grid is uniquely determined by the four quantities. We chose \( q \) such that \( q \cdot l \) is always an integer.

### 4.4.3 Results

The simulation results for \( l = 2 \) and \( 4 \) are given in Table 2 and for \( l = 6 \) and \( 8 \) in Appendix C. One can see from the tables that control of the PWER, in comparison to FWER-control, provides a substantially larger power and larger average proportion of ‘correctly’ chosen subgroups and a larger average effect. It also increases the proportion of ‘falsely’ chosen subgroups. This is 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) 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%. The expected effect RAE is always larger with PWER-control. This difference is determined by the difference in the frequency of choosing a non-empty \( S^\ast \).

Under the global null hypothesis (\( P = 1 \)) the average proportion of ‘falsely’ selected populations equals by theory the one-sided family-wise error rate. With PWER-control at level 2.5% the FWER was found to be between 3.6% and 4.5% for \( l = 2, 4, 6, 8 \). Note that the average proportion of ‘falsely’ selected populations exceeds the level of 2.5% (sometimes substantially) also with FWER-control when there is an effect in some but not all population strata.

In summary, we see that control of PWER substantially increases the chance for a delivery of efficient treatments to the populations while the risk of receiving an inefficient treatment and the percentage of patients that do not benefit from the treatment decisions is increased to a moderate extend and remains comparable to the procedure with FWER-control.

### 5 Extension to simultaneous confidence intervals

We are coming back to the general set-up of Section 2 and 3. Utilizing the duality between (multiple) hypothesis tests and (simultaneous) confidence intervals, the multiple test procedure with control of the PWER, introduced in Section 3, can be extended to confidence intervals for the efficacy parameter \( \theta_i = \ldots = \theta_1 \).
\( \theta(P_i, T_i) \), \( i = 1, \ldots, m \). In this section we will introduce the dual simultaneous confidence intervals and discuss their coverage properties.

To introduce the confidence intervals, let \( \delta = (\delta_1, \ldots, \delta_m) \) be a vector of possible values for \( \theta = (\theta_1, \ldots, \theta_m) \) and consider the corresponding null hypotheses \( H_{\delta}^{i} : \theta_i = \delta_i, i = 1, \ldots, m \). Assume further that \( T_{\theta}^{i} \), \( i = 1, \ldots, m \), are (asymptotically) pivotal test statistics for \( H_{\delta}^{i} \), i.e., the (asymptotic) joint distribution of \( (T_{1}^{\delta_{1}}, \ldots, T_{m}^{\delta_{m}}) \) under \( \theta = \delta \) is the same for all \( \delta \). If \( T_{\theta}^{i} \) decreases in \( \delta_i \) for the given data, then it makes sense to form the one-sided intervals \( C_i = [\tilde{\theta}_i \pm \delta SE_i] \) with an lower bound

\[
\tilde{\theta}_i := \min \{ \delta_i : T_{\theta}^{i} \leq c^* \}
\]

where \( c^* \) is the critical value defined in (3) for \( \theta^* = \delta \). Because \( (T_{1}^{\delta_{1}}, \ldots, T_{m}^{\delta_{m}}) \) is pivotal, the critical value \( c^* \) is independent from \( \delta \). The monotonicity of \( T_{\theta}^{i} \) applies to most (one-sided) tests and is satisfied e.g. for Wald-type test statistics \( T_{\theta}^{i} = (\bar{\theta}_i - \tilde{\delta}_i) / SE_i \) where \( \tilde{\delta}_i \) is an estimate of \( \delta_i \) (usually the maximum likelihood estimate) with an standard error \( SE_i \) that is independent of the parameter value \( \delta \). In this case we obtain \( \tilde{\theta}_i = \tilde{\theta}_i - c^* SE_i \).

Upper confidence bounds can be derived by applying the same principle to the parameter \( -\theta = (-\theta_1, \ldots, -\theta_m) \) and two-sided confidence intervals are obtained by the intersection of the two one-sided intervals. In particular, if the distribution of \( (\tilde{T}_{\theta}^{i}, \ldots, \tilde{T}_{m}^{i}) \) under \( \theta = -\delta \) is the same as the distribution of \( (T_{1}^{\delta_{1}}, \ldots, T_{m}^{\delta_{m}}) \) under \( \theta = \delta \), then two-sided intervals can be directly obtained by applying \( c^* \) to the absolute test statistics \( |T_{\theta}^{i}| \). With Wald-type dual tests we obtain the two-sided intervals

\[
C_i = [\tilde{\theta}_i - c^* SE_i, \tilde{\theta}_i + c^* SE_i].
\]

We finally discuss the coverage properties of the above introduced confidence bounds and intervals. We start with the lower confidence bounds \( \tilde{\theta}_i \). To this end, consider a patient \( P \) that is randomly drawn from \( P \) and let \( I_P \) be the set of indices of the sub-populations \( S_i \) the patient \( P \) belongs to, i.e. \( I_P = \{i : P \in S_i\} \). The set \( I_P \) gives all population efficacy parameter \( \theta_i, i \in I_P \), that are relevant for patient \( P \). Note that \( I_P \) is a random set, because \( P \) is randomly drawn from \( P \). If \( \tilde{\theta}_i \) is the true unknown efficacy parameter, then by the definition (12) we get \( \tilde{\theta}_i > \theta_j \) if and only if \( T_{\theta}^{i} > c^* \). Since the dual tests for \( H_{\delta}^{1}, \ldots, H_{\delta}^{m} \) control the POWER, the (simultaneous) probability that any of the lower confidence bounds \( \tilde{\theta}_j, j \in I_P \) fall above the true \( \theta_j \) is at most \( \alpha \). This gives the coverage property

\[
\mathbb{P}_\theta \left( \tilde{\theta}_j \leq \theta_j \text{ for all } j \in I_P \right) \geq 1 - \alpha
\]

meaning that with a probability of at most \( 1 - \alpha \), for a randomly chosen patient, the lower confidence intervals \( [\tilde{\theta}_j, \infty], i = 1, \ldots, m \), cover all true \( \theta_j \) (in \( S_i \) for \( j \in I_P \)). Hence, equation (12) means to control a kind of average 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 \( P \).

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.
Table 2: Simulation results for \( l = 2 \) and \( l = 4 \). 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 \( \theta = (\theta_1, \ldots, \theta_l) \) that depend on the fraction of true nulls \( q \) and the relative half-range \( \tau \) of the positive \( \theta_i \)'s.

| \( \tau \) | Power correct | false | RAE | Power correct | false | RAE |
|-----------|---------------|-------|-----|---------------|-------|-----|
| \( \tau = 0 \) | 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 |
| \( l = 2 \) | \( q = 1/2 \) | \( q = 1 \) |
| \( \tau = 0 \) | PWER 57.7 | 52.9 | 4.8 | 5.8 | 0 | 0 | 3.6 | 0 |
| \( \tau = 0.4 \) | PWER 52.0 | 47.8 | 4.2 | 5.2 | 0 | 0 | 2.4 | 0 |
| \( \tau = 0.8 \) | PWER 43.0 | 43.0 | 0 | 3.0 | 52.7 | 48.9 | 3.7 | 4.2 |
| \( l = 4 \) | \( q = 0 \) | \( q = 1/4 \) |
| \( \tau = 0 \) | PWER 36.2 | 36.2 | 0 | 2.3 | 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 |
| \( \tau = 0.8 \) | PWER 43.0 | 43.0 | 0 | 3.0 | 52.7 | 48.9 | 3.7 | 4.2 |
| \( \tau = 0 \) | FWER 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 |
| \( \tau = 0 \) | PWER 81.5 | 70.1 | 11.5 | 8.1 | 0 | 0 | 4.2 | 0 |
| \( \tau = 0.4 \) | FWER 75.1 | 64.9 | 10.2 | 7.5 | 0 | 0 | 2.4 | 0 |
6 Discussion

With this paper we have introduced a new multiple type I error rate concept for clinical trials with multiple and possibly intersecting populations that permits for more liberal and more powerful tests than control of the family-wise error rate. It relies on the observation that not all patients and sub-population strata are affected by all test decisions, since not all hypotheses concern all patients or patient strata. By averaging the individually relevant, multiple type I errors over the entire population, we provide control of the probability that a randomly selected patient will be exposed to an inefficient treatment strategy. This average multiple type I error rate, which we call the population-wise error rate (PWER), is more liberal than the family-wise error rate (FWER), because the latter equals or is sometimes even larger than the maximum multiple type I error rate a patient is exposed to.

We would like to recall at this point, that we only consider population-wise claims, i.e. claims on treatment strategies that consist of a treatment and a population the treatment is intended for and for which the average treatment effect is the estimand of interest. This is also the case when going for FWER control. No individual efficacy claims are anticipated here. Error control of patient-wise claims is impossible without sacrificing power or making strong assumptions. However, a population-wise claim can be viewed as a proxy or approximation for individual claims in the target population. Test results from more than a single population augment this information and may be used for more informed and sophisticated 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. In this sense, control of the PWER is a conservative approach which could only be improved by (usually unavailable) information on how treatments will be applied in future medical practice. Note that we do not account for a potential off-label use where a treatment is applied to patients outside its target population.

We have presented a simple and straightforward approach for achieving control of the PWER by an adjustment of critical boundaries and have illustrated the power gain achieved when passing from FWER to PWER control in a number of examples. We have considered the simple situation of multivariate normal distributed test statistics. This situation applies at least asymptotically to a large number of hypothesis tests for which PWER control is then guaranteed asymptotically. The methods and principle introduced here can also be implemented with finite sample distributions like e.g. the multivariate t-distribution (as done in Section 4.4) or be improved via resampling methods. Variance heterogeneity across populations is a general issue for trials with multiple populations that applies similarly to procedures with FWER control (see e.g. Placzek and Friede, 2019). One can say, whenever control of the FWER is possible then control of the PWER is possible as well, since the latter just controls an average of family-error rates. We have also extended the suggested multiple test to simultaneous confidence intervals and showed that these intervals control, for a randomly chosen patient, the probability of a simultaneously correct statement on the parameters that are relevant for this individual.

Control of the PWER requires the knowledge of the prevalences of all disjoint population strata. These may either be obtained from previous studies or may be estimated at the end of the study. This complicates PWER control. We have illustrated in an example with two populations that the estimation of the prevalences does not strongly harm PWER control even with moderate sample sizes. However, more examples with more hypotheses are required to fully explore this issue. At least, PWER control is always guaranteed asymptotically.

Since our procedure simply results in an adjustment of critical values, power calculations and power simulations are straightforward and deviate only minimally from approaches for classical multiple tests, except for the fact that the critical values may depend on the sample via the prevalence estimates. This can be resolved by using a priori estimates of the prevalences based on experience and past studies. The same issues arises from the estimation of the correlation structure of the test statics used for an efficient PWER and FWER control. A miss-specification of the prevalences can be corrected in a mid-trial blinded
In Section 4.1 we have suggested a single-step procedure to control the PWER and one might ask whether this procedure can be uniformly improved by a step-down test because this is the case for single-step tests with FWER control (e.g. Dmitrienko et al., 2009). For instance, in Example 2.2 with two intersecting hypotheses, we may ask whether we can test \( H_2 \) with a smaller critical \( c^*_2 < c^* \) when \( H_1 \) has already been rejected with critical value \( c^* \). One can quickly see that this is not possible. To this end assume that both hypotheses \( H_1 \) and \( H_2 \) are true. Rejection of \( H_2 \) when \( Z_2 \geq c^* \) or \( Z_2 \geq c^*_2 \) with \( Z_1 \geq c^* \), obviously increases the second and third terms in equation (2) of the PWER. Since we have chosen \( c^* \) to be the smallest critical value that satisfies (3), which leads to an PWER equal to \( \alpha \) with continuously distributed \( Z_1 \) (a generic and common situation), we do not control the PWER for any \( c^*_2 < c^* \). We may define PWER-controlling step-down tests with an enlarged \( c^* \) in order to mimic and improve step-down tests with FWER control. However, such procedures do not uniformly improve the single-step test with FWER control and are therefore beyond the scope of this paper. The development of step-down tests with PWER control is a topic of future research.

Single-step procedures have the advantage that they can directly be extended by simple and well behaving simultaneous confidence intervals (SCIs). We have illustrated this in Section 5 for single-step tests with PWER control. An extension to simple and well behaving SCIs is impossible for step-down tests: Compatible SCIs often are non-informative in the sense that they do not provide any additional information to the sheer hypothesis tests (Strassburger and Bretz, 2008; Guildbaud, 2009) and sufficiently informative SCIs are compatible only to a modification of the original step-down test (Brannath and Schmidt, 2014). This justifies the use of single step tests in practice.

We finally remark that an extension of the presented PWER approach to multi-stage and adaptive designs is under development by the authors and will be a topic of future contributions. Multi-stage and particularly flexible designs provide the opportunity for adding or dropping populations at interim analyses based on the unblinded interim data (e.g. Brannath et al., 2009; Wassmer and Brannath, 2016; Placzek and Friede, 2019). In the example of Section 2.2 we may for instance add and enrich the intersection of the two populations for an investigation in a second stage of the study if the efficacy of the treatment is seen at interim only in one of the two populations. Hence, the development of adaptive and sequential designs with PWER control is an interesting and valuable research task.

Acknowledgements

This research was supported by the BMBF under the funding number 01EK1503B.

A Derivation of \( c^*_s \) in Section 4.1

The solutions of the quadratic equation \( 1 - \Phi(c^*_s) = \pi_{1,2}s^2 \) are:

\[
x_{1/2} = \frac{(1 - \pi_{1,2}) \pm \sqrt{(1 - \pi_{1,2})^2 + 4\pi_{1,2}(1 - \alpha)}}{2\pi_{1,2}}.
\]

Since \( \sqrt{(1 - \pi_{1,2})^2 + 4\pi_{1,2}(1 - \alpha)} > (1 - \pi_{1,2}) \) for all \( \pi_{1,2} \in (0, 1] \) it follows that

\[
c^*_s = \Phi^{-1}(x_2) = \Phi^{-1}\left(\frac{(1 - \pi_{1,2}) + \sqrt{(1 - \pi_{1,2})^2 + 4\pi_{1,2}(1 - \alpha)}}{2\pi_{1,2}}\right)
\]

is the only valid solution. To show that \( c^*_s \) is strictly monotonically increasing in \( \pi_{1,2} \), we consider the function \( y = y(\pi_{1,2}) = \Phi(c^*_s) \in (0, 1) \) which satisfies the equation

\[
\alpha = 1 - y + \pi_{1,2}y - \pi_{1,2}y^2.
\]
Taking derivatives w.r.t. \( \pi_{(1,2)} \) on both sides of this equation yields after a rearrangement of terms:

\[
y' = y(1 - y)/\{1 + \pi_{(1,2)}(2y - 1)\}.
\]

Due to \( \pi_{(1,2)}(2y - 1) \geq -1 \) it follows that \( y' > 0 \) for all \( \pi_{(1,2)} \in (0, 1) \).

**B Calculation of the correlation expressions in Section 4.2**

The variance of \( \bar{x}_{T_i} - \bar{x}_{C_i} \), is easily found by exploiting the independence of the individual observations and is given by \( \text{Var}(\bar{x}_{T_i} - \bar{x}_{C_i}) = (2\sigma^2/N)v_i^2 \) with

\[
v_i^2 = (\pi_{(1)}/\pi_i)^2 (2/\pi_{(1)}) + (\pi_{(1,2)}/\pi_i)^2 (3/\pi_{(1,2)}) \quad \text{where} \quad \pi_i = \pi_{(1)} + \pi_{(1,2)}.
\]

We turn to the correlation between \( Z_1 \) and \( Z_2 \) for the case (i) of two different treatments \( T_1 \neq T_2 \). By the independence of means from disjoint cohorts, we calculate

\[
\text{Cov}(Z_1, Z_2) = \frac{\text{Cov}(\bar{x}_{T_1} - \bar{x}_{C_1}, \bar{x}_{T_2} - \bar{x}_{C_2})}{2\sigma^2 v_1v_2/N} = \frac{\text{Cov}(\bar{x}_{C_1}, \bar{x}_{C_2})}{2\sigma^2 v_1v_2/N} = \frac{\pi^2_{(1,2)}}{2\sigma^2 v_1v_2/N} = \frac{3\pi_{(1,2)}}{2\pi_1\pi_2 v_1v_2}
\]

where we used \( n_J = N\pi_J \) for \( J \in \{1, 2\} \) in the last equation. Now, if \( \pi_{(1)} = \pi_{(2)} \) then \( \pi_{(1)} = \pi_{(2)} = (1 - \pi_{(1,2)})/2 \) and \( \pi_1 = \pi_2 = (1 + \pi_{(1,2)})/2 \), and the correlation reduces to

\[
\text{Cov}(Z_1, Z_2) = \frac{6\pi_{(1,2)}}{(1 + \pi_{(1,2)})^2 \left\{ \frac{1 - \pi_{(2)}}{1 + \pi_{(1,2)}} \right\}^2 \left( \frac{4}{1 + \pi_{(1,2)}} \right)^2 \left( \frac{3/\pi_{(1,2)}}{\pi_{(1,2)}} \right)}
\]

\[
= \frac{6\pi_{(1,2)}}{4(1 - \pi_{(1,2)}) + 12\pi_{(1,2)}} = \frac{3\pi_{(1,2)}}{2(1 + 2\pi_{(1,2)})}
\]

For case (ii), where \( T_1 = T_2 = T \) we calculate

\[
\text{Cov}(Z_1, Z_2) = \text{Cov}(\bar{x}_{T_1} - \bar{x}_{C_1}, \bar{x}_{T_2} - \bar{x}_{C_2}) \sqrt{n_1 n_2}/4\sigma^2 = \frac{n^2_{(1,2)}}{4\sigma^2/\sqrt{n_1 n_2}} = \frac{n_{(1,2)}}{\sqrt{n_1 n_2}} = \frac{\pi_{(1,2)}}{\sqrt{n_1 n_2}}
\]

and for \( \pi_{(1)} = \pi_{(2)} \) we obtain \( \text{Corr}(Z_1, Z_2) = 2\pi_{(1,2)}/(1 + \pi_{(1,2)}) \). Obviously, this correlation is greater than the one from case (i) for all \( \pi_{(1,2)} \in [0, 1] \).
### Further simulation results

|      | Power | correct | false | RAE   | Power | correct | false | RAE   |
|------|-------|---------|-------|-------|-------|---------|-------|-------|
| $l = 6$ |       |         |       |       |       |         |       |       |
| $q = 0$ |       |         |       |       |       |         |       |       |
| $\tau = 0$ |       |         |       |       |       |         |       |       |
| 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    |
| $q = 1/6$ |       |         |       |       |       |         |       |       |
| $\tau = 0$ |       |         |       |       |       |         |       |       |
| 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    |
| $q = 2/6$ |       |         |       |       |       |         |       |       |
| $\tau = 0$ |       |         |       |       |       |         |       |       |
| 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  | 6.2   | 87.2  | 75.2    | 12.0  | 87    |
| FWER | 61.9  | 50.8    | 11.1  | 5.3   | 78.9  | 68.5    | 10.4  | 72    |
| $\tau = 0.8$ |       |         |       |       |       |         |       |       |
| PWER | 85.8  | 74.1    | 11.8  | 7.9   |       |         |       |       |
| FWER | 78.9  | 68.5    | 10.4  | 7.2   |       |         |       |       |
| $q = 3/6$ |       |         |       |       |       |         |       |       |
| $\tau = 0$ |       |         |       |       |       |         |       |       |
| PWER | 0     | 0       | 4.2   | 0     |       |         |       |       |
| FWER | 0     | 0       | 2.2   | 0     |       |         |       |       |

19
\[
\begin{array}{cccccccc}
 l = 8 & \text{Power} & \text{correct} & \text{false} & \text{RAE} & \text{Power} & \text{correct} & \text{false} & \text{RAE} \\
 & q = 0 & q = 1/8 & & & q = 2/8 & q = 3/8 & & \\
\tau = 0 & \text{POWER} & 34.0 & 34.0 & 0 & 1.8 & 35.6 & 33.5 & 2.2 & 19 \\
& & 24.3 & 24.3 & 0 & 1.3 & 26.2 & 24.6 & 1.6 & 14 \\
\tau = 0.4 & \text{POWER} & 34.8 & 34.8 & 0 & 1.9 & 36.9 & 34.7 & 2.2 & 21 \\
& & 25.5 & 25.5 & 0 & 1.4 & 27.0 & 25.4 & 1.6 & 15 \\
\tau = 0.8 & \text{POWER} & 37.4 & 37.4 & 0 & 2.1 & 40.0 & 37.7 & 2.3 & 24 \\
& & 27.9 & 27.9 & 0 & 1.6 & 30.4 & 28.7 & 1.7 & 19 \\
\tau = 0.4 & \text{POWER} & 38.0 & 33.5 & 4.5 & 2.2 & 41.3 & 34.2 & 7.1 & 26 \\
& & 28.4 & 25.1 & 3.3 & 1.7 & 31.1 & 25.9 & 5.2 & 20 \\
\tau = 0.8 & \text{POWER} & 39.6 & 35.0 & 4.6 & 2.4 & 43.6 & 36.4 & 7.3 & 29 \\
& & 29.8 & 26.4 & 3.4 & 1.8 & 33.1 & 27.5 & 5.5 & 22 \\
\tau = 0 & \text{POWER} & 44.0 & 39.0 & 5.0 & 2.9 & 49.3 & 41.4 & 7.9 & 35 \\
& & 33.6 & 29.8 & 3.8 & 2.2 & 38.4 & 32.3 & 6.1 & 27 \\
\tau = 0.4 & \text{POWER} & 46.9 & 36.9 & 10.0 & 3.3 & 56.0 & 42.3 & 13.7 & 45 \\
& & 36.7 & 28.9 & 7.8 & 2.6 & 45.9 & 34.7 & 11.2 & 38 \\
\tau = 0.8 & \text{POWER} & 49.5 & 39.1 & 10.5 & 3.6 & 59.7 & 45.5 & 14.3 & 48 \\
& & 39.3 & 31.1 & 8.2 & 2.9 & 49.5 & 37.8 & 11.7 & 41 \\
\tau = 0 & \text{POWER} & 57.4 & 46.0 & 11.5 & 4.4 & 70.4 & 55.3 & 15.1 & 59 \\
& & 46.6 & 37.4 & 9.2 & 3.6 & 60.9 & 48.1 & 12.8 & 51 \\
\tau = 0.4 & \text{POWER} & 72.9 & 54.6 & 18.2 & 6.8 & 96.3 & 83.8 & 12.5 & 96 \\
& & 63.8 & 48.1 & 15.7 & 6.0 & 93.4 & 81.6 & 11.8 & 93 \\
\tau = 0.8 & \text{POWER} & 79.0 & 61.5 & 17.5 & 6.9 & 93.4 & 81.6 & 11.8 & 93 \\
& & 70.4 & 55.1 & 15.3 & 6.1 & 90.8 & 80.2 & 11.5 & 91 \\
\tau = 0 & \text{POWER} & 91.3 & 77.6 & 13.8 & 8.4 & 98.2 & 86.9 & 13.8 & 84 \\
& & 86.9 & 74.2 & 12.7 & 7.9 & 96.3 & 83.8 & 12.5 & 96 \\
\tau = 0 & \text{POWER} & 0 & 0 & 4.5 & 0 & 0 & 0 & 2.3 & 0 \\
\end{array}
\]

References

Brannath, W. and S. Schmidt (2014). A new class of powerful and informative simultaneous confidence intervals. *Statistics in Medicine* 33(19), 3365–3386.

Brannath, W., E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon (2009). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 28(10), 1445–1463.
Bretz, F., T. Hothorn, and P. Westfall (2016). *Multiple Comparisons Using R*. CRC Press.

Collignon, O., C. Gartner, A.-B. Haidich, R. J. Hemmings, B. Hofner, F. Pétavy, M. Posch, K. Rantell, K. Roes, and A. Schiel (2020). Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials. *Clinical Pharmacology and Therapeutics* 107(5), 1059–1067.

Dmitrienko, A., A. Tamhane, and F. Bretz (2009). *Multiple testing problems in pharmaceutical statistics*. CRC Press.

Fletcher, J. I., D. S. Ziegler, T. N. Trahair, G. M. Marshall, M. Haber, and M. D. Norris (2018, June). Too many targets, not enough patients: rethinking neuroblastoma clinical trials. *Nature reviews. Cancer* 18(6), 389–400.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2017). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6.

Glimm, E. and L. Di Scala (2015). An approach to confirmatory testing of subpopulations in clinical trials. *Biometrical Journal* 57(5), 897–913.

Guilbaud, O. (2009). Alternative confidence regions for bonferroni-based closed-testing procedures that are not alpha-exhaustive. *Biometrical Journal* 51(4), 721–735.

Kaplan, R., T. Maughan, A. Crook, D. Fisher, R. Wilson, L. Brown, and M. Parmar (2013). Evaluating many treatments and biomarkers in oncology: A new design. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31.

Liu, K. and X.-L. Meng (2014). Comment: A fruitful resolution to simpson’s paradox via multiresolution inference. *The American Statistician* 68, 17 – 29.

Malik, S. M., R. Pazdur, J. S. Abrams, M. A. Socinski, W. T. Sause, D. H. Harpole, J. J. Welch, E. L. Korn, C. D. Ullmann, and F. R. Hirsch (2014). 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. *Journal of Thoracic Oncology* 9(10), 1443 – 1448.

Placzek, M. and T. Friede (2018). Clinical trials with nested subgroups: Analysis, sample size determination and internal pilot studies. *Statistical Methods in Medical Research* 27(11), 3286–3303.

Placzek, M. and T. Friede (2019). A conditional error function approach for adaptive enrichment designs with continuous endpoints. *Statistics in Medicine* 38(17), 3105–3122.

Strassburger, K. and F. Bretz (2008). Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. *Statistics in Medicine* 27(24), 4914–4927.

Strzebonska, K. and M. Waligora (2019). Umbrella and basket trials in oncology: Ethical challenges. *BMC Medical Ethics* 20.

Sun, H., F. Bretz, O. Gerke, and W. Vach (2016). Comparing a stratified treatment strategy with the standard treatment in randomized clinical trials. *Statistics in Medicine* 35(29), 5325–5337.

Wassmer, G. and W. Brannath (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Springer.

Woodcock, J. and L. M. LaVange (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine* 377(1), 62–70. PMID: 28679092.

21