Methods for Blinded Sample Size Recalculation in Adaptive Enrichment Designs

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Dissertation zur Erlangung des humanwissenschaftlichen Doktorgrades in der Medizin der Georg-August-Universität Göttingen

UNIVERSITÄTSMEDIZIN
GÖTTINGEN UMG

vorgelegt von

Marius Placzek

aus Fritzlar, Deutschland

Göttingen, 2022

Thesis Committee

First Supervisor (Reviewer 1):

Prof. Dr. Tim Friede
Department of Medical Statistics
University Medical Center Göttingen
Georg-August-University Göttingen

Second Committee Member:

Prof. Dr. Heike Bickeböller
Institute of Genetic Epidemiology
University Medical Center Göttingen
Georg-August-University Göttingen

Third Committee Member (Reviewer 2):

Prof. Dr. Thomas Kneib
Spokesperson of the Centre for Statistics
Dean of research at the Faculty of Business and Economic Sciences
Georg-August-University Göttingen

Further members of the examination board

Prof. Dr. Tim Beißbarth
Department of Medical Bioinformatics
University Medical Center Göttingen
Georg-August-University Göttingen

Prof. Dr. Thomas Meyer
Clinic for Psychosomatic Medicine and Psychotherapy
University Medical Center Göttingen
Georg-August-University Göttingen

Prof. Dr. Markus Zabel
Clinic for Cardiology and Pneumology
University Medical Center Göttingen
Georg-August-University Göttingen

Date of the oral examination: 26.10.2022

Dedication and Acknowledgements

First and foremost I would like to thank my supervisor Tim Friede for supporting me throughout the development of this dissertation, especially for our joint work on the three manuscripts presented here. His ideas and input greatly helped improving and publishing those papers. Further thanks go to my second and third supervisors, Heike Bickeböller and Thomas Kneib, for being part of the Thesis Committee. Special thanks also to my colleagues and fellow PhD students, especially Thomas Asendorf and Roland Gera, for the great working atmosphere under which this dissertation was developed. Last but not least, I would like to thank my friends and family for the support throughout the years.

Contents

1 Introduction

1.1 Clinical Trials in Personalized Medicine

In precision medicine the potential heterogeneity of patient populations becomes more accessible through technical and medical innovations. A population might be heterogeneous in terms of different phenotypes or biomarkers, e.g. genetic markers, and different strata (biomarker positive vs biomarker negative) may in turn respond heterogeneously to a certain treatment. A certain proportion of the population might respond not at all or even negatively to the investigated treatment. Personalized medicine and targeted therapies aim to use this information to find tailored treatment for subjects depending on subgroup affiliation, increasing the success rate of treated subjects. Therefore, clinical trials with the ability to reveal an increased treatment benefit in particular subgroups compared to the whole population are in great demand.

For example, take Pulmonary arterial hypertension (PAH), a rare, progressive disorder characterized by high blood pressure in the arteries of the lungs. The exact cause of PAH is unknown and there is no known cure for the disease. It is treatable, though, and therefore efficient clinical trials are vital in treatment development. A World Symposia on PAH revised a clinical classification system.[1] Accordingly, PAH is divided into so-called groups 1 to 5 which on their part contain subgroups as well, c.f. Figure 1.
Group 1 includes patients suffering from idiopathic or non-idiopathic PAH such as familial or associated PAH. Group 2 'Pulmonary hypertension due to left heart diseases' is divided into three sub-groups: systolic dysfunction, diastolic dysfunction and valvular dysfunction. Group 3 'Pulmonary hypertension due to respiratory diseases' includes a heterogeneous subgroup of respiratory diseases like PAH due to pulmonary fibrosis, COPD, lung emphysema or interstitial lung disease for example. Group 4 includes chronic thromboembolic pulmonary hypertension and group 5 regroups PH patients with unclear multifactorial mechanisms.[2] An efficient clinical trial should enable testing in certain subgroups while simultaneously investigating treatment benefits in larger groups or even the full population. The broad term "efficient" implies there should be flexibility to adjust the design during the trial, e.g. sample size adjustments and selection of promising subgroups, as well as testing strategies ensuring error control and a predefined power to detect a treatment benefit if there is a benefiting population. All while keeping a reasonable sample size.

Methodological approaches to the statistical design and analysis of clinical trials investigating such a potential heterogeneity of treatment effects across subgroups were systematically reviewed by Ondra et al. (2016).[3]

However, in many applications, e.g. in oncology or cardiology, there is quite an uncertainty about the amount of heterogeneity or the choice of biomarkers as well as cut-off values for those biomarkers. Together with imprecise knowledge about the size of treatment effects and important nuisance parameters, such as the subgroup prevalences or

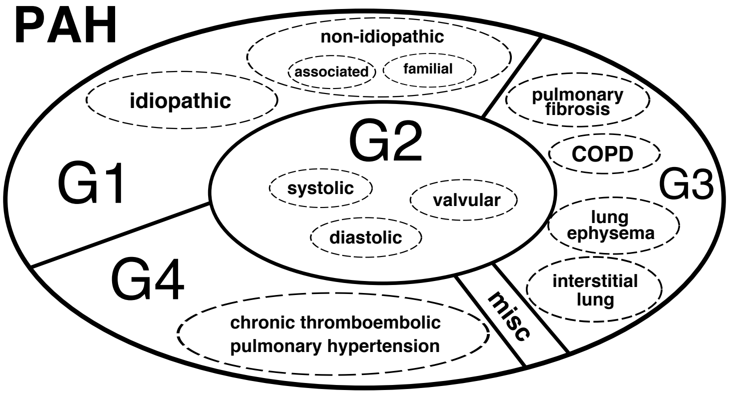

Figure 1: Illustration of the subgroup structure of patients suffering from pulmonary arterial hypertension (PAH).

the variance of the endpoint, sample size determination prior to the study is challenging. Choosing a study design and study size capable to detect a potential treatment effect with a certain power in one or more (sub)populations while using resources optimally is required for ethical and economical reasons. An inadequately small sample size holds the risk of not showing the existing efficacy of a superior treatment. This would waste the employed resources and patients would have been put unnecessarily at risk, hence, - in not showing the benefit - denying other patients access to a superior treatment. On the other hand an unnecessarily large trial wastes resources as well. It puts too many patients at risk and possibly delays the roll out of a superior treatment. Hence, for patients not partaking in the trial access to a more beneficial treatment is delayed as well. Therefore an adequate sample size calculation is important which can only be achieved by reliable assumptions or estimates of the above mentioned nuisance parameters. Often those are obtained from similar studies or pilot studies, i.e. small scale preliminary studies with an exploratory character. For a wider application in clinical research Julious (2016) discusses the role of pilot studies.[4]

Nevertheless, even the best guesses might be wrong, especially concerning the above-mentioned nuisance parameters, e.g. there might be a larger variability in the subgroup or a smaller prevalence than expected. Hence the pre-planned sample size may be too small leading to an underpowered study if not further adjusted in the course of the trial. A large focus of this dissertation will be on those nuisance parameters. Since there are multiple populations of interest more assumptions and even more complex assumptions about those parameters can be made, e.g. differing variances across subpopulations. More parameters implicate more possibilities to missspecify parameters that drive the sample size at the planning stage. To cope with these difficulties, there is a whole class of clinical trial designs covering, among other things, sample size adjustments during an ongoing trial. I will introduce those designs in the following section.

1.2 Adaptive Clinical Trial Designs

“Adaptive or flexible designs are clinical trial designs that use accumulating data to decide on how to modify aspects of a study as it continues, without undermining the validity and integrity of the trial.”[5] Thus, incorrect assumptions at the planning stage, e.g. concerning nuisance parameters or treatment effects, can be detected and corrected during the trial. This means adaptive designs are robust against misspecifications of nuisance parameters at the planning stage. At the same adaptive trials offer a greater flexibility of the ongoing trial and promise a greater efficiency in terms of smaller final sample sizes or “an increased chance of correctly answering the clinical question of interest.”[6] Design adaptations include, but are not limited to, sample size adjustments, early stopping, treatment selection or subgroup selection during an interim analysis (IA). Corresponding adaptive designs are sample size re-estimation designs, group-sequential designs, dose finding designs or adaptive enrichment designs. A broader overview and detailed discussion of those designs is given by Chow and Chang (2008) and Kairalla et al. (2012).[6,7] In the context of the previous example concerning PAH, Grieve et al. (2013) published the results of a workshop on advancing clinical trial design in pulmonary hypertension including a discussion on population enrichment and subgroup analysis.[8] Adaptive enrichment designs and adaptive designs in general have also been suggested as methods for improving clinical trials for cardiovascular diseases in general in a position paper from the Cardiovascular Round Table of the European Society of Cardiology.[9]

From a regulatory point of view the Food and Drug Administration (FDA) discusses adaptive designs in special guidance documents for the industry as well as the European Medicines Agency (EMA) in a reflection paper on methodological issues in confirmatory clinical trials with adaptive designs.[10–12] While approving the advantages, e.g. increased statistical efficiency to a comparable non-adaptive design[13] or same statistical power with a smaller expected sample size,[14] both agencies stress the importance of controlling the chance of erroneous results. Special attention should be paid regarding type I error rate inflation or adaptive design features that might lead to statistical bias in estimation of treatment effects. The FDA created a separate and additional guidance for industry on enrichment strategies for clinical trials providing some general considerations but also giving a classification into prognostic enrichment, i.e. identifying high-risk patients, and predictive enrichment, i.e. identifying more-responsive patients.[15]

In the widely used sample size re-estimation designs regulatory authorities recommend adjustments based on a non-comparative analysis, i.e. “an examination of accumulating trial data in which the treatment group assignments of subjects are not used in any manner in the analysis.”[11] This is also known as a blinded analysis in contrast to an unblinded analysis where treatment group assignments are revealed and used in the analysis. The FDA chose to use the terms comparative/non-comparative analysis since the terms blinded/unblinded might “misleadingly conflate knowledge of treatment assignment with the use of treatment assignment in adaptation algorithms.”[11] Here, I will continue to use the blinded/unblinded framework. In terms of sample size recalculation, a blinded sample size recalculation procedure is based on an early sample size review reestimating nuisance parameters without revealing the treatment group affiliation, c.f. Friede and Schmidli (2010) for an application with count data.[16] An early sample size review can be implemented using an internal pilot study (IPS) design.[17] This means the first part of the trial is employed for improving the sample size calculation, the name IPS indicating its purpose to find better estimates for parameters that drive the sample size. Friede and Kieser (2006) give a review on sample size recalculation in IPS designs.[18] Also, Zucker et al. (1999) compare various procedures concerning IPS designs.[19]

Clinical trial designs with adaptations based on comparative data, i.e. unblinded or unmasked data, include group-sequential designs[20] as well as, with increasing interest, adaptive enrichment trials. Here adaptations are driven by preplanned interim analyses offering the possibility to stop the trial early or select populations that promise an increased treatment effect. Multiplicity issues concerning the type I error rate control are adressed by adjusted critical boundaries or adjusted significance levels at each interim analysis, e.g. using an error spending function,[21] when testing repeatedly or applying a closed testing procedure when testing multiple hypotheses.[22] The latter is computationally more extensive but more flexible and generally more efficient.

Having an interim analysis split the trial in two stages, popular methods ensuring type I error rate control when stagewise selecting (sub)populations and testing the remaining hypotheses are the combination test (CT) approach[23–26] as well as the conditional error function (CEF) approach.[27–30] There are also approaches combining group-sequential designs and subpopulation enrichment.[31–33] Since it often takes a while to fully observe the primary endpoints, interim analyses can be based on early or short-term data, e.g. surrogate endpoints.[34,35]

These procedures, CT approach and CEF approach, require calculating stagewise p-values. With multiple (sub)populations of interest, testing within each stage presents a multiple testing problem again. Besides the classical but inefficient Bonferroni adjustments and the Sidak or Simes tests there are more elaborate options for testing taking into account the correlation between the test statistics.[36–38]

Concerning the interim analysis there are various tools to investigate optimal decision rules,[39] the optimal timing of the interim analysis[40] and overall efficiency and optimality of the design, e.g. utility functions.[41–43]

1.3 Research Questions and Outline

The motivation of this dissertation arose from the joint research project *Biostatistical Methods for Efficient Evaluation of Individualized Therapies (BIMIT)* under grant 05M13MGE BIMIT of the Federal Ministry of Education and Research (BMBF). It was granted in the context of the BMBF call for proposals regarding *Mathematics for Innovations in Industry and Services.* The joint research project was split into three subprojects represented by three research groups in Germany. Part A was handled by the working group in Heidelberg under supervision of the coordinating principal investigator Professor Meinhard Kieser. Here the focus was on methods for interim decisions in adaptive enrichment designs such as optimal decision rules for population selection or the sample size for the interim analysis. Part B was worked on in Bremen under supervision of the principal investigator Professor Werner Brannath dealing with the use of surrogate variables in decision making in adaptive enrichment designs. Part C was located in Göttingen lead by principal investigator Tim Friede. The Göttingen group's task was to investigate blinded sample size recalculation in adaptive enrichment designs. To do so, I split the design into its two main components which I first analyzed separately: On the one hand there is the analysis, sample size determination and blinded sample size recalculation in a multiple subgroups design. On the other hand there is the adaptive enrichment design with subgroup selection at an interim analysis. Both topics individually pose several challenges, e.g. concerning variability across the (sub)populations, blinded reestimation of parameters or distributional properties of the test statistics and each resulted in an individual publication: The first part was published in *Statistical Methods in Medical Research* presenting *Clinical trials with nested subgroups: Analysis, sample size determination and internal pilot studies.* while the second part can be found in *Statistics in Medicine* where we give *A conditional error function approach for adaptive enrichment designs with continuous endpoints.*[30,37] Having analyzed those issues step by step I finally put them together once again presenting *Blinded sample size recalculation in adaptive enrichment designs* published in the *Biometrical Journal.*[44] In all three parts we examine the performance of the proposed methods with simulations in `R`. Methods from the first part - analysis, sample size calculation and blinded sample size recalculation - can be found in the `R` package `spass` which is available on CRAN.[45,46] Methods from the second and third part were reviewed to guarantee reproducible research and are available as supplementary material of the third publication.[44] In the following section I summarize the results and publications in chronological order.

2 Blinded Sample Size Reestimation in Adaptive Enrichment Designs

The following Sections contain the main result of my research on Blinded Sample Size Reestimation in Adaptive Enrichment Designs. The summary is split in three parts which each individually resulted in a paper for publication as described above.

2.1 Clinical Trials with Nested Subgroups

The idea of this first part of the dissertation is to go back from the complex adaptive enrichment design to a simple single-stage design with an overall population and multiple nested subgroups within. We assume that each observation is normally distributed, hence we are dealing with continuous endpoints. Comparing a treatment versus a control we suspect an increased benefit of the treatment in one or more subpopulations. However, we are still interested in simultaneously performing hypothesis test in all subpopulations as well as in the full population since detecting an overall treatment effect is one of the study objectives. Testing multiple hypothesis the family-wise error rate, i.e. the probability to falsely reject at least one hypothesis, has to be controlled. To do so we use the joint multivariate distribution of standardized test statistics for testing intersection hypothesis and then apply the closed testing principle.[22] Each of those Wald-type test statistics corresponds to a population included in the testing. Assuming the variances in each population are known this simplifies to using multivariate normal distributions, c.f. Spiessens and Debois (2010).[36] We extend this approach by allowing for unknown variances giving exact multivariate t-distributions where possible and providing approximations otherwise. Using these results we derive a method for sample size determination prior to the trial which depends on estimates of so-called nuisance parameters. In this multiple subgroups design those parameters are the variances in the populations and the prevalences of the subgroups. If there is no prior knowledge, they have to be estimated or guessed at the planning stage. Consequently they are afflicted with a certain uncertainty and a misspecification leads to inadequately sized studies. To solve this problem we add a sample size review in an internal pilot study to the design.[17] This means, after a prespecified amount of observations is obtained, the nuisance parameters are reestimated based on this early data and the new estimates are plugged in the sample size determination method to calculate an adjusted sample size. This is done in a blinded fashion, i.e. without revealing the treatment group affiliation, as preferred by regulatory authorities.[12, 47]

2.1.1 Statistical Model

First, the statistical model, i.e. the theoretical setting, had to be defined. Since our aim was to analyse trials with subgroups we chose the most challenging subgroup design in terms of dependencies between test statistics, the multiple nested subgroups design. Here we had to deal with subgroups within subgroups and hence test statistics for hypotheses testing for an effect in different subgroups would be highly correlated. We consider a patient population with k nested subgroups. Let $F = S_0$ denote the full population and $S_1, \dots, S_k$ the nested subpopulations

$$S_k \subset S_{k-1} \subset \cdots \subset S_1 \subset S_0 = F.$$

The proportion of subjects in S_i among all subjects in F is the prevalence of belonging to subgroup i and is denoted by τ_i. Since the subgroups are nested we have

$$\tau_1 > \tau_2 > \cdots > \tau_{k-1} > \tau_k.$$

We want to compare an experimental treatment to a control, globally and in each subgroup individually, assuming normal distributed observations. Mean treatment effects are denoted by $\theta_0, \dots, \theta_k$. This means, we assume a treatment effect θ_0 in the full population and a treatment effect θ_k in the smallest subpopulation. In the same manner we want to allow an individual variance $\sigma^2_{S_0}, \dots, \sigma^2_{S_k}$ for each population, assuming same variances in the treatment and control group, i.e. $\sigma^2_{S_i,T} = \sigma^2_{S_i,C} = \sigma^2_{S_i},\ i = 0, \dots, k$. Since smaller subgroups contribute to larger populations in terms of treatment effect and variance due to the nested design, the distribution of each single observation is more clearly represented by using a disjunctive partition of the whole patient population. The ring $R_i = S_i \backslash S_{i+1}$ denotes the set of subjects in population S_i not included in subpopulation S_{i+1}. Therefore, all rings $R_i,\ i = 0, \dots, k-1$, define a disjunctive partition of the whole patient population $R_i \cap R_j = \varnothing, i \neq j,\ F = \cup_i R_i$ (cf. Figure 2). Let X_{ijl} denote the jth subject in ring R_i and treatment group l. Each individual observation is assumed to be normally distributed, i.e.

$$X_{ijC} \sim \mathcal{N}(0, \tilde{\sigma}^2_{R_i}),\ i = 0, \dots, k,\ j = 1, \dots, n_C^{R_i},$$

$$X_{ijT} \sim \mathcal{N}(\tilde{\theta}_i, \tilde{\sigma}^2_{R_i}),\ i = 0, \dots, k,\ j = 1, \dots, n_T^{R_i}.$$

The effects $\tilde{\theta}_i$ and variances $\tilde{\sigma}^2_{R_i}$ for each ring $R_i,\ i = 0, \dots, k$ are chosen in such a way that they can be combined to obtain the assumed effects θ_i and variances $\sigma^2_{S_i}$ of the populations S_i, e.g. the effects via

$$\begin{aligned} \theta_k &= \tilde{\theta}_k \\ \theta_{k-1} &= \left(1 - \frac{\tau_k}{\tau_{k-1}}\right)\tilde{\theta}_{k-1} + \frac{\tau_k}{\tau_{k-1}}\theta_k \\ &\vdots \\ \theta_0 &= (1-\tau_1)\tilde{\theta}_0 + \tau_1\theta_1, \end{aligned}$$

There are $n_C^{R_i}$ subjects in $R_i = S_i \backslash S_{i+1}$ and hence $n_C^{S_i} = \sum_{j=i}^k n_C^{R_j}$ in S_i receive the control. Let $n_C^{S_0} = n_C$ and $n_T^{S_0} = n_T$ be the total number of subjects in the experimental treatment and control group and $n = n_C + n_T$. For unbalanced sample sizes an allocation parameter $a = n_T/n_C$ is defined. This means there are $n_T^{S_i} = a \cdot n_C^{S_i}$ subjects in S_i receiving the treatment.

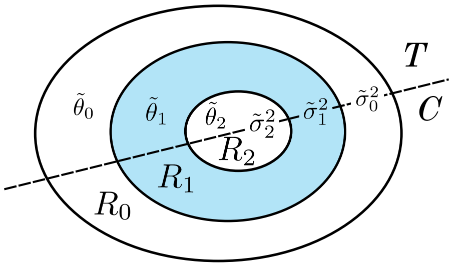

Figure 2: Two nested subgroups within a full population. It is $S_2 = R_2$, $S_1 = R_1 \cup R_2$ and $F = R_0 \cup R_1 \cup R_2$. Effects in the treatment group are denoted by $\tilde{\theta}_i$ in ring R_i. Variances $\tilde{\sigma}_i^2$ in ring R_i are assumed to be the same for treatment and control group.

2.1.2 Hypotheses, Distributional Assumptions and Nuisance Parameters

Having set up the statistical model we aimed to explore methods for hypothesis testing in the multiple nested subgroups design. In such a design not only the global hypothesis in the full population is of interest. We are also interested in elementary hypotheses in each subpopulation. To this end let

$$\begin{aligned} H_0^{\{F\}}\colon \theta_0 &= 0 \\ H_0^{\{S_i\}}\colon \theta_i &= 0,\ i = 1, \ldots, k \end{aligned}$$

denote the null hypothesis of no treatment effect in the full population and in each subpopulation, respectively, and

$$H_0^{\cap_{i \in I} S_i}\colon \theta_i = 0\ \forall i \in I \subseteq \{0, \ldots, k\}$$

the intersection hypothesis that there is no effect in any population of a subset $\cap_{i \in I} S_i$,. For $I = \{0, \ldots, k\}$ this corresponds to the global intersection hypothesis of no treatment effect. Testing will be performed using standardized mean differences $Z^{\{S_i\}}$ for the elementary hypotheses and joint vectors of those test statistics for testing of the intersection hypotheses. To this end we will need to determine the probability distributions of the test statistics which depend on nuisance parameters such as the variances in the populations. Since we are simultaneously testing multiple hypotheses the familywise error rate (FWER) control in the strong sense is an issue which is solved by applying a closed testing procedure.[22] This means, the FWER, i.e. the probability of making at least one type I error in the testing family, is controlled for any configuration of true and non-true null hypothesis. For example, in the case of two subgroups an individual hypothesis, e.g. $H_0^{\{S_2\}}$, is only rejected if all hypotheses relating to intersections that include S_2 can be rejected. Here these are

$$H_0^{F \cup S_1 \cup S_2}, H_0^{S_1 \cup S_2} \text{and } H_0^{F \cup S_2}.$$

Constructing a test for an elementary hypothesis is straightforward since the distribution of the standardized test statistic is either a normal distribution or a t-distribution

$$Z^{\{S_i\}} = \sqrt{\frac{n_{S_i}^P}{a^*}} \frac{\hat{\theta}_{S_i}}{\sigma_{S_i}} \sim N \qquad\qquad Z^{\{S_i\}} = \sqrt{\frac{n_{S_i}^P}{a^*}} \frac{\hat{\theta}_{S_i}}{\hat{\sigma}_{S_i}} \sim t_{n_{S_i}^P},\ i = 0, \ldots, k,$$

depending on whether the variance is given σ_{S_i} or estimated $\hat{\sigma}_{S_i}$. Here, $\hat{\theta}_{S_i}$ denotes the estimated treatment mean difference in population S_i, $a* = 1 + 1/a$ and n_P is the number of all subjects in the control group, $n_P = \sum_{i=0}^{k} n_{S_i}^{P}$. Analogously, tests for the intersection hypothesis are based on the joint distribution of the vector of test statistics corresponding to the subset of populations. For example, for the global intersection hypothesis we need the joint distribution of $\boldsymbol{Z} = (Z^{\{F\}}, Z^{\{S_1\}}, \dots, Z^{\{S_k\}})$. In the case of known variances, since $\boldsymbol{Z}$ is a vector of individually normally distributed test statistics, this is again determined as a multivariate normal distribution, i.e.

$$\boldsymbol{Z} = \begin{pmatrix} Z^{\{F\}} \\ Z^{\{S_1\}} \\ \vdots \\ Z^{\{S_k\}} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{n_P}{a^*}} \frac{\hat{\theta}_F}{\sigma_F} \\ \sqrt{\frac{\tau_1 n_P}{a^*}} \frac{\hat{\theta}_{S_1}}{\sigma_{S_1}} \\ \vdots \\ \sqrt{\frac{\tau_k n_P}{a^*}} \frac{\hat{\theta}_{S_k}}{\sigma_{S_k}} \end{pmatrix} \overset{H_0}{\sim} MN(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Details on the covariance matrix $\boldsymbol{\Sigma} = \mathbb{C}\mathrm{ov}(\boldsymbol{Z})$ for equal and unequal variances and a derivation of its entries can be found in the publication.[37] You will find that for equal variances, the covariance matrix depends only on the prevalences τ_i. Due to the nested structure of the data it is analogous to covariance matrices seen in group-sequential designs.[20] A similar setting, but only for a single subgroup in a full population and equal variances in both populations, was already considered by Spiessens and Debois (2010).[36]

In practice it is uncommon to have such knowledge about the variances in the populations that $\sigma_{S_i}^2$ $i = 0, \dots, k$ are considered fixed and given. More realistic and challenging is the case of unknown variances that have to be estimated via $\hat{\sigma}_{S_i}^2$ $i = 0, \dots, k$. Here, for the determination of the joint distribution of $\boldsymbol{Z}$, we distinguished two scenarios. Firstly, we assumed that the variances are equal across all populations, i.e. $\sigma = \sigma_F = \sigma_{S_1} = \dots = \sigma_{S_k}$, and therefore the complete dataset can be used to estimate the variance σ. It can be shown that under the null hypothesis

$$\boldsymbol{Z} = \begin{pmatrix} Z^{\{F\}} \\ Z^{\{S_1\}} \\ \vdots \\ Z^{\{S_k\}} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{n_P}{a^*}} \frac{\hat{\theta}_F}{\hat{\sigma}} \\ \vdots \\ \sqrt{\frac{\tau_k n_P}{a^*}} \frac{\hat{\theta}_{S_k}}{\hat{\sigma}} \end{pmatrix} \overset{H_0}{\sim} MT_{n_P + n_T - 2(k-1)}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right).$$

follows a multivariate t-distribution with degrees of freedom $df = n_P + n_T - 2(k-1)$ and the same covariance matrix $\boldsymbol{\Sigma}$ as in the case of equal and known variances (see Appendix in Placzek and Friede (2018)[37]). Secondly, we analysed the most complex, but also the most realistic scenario, where the variances are allowed to vary across the subgroups and have to be estimated each individually. This means each component of

$$\boldsymbol{Z} = \begin{pmatrix} Z^{\{F\}} \\ Z^{\{S_1\}} \\ \vdots \\ Z^{\{S_k\}} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{n_P}{a^*}} \frac{\hat{\theta}_F}{\hat{\sigma}_F} \\ \vdots \\ \sqrt{\frac{\tau_k n_P}{a^*}} \frac{\hat{\theta}_{S_k}}{\hat{\sigma}_{S_k}} \end{pmatrix}$$

is individually univariate t-distributed, each with different degrees of freedom corresponding to the variance estimator. Here, the joint vector of test statistics $\boldsymbol{Z}$ is not multivariate t-distributed. Its distribution is unknown. However, for $n \to \infty$ each entry of $\boldsymbol{Z}$ is asymptotically normally distributed and $\hat{\sigma}_{S_i}$ converges in probability to σ_{S_i}. Therefore we have at least that $\boldsymbol{Z}$ is asymptotically multivariate normally distributed

$$\boldsymbol{Z} \mathrel{\dot{\sim}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

with the covariance matrix as in the previous cases. Since in practice extremely large sample sizes are not feasible, we considered and analyzed finite approximations to perform testing including the normal approximation, a liberal t-approximation and a conservative t-approximation. Here, "liberal" means testing using this approximation might have a larger type I error than the nominal significance level. On the other hand, the expression "conservative" stresses that the type I error rate is controlled but the nominal significance level might not be fully exhausted. For the normal approximation we performed testing using the multivariate normal distribution as in the case of given variances but replaced the covariance matrix by an estimator $\hat{\boldsymbol{\Sigma}}$ plugging in the estimates of the variances. The conservative t-approximation uses a multivariate t-distribution with degrees of freedom corresponding to the variance estimator of the smallest subgroup S_k, i.e. $df_{cons} = n_P^{S_k} + n_T^{S_k} - 2$, while the liberal t-approximation is constructed with the degrees of freedom corresponding to the variance estimator of the full population, i.e. $df_{lib} = n_P + n_T - 2(k-1)$. We considered these choices of degrees of freedom since these are the two extremes. This means values of df smaller than df_{lib} would lead to a more conservative and values larger than df_{cons} to a more liberal test procedure. So any approximation using MT_{df}, $df_{cons} \leq df \leq df_{lib}$ lies in between these two. Note that by construction the conservative t-approximation controls the family-wise type I error rate when using equicoordinate quantiles to perform testing as we have shown in Placzek and Friede (2018).[37] We took a look at the performance under the null hypothesis. To do so, we simulated type I error rates in a design with one subgroup increasing the sample size and subgroup prevalence (Figure 3). As expected for large sample sizes the methods converge to the one-sided nominal level of 0.025, since they are asymptotically exact, for smaller sample sizes they behave as constructed, liberal or conservative. The normal approximation is even more liberal than the t-approximation. Generally, the larger the subgroup (increasing prevalence), the better the approximation. The additional nuisance parameter τ is estimated in the simulations by the amount of subjects in the subpopulation among all subjects. Assuming it as fixed and given does not change the results meaningfully (see publication for a comparison).[37]

An alternative approximation was given by Graf et al. (2019).[38] They use a multivariate normal distribution as approximative distribution for $\boldsymbol{Z}$, calculate one single equicoordinate quantile for all entries of $\boldsymbol{Z}$ and then use univariate t-distributions with degrees of freedom corresponding to the number of subjects per subgroup in order to transform this quantile into individual critical values for each subpopulation. I will discuss it as a comparator in detail when summarizing the second publication.

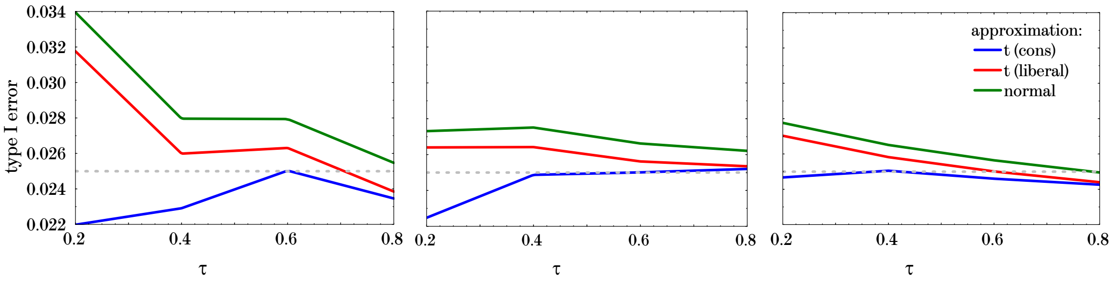

Figure 3: Type I error rates for designs with one subgroup. Comparison of the conservative (MT_{2n^S-2}), the liberal (MT_{2n-4}) and the normal approximation. The prevalence τ is estimated. Number of simulation runs $n_{sim} = 100,000$.

2.1.3 Sample Size Determination and Recalculation

In order to calculate a sample size prior to a study several assumptions have to be made, e.g. about the nuisance parameters, here the variances and prevalences, and about the treatment effects $\theta_0, \ldots, \theta_k$ which we would like to detect. Throughout our work we defined power as the probability to reject at least one false null hypothesis. This is called the disjunctive power[48](c.f. Senn and Bretz (2007)[49]). This means, when we calculate the sample size we have to look at the rejection of the global intersection hypothesis which implies the rejection of at least one elementary null hypothesis. Hence, we are again interested in the complete vector of standardized test statistics. As we have seen in the previous Section the distribution of the joint vector of test statistics under the null hypothesis depends on the nuisance parameters. Naturally, this holds for the distribution under the alternative. Additionally, due to the treatment effects under the alternative, a non-centrality parameter $\boldsymbol{\delta} \in \mathbb{R}^{k+1}$ has to be introduced

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{k+1})' = \left(\sqrt{\frac{n_P}{a^*}}\frac{\theta_0}{\sigma_F}, \sqrt{\frac{n_P\tau_1}{a^*}}\frac{\theta_1}{\sigma_{S_1}}, \ldots, \sqrt{\frac{n_P\tau_k}{a^*}}\frac{\theta_k}{\sigma_{S_k}}\right)'.$$

In Placzek and Friede (2018) we present the procedure for sample size determination for each scenario discussed in the previous section, namely known variances, unknown but equal variances and unknown and unequal variances.[37] As an example I will describe the method for the t-approximation in the case of unknown and unequal variances. Here, under the alternative $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_k)'$, the joint distribution of $\boldsymbol{Z}$ is a multivariate t-distribution with noncentrality parameter $\boldsymbol{\delta}$ and $df = n_P + n_T - 2(k-1)$ (liberal approximation) or $df = n_P^{S_k} + n_T^{S_k} - 2$ (conservative approximation),

$$\boldsymbol{Z} \overset{\cdot}{\sim} MT_{n_P+n_T-2(k-1)}(\boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}}).$$

Note that the alternative also introduces a shift to the covariance matrix $\boldsymbol{\Sigma}$ which is denoted by $\tilde{\boldsymbol{\Sigma}}$ and can be found in Placzek and Friede (2018).[37] We define $t_{\mathbf{0},\boldsymbol{\Sigma},df,1-\alpha}$ as the $(1-\alpha)$-equicoordinate quantile of the distribution $MT_{df}(\mathbf{0}, \boldsymbol{\Sigma})$ under the null hypothesis $H_0^{\{\cap_{i=0}^k S_i\}}$, i.e. for $\boldsymbol{X} = (X_0, \ldots, X_k)' \sim MT_{df}(\mathbf{0}, \boldsymbol{\Sigma})$ it holds that

$$P\left(\bigcap_{l=0}^{k} X_l \leq t_{\mathbf{0},\boldsymbol{\Sigma},df,1-\alpha}\right) = 1 - \alpha.$$

Now, if $\boldsymbol{G}_{\boldsymbol{\delta},\tilde{\boldsymbol{\Sigma}},df}$ denotes the distribution function of $MT_{df}(\boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}})$ and $n = n_P + n_T$ we can find the initial sample size N_0 required to achieve a power of $1 - \beta$ via

$$N_0 = \min n \text{ s.t. } 1 - \boldsymbol{G}_{\boldsymbol{\delta},\tilde{\boldsymbol{\Sigma}},df}(\boldsymbol{t}_{\mathbf{0},\tilde{\boldsymbol{\Sigma}},df,1-\alpha}) \geq 1 - \beta. \tag{2.1}$$

This can be done using a search algorithm. Equicoordinate quantiles and multivariate normal and t-distributions can be calculated using the `R` packages `multcomp` and `mvtnorm`.[50, 51] We have checked the performance of this sample size calculation procedure for exactly this scenario, unknown and unequal variances, using simulations. The results show that the proposed method reaches the nominal power of 90% for all approximations. There is a small loss in power when analysing two nested subgroups instead of one subgroup only and if additionally the prevalence has to be estimated. Throughout the simulations the conservative approximation requires slightly more subjects than the other approximations in terms of mean calculated sample size which is reasonable. The complete result tables are available in the first paper.[37]

Sample size calculations always depend on the nuisance parameters. Therefore misspecifications automatically lead to inadequately sized studies, e.g. an overestimation of variances leads unnecessary large sample sizes while an underestimation leads to insufficient sample sizes. A solution is a sample size review and adjustment during the ongoing trial which we considered next. To do so we considered blinded sample size reestimation in an Internal Pilot Study Design,[17] i.e. after the initial sample size N_0 is determined, patients are recruited until a predefined portion $n_1 = t \cdot N_0$ of subjects have entered the trial. Values between 0.3 and 0.5 are not uncommon for t. These n_1 observations are used to reestimate the nuisance parameters $\tau_1, \ldots, \tau_k$ and $\sigma_F^2, \sigma_{S_i}^2, \tau_i,\ i = 1, \ldots, k$ preferable without unblinding the treatment allocation satisfying regulatory concerns. To reestimate the variances in such a way we used so-called "lumped variance estimators"[19]

$$\widehat{\sigma}_F^2 = \frac{1}{n_1 - 1} \sum_{i=0}^{k} \sum_{j=1}^{n_1^{R_i}} \sum_{l=1}^{2} (X_{ijl} - \bar{X}_{i\cdot\cdot})^2 \tag{2.2}$$

$$\widehat{\sigma}_{S_i}^2 = \frac{1}{n_1^{R_i} - 1} \sum_{s=i}^{k} \sum_{j=1}^{n_1^{R_l}} \sum_{l=1}^{2} (X_{sjl} - \bar{X}_{s\cdot\cdot})^2,\ i = 1, \ldots, k, \tag{2.3}$$

with

$$\bar{X}_{i\cdot\cdot} = \frac{1}{n_1^{R_i}} \sum_{j=1}^{n_1^{R_i}} \sum_{l=1}^{2} X_{ijl},\ i = 0, \ldots, k.$$

The prevalences are estimated as follows

$$\widehat{\tau}_i = \frac{n_1^{S_i}}{n_1}, \; i = 1, \dots, k. \tag{2.4}$$

Recapitulate the notation from the beginning splitting F in disjunct rings R_i and define $n_1^{R_i}$ as the number of subjects in ring R_i at the sample size review. Then this is just calculating the one sample variance in each ring and combining them together to obtain a blinded variance estimator for each subpopulation. The prevalence at the sample size recalculation is again estimated as ratio of number of patients in the subgroup and total number of patients at the blinded review. These new estimates are then treated as "new prior information" on the nuisance parameters and simply plugged in the previously described sample size calculation procedure to find the final sample size N. The remaining $n_2 = N - n1$ subjects are recruited and the final analysis is performed using all N observations.

To assess the performance of the proposed sample size determination and recalculation procedures in combination with methods for the analysis of a nested subgroups design we simulated the power as well as type-I error rates for designs with an internal pilot study. I will here summarize the results for the most interesting case, unknown and unequal variances. The focus of the first part of the simulation was on the power, mean recalculated sample size and variability of the recalculated sample size. We wanted to evaluate the impact of the choice of degrees of freedom for the multivariate t-distribution during the blinded review on these outcomes. To this end we chose a one subgroup design which was to be analysed at the final analysis using the conservative t-approximation, i.e. a multivariate t-distribution with degrees of freedom depending on the number of subjects in the subgroup. The alternative was generated in the subgroup $\theta_S = 1$ while there was no effect in the complement. To create various initial sample sizes the assumptions on the variance in the subgroup were varied between 0.78 and 1.11 resulting in $40 - 75$ subjects per treatment group. Since the true variance in the subgroup is 1.3 a sample size review is reasonable and simulated at timepoints $0.33, 0.5, 0.66$. Figure 4 shows the characteristics against the number of subjects in the subgroup at the blinded review, from left to right, power, mean recalculated sample size and the standard deviation of the recalculated sample size. Red lines depict the described sample size recalculation method, based on the number of subjects in the subgroup at the final sample size, according to the conservative t-approximation. Due to small sample sizes at the blinded review and therefore possibly poor variance estimates we added two modifications following Zucker et al. (1999).[19] They suggest using degrees of freedom depending on the size of the blinded review. Hence, blue lines represent a method that uses degrees of freedom with respect to n_1^S for the recalculation procedure, while black lines use degrees of freedom with respect to n_1. Concerning the power only the method that uses $df(n_1^S)$ reaches the nominal power of 90% throughout the simulations while the other are quite conservative at small sample sizes and need at least $25 - 30$ subgroup subjects at the blinded review to attain the nominal power. The mean recalculated sample sizes show the price that the method using $df(n_1^S)$ (blue line) pays with extremely large sample sizes for small blinded

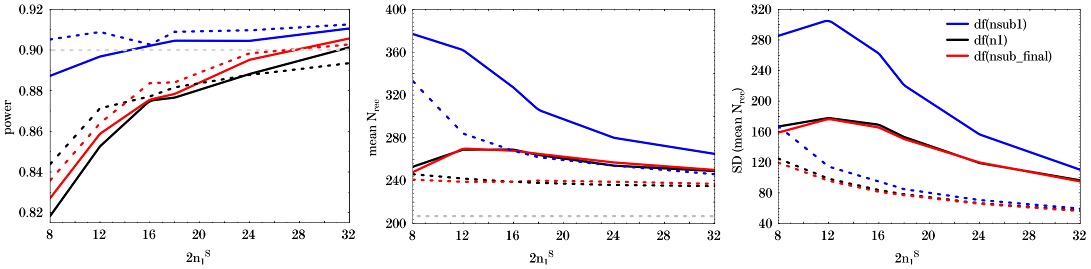

Figure 4: Simulated power, mean and standard deviation of the recalculated sample size of three approaches based on different degrees of freedom of the multivariate t-distribution approximation at the recalculation step for a design with one subgroup. Degrees of freedom are chosen with respect to the number of subjects in the subgroup at the final analysis (red), the number of subjects at the blinded review (black) or the number of subjects in the subgroup at the blinded review (blue). The setting of a fixed prevalence by design (dashed line) is included as well as the case where the prevalence has to be estimated (solid line).

reviews. Both other methods have an increased recalculated sample size compared to the fixed design without recalculation (dashed grey line). Further simulations show that this increase is independent of the true sample size needed in a fixed design which is in line with findings by Friede and Kieser (2011).[52] The rightmost panel of Figure 4 also reveals that the variability of the method depending on the number of subjects in the subgroup at the sample size determination is way larger compared to the other methods. As expected it decreases with increasing number of subjects at the recalculation. Additionally, having to estimate the prevalence τ results in larger recalculated sample sizes and variability (solid lines). Summing up, a sample size review with a small number of subject leads to a decreased power at the final sample size except an unethically large sample size is accepted. There should be at least 25 subjects in the smallest subgroup when performing a sample size recalculation based on degrees of freedom depending on the final subgroup size n^S or the number of subjects at the sample size review n_1.

We assessed the type I error rates of the three methods for sample size recalculation shown in Figure 4 in another similar simulation setting. The final analysis is still performed using the conservative t-approximation. Here we changed the misspecification of the variances prior to the study, now overestimating the true variance in the subgroup. Under the null hypothesis we assumed variances $\sigma^*_F = \sigma^*_S = 1$ while the true variance in the subpopulation was actually $\sigma_S = 0.8$. Varying the subgroup size $\tau = 0.2, 0.3, 0.5$ and simultaneously the initial sample sizes $N_0 = 125, 82, 50$ we reported the mean recalculated sample sizes ($\hat{N}$), the variability of the sample sizes (SD) and the type I error rates ($\hat{\alpha}$) for different timepoints of the blinded review $t = 0.2, 0.4, 0.6, 0.8$ (Table 1). All three methods control the type I error rate and are a bit conservative due to the conservative t-approximation. The method performing its sample size recalculation depending on $df(n_1^S)$ recalculates the largest sample sizes with the highest variability as already seen in the power simulations. All methods recalulate smaller sample sizes than initially planned since the true variance in the subgroup is smaller than assumed at sample

Table 1: Type I error rates for a one subgroup design with internal pilot study. The number of simulation runs is $n_{sim} = 100\,000$.

			$df = n_1^S - 1$			$df = n_1 - 2$			$df = n^S - 2$		
τ	N_{init}	t	$\hat{N}$	SD	$\hat{\alpha}$	$\hat{N}$	SD	$\hat{\alpha}$	$\hat{N}$	SD	$\hat{\alpha}$
0.2	125	0.2	111	50	0.0237	82	37	0.0234	85	35	0.0244
		0.4	90	27	0.0232	81	24	0.0223	85	24	0.0231
		0.6	89	16	0.0233	86	14	0.0231	88	15	0.0236
		0.8	102	6	0.0235	102	4	0.0235	102	5	0.0230
0.3	82	0.2	72	32	0.0230	56	25	0.0232	56	23	0.0231
		0.4	61	19	0.0238	54	16	0.0241	56	16	0.0234
		0.6	59	11	0.0243	57	9	0.0245	59	10	0.0251
		0.8	68	4	0.0240	67	3	0.0237	68	3	0.0236
0.5	50	0.2	43	19	0.0254	35	15	0.0252	33	13	0.0258
		0.4	35	10	0.0262	33	9	0.0254	33	9	0.0253
		0.6	35	6	0.0246	34	5	0.0242	35	5	0.0235
		0.8	41	2	0.0238	41	1	0.0250	41	1	0.0243

size planning. The larger the subgroup the smaller is the observed variation of recalculated sample sizes SD. The same is true for later timepoints of the sample size review.

In summary, we presented methods for planning and analyzing a trial with normally distributed data. This includes sample size determination as well as blinded sample size recalculation in an internal pilot study. We focused on the nuisance parameters, e.g. the variances in the (sub)populations, and their impact on the distribution of the test statistics. We gave exact distributions where possible and suggested approximations otherwise. Performance was assessed via simulations which showed that for type I error rate control at all times the conservative t-distribution approximation should be used. The proposed sample size recalculation procedures do not inflate the type I error rate. We have seen that at least $20 - 25$ subjects in the smallest subgroup are needed at the timepoint of the sample size review in order to recalculate a reasonable sample size reaching the desired power. This is in line with Sandvik et al. (1996)[53] and Birkett and Day (1994)[54] who found that the minimal number of degrees of freedom should be 20. For early sample size review, hence small numbers of subjects available, we gave slight modifications in the recalculation procedure using degrees of freedom based on the number of subjects at the blinded review rather than the projected final sample size.

2.2 Adaptive Enrichment Designs

In the second part we consider methods for adaptive enrichment designs. Those are designs with one or more interim analyses at which decisions on design adaptations can made without inflating the family-wise type I error rate. This includes sample size adjustments, population selection or alterations of the further testing strategy. If the interim analysis suggests there is an increased treatment benefit in a particular subgroup and it is decided to reallocate the stage-two sample size to this subgroup only, this is called an enrichment. There are many different aspects discussed in literature reaching from adaptive designs with subgroup selection[25,29] or population enrichment trials[26,55] over estimation in single- and multstage designs that select subgroups[56] to combinations of group-sequential and subpopulation enrichment designs.[31,32] Here we focus on the methods used to combine two stages of an adaptive enrichment design while still controlling the overall type I error rate. Popular methods are based on the combination test (CT)[23] or the conditional error function (CEF) approach.[27] Especially in settings with one subgroup the CT approach[24,25] as well as the CEF approach[28,55] have already been explored. Using our previous results on clinical trials with nested subgroups, in particular transferring the distributional properties of the (vector of) test statistics, we extend the CEF approach to an adaptive enrichment design with multiple subgroups.

We kept to the framework from the previous section, i.e. multiple nested subgroups and normally distributed outcomes, and wanted to construct an adaptive enrichment design for multiple nested subgroups with potentially unknown and unequal variances. Therefore we transferred our results on the distributional properties of the test statistics, under different scenarios concerning the variances in the populations (known/unknown, equal/unequal), to the conditional error function approach.

The concept of the CEF approach is as follows. An interim analysis is performed when recruitment has reached a predefined number of subjects, e.g. half of the initially planned sample size. At this interim analysis the first stage test statistics are calculated and according to a decision rule it is decided which populations are carried to the next stage to be tested at the final analysis. For each planned hypothesis test a conditional error CE is calculated. It is defined as the probability to reject the test at the final analysis given the observed stage one data available at the interim analysis. Recruitment continues according to the decisions for dropping or enriching certain populations. At the final analysis second-stage p-values q are calculated and the corresponding hypothesis is rejected if $q \leq CE$, c.f. Koenig et al. (2008).[57]

Let $\mathscr{H}_1$ denote the set of hypothesis planned to be tested at the final analysis and $H_I = H_0^{\bigcap_{i \in I} S_i}$. At the interim analysis for each test of $H_I \in \mathscr{H}_1$ we calculate the conditional error

$$CE_I = P_{H_I}(\max_{i \in I} Z^{\{S_i\}} \geq d_s | z_{(1)}^{\{S_i\}}, i \in I), \tag{2.5}$$

where $z_{(1)}^{\{S_i\}}$ is the first-stage observed test statistic in population S_i. The critical value

d_s is a Dunnett-type critical boundary[58] and $s = |I|$. A selection rule that we will later apply in the simulations, the ε rule, which was already used in treatment selection[59] and multi-arm designs,[60] was adopted to subgroup selection by Friede et al. (2012).[28] The idea is to choose a distance $\varepsilon \geq 0$ and then continue with all populations that have test statistics within ε-range of the maximum test statistic at the interim analysis, i.e. all S_i with

$$z_{(1)}^{\{S_i\}} \geq \max_j(z_{(1)}^{\{S_i\}}) - \varepsilon. \tag{2.6}$$

Assume we have decided on certain populations for the second stage and the remaining subjects are recruited accordingly. Let $\mathscr{H}_2$ denote the set of the hypothesis left for the final analysis and I_2 the corresponding set of indices of subgroups involved. Second-stage p-values are calculated as

$$q_I = P_{H_I}(\max_{i \in I_2} Z_i \geq z_{I_2}^{\max} | z_{(1)}^{\{S_i\}}, i \in I_2), \tag{2.7}$$

where $z_{I_2}^{\max}$ is the actually observed value of $\max_{i \in I_2} Z_i$. Finally, H_I is rejected in the final analysis if $q_I \leq CE_I$. Friede et al. (2012) showed in a one subgroup design with equal and known variances that these probabilities can be easily computed using a multivariate normal distribution.[28] However, assuming unknown and unequal variances and multiple subgroups, determining CE_I and q_I is tedious. Certainly, the derived distributional properties of the joint vector of test statistics $\boldsymbol{Z}$ and the suggested approximations remain valid, but to evaluate (2.5) and (2.7) we need the conditional distributions of $\boldsymbol{Z}$ given the first-stage data.

Since it is difficult to determine the conditional distribution of this multivariate vector, we proposed an efficient procedure for simulating conditional distributions. For details see Placzek and Friede (2019).[30] The idea is, given a particular stage-one data set, to calculate the first-stage means and variances for the treatment and the control group. Then we generate n_{dsim} stage-two data sets. For each of these data sets we calculate second-stage means and variances in the same manner, and combine the means and variances of the two stages for each treatment group seperately using the stage-wise sample sizes as weights. Then, the overall treatment differences and variances for each population are determined resulting in n_{dsim} vectors of final test statistics given this stage-one data set $\boldsymbol{Z}_1^*, \boldsymbol{Z}_2^*, \ldots, \boldsymbol{Z}_{dsim}^*$. Suppose we wanted to calculate the conditional error for the rejection of the global intersection hypothesis assuming a scenario with unknown and unequal variances. With the conservative t-approximation, let $c = t_{\mathbf{0},\mathbf{\Sigma},df,1-\alpha}$ denote the $(1-\alpha)$-equicoordinate quantile of $MT_{df}(\mathbf{0}, \mathbf{\Sigma})$ with $df = n_P^{S_k} + n_T^{S_k} - 2$. According to (2.5) we have to calculate

$$CE_{H_0^{\cap_i S_i}} = P_{H_0^{\cap_i S_i}}(\max_{i=0,\ldots,k} Z^{\{S_i\}} \geq c | z_{(1)}^{\{S_i\}}, i = 0, \ldots, k), \tag{2.8}$$

which is now easily estimated by generating n_{dsim} stage-two data sets under $H_0^{\cap_i S_i}$ and counting

$$\#\{\max \boldsymbol{Z}_i^* \geq c \, i = 1, \ldots, n_{dsim}\}/n_{dsim}. \tag{2.9}$$

This procedure can be applied whenever multivariate conditional distributions have to be evaluated. We used it in the extensive simulations assessing the performance of the extended CEF approach that I will summarize next.

An overview of the simulation scenarios considered is given in Table 2. Since the conditional distributions are simulated no matter what scenario is considered we also showed simulation results for the case where we know the exact multivariate distribution of the vector of test statistics $\boldsymbol{Z}$, i.e. unknown but equal variances. As a comparison to the CEF approach we additionally present results for a combination test. The combination test approach, c.f. Brannath et al. (2009) or Friede et al. (2012),[24,28] is based on combining stage-one and stage-two p-values using a combination function.[23] Here the inverse normal combination function is used, e.g. let p_1 denote a first-stage p-value and p_2 the corresponding second-stage p-value, then

Table 2: Overview of simulation scenarios considered for two nested subgroups. Type I error rates and power values were simulated with 100,000 and 10,000 replications, respectively. τ_i, $i = 1, 2$, denote the prevalences of the subgroups, σ_j^2 the variances and θ_j the effects in the smallest subgroup and its complement, $j = S_2, \bar{S}_2$. The number of subjects from the control group initially planned in the full population is given by n_C.

	equal variances		unequal variances	
	type I error	power	type I error	power
τ_1	$0.4, 0.6$	$0.4, 0.6$	0.4	0.4
τ_2	$0.2, 0.3$	$0.2, 0.3$	0.2	0.2
$\sigma^2_{S_2}$	1	1	2	2
$\sigma^2_{\bar{S}_2}$	1	1	1	1
θ_{S_2}	0	1	0	1.5
$\theta_{\bar{S}_2}$	0	0	0	0
n_C	$50, 60, \dots, 90$	$50, 60, \dots, 90$	$60, 70, \dots, 100$	$60, 70, \dots, 100$

$$p = Comb(p_1, p_2) = 1 - \Phi\left(w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)\right)$$

combines the evidence from both stages in a single final p-value. The weights w_1, w_2 have to fulfill $w_1^2 + w_2^2 = 1$ and are usually chosen depending on the sample sizes, here $w_i = \sqrt{n_i/n}, i = 1, 2$ and $n = n_1 + n_2$. P-values are obtained testing intersection hypotheses using the previously presented multivariate distributions while elementary hypothesis are tested using a closed testing procedure, the same way testing is performed with the CEF approach. I will skip the one subgroup scenario here and directly discuss Figure 5 which shows the results for a design with two nested subgroups and unknown but equal variances. This corresponds to columns $1 - 2$ in Table 2. The left two panels show type I error rates for three methods, namely the CEF approach and the CT approach, both using the exact multivariate t-distribution at the final analysis, and additionally, the normal approximation, plugging in the estimated variances and using the multivariate normal distribution. In the process, conditional distributions were

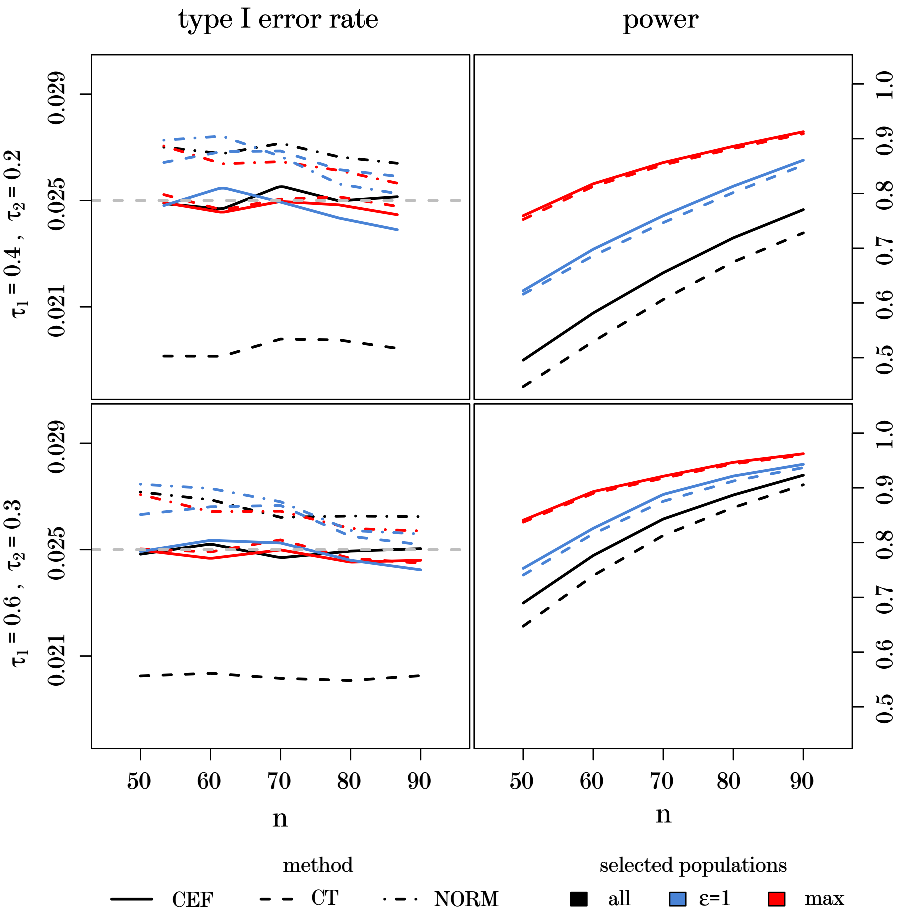

Figure 5: Type I error rate and power for an adaptive enrichment design with two nested subgroups and equal variances as a function of the sample size in the treatment group.

simulated with $n_{dsim} = 10,000$ samples. Smaller subgroup sizes are shown in the top panels while larger subgroups are analyzed in the bottom panels. We considered three different selection rules at the interim analysis which were implemented using the ε-rule choosing $\varepsilon = 0, 1, \infty$. Note that for $\varepsilon = 0$ there is always only the population with the maximum test statistic selected at interim. For $\varepsilon = \infty$ always all populations are carried to the next stage while $\varepsilon = 1$ leads to varying decisions at interim depending on the observed test statistics, selecting one, two or all populations. Reported are rejection probabilities for the rejection of at least one individual hypothesis. For both settings of subgroup sizes and across all sample sizes (x-axis) and selection rules at interim the CEF approach contains the nominal level of 0.025. For the CT approach this is true only using the selection rule “max”. For the other two selection rules there is a slight decrease in type I error rate which can be explained by the non-consonance of the CT approach, i.e. an intersection hypothesis may be rejected without rejecting a corresponding elementary hypothesis. This decrease diminishes with increasing subgroup sizes due to the higher correlation of the test statistics. The normal approximation is too liberal in all cases improving with increasing sample sizes and increasing subgroup sizes. Therefore, we did not include the normal distribution when showing power results. Here, both CEF and CT approach have the highest power when choosing the most promising population at the interim analysis. This is reasonable since the alternative is generated in the smallest subgroup. Hence, correctly selecting only the smallest subgroup consequently increases the power. As expected always continuing with all populations and never enriching the design has the lowest power. Choosing populations based on $\varepsilon = 1$ and therefore sometimes performing an enrichment gives rejection rates in between those two extremes.

For the scenario with two nested subgroups and unknown and unequal variances we considered three different methods. Since the joint distribution of the vector of test statistics $\boldsymbol{Z}$ has to be approximated, we simulated the CEF approach and the CT approach in combination with the conservative multivariate t-approximation and the normal approximation. Additionally, as a third comparison, we included an approximation given by Graf et al. (2019).[38] For testing of intersection hypotheses, instead of using one equicoordinate quantile c they calculate individual critical values c_i corresponding to each entry of $\boldsymbol{Z}$. To do so they start similarly: A multivariate normal distribution approximation

$$\boldsymbol{Z} \dot{\sim} \mathcal{N}\left(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}\right)$$

with the same covariance matrix $\tilde{\boldsymbol{\Sigma}}$ as previously described is used and Pocock type boundaries[61] are applied to calculate an equicoordinate quantile c_α as the same critical value for all subpopulations. Here, to account for the unknown variances, they take it a step further and transform this critical value based on univariate t-distributions with degrees of freedom depending on the subgroup sizes following Jennison and Turnbull (1999),[20]

$$c_i = \Psi^{-1}_{(a+1)n_P^{S_i}-2}(\Phi_{0,1}(c_\alpha)), \ i = 0, \dots, k.$$

Here Ψ_{df} and $\Phi_{0,1}$ denote the distribution functions of a univariate t-distribution with df degrees of freedom and the standard normal distribution. Since $df = (a+1)n_P^{S_i} - 2$ is different for each (sub)population this leads to individual critical values. Hence, the calculation of the conditional error for this method resolves to

$$CE_I = 1 - P_{H_I}(Z^{\{S_i\}} < c_i, i \in I | z_{(1)}^{\{S_i\}}, i \in I). \tag{2.10}$$

Graf et al. (2019) call this method the corrected t-test,[38] we referred to this as the univariate t-approximation. In Figure 6 type I error rates and power for the same three selection rules at interim are shown for a subgroup sizes $\tau_1 = 0.4$ and $\tau_2 = 0.2$ (columns 3-4 in Table 2). The best performing method is the univariate t-approximation using the CEF approach. Although showing a minimally inflated type I error rate when always selecting the population with the maximum test statistic at the interim analysis (bottom panels) it exhausts the alpha more fully than the conservative t-approximation.

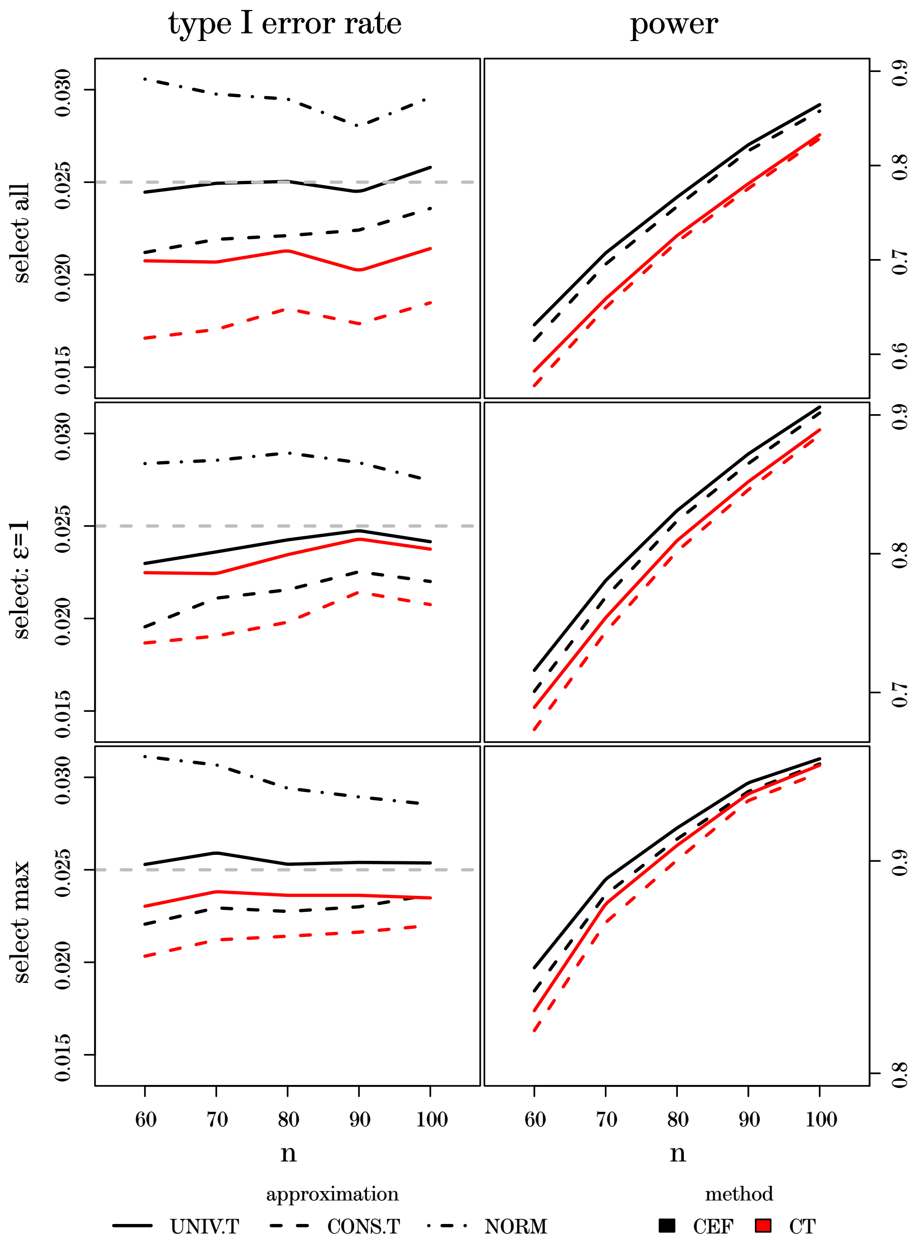

Figure 6: Type I error rate and power for an adaptive enrichment design with two nested subgroups and unequal variances as a function of the sample size in the treatment group.

The normal approximation is once again too liberal. The CT approach performs slightly worse than the CEF approach due to its non-consonance as in the scenario with equal variances. Power-wise the results transfer as well. Selecting the most promising population leads to a more powerful design if there is indeed an increased benefit in a specific population. Since we generated the effect in S_2 the design with selection rule “max” shows the highest power.

To sum up, we have seen that in both scenarios, unknown and equal or unequal variances, the normal approximation which is often used in practice is too liberal. Instead one of the proposed methods should be used, which incorporate the fact that the variances are estimated, by using multivariate or univariate t-distributions. Generally, the CEF approach slightly outperforms the CT approach. When performing an enrichment after the interim analysis the power is increased notably if the correct subpopulation is selected.

2.2.1 Nonoverlapping Subgroups and Multistage Designs

In Placzek and Friede (2019) we also gave extensions to our initially considered statistical design.[30] Instead of solely nested subgroups, we described how our results transfer to nonoverlapping subgroups. Nonoverlapping subgroups within a full population means $S_i \cap S_j = \emptyset$, for all $i \neq j$. Hence, the standardized test statistics $Z^{\{S_i\}}$ and $Z^{\{S_j\}}$ are independent. There only remains a correlation between $Z^{\{F\}}$ and $Z^{\{S_i\}}$, for all $i = 1, \ldots, k$, that depends on the prevalence τ_i and the variances σ_F and σ_{S_i}. Therefore, using the methods proposed here can easily be applied to nonoverlapping subgroups by adjusting the covariance matrix $\mathbf{\Sigma}$ accordingly. Distributions and approximations remain the same.

Concerning the interim stops we presented some advice on how to plan a multistage design. Rather than only one interim analysis we considered designs with multiple data looks. Here we distinguished two strategies. First, we assumed that at a data look only subgroup selection and no testing can be performed. This means the trial always ends at the final analysis after all stages are completed. The second case included testing at each data look and consequently options to stop the trial for futility or efficacy early.

The idea in the first case, when testing only at the final analysis, is to treat data as if there were only two stages at each data look. Conditional errors are calculated conditioning on all data prior to this particular data look but according to a test at the final analysis using the significance level of the conditional error of the previous data look. Assume there are $k-1$ interim analysis and a final analysis at the end, i.e. there are k data looks. Let $CE_0 = \alpha$ denote the significance level. At each data look $l < k$ a set of populations are selected to continue the trial. For each corresponding hypothesis the conditional error CE_l is calculated as the probability to reject the hypothesis at the final analysis given the data observed in stages $1, \ldots, l$ when testing to the level of the previous conditional error CE_{l-1}. At the final analysis, the hypothesis is rejected if the corresponding p-value q is smaller than the conditional error CE_{k-1} of the last interim analysis .

In the second case we additionally allowed for hypothesis testing at each data look enabling the option to stop for futility or efficacy. For the analysis and type I error rate control throughout the whole trial we applied a similar idea by Müller and Schäfer (2001) considering data from stages before and after an interim analysis as separate independent trials.[27] Evidence is then combined using group sequential testing.[14,62] For example one could plan a two stage adaptive design spending a prespecified amount α_1 of the significance level at the first data look. After testing and selecting populations at this interim analysis it is decided not to stop for futility or efficacy but to extend the design by another interim analysis. Since data before and after data look one is treated separately, for hypotheses of interest the conditional error CE_1 is calculated which is the probability to reject the hypothesis at the final analysis given the stage-one data when testing to the remaining significance level. The next part of the trial can then be planned using a group-sequential design with respect to the significance level CE_1. This procedure can be repeated iteratively leading to a design with multiple stages still controlling the type I error rate. For more details see Placzek and Friede (2019).[30] The same can be applied to the CT approach resulting in recursive combination tests.[63]

2.3 Incorporating BSSR into the Adaptive Enrichment Design

Finally, having discussed the analysis of multiple subgroups designs and adaptive enrichment designs, our aim was to give a procedure incorporating the methods for blinded sample size calculation into an adaptive enrichment design with multiple subgroups and normally distributed endpoints. This means the statistical model remains unchanged but at the planning stage of a trial additional decisions have to be made: For the blinded sample size review we install an internal pilot study; hence, a timepoint for the blinded review has to be chosen. This is usually done by specifying an amount of observed patients available, e.g. $p = 30\%$ of the initial sample size N_0. Additionally, there is a stop for an interim analysis. Here the timepoint is specified depending on the recalculated sample size obtained from the blinded review, e.g. the interim analysis is performed when half of those patients are observed (c.f. Figure 7). Further decisions deal with options for design adaptations such as possibilities for early stopping or enrichment of the design. Based on that a testing strategy is chosen. Besides those advanced features the usual assumptions for the initial sample size calculation have to be made. This includes assumptions on treatment effects as well as nuisance parameters.

We will break the following steps down using a one subgroup design as an example.

- Planning stage: Use assumptions on nuisance parameters/testing strategy/design adaptations to calculate initial sample size N_0.
- Blinded review: Reestimate nuisance parameters using $p \cdot N_0$ observations and recalculate final sample size N.
- unblinded interim analysis: After $t \cdot N$ patients are observed carry out unblinded analysis and decide on further course of the trial, e.g. drop populations and enrich trial.
- Final analysis: Perform hypothesis tests in the remaining populations with total number of observed patients.

At the planning phase we begin by defining the statistical model. Since there are different distributional implications depending on the restrictions concerning the variances in the subpopulation and the full population as we have seen in Section 2.1, we have to state which scenario is chosen here. Assume we put no restrictions on the variances in the populations, i.e. we allow for unknown and unequal variances across the (sub)populations, as this is close to reality. This implies that our testing strategy is based on approximations of multivariate distributions of vectors of standardized test statistics. We further specify the timepoints p of the blinded review and t of the interim analysis and choose an adaptive method to combine evidence from the two stages. Suppose we use the CEF approach without an option for early stopping at the interim analysis but with the possibility to select a promising population and enrich the trial in the second stage.

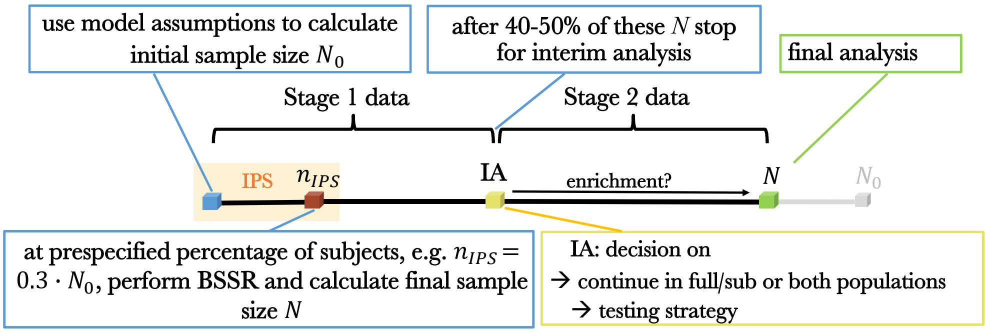
use model assumptions to calculate initial sample size N_0
after 40-50% of these N stop for interim analysis
final analysis
Stage 1 data
Stage 2 data
IPS
n_{IPS}
IA
enrichment?
N
N_0
at prespecified percentage of subjects, e.g. $n_{IPS} = 0.3 \cdot N_0$, perform BSSR and calculate final sample size N
IA: decision on
→ continue in full/sub or both populations
→ testing strategy

Figure 7: Combining blinded sample size recalculation in an internal pilot study design and an adaptive enrichment procedure with an interim analysis. In this example there is one subpopulation inside a full population and a two-stage design shall reveal whether there is an increased treatment benefit in the subpopulation while still simultaneously examining the full population.

To calculate the initial sample size we need prior knowledge or assumptions on the variances and prevalence of the subgroup, define a one-sided significance level $\alpha = 0.025$ and a power $1 - \beta = 0.8$ to detect an anticipated alternative $\boldsymbol{\theta}$. We use the sample size formula from Equation (2.1). Therefore, we have to decide which approximation should be employed as this determines the degrees of freedom of the multivariate t-distribution. We choose the conservative approach, i.e. $df = 2n_S - 2$. Note that we assume a 1:1 allocation between treatment and control group. The iterative search algorithm yields an initial sample size N_0.

At the blinded review, i.e. after $p \cdot N_0$ observations, the variances and prevalence are reestimated based on the data available. This is done in a blinded fashion, c.f. Equations (2.2)-(2.4). The new estimates are once again plugged in the original sample size formula (2.1) and the final sample size N is obtained. This new sample size may be larger or smaller than the initially calculated N_0 depending on whether the nuisance parameters were over- or underestimated at the planning stage. Since the timepoint of the interim analysis depends on N the interim analysis is now performed later or earlier than initially planned (with N_0).

The interim analysis is performed with $t \cdot N$ observations. Here treatment group affiliations are revealed, unblinded estimators for the nuisance parameters as well as the treatment effects obtained and standardized test statistics calculated. Based on those test statistics it is decided in which populations testing will be continued. A selection rule that we already discussed is the ε rule, c.f. Equation (2.6). The ε rule is capable of covering all possible paths, i.e. continue with both populations, only with the full population or only with the subpopulation. In the last case second-stage recruitment can focus exclusively on patients from the subgroup, hence, enriching the trial. For each hypothesis the conditional error is calculated as described in (2.5). This can be implemented using Monte Carlo simulations of the conditional distribution of the (vector of) test statistics as described in (2.8) and (2.9).

According to the decisions at the interim analysis stage-two recruitment is conducted. At the final analysis stage-two p-values are calculated for each of the remaining hypothesis, once again employing Monte Carlo simulations of the conditional distributions given the stage-one data that are needed in (2.7). A hypothesis is then rejected if the stage-two p-value is smaller than the corresponding conditional error that was calculated at the interim analysis. Following the closed testing principle, an individual hypothesis concerning the full population or the subpopulation is only rejected if the global intersection hypothesis is rejected and the elemental hypothesis is rejected as well (both tested with significance level α).

In simulations we assessed various aspects of this procedure including power, type I error and variability of final sample sizes, as well as (optimal) timepoints of the blinded review and the interim analysis. We included different selection rules at the interim analysis and a comparison to a design without BSSR.

In Section 5, we demonstrated the advantages the combination of BSSR and adaptive enrichment brings along performing some simulations in a one subgroup design. First, we compared the power to reject the intersection hypothesis for four variations of the procedure in the following scenario: We generated a subgroup with prevalence $\tau = 0.4$ inside a full population. The true treatment effect is $\Delta_S = 0.75$ in the subgroup and zero in the complement of the subgroup. For initial sample size calculation we assumed that the variances in both populations equal 1 while the true variance in the subpopulation was varied, i.e. $\sigma_S = 0.8, 1, \ldots, 1.6$ on the x-axis. This means at the planning stage we had an intended misspecification of this nuisance parameter in almost all cases. The blinded review was planned at the timepoint of having observed 30% of the initial sample size while the interim analysis should take place after observing 50% of the recalculated sample size. Throughout the simulations we assumed unknown and unequal variances. The prevalence, i.e. the subgroup size, is assumed to be fixed and known, since we found in the first publication[37] that estimating this parameter additionally does not notably change the simulation results. As adaptive method we chose the CEF approach employing the univariate t-approximation. We compared a method without BSSR and three strategies with BSSR but different decision rules at the interim analysis, namely always selecting both populations, selecting the population with the maximum test statistic at interim and always selecting the subpopulation. The last rule represents the theoretical strategy always choosing correctly since the true treatment effect is indeed only in the subpopulation. Figure 8 shows the power curves (left panel) obtained from $10,000$ simulation runs with significance level $\alpha = 0.025$ and nominal power of $1 - \beta = 80\%$. The robustness against misspecifications that is added by implementing a BSSR in an internal pilot study is striking. The method without a blinded review is over- or underpowered in those cases where the assumptions on the variance in the subgroup is wrong while the other three methods contain the nominal power with only slight losses. Those are discussed in more detail in the third publication.[44] Note that there is an additional adjustment of the sample size in case of an enrichment of the subpopulation which can already be planned at the blinded review. Since we are enriching the trial with patients from the promising population we may decrease the sample size while maintaining the desired power. This explains the differences in the final sample sizes (right panel). The method without BSSR keeps the initially calculated sample size fixed in all cases. For the methods with blinded review the sample size increases with increasing variability in the subgroup. Generally, the strategy with no enrichment has the largest sample sizes while both enrichment strategies are able to save 25-60 patients in terms of mean recalculated sample sizes. Comparing those two, the method that always chooses the subpopulation shows slightly smaller sample sizes. This is not surprising since in this example only patients in the subgroup benefit from the treatment, hence, always enriching S leads to a more efficient design. We obtained analogous results for the variability of the mean recalculated sample sizes, c.f. Placzek and Friede (2022).[44]

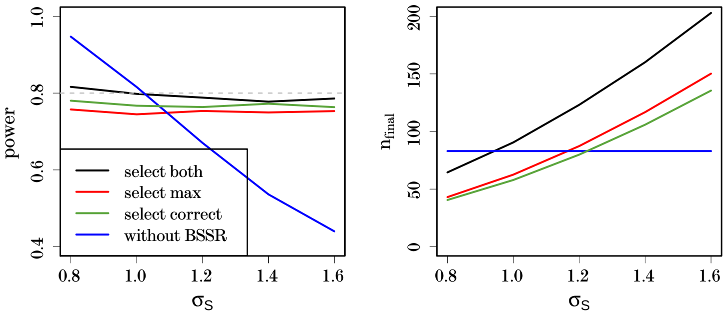

Figure 8: Comparison between methods combining BSSR and adaptive testing strategies (black, red, green) and a strategy without BSSR (blue). Testing strategy in all cases is based on the conservative approximation of the multivariate t-distribution.

In the previous simulation scenario we only considered an effect in the subpopulation. To complement these power and sample size simulations we additionally simulated a scenario with effects in both full population and subpopulation, i.e. $\Delta_F = 0.5$ and $\Delta_S = 0.5$, holding on to the remaining settings. The results do not differ much from the ones seen in Figure 8, except with effects in both populations there is almost no difference between the two selection rules. However, those results underline the integrity of the proposed procedure and can be found in the third publication.[44]

Next, we briefly simulated type I error rates to demonstrate FWER control. Since BSSR and adaptive enrichment methods are applied independently in the combined procedure, type I error rate control follows directly from FWER control of both independent components, cf. Placzek and Friede (2017, 2019).[30,37] The adaptive testing strategy applied in the simulations is the conditional error function approach. For both selection rules, deciding at an interim analysis whether to continue testing in both populations or only in the population with the maximum test statistic at interim, the type I error is controlled across all scenarios.[30] Rejection rates are even a bit conservative which is inherited by

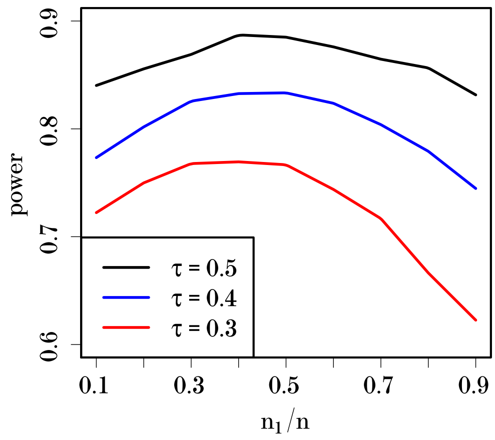

Figure 9: Power of the conditional error function approach for different timepoints of the interim analysis (x-axis) varying the subgroup prevalence.

the conservative approach used at the BSSR.

Finally, we took a look at the timepoints of the blinded review and the interim analysis and how those influence the performance of the procedure. First, we simulated an adaptive enrichment design without BSSR varying the timepoint of the interim analysis in order to find an optimal timepoint for population selection and enriching of the trial. For three different subgroup sizes we assessed the power of the CEF approach which selects the population with the maximum test statistic at interim. Timepoints ranged from 10% to 90% of the planned sample size. There was no adjustment of the sample size. Figure 9 suggest an optimal timepoint of $40-50\%$ of the final sample size.

Since recalculating the sample size automatically changes the amount of patients representing $40-50\%$ of the final sample size we wanted to investigate the benefit of adjusting the timepoint of the interim analysis based on the recalculated sample size coming from a blinded review. Therefore we compared such a strategy (two-stop strategy) with a design where the sample size review and the interim analysis are simultaneously performed at one timepoint (one-stop strategy). This means, while in the first case a sample size review is performed after 30% of patients have been observed and then the interim analysis takes place after 50% of the recalculated sample size, in the second case both, sample size recalculation and interim analysis, are conducted at 50% of the initially calculated sample size missing out on optimizing the timepoint of the interim analysis with respect to a sample size adjustment. Once again, we simulated a one subgroup design and compared power, mean recalculated sample size, variance of the sample size and selection probabilities of the subgroup between these two approaches. We found that, while quite similar in terms of power and mean recalculated sample size with the one-stop strategy even a bit favorable, concerning the enrichment aspect of the design, i.e. with a selection rule allowing selection of the subgroup, the two-stop strategy outperforms its competitor in terms of variability of the mean recalculated sample size and in terms of selection probabilities of the subgroup. Here, timing the IA based on the results of the blinded review leads to a smaller standard deviation of the recalculated sample size and a higher probability to correctly select the subgroup compared to the procedure with a simultaneous sample size review and IA. This behavior intensifies for increasing variability in the subgroup and associated larger total sample sizes. More details as well as corresponding graphics can be found in Placzek and Friede (2022).[44]

To summarize the part on methods for blinded sample size recalculation and subgroup selection in adaptive enrichment designs, we have seen how the individual results from Sections 2.1 and 2.2 can be combined to construct a robust, flexible and efficient procedure. However, the advanced complexity with the increasing number of parameters requires more elaborate planning of the trial and a diligent execution of the study plan.

3 Discussion

In the context of the BIMIT project I tackled the analysis of subgroup designs. More specifically, I considered blinded sample size recalculation and the selection of a subgroup in an adaptive enrichment design as this was the focus of package C carried out in Göttingen. Since such a sophisticated design requires a solid statistical foundation we thoroughly examined the underlying statistical model and established exact methods or approximations for testing hypothesis in a multiple (nested) subgroups design comparing a treatment versus a control in the full population and the subpopulations simultaneously. Here, other than in common practice, we put no restrictions on the variances in the populations, i.e. additional to the cases of known or equal variances we included the scenario of unknown and unequal variances. However, we restricted our research to normally distributed outcomes. We use standardized test statistics and their (joint) distributions to test intersection hypotheses and individual hypothesis and then apply the closed testing principle in order to control the type I error rate. A sample size determination is performed using equicoordinate quantiles of the multivariate distribution of the vector of test statistics and an interative algorithm. We analyzed the proposed procedures in simulations which confirmed the validity concerning power and type I error rate. Next, we added a method for blinded sample size recalculation in an internal pilot study giving blinded estimators for the variances and prevalences. In further simulations we found no notably inflated type I error when using blinded sample size adjustments. However, we found a lower boundary for the size of the internal pilot study, i.e. a minimum of 20-25 subjects in the smallest subgroup. Otherwise the target power cannot be achieved. Though we presented additional adjustments for the BSSR in the IPS handling smaller sample sizes, this comes at a high cost in terms of the final sample size.

With the same statistical model as the basic framework we went on to adaptive enrichment designs. Transferring our results on the different scenarios accounting for uncertainty in the variance estimation we extended the conditional error function approach to a more general setting, i.e. multiple subgroups and standardized test statistics with estimated variances. Subgroup selection at the interim analysis was based on different rules including always continuing with all populations, only with the most promising or with a combination of populations that lay in between those two extremes (ε-rule). In the last two cases an enrichment of the promising populations can be performed. Calculation of the conditional error and second stage p-values relies on conditional multivariate distributions which we generated in Monte Carlo simulations. In an extensive simulation study we compared the CEF approach with the combination function approach. In the case of unknown and unequal variances we analyzed three different approximations, namely the conservative and liberal multivariate t-approximation as well as the univariate t-approximation. Generally the CEF approach slightly outperforms the CT approach in terms of type I error rate and power. This becomes more pronounced with an increasing number of subgroups. From the three approximations the univariate t-approximation is the best performing, slightly better exhausting the type I error and with a somewhat better power. Naturally, if an enrichment is conducted and there is indeed an increased treatment effect in the selected subpopulation, the overall power is significantly increased.

Finally, in a third publication, we combined both BSSR and adaptive enrichment and gave a full procedure including sample size calculation, blinded sample size recalculation, (sub)population selection in an interim analysis, subgroup enrichment and testing at a final analysis. We discussed model assumptions and optimal timepoints for the interim analysis and showed that in certain scenarios it is beneficial to perform the BSSR prior to the interim analysis and to adjust the timepoint of the interim analysis based on those results. We saw that the full procedure unites the benefits of both individual methods, robustness against misspecifications and flexibility in terms of design adaptations without sacrificing power. Type I error rate control follows from FWER control in the strong sense of both the BSSR methods and the adaptive enrichment methods individually, since those are applied independently in our design.

Revisiting the example in pulmonary arterial hypertension from the Introduction, the advantages of those adaptive designs with subgroup analysis are striking. They offer a flexible and effective way to identify groups of patients, and even subgroups within those groups, which benefit more from a treatment while keeping the sample size needed at a minimum. In PAH patients there were already some achievements toward personalized medicine, e.g. it was found that most drugs for PAH seem to be more efficient in patients with idiopathic PAH than in those with non-idiopathic PAH, c.f. Figure 1. On the other hand, PAH patients with connective tissue disease consistently showed a lesser response to treatment.[64] Further directions of targeted therapies are the investigation of proteins to develop personalized proteomics of PAH as well as genetic studies to identify subgroups.[65,66] Those can then be used to apply the study design proposed here in a future clinical trial.

We published methods for the analysis, sample size calculation and blinded sample size recalculation in an `R` package `spass` on the Comprehensive R Archive Network (CRAN).[45,46] Additionally, methods for the combination of BSSR and Adaptive Enrichment Designs are available as supplementary material of the third publication.[44] Those packages also contain the simulation code that was used to obtain the results in this summary. Another powerful `R` package for the design and analysis of confirmatory adaptive group sequential designs is `rpact` by Wassmer and Pahlke (most recent update August 2022).[67] It provides power and sample size calculation as well as simulation and analysis tools for adaptive designs with interim analyses including enrichment designs, not only for continous endpoints but also for binary and survival endpoints (hazard ratios). Implemented are the methods by Wassmer and Brannath (2016).[68]

We assessed performance of the procedures using the most common metrics for adaptive designs reporting type I error rates, power, expected total sample size and variability of the sample size in various simulation scenarios. One might argue that trial duration should also be considered a metric of interest, since both adaptations, BSSR and adaptive enrichment, can increase or decrease trial duration. Obviously, there is a linear relation between final sample size and trial duration. Hence, increasing or decreasing the sample size via a BSSR will increase or decrease trial duration. If at interim it is decided to continue recruiting only from a specific subpopulation in the second stage, this might as well increase trial duration since recruitment might be slower. This, however, can be compensated by fewer subjects needed to achieve the pre-planned power, therefore decreasing trial duration. Naturally, multiple interim looks (BSSR and interim analysis) potentially increase costs in terms of time and work for a statistician. On the other hand those hold benefits as well, e.g. a more continous monitoring and data cleaning as well as the statistician getting familiar with the data set early on. In the long run this improves data quality and saves time at interim and final analysis. Friede et al. (2019) assess the operating characteristics, including trial duration, of a BSSR procedure in an event-driven trial and compare them with those of a fixed sample size design.[69]

A new understanding of the type I error in designs with multiple populations was presented by Brannath et al. (2020) who suggest a criterion ignoring multiplicity adjustments if disjoint subpopulations are considered and controlling the average multiple type I error rate for intersecting subpopulations. This is the probability that a randomly selected patient received an inefficient treatment. They call it the population-wise error rate.[70]

Type I error rate control should also be kept in mind when thinking about unblinded sample size recalculation. Naturally, one might want to perform an additional sample size adjustment at the interim analysis in an unblinded fashion, since data is unblinded at that timepoint anyway. Here, we have to differentiate between two ways of doing so. On the one hand such an unblinded sample size recalculation can be based solely on unblinded estimates of the nuisance parameters, e.g. variances, not taking into account observed treatment effects. There are two major drawbacks: First, unblinded sample size recalculation, other than its blinded counterpart, does inflate the type I error rate. Secondly, although counterintuitive, Friede and Kieser (2013) compared sample size reestimation based on blinded and unblinded variance estimators and showed that the unblinded method does not guarantee that the nominal power is attained.[71] The blinded methods turn the obvious disadvantage of using biased estimators in case of treatment group differences into an advantage by compensating a small power loss introduced by recalculating the sample size in an ongoing trial through a slightly increased sample size. That is why we kept using the blinded one-sample variance estimator in all three publications experiencing the same results. On the other hand an unblinded sample size recalculation might additionally incorporate observed treatment effects. This, however, can inflate the type I error to more than two times the size than the nominal level and statistical adjustment is necessary.[72] Appropriate methods were summarized and reviewed by Wassmer (2000) including variance spending, alpha spending and conditional error function approaches.[72–74] Needless to say, the conditional error function approach, as discussed here, allows for effect based sample size adjustments while controlling the type I error rate.

Blinded sample size recalculation and adaptive designs remain relevant topics in ongoing research. BSSR in various designs including multitreatment crossover trials or stepped-wedge cluster randomized trials were considered by Grayling et al. (2018)[75,76] while Harden and Friede (2020) considered BSSR in multicenter randomized controlled clinical trials based on noncomparative data.[77] In the field of adaptive designs Friede et al. (2020) present a framework on adaptive seamless designs combining phase II and phase III characteristics such as treatment or subgroup selection and confirmatory testing.[35] Here interim analyses are informed by either the primary outcome or an early outcome. Additionally they implemented an extension of the `R` package `asd`[78] to include adaptive enrichment designs.

In our publications we considered normally distributed endpoints. However, the principle and general idea of BSSR as well as adaptive enrichment designs can be transferred to other endpoints, e.g. binary, survival or other event-based outcomes. Here, we had to deal with variances and prevalences as nuisance parameters, standardized mean differences as test statistics and (multivariate) normal or t-distributions. Of course, those parameters change depending on the distribution of the endpoint, but the methods and procedure in general remain the same. For example, Asendorf et al. (2019) proposed methods for BSSR in clinical trials with longitudinal negative binomial counts.[79] They include the overall rate and shape parameter as nuisance parameters working with a negative binomial distribution.

4 References

[1] Simonneau G, Robbins IM, Beghetti M, et al. Updated Clinical Classification of Pulmonary Hypertension. *Journal of the American College of Cardiology.* 2009;54(1, Supplement):S43 - S54. Proceedings of the 4th World Symposium on Pulmonary Hypertension.

[2] Simonneau G, Gatzoulis MA, Adatia I, et al. Updated clinical classification of pulmonary hypertension. *Archives of the Turkish Society of Cardiology.* 2014;42 Suppl 1:45–54.

[3] Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics.* 2016;26(1):99–119.

[4] Julious SA. Pilot Studies in clinical research. *Statistical Methods in Medical Research.* 2016;25(3):995-996.

[5] Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive Designs in Clinical Drug Development - An Executive Summary of the PhRMA Working Group. *Journal of Biopharmaceutical Statistics.* 2006;16(3):275-283.

[6] Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities. *Trials.* 2012;13(145).

[7] Chow SC, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet Journal of Rare Diseases.* 2008;3(11).

[8] Grieve AP, Chow S, Curram J, et al. Advancing clinical trial design in pulmonary hypertension. *Pulmonary Circulation.* 2013;3(1):217-225.

[9] Jackson N, Atar D, Borentain M, et al. Improving clinical trials for cardiovascular diseases: a position paper from the Cardiovascular Round Table of the European Society of Cardiology. *European Heart Journal.* 2016;37(9):747-754.

[10] Food and Drug Administration (FDA) . Adaptive Designs for Medical Device Clinical Studies: Guidance for Industry and Food and Drug Administration Staff, `https://www.fda.gov/media/92671/download`, Accessed: 2022-08-29; 2016.

[11] Food and Drug Administration (FDA) . Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry, `https://www.fda.gov/media/78495/download`, Accessed: 2022-08-29; 2019.

[12] (EMEA) European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design, Published online, Accessed: 2022-09-14; 2007.

[13] Chow SC, Chang M. *Adaptive Design Methods in Clinical Trials.* Chapman and Hall/CRC Biostatistics Series; 2011.

[14] Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics.* 1999;55(3):853-857.

[15] Food and Drug Administration (FDA) . Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products, `https://www.fda.gov/regulatory-information/search-fda-guidance-documents`, Accessed: 2022-09-14; 2019.

[16] Friede T, Schmidli H. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Statistics in Medicine.* 2010;29(10):1145–1156.

[17] Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine.* 1990;9(1-2):65–72.

[18] Friede T, Kieser M. Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal.* 2006;48(4):537–555.

[19] Zucker David M, Wittes Janet T, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine.* 1999;18(24):3493-3509.

[20] Jennison C, Turnbull B. *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman and Hall/CRC; 1999.

[21] Demets David L., Lan K. K. Gordon. Interim analysis: The alpha spending function approach. *Statistics in Medicine.* 1994;13(13â14):1341-1352.

[22] Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63(3):655-660.

[23] Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994;50(4):1029–1041.

[24] Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine.* 2009;28(10):1445–1463.

[25] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics.* 2011;10(4):347–356.

[26] Wassmer G, Dragalin V. Designing Issues in Confirmatory Adaptive Population Enrichment Trials. *Journal of Biopharmaceutical Statistics.* 2015;25(4):651–669.

[27] Mueller HH, Schaefer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics.* 2001;57(3):886–891.

[28] Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine.* 2012;31(30):4309–4320.

[29] Stallard N, Hamborg T, Parsons N, Friede T. Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics.* 2014;24(1):168–187.

[30] Placzek M, Friede T. A conditional error function approach for adaptive enrichment designs with continuous endpoints. *Statistics in Medicine.* 2019;38(17):3105-3122.

[31] Rosenblum M, Luber B, Thompson RE, Hanley D. Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine.* 2016;35(21):3776–3791.

[32] Rosenblum M, Qian T, Du Y, Qiu H, Fisher A. Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics.* 2016;17(4):650-662.

[33] Cui Lu, Zhang Lanju. On the efficiency of adaptive sample size design. *Statistics in Medicine.* 2019;38(6):933-944.

[34] Brückner M, Burger HU, Brannath W. Nonparametric adaptive enrichment designs using categorical surrogate data. *Statistics in Medicine.* 2018;37(29):4507-4524.

[35] Friede T, Stallard N, Parsons N. Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. *Biometrical Journal.* 2020;62(5):1264-1283.

[36] Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials.* 2010;31(6):647–656.

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

[38] Graf AC, Wassmer G, Friede T, Gera RG, Posch M. Robustness of testing procedures for confirmatory subpopulation analyses based on a continuous biomarker. *Statistical Methods in Medical Research.* 2019;28(6):1879-1892.

[39] Krisam Johannes, Kieser Meinhard. Decision Rules for Subgroup Selection Based on a Predictive Biomarker. *Journal of Biopharmaceutical Statistics.* 2014;24(1):188-202.

[40] Benner L, Kieser M. Timing of the interim analysis in adaptive enrichment designs. *Journal of Biopharmaceutical Statistics.* 2018;28(4):622-632.

[41] Graf AC, Posch M, Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal.* 2015;57(1):76–89.

[42] Ondra T, Jobjoernsson S, Beckman RA, et al. Optimizing trial designs for targeted therapies. *PLOS ONE.* 2016;11(9):1-19.

[43] Ondra T, Jobjoernsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Statistical Methods in Medical Research.* 2019;28(7):2096-2111.

[44] Placzek M, Friede T. Blinded sample size recalculation in adaptive enrichment designs. *Biometrical Journal.* 2022;00:1-16. https://doi.org/10.1002/bimj.202000345.

[45] Asendorf T, Gera R, Islam S, Harden M, Placzek M. spass - Study Planning and Adaptation of Sample Size (2020). R package version 1.3 — For new features, see the 'Changelog' file (in the package source).

[46] R Core Team . *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria; 2018.

[47] Harmonisation E9 Expert Working Group International Conference. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. *Statistics in Medicine.* 1999;18(15):1905–1942.

[48] Hothorn Torsten, Bretz Frank, Westfall Peter. Simultaneous Inference in General Parametric Models. *Biometrical Journal.* 2008;50(3):346-363.

[49] Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics.* 2007;6(3):161-170.

[50] Hothorn T, Bretz F, Westfall P. multcomp - Simultaneous Inference in General Parametric Models (2022). R package version 1.4-20 — For new features, see the 'Changelog' file (in the package source).

[51] Genz A, Bretz F, Miwa T, Mi X, Hothorn T. mvtnorm - Multivariate Normal and t Distributions (2021). R package version 1.1-3 — For new features, see the 'Changelog' file (in the package source).

[52] Friede T, Kieser M. Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics.* 2011;10(1):8-13.

[53] Sandvik L, Erikssen J, Mowinckel P, Rødland EA. A method for determining the size of internal pilot studies. *Statistics in Medicine.* 1996;15(14):1587-1590.

[54] Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Statistics in Medicine.* 1994;13(23-24):2455-2463.

[55] Mehta C, Schäfer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine.* 2014;33(26):4515–4531.

[56] Chiu YD, Koenig F, Posch M, Jaki T. Design and estimation in clinical trials with subpopulation selection. *Statistics in Medicine.* 2018;0(0):1-18.

[57] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine.* 2008;27(10):1612–1625.

[58] Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association.* 1955;50(272):1096-1121.

[59] Kelly PJ, Stallard N, Todd S. An Adaptive Group Sequential Design for Phase II/III Clinical Trials that Select a Single Treatment From Several. *Journal of Biopharmaceutical Statistics.* 2005;15(4):641-658.

[60] Friede T, Stallard N. A Comparison of Methods for Adaptive Treatment Selection. *Biometrical Journal.* 2008;50(5):767-781.

[61] Pocock SJ. Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika.* 1977;64(2):191-199.

[62] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics.* 1999;55(4):1286–1290.

[63] Brannath W, Posch M, Bauer P. Recursive Combination Tests. *Journal of the American Statistical Association.* 2002;97(457):236–244.

[64] McLaughlin VV, Badesch DB, Delcroix M, et al. End points and clinical trial design in pulmonary arterial hypertension. *Journal of the American College of Cardiology.* 2009;54(1 Suppl):97–107.

[65] Colvin KL, Yeager ME. Proteomics of pulmonary hypertension: could personalized profiles lead to personalized medicine?. *Proteomics - Clinical Applications.* 2015;9(1-2):111–120.

[66] Loyd JE. Pulmonary arterial hypertension: insights from genetic studies. *Proceedings of the American Thoracic Society.* 2011;8(2):154–157.

[67] Wassmer G, Pahlke F. rpact - Confirmatory Adaptive Clinical Trial Design and Analysis (2022). R package version 3.3.1 — For new features, see the 'Changelog' file (in the package source).

[68] Wassmer G, Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials.* Springer; 2016.

[69] Friede T, Pohlmann H, Schmidli H. Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis. *Pharmaceutical Statistics.* 2019;18(3):351-365.

[70] Brannath W, Hillner C, Rohmeyer K. A liberal type I error rate for studies in precision medicine. 2020;. arXiv:2011.04766.

[71] Friede T, Kieser M. Blinded sample size re-estimation in superiority and noninferiority trials: bias versus variance in variance estimation. *Pharmaceutical Statistics.* 2013;12(3):141-146.

[72] Proschan MA, Hunsberger SA. Designed Extension of Studies Based on Conditional Power. *Biometrics.* 1995;51(4):1315–1324.

[73] Wassmer G. Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers.* 2000;41(3):253–279.

[74] Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine.* 2003;22(6):971-993.

[75] Grayling MJ, Mander AP, Wason JMS. Blinded and unblinded sample size reestimation in crossover trials balanced for period. *Biometrical Journal.* 2018;60(5):917-933.

[76] Grayling MJ, Mander AP, Wason JMS. Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometrical Journal.* 2018;60(5):903-916.

[77] Harden M, Friede T. Sample size recalculation in multicenter randomized controlled clinical trials based on noncomparative data. *Biometrical Journal.* 2020;62(5):1284-1299.

[78] Parsons N, Friede T, Todd S, et al. An R Package for Implementing Simulations for Seamless Phase II/III Clinical Trials Using Early Outcomes for Treatment Selection. *Computational Statistics and Data Analysis.* 2012;56(5):1150-1160.

[79] Asendorf T, Henderson R, Schmidli H, Friede T. Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends. *Statistics in Medicine.* 2019;38(9):1503-1528.