Aus dem Institut für Medizinische Biometrie und Informatik
des Universitätsklinikums Heidelberg
(Geschäftsführender Direktor: Prof. Dr. Meinhard Kieser)
Abteilung für Medizinische Biometrie
(Institutsdirektor: Prof. Dr. Meinhard Kieser)

Incorporation of Historical Two-Arm Data in Clincal Trials with Binary Outcome

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Manuel Feißt
aus
Herbolzheim
2021

Dekan: Prof. Dr. Hans-Georg Kräusslich
Doktorvater: Prof. Dr. Meinhard Kieser

Contents

List of Abbreviations and Symbols

- α: type I error rate
- α_0: level of significance
- $1 - \beta_0$: statistical power
- c: number of responders in the control group
- c_H: number of responders in the historical control group
- δ: parameter that determines the amount of historical data that is incorporated in the new trial
- $\Delta = \pi_T - \pi_C$: rate difference between true proportion of responders of the treatment and true proportion of responders of the control arm
- H_0: null hypothesis
- H_1: alternative hypothesis
- n_C: number of patients in the control group
- n_{CH}: number of patients in the historical control group
- n_T: number of patients in the treatment group
- n_{TH}: number of patients in the historical treatment group
- p_C: proportion of responders in the control group
- p_{CH}: proportion of responders in the historical control group
- p_T: proportion of responders in the treatment group
- p_{TH}: proportion of responders in the historical treatment group

- π_C: true proportion of responders in the control arm
- π_T: true proportion of responders in the treatment arm
- t: number of responders in the treatment group
- t_H: number of responders in the historical treatment group

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Clinical trials often examine the efficacy of new experimental treatments versus placebo or, if available, the current gold standard therapy. The classic design of such a study is the comparison of a new treatment with one established or placebo treatment, a so-called two-arm clinical trial. An important aspect is the choice of the number of patients to be included. For ethical and economic reasons, the sample size should be neither too high nor too low. Including too many patients does not only consume excessively high resources in terms of time, costs, and personnel, but also unnecessarily exposes patients to a study burden or (potentially) ineffective therapy. On the other hand, it is necessary to collect an adequate number of patients in order to achieve a sufficiently high probability for detecting a positive treatment effect, i.e. a high statistical power. Thereby, the number of patients actually needed to be recruited in the trial depends not only on the size of the effect to be revealed but rather on the characteristics of the variables by which the study objective will be assessed, the so-called primary endpoints of the trial. In order to keep the complexity of the study low, often only one endpoint is determined to be the primary endpoint in the respective clinical trial. A main characteristic of this primary endpoint is its scale, that may range from binary outcome (e.g. response vs no response to a treatment) up to continuous outcome (e.g. biomarker values). In many trials, binary outcomes are used as the primary endpoint, also due to their simple construction with respect to a clear interpretation of the study objective. However, the use of binary data is in general associated with an increased sample size required for the study, as binary data is the type of data that comprises the least information. In summary, it is generally more challenging to achieve sufficient power in a clinical study with a binary endpoint than with other

endpoints because the sample size required is significantly increased.

Especially in the field of rare diseases, it is a particular difficulty to conduct an adequately powered study, as the available number of patients for a clinical trial is severely limited. One way to increase the power of such studies is to incorporate existing information from previous studies, so-called 'historical data'. In the literature, methods for the integration of historical data into the control group have mainly been investigated so far. Promising (both frequentist and Bayesian) methods have been developed, studied, and compared against each other (Viele et al., 2014; Chen et al., 2000, 2011; Duan et al., 2006; Hobbs et al., 2011; Rietbergen et al., 2011; Schmidli et al., 2014). Viele et al. (2014) gave an comprehensive overview over existing methods. In comparison to the exclusive use of historical control data, there is only sparse literature on the integration of historical data from both arms, i.e., 'borrowing' information from both intervention and control arm (Gamalo-Siebers et al., 2017; Weber et al., 2018). This may be due to the fact that historical control data are more readily available, especially if the control group of the new study is a placebo group. However, incorporating historical control data does reduce the required sample size solely in the control arm of the new trial.

Nevertheless, there are situations in which the inclusion of historical data from both treatment arms can be useful, e.g. in the field of rare diseases, where the number of patients required in the currently planned study may be reduced or, *vice versa*, the power of the trial may be increased. Information may also be available for these scenarios from studies already performed, e.g. from previous pilot studies or trials where the new primary endpoint has already been assessed as a secondary endpoint. In both cases (control arm and two-arm borrowing), however, the question arises as to which requirements historical data must meet in order to be suitable for integration into a new study.

One of these requirements could be that large heterogeneity between the results of historical and current data could prove to be an obstacle for the successful integration of the historical data. Therefore, this should be penalized by the model that comprises the merging of the two data sources. For the case of control arm borrowing, it has already been shown that this heterogeneity results in an inflation of the type I error rate. However, it is not clear whether this result also applies to the case of a two-arm incorporation of historical data. Nevertheless, control of this quantity is of particularly interest since it is a main requirement for its suitability in the regulatory context (ICH E9 expert working group, 1999). As countermeasures, approaches which do not include the full information from the historical data but rather downweight it by a factor, have been proposed. In particular, the so-called 'power prior' approach uses this method, where the scaling factor controlling the downweighting can be

handled in a straightforward way (Ibrahim et al., 2015). However, the optimal determination of the scaling factor is currently still under discussion (Gravestock et al., 2017).

1.2 Objectives and Structure of the Present Work

In this thesis, the integration of historical experimental and control data in the planning and evaluation of two-arm clinical trials with a binary endpoint will be investigated. For sample size calculation based on binary endpoints, the resulting sample size does not only depend on significance level, power, and the assumed treatment effect, but is additionally influenced by the true response rate of the control group. This parameter mainly influences the type I error rate inflation in case that integration of historical data is only done for the control arm. Therefore, the appropriate handling of this nuisance parameter is of particular importance. Furthermore, these considerations should be integrated into a framework that allows the control of the type I error rate at the nominal significance level. Finally, it will be investigated whether and how this approach can lead to a benefit in terms of increased power or *vice versa* to a reduced sample size for a new clinical trial.

The outline of this thesis is as follows. In Chapter 2, the statistical methods are introduced. In particular, the general framework and notation is presented in Section 2.1. Afterwards, different general approaches for incorporation of historical data are compared and assessed for its suitability for the two-arm case. Then, Section 2.3 introduces the power prior framework and the application to binary data is examined. Following this, the subsequent analysis is presented in a frequentist (Section 2.4) and Bayesian (Section 2.5) framework. In Section 2.6, the operating characteristics in terms of type I error rate, power and the resulting rejections regions are examined. Based on these considerations, three approaches to handle and control the incorporation of historical data are introduced in Section 2.7 and their application to a sample size calculation procedure is presented in Section 2.8. Afterwards, some practical recommendations are given in Section 2.9 in order to simplify and thus, accelerate the presented calculation procedures. In Chapter 3, results from investigations and examples are presented. In detail, the performance of the proposed approaches is examined by systematic investigations on various scenarios in Section 3.1 and is illustrated in more detail by means of a clinical trial example in Section 3.2. Finally, the proposed methods and the corresponding results are evaluated, compared, and discussed in Chapter 4.

Comment: Major parts of the content of this work have already been pub-

lished (Feißt et al., 2020). This publication has been written by myself but includes comments and corrections from the co-authors.

Chapter 2

Methods

2.1 Framework

The framework supposed in this thesis is given by a two-arm clinical trial with binary outcome. The primary analysis is a two-sided test at a significance level α_0 assessing the test problem

$$H_0 : \pi_C = \pi_T \quad versus \quad H_1 : \pi_C \neq \pi_T,$$

where π_C denotes the true response rate for the control arm and π_T the true response rate for the treatment arm.

In the following, higher rates indicate a preferable outcome, e.g. response to treatment, nevertheless the two-sided test framework is maintained. In the further course of this thesis it will be elaborated why this framework is preferred.

Furthermore, the existence of historical data from a single historical two-arm trial with binary data in the following form is assumed: n_{CH}, c_H, and $n_{CH} - c_H$ denote the number of patients, responders, and non-responders in the historical control arm, respectively. Similarly, n_{TH}, t_H, and $n_{TH} - t_H$ denote the number of patients, responders, and non-responders in the historical treatment arm, respectively. Analogously, the data of the new clinical trial is denoted by n_C, c, and $n_C - c$ as number of patients, responders and non-responders in the control arm of the new trial and, similarly, n_T, t, and $n_T - t$ as number of patients, responders, and non-responders in the new treatment arm. This is depicted in Table 2.1.

As in the work by Viele et al. (2014), the investigations in this thesis are limited to the case where the historical data is fixed and, therefore, the performance characteristics (e.g. type I error rate and power) are conditional on fixed historical response rates. This refers to the case in which the data of an already

Table 2.1: Fourfold table of a clinical trial with binary outcome.

	Control arm	Treatment arm
Responders	c	t
Non-responders	$n_C - c$	$n_T - t$
Total	n_C	n_T

completed historical trial is incorporated into a new trial.

Thus, in the following, the integration of existing historical experimental and control data in the planning and evaluation of two-arm clinical trials with a binary endpoint will be investigated.

2.2 Methods for Borrowing Historical Data

Viele et al. (2014) presented different methods of incorporating historical control data into a new trial. In detail, they outlined six approaches:

1. Separate analysis: historical data are ignored, standard analysis of current study data.

2. Pooling: incorporating the whole historical information as if they had been observations of the new trial.

3. Single arm trial: the cutoff-rate 'to beat' in the treatment arm is deduced from the historical data (no control arm in the current study).

4. Test-then-pool: first a test is performed whether the historical control data are comparable to the current control data and then the data is pooled.

5. Power priors: the amount of historical data incorporated in the current study is controlled by a parameter.

6. Hierarchical modeling: assuming a distribution across historical studies and current study and measuring the variation across studies. A larger variation leads to a smaller amount of historical data incorporated into current study.

Viele et al. (2014) focused their presented approaches solely on incorporating historical control data. Therefore, in the next step, the suitability of these methods for incorporation of historical two-arm data in the framework of this thesis is examined, respectively. Since the approaches which are developed in this thesis should provide a framework that can be used already in the planning phase of a new study, approaches that are based on the availability of current study data will not be suitable for further investigations.

1. Separate analysis: analysis solely based on the current study data.
   Suitable for two-arm data.

2. Pooling: pooling historical and current study data as they are from one study.
   Suitable for two-arm data.

3. Single arm trial: as the focus is on two-arm randomized controlled trials in this study, this approach is not suitable for this thesis.
   Not suitable for the framework considered in this thesis.

4. Test-then-pool: Viele et al. suggest to test the comparability of historical and current study data. Therefore, the data of the new study has to be available. However, in the current framework of this thesis (integration of historical data in the planning and evaluation of two-arm clinical trials) this assumption is not met.
   Not suitable for the framework considered in this thesis.

5. Power priors: this approach is presented in more detail in Section 2.3. The parameter that controls the amount of historical data integrated in the current study can be predetermined without knowledge of the current study data. Thus, this approach fits to the framework considered in this thesis.
   Suitable for two-arm data.

6. Hierarchical modeling: as in the test and pool approach, in this approach the current study data has to be available and this assumption is not met. Furthermore, building a hierarchical model based on only two or few studies has been shown to perform rather poorly (Seide et al., 2019).
   Not suitable for the framework considered in this thesis.

In summary, the suitable methods range from a separate analysis (ignoring the historical data, standard analysis) to pooling (incorporating the whole historical information as if they had been observations of the new trial). One method that presents a compromise between separating and pooling is the so-called 'power prior' approach, a Bayesian approach first introduced by Ibrahim et al. (2000). Thereby, a parameter controls the amount of historical information that is borrowed from the historical data and ranges from 0 (no borrowing, separate analysis) to 1 (incorporation of whole information, pooling). Thus, the power prior approach already comprises the separating and the pooling approach. Therefore, in the following, this work is focusing on the power prior approach and its application to the integration of historical two-arm data in the planning and evaluation of two-arm clinical trials with binary outcome.

2.3 The Power Prior

The power prior approach is a Bayesian approach, i.e. its idea is developed based on the principles of Bayesian statistics. In Bayesian statistics, for estimating a parameter of interest, the 'prior' knowledge about the parameter is updated by collected data into a 'posterior' knowledge. Considering probability distributions, thereby a prior distribution illustrating the prior knowledge about the parameter (which can also be a non-informative flat distribution, e.g. the uniform distribution) is combined with a distribution derived from the collected data, the so called likelihood function, to a posterior distribution illustrating the posterior knowledge about the parameter of interest. This posterior distribution can again be used as a prior distribution for a new data collection. This reflects the idea of the power prior approach by updating a prior distribution at first by initially collected data (e.g. historical data) and afterwards updated with currently collected data (e.g. data of a current study). Thereby, in order to limit or control the 'influence' of the initially collected data on the posterior distribution, their likelihood function is depending on a so called 'power parameter', in the following denoted as δ. As the name suggest, the likelihood of the initially collected data is thereby provided with an exponent δ.

In the following, this principle is presented in a mathematical framework in the context of this thesis. Thus, an initial prior f_0 of a treatment effect Δ (parameter of interest) is updated by the likelihood L based on the historical data x_H (initially collected data), raised to the power of a weight $\delta \in [0, 1]$:

$$f_H\left(\Delta|\, x_H, \delta\right) \propto \mathrm{L}_H\left(\Delta|\, x_H\right)^{\delta} f_0(\Delta),$$

where f_H denotes the power prior. Note that 'proportional to' means that the resulting distribution has to adjusted by a constant to get a standardized distribution (with probabilities ranging from 0 to 1). The principle of the weighting is straightforward: When $\delta = 0$, the likelihood factor becomes 1, and thus only the initial prior is used, corresponding to a separate analysis in which the historical data are not used at all. Similarly, when $\delta = 1$, the likelihood factor is fully used (not downweighted) and therefore, it is the same as the usual Bayesian updating process of the initial prior, but now for the power prior as the initial prior. This corresponds to a situation of complete pooling historical and new data.

Updating the power prior f_H by the likelihood L based on the data of the new study x, the posterior distribution f is proportional to the power prior and the likelihood of the new data L:

$$f\left(\Delta|\, x,\, x_H, \delta\right) \propto L\left(\Delta|\, x\right) f_H\left(\Delta|\, x_H, \delta\right) \propto L\left(\Delta|\, x\right)\, \mathrm{L}_H\left(\Delta|\, x_H\right)^{\delta} f_0(\Delta).$$

Table 2.2: Fourfold table of a clinical trial with binary outcome and incorporated historical data.

	Control arm	Treatment arm
Responders	$c + \ \delta c_H$	$t + \ \delta t_H$
Non-responders	$(n_C - c) + \delta(n_{CH} - c_H)$	$(n_T - t) + \delta(n_{TH} - t_H)$
Total	$n_C + \delta n_{CH}$	$n_T + \delta n_{TH}$

In the case of two-arm trials with binary outcome, one is often interested in the posterior distribution of the true effect measured in terms of the rate difference $\Delta = \pi_T - \ \pi_C$ based on binomial likelihoods. Of course, also odds ratio or risk ratio might be the summary measure of interest in such situations. In this thesis the focus lies on the rate difference, however, the methods developed in this thesis can be straightforwardly adapted to other summary measures of interest.

When working with binary outcomes, the standard method in the corresponding Bayesian framework is to work with Beta distributions, that lead to a so-called Beta-binomial model. Using Beta distributions as prior distributions along with a binomial likelihood of the binary data, also leads to a Beta posterior distribution. A framework where the posterior distribution is in the same distribution family as the prior probability distribution is called a conjugate analysis. Conjugate analyses can be considered as worthwhile to achieve, since their handling is straightforward. For Beta(α, β) distributions (with parameters α and β), the determination of the parameters α and β in the case of binary outcome as is follows: If there are c responders within the n_C patients in the control arm, the initial prior Beta($\alpha_0, \ \beta_0$) of the trial arm is updated to the posterior distribution following a Beta($\alpha_0 + c, \ \beta_0 + n_C - c$) distribution. Similarly, if the initial prior is updated with the power prior likelihood, the historical data are simply downweighted with the factor δ to the power prior following a Beta($\alpha_0 + \ \delta c_H, \ \beta_0 + \delta(n_{CH} - c_H)$) distribution, where c_H denotes the number of responders within n_{CH} patients in the historical control arm. Thus, the posterior distribution of the control arm based on the power prior has the form of a Beta($\alpha_0 + c + \ \delta c_H, \ \beta_0 + (n_C - c) + \delta(n_{CH} - c_H)$) distribution.

2.4 Frequentist Analysis

Based on the principle of the simple form of the beta distribution, the power prior approach for binary outcomes from Section 2.3 can be transformed straightforwardly to a frequentist fourfold table with subsequent analysis (Zaslavsky, 2013). One simply adds the weighted historical data to the respective cell of the fourfold table (see Table 2.2). Thereby, $n_C, \ c$, and $n_C - c$ are the number of

patients, responders and non-responders, respectively, in the new control arm, and $n_t,\ t$, and $n_t - t$ are the number of patients, responders and non-responders in the new treatment arm. $\delta \in [0, 1]$ determines the amount of historical data that is incorporated in the new trial, $n_{CH},\ c_H$, and $n_{CH} - c_H$ are the number of patients, responders and non-responders, respectively, in the historical control arm, and $n_{TH},\ t_H$, and $n_{TH} - t_H$ are the number of patients, responders and non-responders in the historical treatment arm. In addition, if there is initial prior information in the form of a prior distribution, it can again be similarly added to the respective cell. Note that in the following, the investigations are limited to the case where there is no initial prior information, i.e., a vague, non-informative prior is used.

After determination of δ (which will be intensively evaluated in the further sections of this thesis) the fourfold table (Table 2.2) can be statistically analyzed, i.e. the test problem from Section 2.1 can be evaluated. A plethora of statistical procedures have been developed with regard to hypothesis testing in four-fold tables. Common procedures are the chi-square test, Fisher's exact test, the z-test (proportion test) as well as exact unconditional tests (e.g. the Fisher-Boschloo test). For a specific clinical trial application, the use of the latter class of tests could be more favorable (Lydersen et al., 2009). However, their use is involved with a larger computational effort and therefore, they are not suitable for extensive systematic investigations as performed in Chapter 3. Lydersen et al. (2009) recommend to refrain from the use of Fisher's exact test due to the fact that this test is too conservative. In addition, for a two-sided test problem applied to a $2x2$ fourfold table, the chi-squared test and the z-test are equivalent (since the z-test statistic corresponds to the square root of the test statistic of the chi-squared test (Fagerland et al., 2017)). However, the z-test and the chi-squared are merely approximate tests, thus, for small sample sizes, the true type I error rate occasionally exceeds the nominal significance level. Nevertheless, the further systematic investigations of this thesis are based on the chi-square test procedure, as the computational effort over the wide range of parameter settings is more feasible than for exact unconditional tests.

2.5 Bayesian Analysis

Following the Bayesian methodology, the method at hand is to work with using Beta distributions (following Section 2.3). However, analytical computation of the posterior distribution of $\Delta = \pi_T - \ \pi_C$, i.e. the difference of two beta-distributed parameters, is not trivial (because no conjugated model is available) (Kawasaki and Miyaoka, 2012; Lee, 2004; Howard, 1998; Altham, 1969; Nurminen and Mutanen, 1987). Nevertheless, this problem can be solved by perform-

ing Monte Carlo simulation (Chen et al., 2012) to obtain the empirical posterior distribution.

Seen from another perspective, in order to achieve a conjugate analysis, in the two-arm framework there exists a further Bayesian approach based directly on the difference $\Delta = \pi_T - \pi_C$.

This approach directly models the difference $\Delta = \pi_T - \pi_C$ as the parameter of interest and assumes underlying normal distributions for prior and likelihood (achieving a conjugate analysis). Therefore, the posterior distribution $f(\Delta|x, x_H, \delta)$ for Δ based on the historical data x_H, the data of the new study x, and the power parameter $\delta \in [0, 1]$ from the power prior approach is given as follows (using the same notation as in Section 2.3):

$$f\left(\Delta|\, x,\, x_H, \delta\right) \;\propto L\left(\Delta|\, x\right)\, \mathrm{L}_H\left(\Delta|\, x_H\right)^{\delta} f_0(\Delta).$$

Based on the assumption of normality it follows

$$L\left(\Delta|\, x\right) \sim N\left(p_T - p_C,\; \frac{p_T\left(1-p_T\right)}{n_T} + \frac{p_C\left(1-p_C\right)}{n_C}\right),$$

where $p_T = t/n_T$ and $p_C = c/n_C$. Similarly:

$$\mathrm{L}_H\left(\Delta|\, x_H\right) \sim N\left(p_{TH} - p_{CH},\; \frac{p_{TH}\left(1-p_{TH}\right)}{n_{TH}} + \frac{p_{CH}\left(1-p_{CH}\right)}{n_{CH}}\right),$$

where $p_{TH} = t_H/n_{TH}$ and $p_{CH} = c_H/n_{CH}$. Again, prior knowledge (e.g. from further historical trials) can be included straightforwardly through the initial prior $f_0(\Delta)$. However, as mentioned above, in this thesis the focus lies on the case where historical information from only one historical trial is available and thus in the following a non-informative prior (reflecting the case that there is no initial knowledge) is used.

2.6 Operating Characteristics

By integrating historical data, the operating characteristics (e.g. type I error rate, power) of the current trial may be altered. In the following, the influence of the historical data on important operating characteristics is examined, i.e. the type I error rate, the power, and the rejection regions.

2.6.1 Type I Error Rate

The type I error rate is the probability of falsely rejecting the null hypothesis H_0. As a result of integrating historical data, the type I error rate may change. Thus,

it may be inflated and thereby violate the nominal significance level which leads to the urgency to control the amount of inflation. In the frequentist framework, the actual type I error rate can be determined by calculating the proportion of fourfold tables that reject the null hypothesis weighted by their probability of occurrence under the null hypothesis $\pi_T = \pi_C$:

$$\sum_{c=0}^{n_C} \sum_{t=0}^{n_T} P\left(c, n_C \mid \pi_C\right) \cdot P\left(t, n_T \mid \pi_T = \pi_C\right)$$
$$\cdot I\left(P\left(c, n_C, t, n_T, \delta, c_H, n_{CH}, t_H, n_{TH} \mid \pi_T = \pi_C\right) < \alpha_0\right), \quad (1)$$

where $P\left(x, n|\pi\right)$ denotes the probability for x success in n trials by a binomial success probability of π and I is the indicator function (i.e. a function that is 1 if the condition in brackets is fulfilled and 0 otherwise). Note that from the formula of the type I error rate it follows that the type I error rate is depending on the true control proportion π_C for this binary framework.

In case that the borrowing of historical data is limited to the control arm, a type I error rate inflation occurs if the observed historical control rate notably differs from the true control rate π_C (see e.g Viele et al. (2014)). For two-arm borrowing, the type I error rate is rather independent from the difference between the observed historical and the true rate π_C which is equal to π_T under the null-hypothesis. Figure 2.1 depicts the actual type I error rate for the scenario where it is assumed that there are:

- 65 responders within 100 patients in the historical control arm,
- 75 responders within 100 patients in the historical treatment arm,
- 200 patients in both the new control and the new treatment arm,
- a significance level of α_0=0.05,

considering both the approach of control arm borrowing and two-arm borrowing with values of $\delta = 0$, 0.2, 0.4, 0.6, 0.8 and 1 (depicted in a color spectrum ranging from blue to red). Note that the 'waves' in the type I error curves are due to the character of the chi-squared distribution, since the actual type I error rate is dependent on the true control proportion π_C and on the sample size n (see Figure A.1 in the appendix).

For the control arm borrowing approach (Figure 2.1 top), one can see that the more the observed historical control rate (i.e. number of responses divided by number of patients in the control arm) differs from the true control rate π_C, the higher the type I error rate inflation is. Similarly, the smaller the amount of borrowing δ, the smaller the type I error rate inflation is. For δ= 0, the type I error rate is about 0.05, which corresponds to the nominal significance

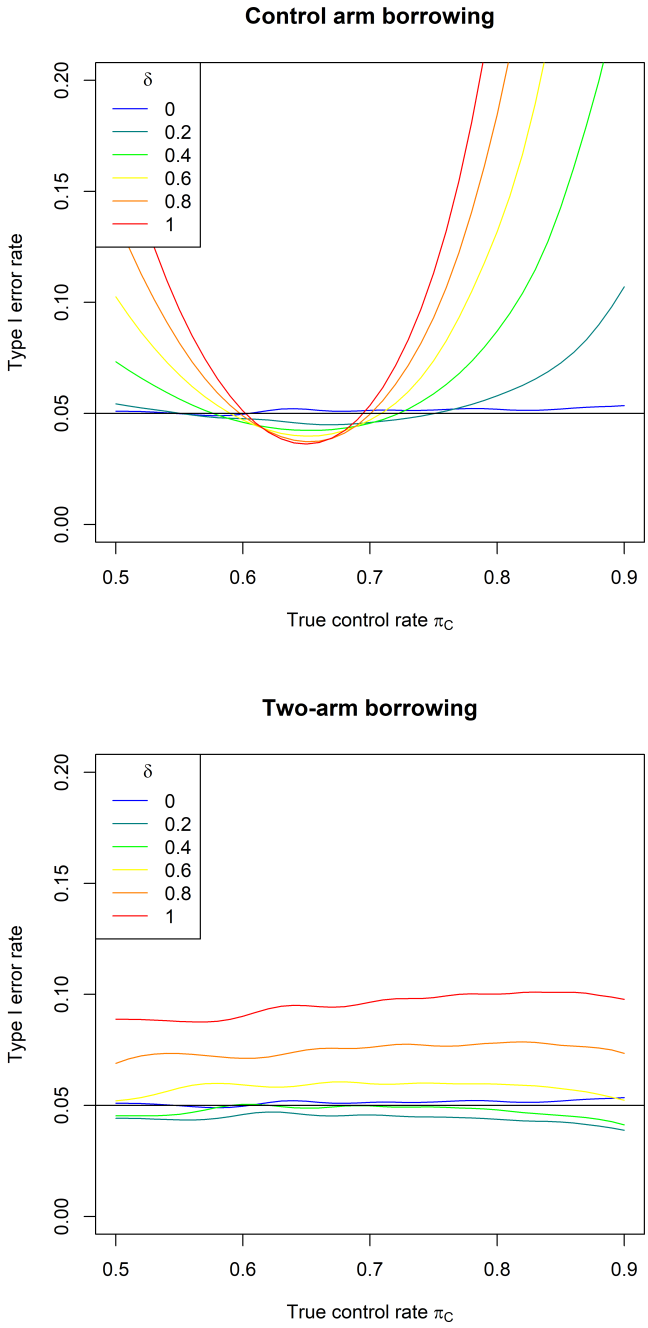

Figure 2.1: Actual type I error rate for control arm borrowing (top) and for two-arm borrowing (bottom) depending on true control rate $\pi_C \in [0.5, 0.9]$ using various values of δ.

level α_0. For the two-arm borrowing approach (Figure 2.1 bottom), the type I error rate is mainly influenced by δ but not by the true control response rate π_C. For δ= 0.2 and 0.4, the actual type I error rate is completely below the nominal significance level. Actually, instead of the difference between historical observed data and the true parameters, the observed historical difference is the main factor influencing the inflation of the type I error rate in case of two-arm borrowing. In Figure 2.2 (top), the type I error rate for increasing δ (0 to 1, on the x-axis) is displayed for

- an increasing historical difference ranging from 0 to 0.3 (in intervals of 0.05, depicted in a color spectrum ranging from blue to red),
- 65 responders within 100 patients in the historical control arm,
- 65+x (x ranging from 0 to 30 in intervals of five) responders within 100 patients in the historical treatment arm,
- 200 patients per arm in the new trial,
- a fixed true control response rate of $\pi_C = 0.65$,
- a significance level of $\alpha_0 = 0.05$.

It can be observed that for every scenario the type I error functions (i.e. the actual type I error rate depending on the amount δ of historical data that is included) are nearly convex and show some values below the significance level. Thus, it may be concluded that for every scenario there exists a $\delta > 0$ such that the significance level is controlled at $\alpha_0 = 0.05$. For small observed historical differences, even full borrowing is possible while controlling the type I error rate at the time. The maximal value of δ which still ensures type I error rate control mainly depends on the rate difference between treatment groups observed in the historical study: the larger the difference, the smaller is the maximal δ.

Regarding the Bayesian analysis, the definition of the type I error rate is based on continuous distributions (instead of discrete count data, as in the frequentist fourfold approach). Thus, the type I error for the assessment of the test problem:

$$H_0 : \Delta = \pi_T - \pi_C = 0 \quad \text{versus} \quad H_1 : \Delta \neq 0,$$

has to be based on integrals in stead of sums and therefore, can be calculated

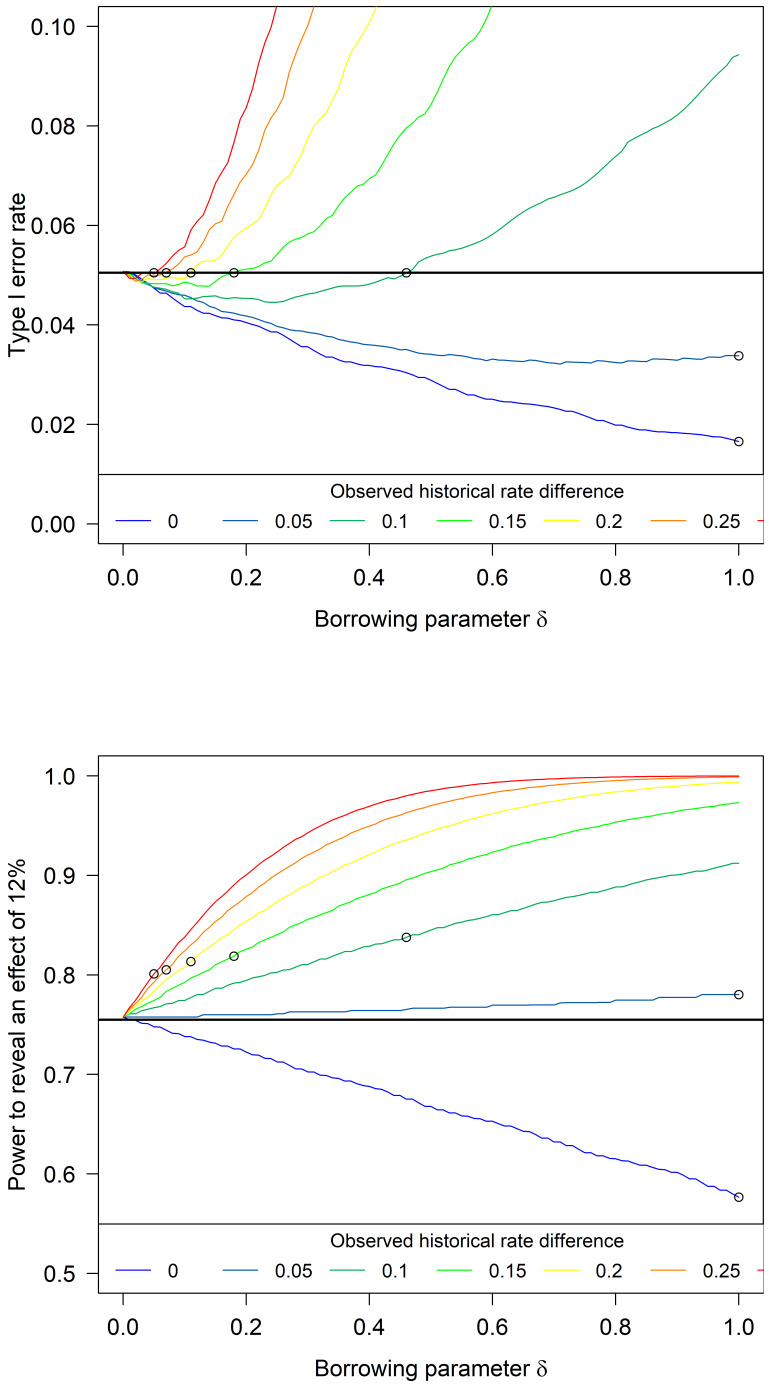

Figure 2.2: Top: Actual type I error rate depending on borrowing parameter $\delta \in [0, 1]$ for various observed historical rate differences. Bottom: Power to reveal an effect of $\Delta = 0.12$ depending on borrowing parameter $\delta \in [0, 1]$ for various observed historical rate differences. The dots identify $\delta^* = max_{\alpha \leq 0.05}\{\delta : \delta \in [0, 1]\}$, the maximum value of δ controlling the type I error rate α at the nominal significance level of $\alpha_0 = 0.05$.

as:

$$\int_{-\infty}^{\infty} N\left(0, \frac{\pi_C n_C + \pi_C n_T}{n_C + n_T}\right)(x)$$
$$\cdot(I\left(P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, p_{CH}, p_{TH}\right) > 0.975\right)$$
$$+ I\left(P\left(\Delta < 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, p_{CH}, p_{TH}\right) > 0.975\right))dx,$$

where

$$P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, p_{CH}, p_{TH}\right)$$
$$= \int_0^{\infty} N\left(\Delta, \frac{\pi_C n_C + \pi_C n_T}{n_C + n_T}\right)(x) \cdot N\left(\widehat{\Delta}_H, \frac{p_{CH} n_{CH} + p_{TH} n_{TH}}{n_{CH} + n_{TH}}\right)^{\delta}(x) d\Delta$$

with $\widehat{\Delta}_H = p_{CH} - p_{TH} = \frac{c_H}{n_{CH}} + \frac{t_H}{n_{TH}}$ and $N\left(\mu, \sigma^2\right)(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ denoting the density function of the normal distribution with mean μ and variance σ^2.

Compared to the frequentist framework, this Bayesian framework produces very similar results regarding type I error rate and δ^*. Systematic comparisons between these two frameworks can be found in Chapter 3.

2.6.2 Power

The aim of incorporating historical data is that the use of additional information leads to a benefit, i.e. an increase in power or, *vice versa*, a reduction of the required sample size for the new trial. Thus, just like the type I error rate, the power is an operating characteristic of main interest.

Despite the fact that a two-sided test problem is considered, in this thesis, the power of an effect pointing in the direction of the observed effect in the historical data is evaluated. This is due to the fact that a power increase by simultaneously controlling the type I error rate is in contradiction with the construction of uniformly most powerful tests in the theory of mathematical statistics, since the likelihood ratio test (which is equivalent to the chi-square test in the case of a fourfold table) is the uniformly most powerful test according to the Neyman-Pearson lemma (Kopp-Schneider et al., 2020) (Witting, 1985). This fact is discussed more detail in the further course of this thesis and especially in Chapter 4.

Similarly to the case of calculating the type I error rate (see Equation 1, Subsection 2.6.1), in the frequentist analysis the power calculation is based on the proportion of fourfold tables that reject the null hypothesis, weighted by their probability of occurrence, but now assuming that $\pi_C \neq \pi_T$, i.e. that there is a clinically relevant effect $\pi_T - \pi_C > 0$ to be detected.

Accordingly, based on the fourfold table approach, the power amounts to

$$\sum_{c=0}^{n_C}\sum_{t=0}^{n_T} P\left(c, n_C \mid \pi_C\right) \cdot P\left(t, n_T \mid \pi_T\right)$$
$$\cdot I(P\left(c, n_C, t, n_T, \delta, c_H, n_{CH}, t_H, n_{TH} \mid \pi_T = \pi_C\right) < \alpha_0) \ , \quad (2)$$

where again, $P\left(x, n|\pi\right)$ denotes the probability for x success in n trials by a binomial success probability of π and I is the indicator function.

As in the case of the type I error rate, the observed historical difference is the main factor influencing the power. Figure 2.2 (bottom) shows the power for increasing δ (0 to 1, on the x-axis) and increasing historical difference (0 to 0.30, in intervals of 0.05, depicted in a color spectrum ranging from blue to red) for the same scenarios as in the previous subsection (based on true rates of $\pi_C = 0.65$ and $\pi_T = 0.77$; accordingly, the initial power without borrowing amounts nominal to 0.758). The dots identify δ^*, the maximum value of δ by controlling the type I error rate at the significance level of $\alpha_0 = 0.05$ (for a fixed $\pi_C = 0.65$). Therefore, the 'price' that has to be paid when borrowing full information can be quantified: There is no or merely a slight increase in power for the scenarios where the type I error rate is always controlled up to $\delta^* = 1$ (observed historical difference of 0 or 0.05). *Vice versa*, for larger observed historical differences (0.25, 0.3), δ^* gets smaller, i.e., fewer information can be borrowed and, thus, the gain in power becomes smaller as well. Therefore, the incorporation of historical data while at the same time controlling the type I error seems the most beneficial in case of moderate observed historical rate differences (0.05-0.2).

Similarly as for the Bayesian calculation of the type I error rate, the formula for the Bayesian calculation of the power can be derived (see Section 2.5):

$$\int_{-\infty}^{\infty} N\left(d, \frac{\pi_C n_C + \pi_C n_T}{n_C + n_T}\right)(x)$$
$$\cdot(I\left(P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, \pi_T, p_{CH}, p_{TH}\right) > 0.975\right)$$
$$+ I\left(P\left(\Delta < 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, \pi_T, p_{CH}, p_{TH}\right) > 0.975\right))dx,$$

where

$$P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T, \pi_C, \pi_T, p_{CH}, p_{TH}\right)$$
$$= \int_0^{\infty} N\left(\Delta, \frac{\pi_C n_C + \pi_T n_T}{n_C + n_T}\right)(x) \cdot N\left(\widehat{\Delta}_H, \frac{p_{CH} n_{CH} + p_{TH} n_{TH}}{n_{CH} + n_{TH}}\right)^{\delta}(x) d\Delta$$

with $\widehat{\Delta}_H = p_{CH} - p_{TH} = \frac{c_H}{n_{CH}} + \frac{t_H}{n_{TH}}$ and $N\left(\mu, \sigma^2\right)$ denoting the density function

of the normal distribution with mean μ and variance σ^2. In addition, the effect size to detect is denoted by d.

Note that in this section the definition of the power was based only on d (and not on $|d|$). Thus, only the power in direction in favor of the observed historical effect is calculated. For calculating the two-sided power the formulas in this section have to adapted by basing the formula on $|d|$ instead of d.

2.6.3 Rejection Regions

Another operating characteristic of interest is how the incorporation of historical data influences the rejection regions of the corresponding statistical test. As its name suggests, the rejection region covers the area of observed effects where the null hypothesis is rejected, i.e. the amount of values for which the p-value of the corresponding statistical test falls below the nominal significance level α_0. In a clinical trial with binary outcome, the location of this region depends on the observed rates in the control arm (p_C) and in the treatment arm (p_T), and the corresponding sample sizes (n_C, n_T), respectively. The combinations of these values where the p-values firstly fall below the significance level α_0 is called boundary of the rejection region. In Figure 2.3 the boundary of the rejection region of a chi-square-test with $n_C = n_T = 200$ for various combinations of c (number of responses in the control arm, x-axis) and t (number of responses in the treatment arm, y-axis) are depicted for several values of the power parameter δ. In addition, it is assumed that there is data from a historical trial available, i.e. the same as in Subsections 2.6.1. and 2.6.2 (65 responses within 100 patients in the control arm and 75 responses within 100 patients in the treatment arm). The areas above the respective upper lines and the areas below the respective lower lines per color are the rejection regions.

It can be seen that by integration of historical data the upper boundaries of the rejection region decrease, as well as the lower boundaries. However, the lower boundaries decrease to a greater extent than the upper limits. Thus, as a consequence, by integration of historical data the rejection regions are becoming smaller. Therefore, integrating historical data is only beneficial when focusing on the benefit for revealing an effect in direction of the effect observed the historical data. This fact and its consequences are discussed in further detail in the Chapter 4.

2.7 Determination of Power Parameter

The choice of the value for the weighting parameter δ is a topic which is currently still under discussion. Methods range from determination (e.g. by an expert)

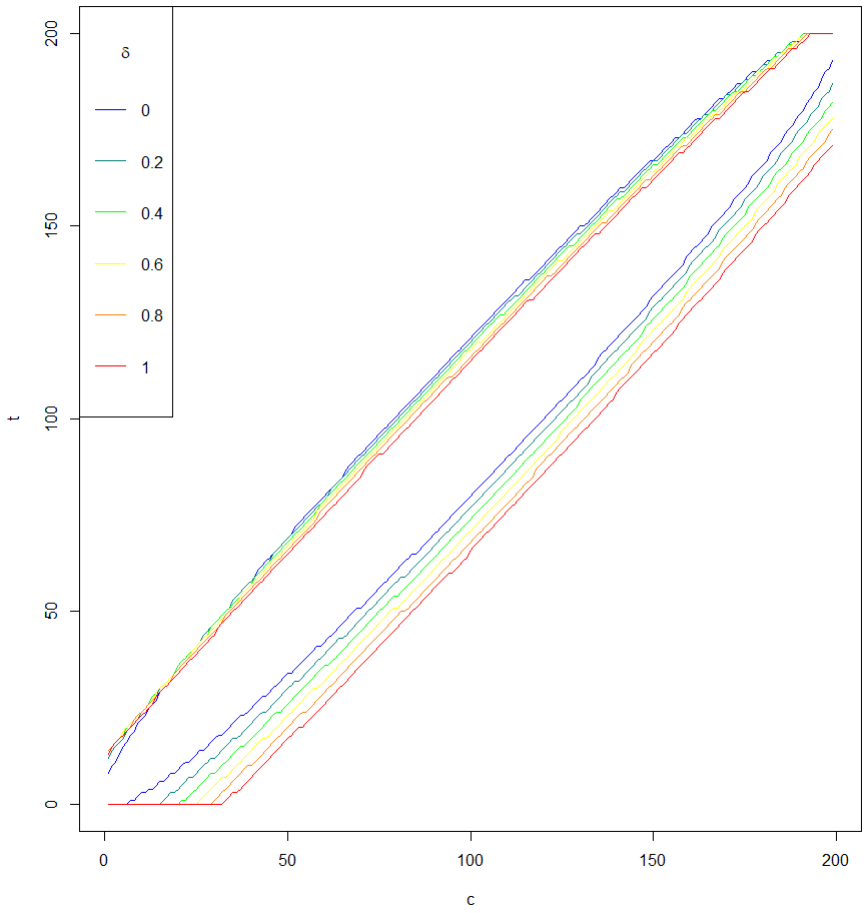

Figure 2.3: Boundary of the rejection regions for several values of δ, c denotes the number of responses in the control arm, t denots the number of patients in the treatment arm.

to treating δ as an unknown parameter (e.g. fully Bayes approach (Gravestock et al., 2017)) or estimation based on the data of the historical and the new study (e.g. empirical Bayesian approach (Gravestock et al., 2017)). Nevertheless, all methods have in common that they aim to reduce the occurring type I error rate inflation, i.e. when the type I error rate exceeds the nominal significance level α_0. Furthermore, in application in a clinical trial it is of main interest to control the type I error rate. In the ideal case the type I error rate can be controlled below the nominal significance level.

2.7.1 Global Approach

A possible interpretation of Subsection 2.6.1 may be that for every scenario there exists a value $\delta^* > 0$ that determines the maximal δ which still guarantees type I error rate control. Furthermore, from Formula (1) (Subsection 2.6.1) it also follows that the type I error rate depends on the true control response rate π_C, which can be regarded as a so-called 'nuisance' parameter. Nuisance parameters are parameters which one primarily does not intend to estimate, but need to be considered nonetheless (Fagerland et al., 2017). Compared to the case of control arm borrowing, in the case of two-arm borrowing, the unknown value of π_C has a rather small impact on the type I error rate (see Figure 2.1 bottom). However, type I error rate control needs to be ensured for all possible values of π_C because δ^* depends on π_C, and since δ^* determines the amount of historical data integrated into the fourfold table on which the type I error calculation is based, the resulting type I error rate consequently also depends on π_C. Therefore, δ^* can be determined by calculating the maximal δ for which the actual type I error rate α falls below a predetermined value α_0 for all values of π_C, and then taking the minimum of all maximal δ . This approach is valid due to the convexity of the type I error functions (see also Figure 2.2 (top)):

$$\delta^* = \min_{\pi_C \in [0,1]} \max\{0 < \delta \leq 1 \mid \alpha(\delta \mid c_H, n_{CH}, t_H, n_{TH}, n_C, n_T, \pi_C) < \alpha_0\} \ . \quad (3)$$

In the following, this procedure is referred to as the 'global approach' for the determination of δ^*.

Thereby, the convexity of the type I error functions in terms of δ is a crucial assumption for this 'minimax' approach. In the case of a normally distributed test statistic (which is closely related to the test statistic of the chi-square test, due its relationship to the z-test (Fagerland et al., 2017)), it is possible to prove the convexity of the type I error functions.

For this purpose, a normally distributed test statistic is considered. Note that in this thesis the presented calculations were based on the chi-square test

statistic. Since the test statistic of a test based on a chi-square distribution can be converted into a normal distribution by taking its square root and adding a sign function depending on the direction of the treatment effect, the result obtained for the normal distribution can be transferred to the chi-square test setting. However, the test statistic of the chi-square test is only approximately chi-square distributed and thus, the type I error curves that can be seen in Figure 2.2 (top) are only approximately convex.

The proof is divided into three parts: At first, the convexity of a standard normally distributed test statistic modified by the inclusion of historical data by increasing δ is proven. At second, it is proven that if a two-dimensional function is convex in one variable, then the one-dimensional function resulting from integration over the other variable is convex, too. Based on these results, the convexity of the type I error rate in terms of δ is proven in the third part.

Part I:

It is assumed the random variable T to be a standard normally distributed test statistic, which becomes shifted due to the inclusion of historical data. Furthermore, it is assumed that the test statistic of the historical data follows a $N\ (\widehat{\mu}, \widehat{\sigma}^2)$ distribution, with $\widehat{\mu}, \widehat{\sigma}^2$ estimated from the historical data. Based on the concept of the power prior (with a uniform initial prior) it follows:

$$T\left(x, \delta\right) \propto \frac{1}{\sqrt{2\pi}} \exp(-x^2) \cdot \frac{1}{\sqrt{2\pi}^{\delta}} \exp\left(\frac{-\delta(x-\widehat{\mu})^2}{2\widehat{\sigma}^2}\right)$$

as the distribution function of the test statistic T. To prove the convexity in δ, it is assumed x to be a constant and consider the function

$$T\left(\delta\right) = \frac{a}{b^{\delta}} \exp\left(-\delta c\right) \ ,$$

where a, b and c are positive constants. Convexity can be shown by the positivity of the second derivative of $T\left(\delta\right)$. The second derivative T'' of $T\left(\delta\right)$ based on the chain rule and product rule is

$$T''(\delta) = \frac{a}{b^{\delta}} \exp\left(-\delta c\right) \ \left(c^2 + c \cdot \ln(\mathrm{b}) \ + \frac{1}{4}(\ln(\mathrm{b}))^2\right) = \frac{a}{b^{\delta}} \exp\left(-\delta c\right) \ (c + \frac{1}{2}\ln(\mathrm{b}))^2$$

which is positive since each of the factors is positive.

Part II:

In this part, it is proven that if $f(x, y)$ is a function that is convex in y for all $x \in [a, b]$, then $\int_a^b f\left(x, y\right) dx$ is also convex in y.

To prove this it is defined: $g\left(y\right) := \int_a^b f\left(x, y\right) dx$ for $a, b \in \mathbb{R}$. A further

definition of the convexity of a function g is given by showing that:

$$g\left(\lambda y_1 + (1-\lambda)\, y_2\right) \leq \lambda g\left(y_1\right) + (1-\lambda)\, g(y_2)$$

for all real $y_1,\ y_2\ \in \mathbb{R}$ and $0 \leq\ \lambda \leq 1$. It follows from the convexity of f and the monotonicity of the integral that

$$\begin{aligned} & g\left(\lambda y_1 + (1-\lambda)\, y_2\right) \\ &= \int_a^b f\left(x, \lambda y_1 + (1-\lambda)\, y_2\right) dx \\ &\leq \int_a^b \lambda f\left(x, y_1\right) + (1-\lambda)\, f\left(x, y_1\right) dx \\ &= \lambda \int_a^b f\left(x, y_1\right) dx + (1-\lambda) \int_a^b f\left(x, y_2\right) dx \\ &= \lambda g\left(y_1\right) + (1-\lambda)\, g(y_2)\,. \end{aligned}$$

Part III:

At last, the results of part I and part II are used to show the convexity of the type I error rate in δ. Based on the test statistic T, the type I error rate in terms of δ is

$$Type\ I\ error(\delta) = \int_{-\infty}^{q} T(x, \delta) dx + \int_r^{\infty} T(x, \delta) dx$$

with q and r denoting the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$- quantile of the standard normal distribution, respectively. Note that the improper integrals are integrable since the integrand is a probability distribution. Each of the terms on the right hand side is convex due to the results of part I and part II. Thus, due to the fact that the sum of two convex function is also convex, the type I error rate is convex in δ. q.e.d.

2.7.2 Local Approach

Since especially for the chi-square tests the actual type I error rate is sensitive to very small or large values of π_C (see Figure A.1 and Fagerland et al. (2017))), it might be reasonable that control of the type I error rate does not need to be ensured for the whole range of π_C. One possible way to achieve this is the Berger and Boos procedure (Berger and Boos, 1994). Berger and Boos propose to control the type I error rate only in a $1 - \gamma$ confidence interval of a nuisance parameter (π_C), which is built based on the historical control data. However,

Table 2.3: δ^*(the maximum value of δ controlling the type I error rate) for various values of γ (from the procedure of Berger and Boos (1994)).

γ	$1-\gamma$ confidence interval for π_C	δ^*
0 (global approach)	$[0;1]$	0.41
0.00001	$[0.42;0.84]$	0.4
0.0001	$[0.45;0.82]$	0.4
0.0005	$[0.47;0.80]$	0.39
0.001	$[0.48;0.79]$	0.39
0.002	$[0.49;0.79]$	0.36
0.01	$[0.51;0.77]$	0

to guarantee global control of the significance level by α_0, it has to be locally adjusted to $\alpha_0 - \gamma$. Then, δ^* can be determined by calculating the maximum δ in the $1-\gamma$ confidence interval of π_C based on a local significance level of $\alpha_0 - \gamma$:

$$\delta^* = \min_{\pi_C \in [a,b]} \max\{0 < \delta \leq 1 \mid \alpha(\delta \mid c_H, n_{CH}, t_H, n_{TH}, n_C, n_T, \pi_C) < \alpha_0 - \gamma\} \ , \quad (4)$$

where $[a,b]$ is the respective $1-\gamma$ confidence interval and γ has to be prespecified. Since the type I error functions for π_C are nearly flat (Figure 2.1), a quite small value can be chosen for γ, or, *vice versa*, a relatively wide confidence interval. For example, Lydersen et al. found that $\gamma = 0.0001$ is approximately optimal under rather general conditions (Lydersen et al., 2012). In the following, the impact of various values of γ on the resulting value of δ^*, is investigated. Hereby the following scenario is examined:

- 65 responders within 100 patients in the historical control arm,
- 75 responders within 100 patients in the historical treatment arm,
- 200 patients per arm in the new trial.

The results are depicted in Table 2.3. Despite the fact that the range of the confidence intervals are very similar for $0.0001 \leq 0.01$, only the values $\gamma \leq 0.001$ resulted in a similar large δ^* while decreasing for higher values. Generally, larger values of γ lead to smaller δ^* (type I error functions for π_C have to lie below a stricter local significance level of $\alpha_0 - \gamma$) and smaller values of γ lead to unnecessarily broad confidence intervals including rather unrealistic (i.e. large difference between observed historical rates and true rates) or too sensitive (i.e. extremely small or large values of π_C) scenarios. Thus, the recommendation of $\gamma = 0.0001$ seems to be legitimate. Therefore, in the presented framework as well as in the Results (see Chapter 3), a value of $\gamma = 0.0001$ is considered. However, a uniformly optimal value of γ for all scenarios does not exist but depends on

the specific situation. Further, more systematic considerations were conducted and confirmed the reasonable choice of the value 0.0001; those considerations can be found in Appendix A.2.

In the following, this procedure is referred as the 'local approach' for the determination of δ^*.

2.7.3 Independent Approach

To avoid conflicts with the nuisance parameter especially for very low and very high values of π_C, not only is it possible the reduce the range of values for which the type I error rate has to be controlled (local approach) but also the approach for the estimation of δ^* can be made independent of π_C. This can be done by a so-called variance-stabilizing transformation (e.g. arcsine transformation (Yu, 2009)) resulting in a constant variance for all values of a nuisance parameter. Especially for the presented Bayesian framework from Section 2.5, where one works directly with the difference of the proportions $\Delta = \pi_T - \pi_C$, it could be desirable and elegant to get rid of the dependency of the nuisance parameter π_C. Therefore, a variance-stabilizing transformation by the function $f(p) = \arcsin\sqrt{p}$ was applied. This results in a constant variance which is only depending on the sample sizes of the trials. Subsequently, calculation of the true type I error rate α can be obtained by:

$$\int_{-\infty}^{\infty} N\left(0, \frac{n_C + n_T}{2n_C n_T}\right)(x) \cdot \left(I\left(P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T\right) > 0.975\right)\right.$$
$$\left. + I\left(P\left(\Delta < 0 \mid n_{CH}, n_{TH}, n_C, n_T\right) > 0.975\right)\right) dx$$

and

$$P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T\right)$$
$$= \int_0^{\infty} N\left(\Delta, \frac{n_C + n_T}{2n_C n_T}\right)(x) \cdot N\left(\widehat{\Delta}_H, \frac{n_{CH} + n_{TH}}{2n_{CH} n_{TH}}\right)^{\delta}(x) d\Delta$$

again, with $\widehat{\Delta}_H = p_{TH} - p_{CH}$ and $N\left(\mu, \sigma^2\right)$ denoting the density function of the normal distribution with mean μ and variance σ^2. Based on this definition of the type I error rate, δ^* can be determined (independent from π_C) by

$$\delta^* = \max\{0 < \delta \leq 1 \mid \alpha(\delta \mid c_H, n_{CH}, t_H, n_{TH}, n_C, n_T) < \alpha_0\} \ . \ (5)$$

Similarly to the definition of the type I error rate, the formula for the power

can be obtained by:

$$\int_{-\infty}^{\infty} N(d, \frac{n_C + n_T}{2n_C n_T})(x) \cdot (I\left(P\left(\Delta > 0 \mid n_{CH}, n_{TH}, n_C, n_T\right) > 0.975\right) + I\left(P\left(\Delta < 0 \mid n_{CH}, n_{TH}, n_C, n_T\right) > 0.975\right))dx$$

with the same notations as above and where d denotes the assumed effect size. In the following, this procedure is referred as the 'independent approach' for the determination of δ^*.

Note that the use of a variance-stabilizing transformation is arguable, since especially for the very high and the very low values of π_C the transformation is inexact (Warton and Hui, 2011). Therefore, the benefits by using this method should be treated with caution and the results presented in this thesis will be interpreted accordingly (see Chapters 3 and 4).

2.8 Sample Size Calculation

When planning a new trial, one usually aims to achieve a prespecified power $1 - \beta_0$, e.g. 0.8 or 0.9 as the probability to reveal an assumed effect with effect size Δ. As it follows from Subsection 2.6.2, the proposed methods can be employed to yield an increased power. *Vice versa*, they can also be used to reduce the sample size required to achieve a prespecified target value for the power. For the presented framework, an integration of historical data is not sensible in all scenarios, as it mainly depends on the observed rate difference of the historical data (see Subsection 2.6.2) whether there is an advantage or not. If in a respective scenario an increase of power can be achieved, it follows that there further exist combinations of δ^*, n_C, and n_T , with n_C and n_T smaller than derived from the initial sample size calculation (without incorporation of historical data but based on the same assumptions). To identify the optimal combination, i.e. those resulting in the largest reduction in sample size (compared to the initial sample size), the combination of n_C, n_T and δ^* with the smallest n_C and n_T has to be determined. Note that in general this most beneficial combination should be accompanied by the largest possible δ^* (by simultaneously controlling the type I error rate). In summary, for performing a sample size calculation based on

- a predefined significance level,
- a predefined power to reveal an assumed effect size of Δ,
- for given historical data,

one has to find the value of δ^* that results in the smallest sample size n_C and n_T in a 'new' trial.

However, searching for this combination would be based on a large computational effort since for every combination in the grid of n_C, n_T, and δ^*, the respective true type I error rate and power have to be calculated. In order to reduce this effort and to accelerate the calculation processes, some practical recommendations are therefore given in the following section.

2.9 Practical Considerations

The methods proposed in the previous sections are particularly computationally intensive, especially the 'grid search' over a huge amount of parameter combinations in the sample size calculation procedure requires a rather large computational effort.

In general, the calculations presented in this work are based on extensive exact calculations of the type I error rate and the power (e.g. Equation (1), Subsection 2.6.1). Thereby, a double sum over the sample sizes of a new trial is considered, which increases by the square of the sample size. Therefore, for large sample sizes, a determination of the type I error rate and power could be accomplished more efficiently by simulations to reduce the computation effort.

In the following, some further practical aspects, methods, and algorithms are presented, which can reduce the computation effort or improve the estimation for the above introduced calculation methods.

2.9.1 Determination of Power Parameter

To find the maximal δ for the amount of π_C, the nested intervals procedure (which works due to the convexity of the type I error functions, see Subsection 2.7.1) can be used. The nested intervals procedure restricts the value of interest (in this case: the root of a function) in a sequence of nested intervals with decreasing width. Thus, the value of δ^* can be determined with arbitrarily precise accuracy. On the one hand, this method considerably decreases the number of computations, and on the other hand the computational effort can be controlled by determining the accuracy of the estimation.

Furthermore, for the local approach the determination of a confidence interval is not straightforward, since there are many different approaches to estimate a confidence interval for a binomial proportion. Methods range from a normal approximation interval to exact methods (Brown et al., 2001). In the case of the local approach the use of the Clopper-Pearson confidence interval may be preferable since it always fulfills the coverage criterion and thus guarantees the main-

tenance of the confidence level (Fagerland et al., 2017). The Clopper-Pearson confidence interval is an exact interval for estimating a binomial proportion and its limits can be calculated based on Beta distributions.

2.9.2 Algorithm for Sample Size Calculation

Finding the value of δ^* for the sample size calculation procedure from Section 2.8 via a 'grid search' requires a rather large computational effort, since for every possible combination of n_C, n_T, and δ^* the type I error rate and the power have to be calculated, respectively. Therefore, the following algorithm may be useful to find this combination in a less time-consuming way, based on predetermined values for significance level α_0, power $1 - \beta_0$, and π_C and π_T:

- Step 1: calculate n_C and n_T based on the predetermined parameter values (standard sample size calculation) and opt for the local approach, the global approach or the independent approach (see Subsections 2.7.1, 2.7.2 and 2.7.3).
- Step 2: calculate δ_0^* as in Equations (3), (4) or (5) (depending on the chosen approach) based on c_H, $n_{CH,}$, t_H, n_{TH}, n_C, n_T, and π_C. By the integration of the historical data, it has to be checked if an increase in power is obtained. If not, then stop, as the integration of historical data does not yield any benefit.
- Step 3: decrease n_C and n_T until the smallest values are reached such that power still lies above the predefined power level $1 - \beta_0$.
- Step 4: calculate δ_1^* as in (3), (4) or (5) (depending on the chosen approach) based on c_H, $n_{CH,}$, t_H, n_{TH}, n_C, n_T, and π_C and on the new n_C and n_T from Step 3.
- Step 5:
  - If $\delta_1^* > \delta_0^*$, obtain an increase in power as in Step 2 and go back to Step 3 with $\delta_1^* = \delta_0^*$ and with n_C and n_T from Step 4.
  - If $\delta_1^* \leq \delta_0^*$, stop. n_C, n_T (from Step 4), and δ_0^* represent the preferable combination.

This algorithm works since for decreasing sample sizes n_C and n_T (Step 3) the type I error rate decreases as well. This phenomenon occurs only if the type I error rate is slightly below the nominal significance level. This is due to the fact that for increasing n_C and n_T and fixed δ^*, the 'weight' of the historical data decreases and, therefore, the type I error rate approaches the

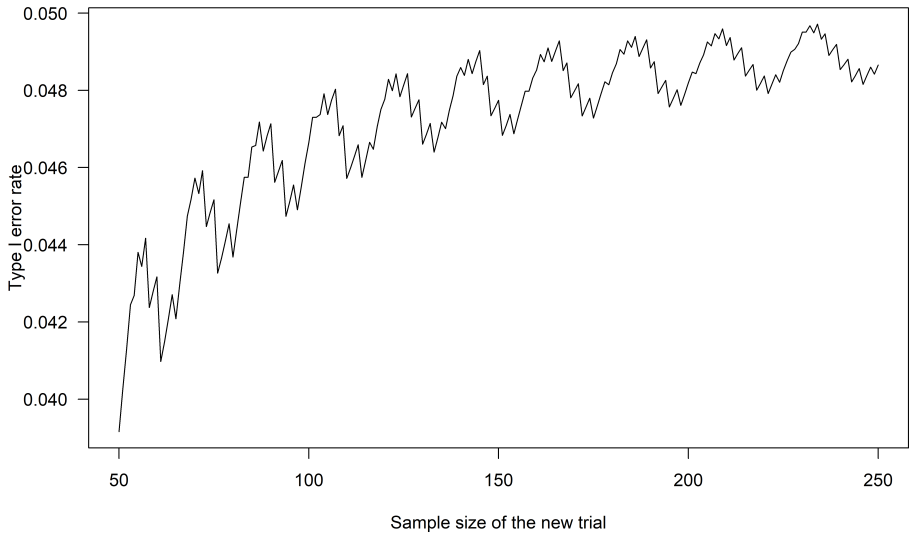

Figure 2.4: Actual type I error rate depending on the sample size of the new trial for $c_H = 65$ responders within $n_{CH} = 100$ patients in the historical control arm, $t_H = 75$ responders within $n_{TH} = 100$ patients in the historical treatment arm, a fixed $\pi_C = 0.65$ and a fixed borrowing parameter $\delta = 0.4$.

nominal significance level. This fact is depicted in Figure 2.4 for the scenario $c_H = 65$ responders within $n_{CH} = 100$ patients in the historical treatment arm, $t_H = 75$ responders within $n_{TH} = 100$ patients in the historical control arm, fixed $\pi_C = 0.65$, and fixed $\delta = 0.4$. Note that the 'mountains and valleys' are due to the discrete character of the chi-square test statistic.

Note that if in the algorithm $\delta_1^* > \delta_0^*$ (Step 5) and one therefore has to go back to Step 3, this is regarded as an additional algorithm step in the following chapter, in which the number of steps required by the algorithm is evaluated.

Furthermore, for the independent approach this algorithm simplifies, because the calculation of δ_0^* and δ_1^* (Step 2 and Step 4) is independent from π_C and therefore the computational effort is strongly reduced.

The code based on the programming language R (Ihaka and Gentleman, 1996) for the presented sample size calculation procedure for each of the three approaches presented in this thesis can be found in Appendix B. Note that due to practical reasons (the function chisq.test from R produces the result 'NA' for zeros in the fourfold tables), the sums in the code which are based on Equations (1, see Subsection 2.6.1) and (2, see Subsection 2.6.2) start at 1 instead of 0. Nevertheless, this aspect is negligible, since omitting these scenarios does not have any impact on the results presented in this thesis.

Chapter 3

Results

Comment: The results presented below differ in some instances slightly from those already published in Feißt et al. (2020) (mainly regarding the independent approach). This is due to recently optimized calculations implemented in the programming code. However, this does not affect the conclusions made in this publication.

3.1 Systematic Investigations

Based on the considerations in the previous chapter, it can be assumed that for every scenario a value δ^* can be determined so that a certain amount of two-arm historical data (represented by δ^*) from a previously performed clinical trial can be incorporated in a current two-arm trial while simultaneously controlling the type I error rate at the nominal significance value. However, it is not clear so far whether this integration of historical information may result in a benefit in terms of an increase in power or, *vice versa*, a sample size reduction, since this depends on the shape of the observed historical data. To detect factors that favor a successful incorporation (i.e. type I error rate control and sample size reduction), systematic investigations are performed in the following for the above proposed global, local, and independent approach. The main outcome of interest is determined as the percent sample size saved compared to the sample size required without integration of historical data. Thereby, the most beneficial (in terms of saved sample size, compare Section 2.8) combination of the parameters δ^*, n_C, and n_T is calculated. To reduce the computation effort, the algorithm presented in Subsection 2.9.2 is used.

3.1.1 Setup

In the following, the setup of the systematic investigations is defined. Possible influence factors are varied and the resulting values of δ^* and proportion of sample size saved are considered. The setup of considered parameter values is as follows:

- True control proportion $\pi_C = 0.1,\ 0.2,\ 0.3,\ 0.4$,
- Effect size $\Delta = \pi_T - \ \pi_C = 0.1,\ 0.15,\ 0.2,\ 0.25$,
- Responses in historical control and treatment arm $c_H = 20;\ t_H = 21,\ 22, \ldots,\ 50$,
- Sample size of the historical trial $n_{CH} = n_{TH} = 100$,
- Nominal significance level $\alpha_0 = 0.05$,
- Power $1 - \beta_0 = 0.8$,
- Berger and Boos parameter (solely for the local approach) $\gamma = 0.0001$.

Thus, a total of 480 scenarios (4 different values for π_C, 4 different values for Δ, 30 different values for t_H) were included in the investigation. Note that due to the symmetry of the binomial distribution, only response rates smaller than 0.5 are investigated; corresponding results for response rates larger than 0.5 would be identical. Variation of further parameters was not examined due to the following reasons:

- Number of responses in the historical control arm c_H: only the difference between c_H and π_C is of interest, which is already reflected by the variation of π_C.
- Sample sizes of the historical trial n_{CH}, n_{TH}: the amount of historical data that is borrowed remains always the same, due to an anti-proportional relationship between the historical sample size and the borrowing parameter; e.g., if $n_{CH} = n_{TH} = 100$ and $\delta^* = 0.25$ then, if $n_{CH} = n_{TH} = 50$, a value of $\delta^* = 0.5$ is obtained. However, there are scenarios where the amount of historical data will be limited (e.g. due to a low number of patients in the historical study and δ^* is limited to 1). However, it follows from Figure 2.2 that the most beneficial scenarios are not those accompanied with larger δ^* and therefore the limitation of the historical data to a 'realistic' value of $n_{CH} = n_{TH} = 100$ seems legitimate.
- $n_{CH} \neq n_{TH}$ and $n_C \neq n_T$: in most cases, unbalanced designs are not of interest since the largest power (and therefore the largest reduction in

sample size) is achieved for a balanced design. The more unbalanced the scenarios are, the larger the inflation of the type I error rate is (since the considered situation becomes more and more similar to the case of one-arm (control arm) borrowing) and thus, less sample size can be saved (see Figure 2.1 top). Nevertheless, there are situations in clinical trial practice where unbalanced designs are used or where unbalanced data of the sample sizes of historical trials are unbalanced. The presented approaches are also able to deal with this case (see Section 3.2).

- The Berger and Boos parameter γ (solely for the local approach): for the main investigations as defined above, γ is set to $\gamma = 0.0001$ as proposed in Subsection 2.7.2.

- Significance level α_0 and power $1-\beta_0$: the proposed approaches work in the same way for (realistic) values other than $\alpha_0 = 0.05$ and $1 - \beta_0 = 0.8$. No fundamental difference are expected for different parameter values since changing these characteristics would linearly transform the value space of the outcome parameters but not change the fact that a certain scenario is more beneficial than another. In addition, it would become difficult to compare the results for different values of α_0 and β_0 if they would be altered.

Altogether, 480 scenarios for each of the different approaches (global, local, and independent) are considered and compared in the following. For each of these three approaches, every calculation step where the maximum of the actual type I error rate has to be found is based on an amount of 101 equidistant values of π_C, respectively. Similarly, also the amount of attainable values for δ^* is restricted to a number with two decimal places. Note that for the local approach in the scenarios including $\pi_C = 0.4$, this value was not located in the respective $1 - \gamma$ confidence interval for π_C calculated based on the historical control data (e.g. the Clopper-Pearson confidence interval for $c_H = 20$ and $n_{CH} = 100$ is $[0.075; 0.387]$). Therefore, for these scenarios, no calculations were performed. However, in order to obtain the same amount of scenarios for the three approaches, respectively, for these scenarios, the historical data was modified by adding 10 responses in each historical treatment arm (i.e. $c_H = 30$ and $t_H = 31, ..., 60$).

3.1.2 Global Overview

Summarizing all 1440 scenarios for the three approaches, up to 22.2% of the sample size could be saved by integration of historical data. However, in 14.2%

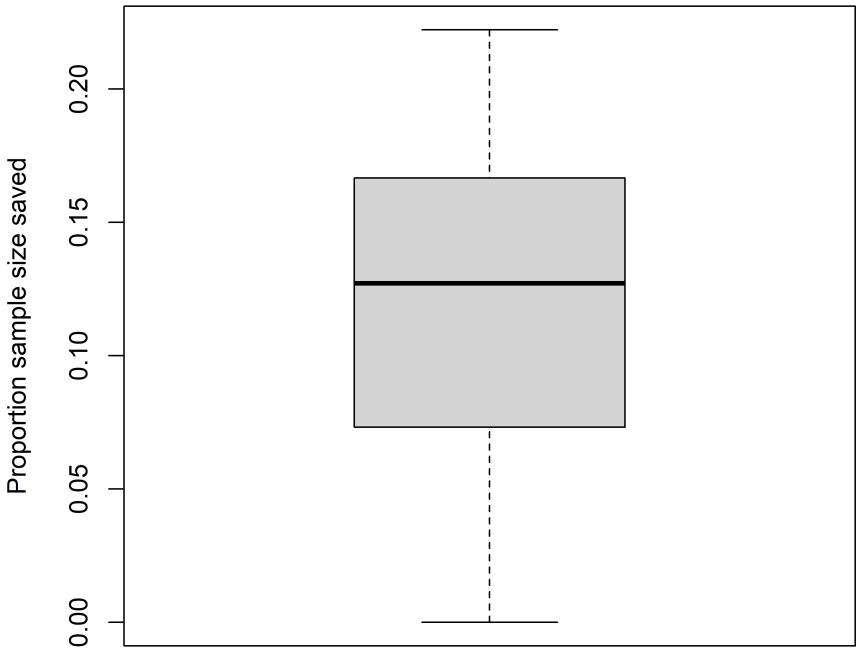

Figure 3.1: Boxplot of sample size saved accumulated over all scenarios and approaches.

of the scenarios, no benefit at all from incorporating historical data could be observed. In more than 50% of the scenarios, the sample size reduction amounted to at least 12.7% (median). Figure 3.1 shows a boxplot of the proportion of the saved sample size, the whiskers range from the minimum (no sample size saved) up to the maximum (22.2%).

To reduce the computational effort, the algorithm presented in Subsection 2.9.2 was used. In some cases the algorithm needed up to 5 steps to converge. In most cases, however, only one or two steps were required indicating a fast convergence of the algorithm in the majority of the scenarios (see Figure 3.2).

3.1.3 Variation in Parameter Values

In this subsection, the influence of difference parameters on the results of the systematic investigations are evaluated. The results presented in this subsection

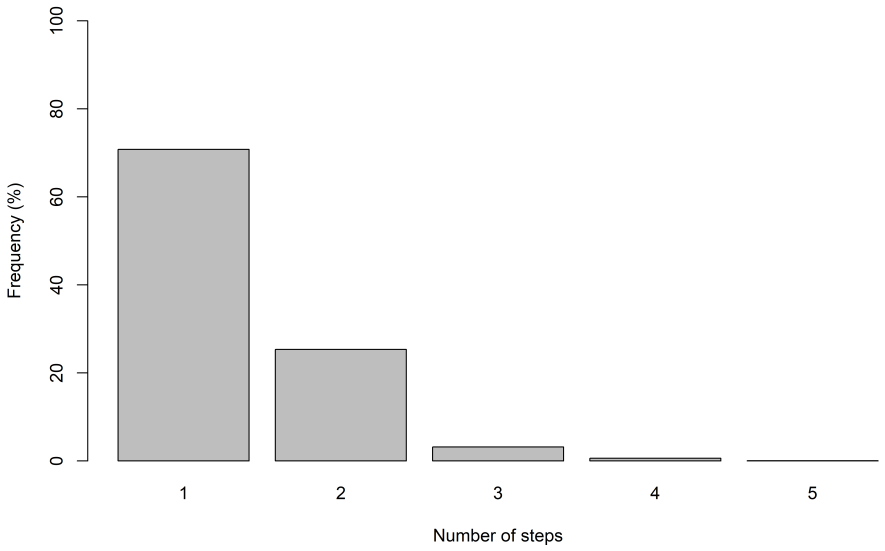

Figure 3.2: Relative frequencies of the number of steps until convergence of the algorithm described in Subsection 2.9.2.

are accumulated over all three approaches and over all scenarios not differentiated, respectively. As already stated above, the main influencing factor on the increase in power and, *vice versa*, the proportion of sample size saved seems to be the observed historical rate difference. Figure 3.3 illustrates the dependency of the outcome variable 'proportion of sample size saved' on the observed historical rate difference. It can be seen that for small differences none or only few reduction in sample size is achieved; the most benefit is achieved when the differences ranges from 0.08 to 0.18. The benefits slowly decreases for increasing rate differences. In the range of 0.08 to 0.15, a minimum benefit of at least 7.5% can be achieved. The parameter δ^* substantially determines the amount of historical data that is included into the new trial. Its dependence on the observed historical rate difference is depicted in Figure 3.4.

For differences of 0.01 to 0.06, the complete historical data is incorporated. With increasing historical rate difference, δ^*, representing the amount of incorporated historical data included into the new trial, is decreasing. For a given historical rate difference, the resulting range of admissible values for δ^* is rather small, which is illustrated by the short range of most boxplots depicted in Figure 3.4.

Figure 3.5 depicts the proportion of sample size saved depending on the resulting value of δ^* (rounded to one decimal place).

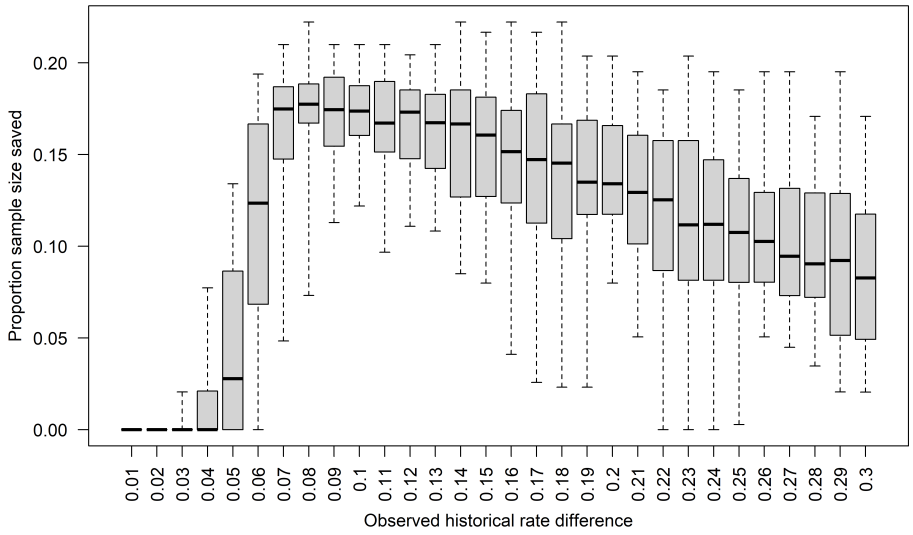

Figure 3.3: Boxplots of the proportion of saved sample size for different values of the observed historical rate difference.

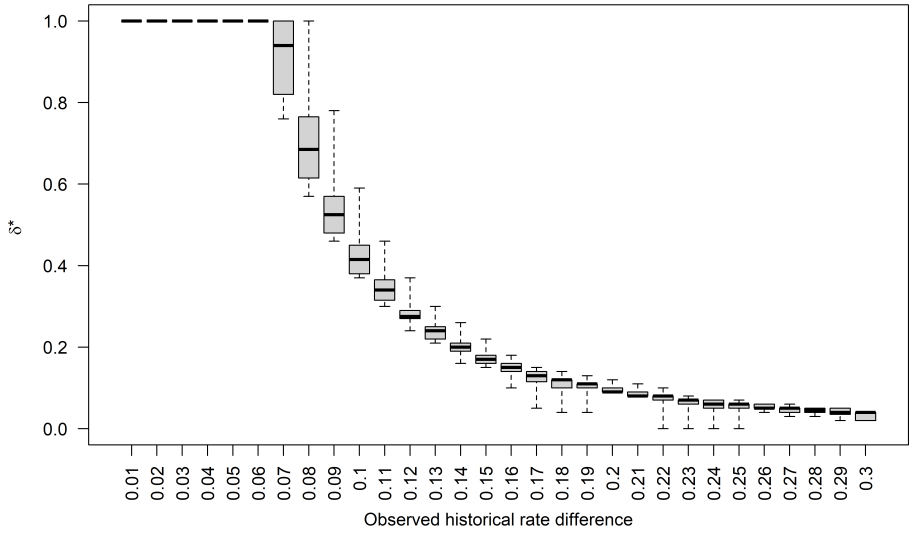

Figure 3.4: Boxplots of δ^* for different values of the observed historical rate difference.

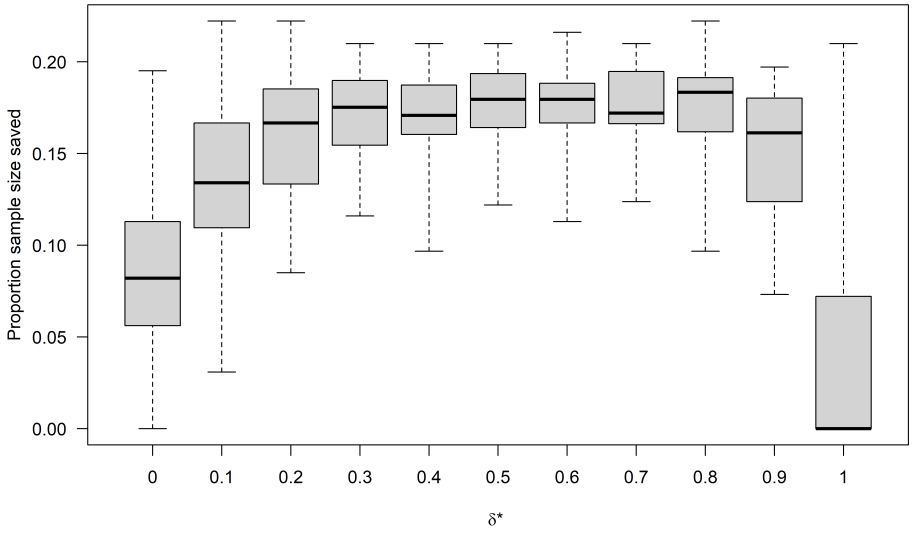

Figure 3.5: Boxplots of proportion of sample size saved for various values of δ^* (rounded to one decimal place).

Table 3.1: Initial sample sizes in the sample size calculation procedure for the different values of π_C and Δ.

$\pi_C \setminus \Delta$	0.1	0.15	0.2	0.25
0.1	195	97	60	41
0.2	292	137	81	54
0.3	356	162	93	60
0.4	388	173	97	62

It can be seen that the most beneficial combinations are those with $0.15 \leq \delta^* \leq 0.84$. However, for every decimal class of δ^* there exist scenarios where a high amount of samples size can be reduced. Nevertheless, for values of δ^* near 1, there is only low benefit in the majority of scenarios.

For varying true control proportion π_C, the proportion of saved sample size remains nearly the same, which can be seen in Figure 3.6. For an increasing true effect size Δ, the reduction in sample size is slightly more pronounced. The results are shown in Figure 3.7 and the different initial sample sizes (deduced in the first step of the calculation based on sample size calculation for a chi-square test) can be found in Table 3.1.

Figure 3.8 shows that the proportion of sample size saved slightly decreases if the observed historical control rate p_{CH} differs from the true control rate π_C. Note that this discrepancy was the main influence factor on the actual type I

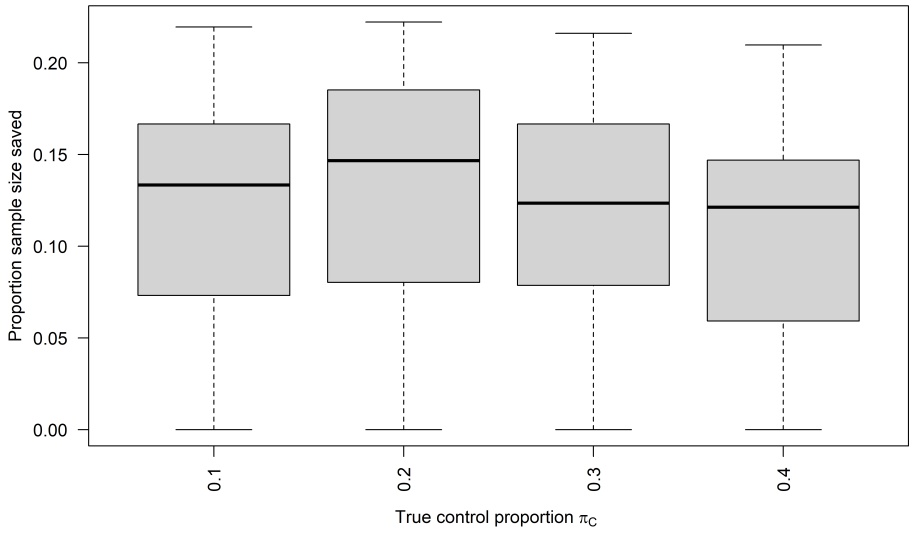

Figure 3.6: Boxplots of the proportion of sample size saved for different values of the true control rate π_C.

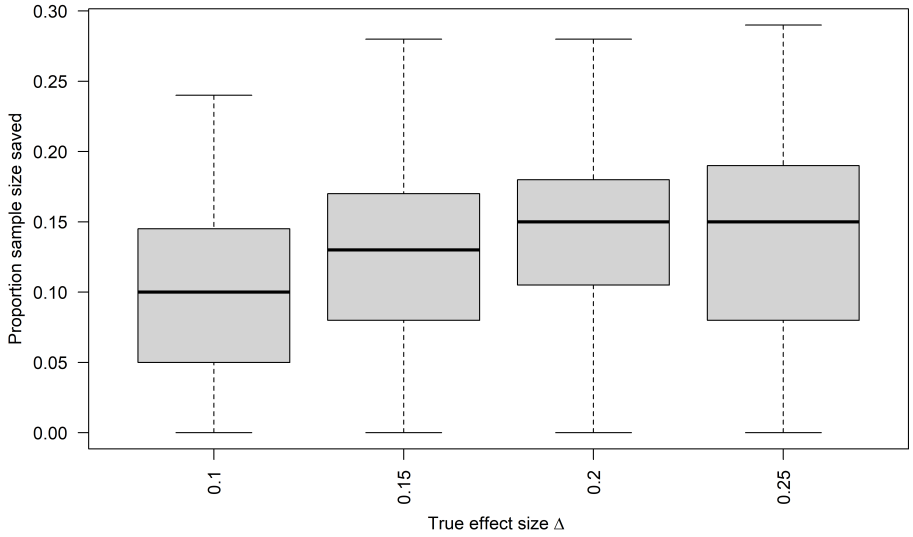

Figure 3.7: Boxplots of the proportion of sample size saved for different values of the true effect size Δ.

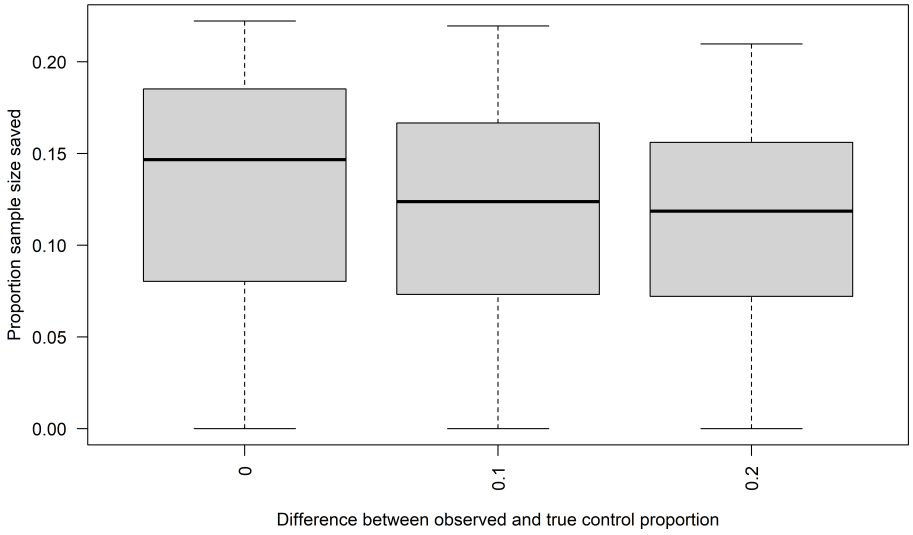

Figure 3.8: Boxplots of the proportion of sample size saved for several differences between observed historical control rate p_{CH} and true control rate π_C.

error rate in the case of one-arm borrowing (see Figure 2.1). It can be seen that, also in the two-arm borrowing approach, this inconsistency (between observed control proportion) is slightly penalized in form of an inflation of the type I error rate and *vice versa* in less sample size saved. However, these effects do not occure in a comparable amount as in the one-arm borrowing. This will be further discussed in the Chapter 4.

In Figure 3.9, the influence of the difference between observed historical rate difference $p_{TH} - p_{CH}$ and true effect size Δ on the sample size reduction is displayed. For an increasing difference between observed historical and true effect size, the proportion of sample size saved slowly decreases (from a difference of 0 to 0.11). Subsequently, it decreases to a greater extent for differences of 0.12 to 0.2. For differences larger than 0.2, the proportion of sample size saved is constantly equal to 0. This phenomenon can be explained by the fact that an difference larger than 0.2 induces both a historical rate difference of 0 to 0.05 and an effect size of 0.25. For these scenarios, the incorporation of historical data is not accompanied with a benefit and thus no sample size can be saved.

3.1.4 Comparison of Considered Approaches

Figure 3.10 displays the proportion of sample size saved for the three approaches (i.e. the global approach (2.8.1), the local approach (2.8.2), and the independent

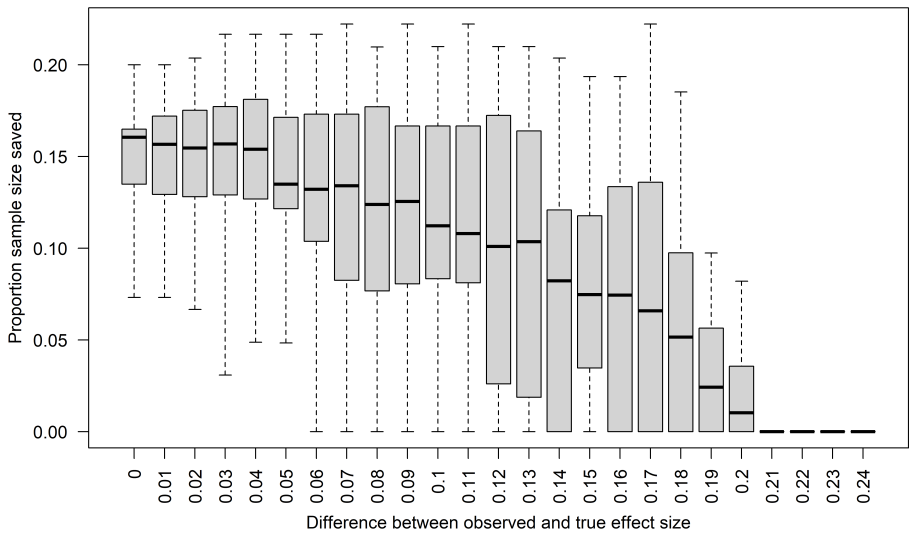

Figure 3.9: Boxplots of the proportion of sample size saved with respect to various differences between observed historical control rate difference $p_{TH} - p_{CH}$ and true control effect size Δ.

approach (2.8.3)). The benefit for the local approach is slightly higher than that for the global approach, whereas the benefit for the independent approach is slightly higher than that for the two remaining approaches. The median and maximum proportions of samples size saved increase from 11.0% and 21.0% in the global approach over 12.4% and 22.2% in the local approach up to 15.1% and 22.2% in the independent approach. The proportion of scenarios without any benefit slightly ranges from 14.4% over 15% to 13.1%, meaning that the amount of scenarios without any benefit remain nearly the same. In detail, nearly all scenarios that achieve no benefit in the independent approach are also scenarios with no benefit in the local and global approach.

Figure 3.11 compares the results of the different approaches directly by displaying a boxplot of the difference in samples size saved in the corresponding scenarios. The results in terms of proportion of sample size saved are very similar for the local and global approach, i.e., the majority of the scenarios (82.1%) differ by only 3% or less (in 46.3% they are equal). However, there are several scenarios where the difference is larger (up to 11%). Furthermore, there are generally more scenarios in which the local approach performs better than the global approach instead of the reverse (32.7% vs 21.0%). Comparing the independent approach to the global approach and the local approach, it performs

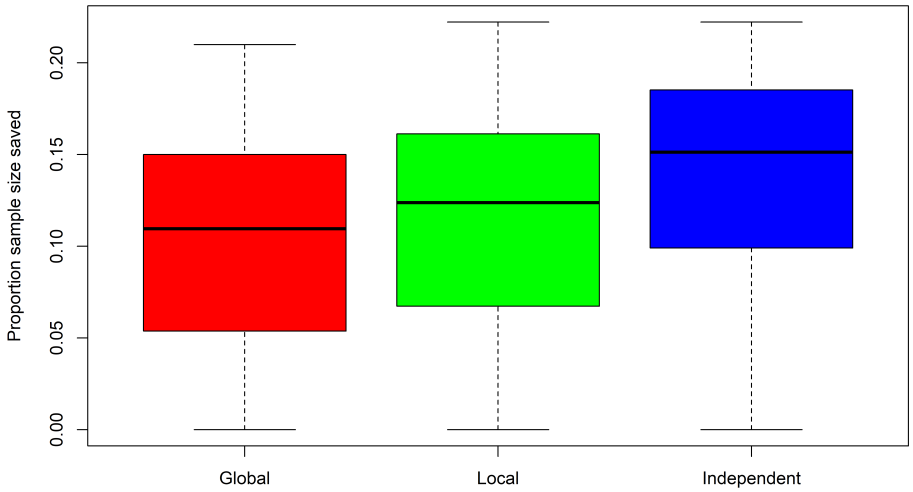

Figure 3.10: Proportion of sample size saved for the global (red), local (green) and independent (blue) approach.

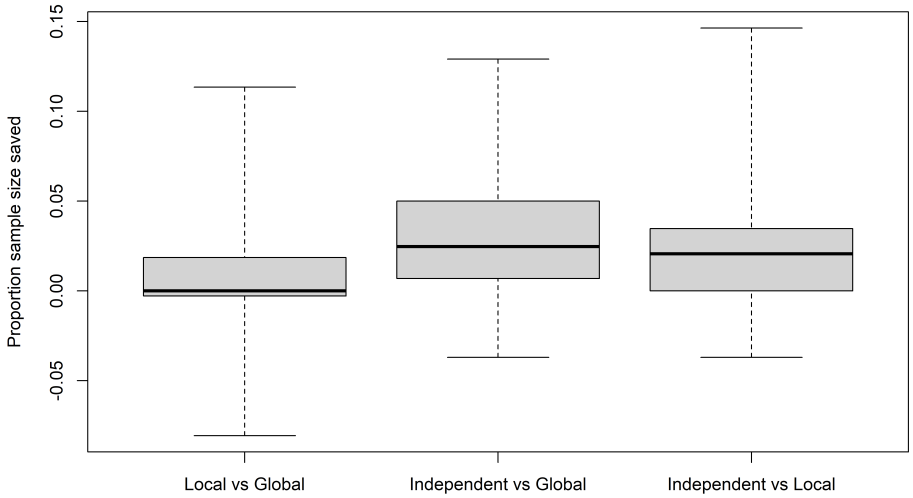

Figure 3.11: Boxplots of the difference in the proportion of saved sample size saved between the global, local, and independent approach.

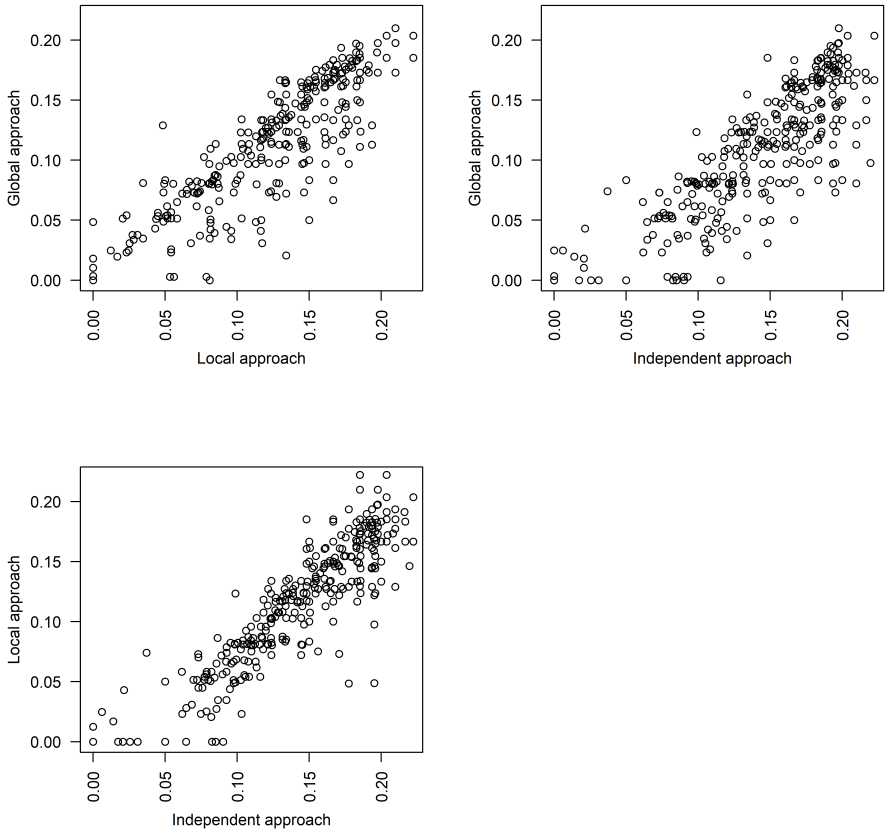

Figure 3.12: Scatterplots of the proportion of sample size saved between the three approaches. Top left: local (y-axis) vs global (x-axis) approach. Top right: independent (y-axis) vs global (x-axis) approach. Bottom left: independent (y-axis) vs local (x-axis) approach.

better: In only 3.8% and 5.0% of the scenarios, the global approach and the local approach are superior, while in 76.7% and 73.8% of all considered scenarios, the independent approach shows better results in terms of proportion of sample size saved, respectively.

Figure 3.12 shows scatter plots of the proportion of sample size saved compared between the respective approaches; each dot illustrates a separate considered scenario. The results shown in Figure 3.12 confirm the findings displayed in Figure 3.11: The results are similar with slight advantages for the local and particularly for the independent approach. Furthermore, these advantages are independent from the amount of sample size saved. Figure 3.13 and 3.14 illustrate the resulting values for δ^* for the three different approaches via boxplots,

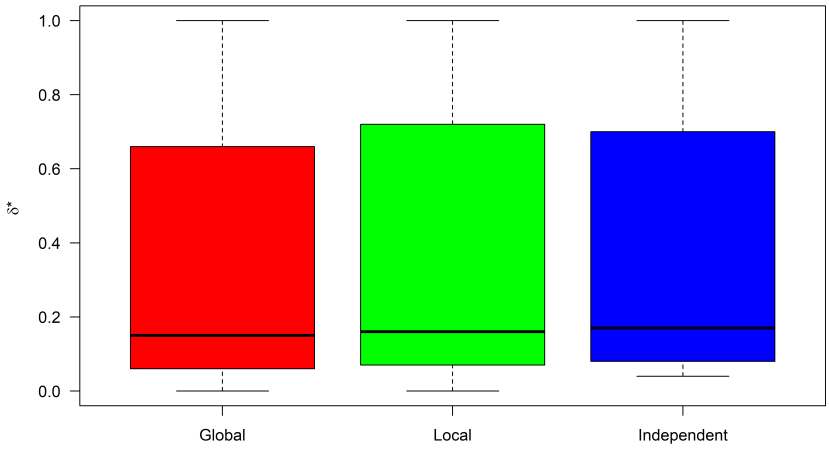

Figure 3.13: Boxplots of the attained values of δ^* in all scenarios for the three considered approaches.

and, respectively, via comparative scatterplots. In summary, the resulting values of δ^* are very similar for the three approaches. However, for the local approach, there are a few scenarios that show higher values for δ^* than the global and the independent approach. After identifying these scenarios, it was found that these are scenarios where the maximum of the type I error function (in dependence of the true control proportion π_C) is located outside of the respective confidence interval of π_C and, therefore, the use of the local approach is accompanied with a larger benefit in terms of amount of incorporated historical data. Thus, the resulting values of δ^* for the global and independent approach are nearly identical.

In the following, the influence of the different parameters on the performance of the three different approaches is evaluated. At first, the proportion of sample size saved in dependence of the observed historical rate difference is depicted separately for the three approaches (see Figure 3.15). All approaches show the same behavior, i.e., there is no or less sample size saved for small difference. The highest proportion in reduced sample size occurs for moderate differences (e.g. from 0.08 to 0.15), and there is a decreasing benefit for an increasing rate difference. However, up to a difference of 0.3 (highest value for this parameter), the decrease of the benefit for the global and local approaches advances to the case of no benefit, but the independent approach still gives a benefit of at least 7% of proportion of sample size saved.

A similar behavior of the three approaches for different values of the observed

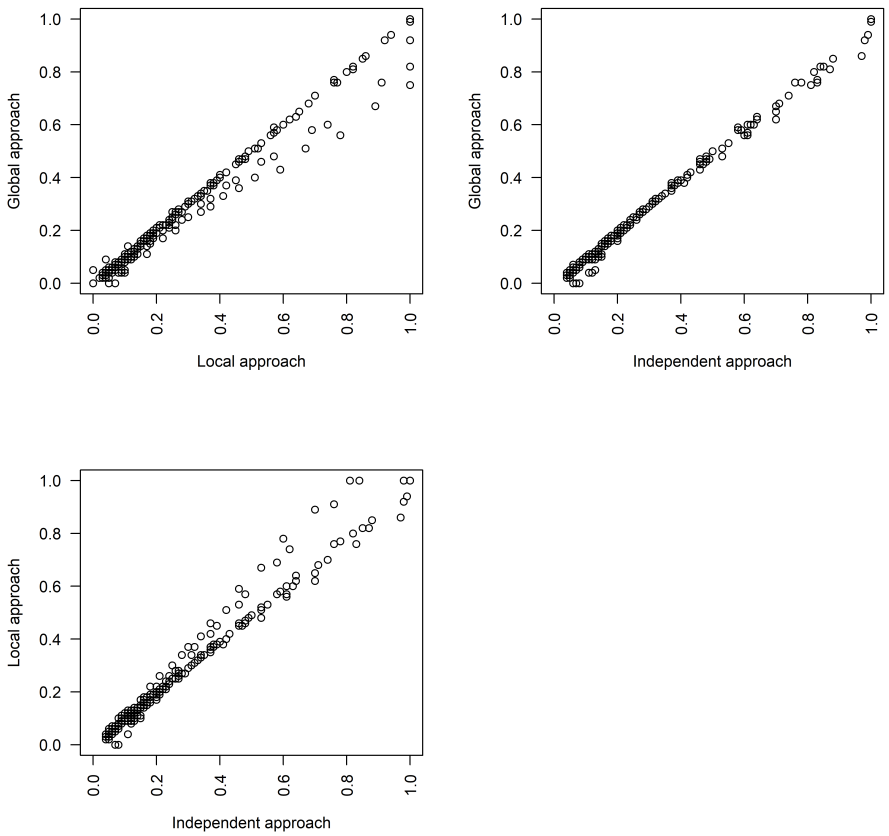

Figure 3.14: Scatterplots comparing the resulting δ^* between the three approaches. Top left: local (y-axis) vs global (x-axis) approach. Top right: independent (y-axis) vs global (x-axis) approach. Bottom left: independent (y-axis) vs local (x-axis) approach.

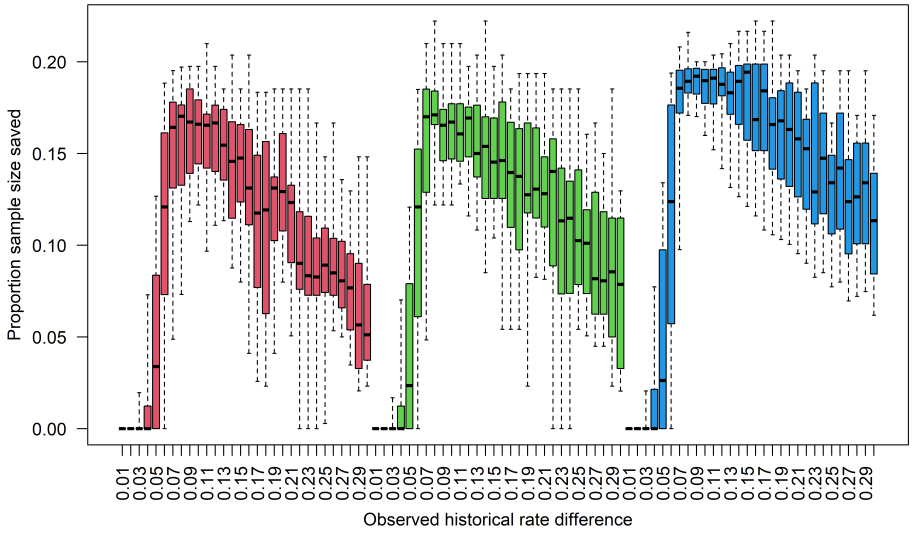

Figure 3.15: Proportion of sample size saved for the three different approaches (red=global, green=local, blue=independent) for varying observed historical difference (from 0.01 to 0.3, respectively).

historical rate difference can also be seen in Figure 3.16 and Figure 3.17. In Figure, 3.16 the resulting value of δ^* for the three approaches in dependence of this rate difference is depicted. For the global and local approach there exist scenarios for a rate difference of 0.22 to 0.25 where the resulting value of δ^* is $\delta^* = 0$ and thus, incorporation of historical data is not beneficial. In Figure 3.17, the resulting sample size reduction for various values of δ^* (rounded to one decimal place) is depicted. Again, no large differences to the general trend (independent approach is most beneficial) can be found.

Figure 3.18 shows the proportion of sample size saved for varying true control proportion π_C separated for the three approaches. As for the combining results, each of the approaches show a slight decrease in sample size saved for increasing π_C. Figure 3.19 displays the proportion of sample size saved for varying true effect size Δ separated for the three approaches. For the global approach, there is only a slight increase, if any at all, whereas for the global approach there is an increase in sample size saved for increasing Δ.

The proportion of sample size saved for varying difference between true and observed historical control proportion is shown in Figure 3.20. As in Subsection 3.1.3, the proportion of sample size saved is decreasing for an increasing difference. This decrease is more pronounced for the global and the local approach.

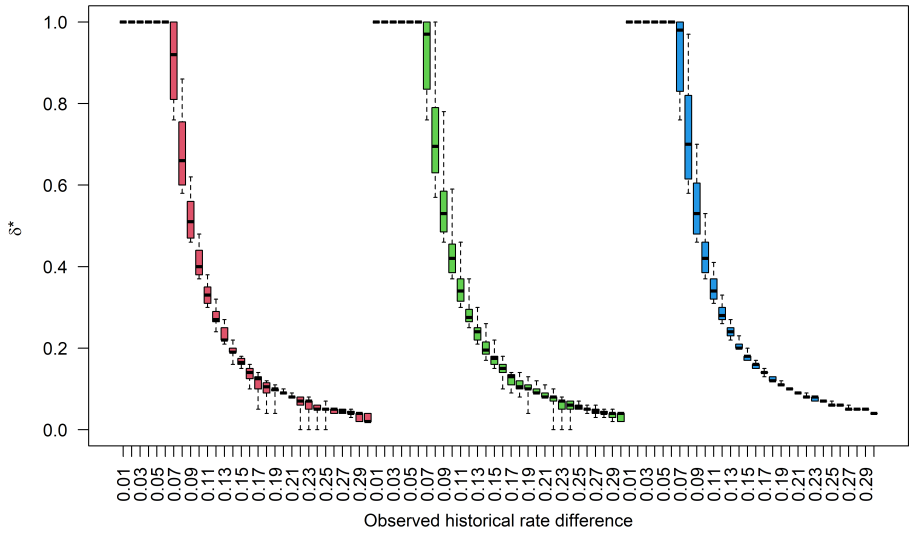

Figure 3.16: Resulting value of δ^* for the three different approaches (red=global, green=local, blue=independent) for varying observed historical difference (from 0.01 to 0.3, respectively).

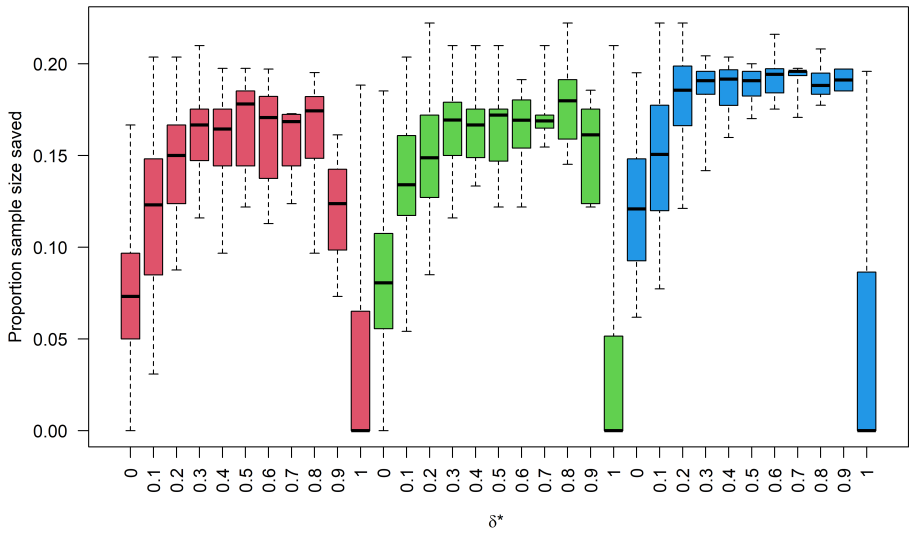

Figure 3.17: Proportion of sample size saved for various values of δ^* (range 0-1, rounded to one decimal place) for the three different approaches (red=global, green=local, blue=independent).

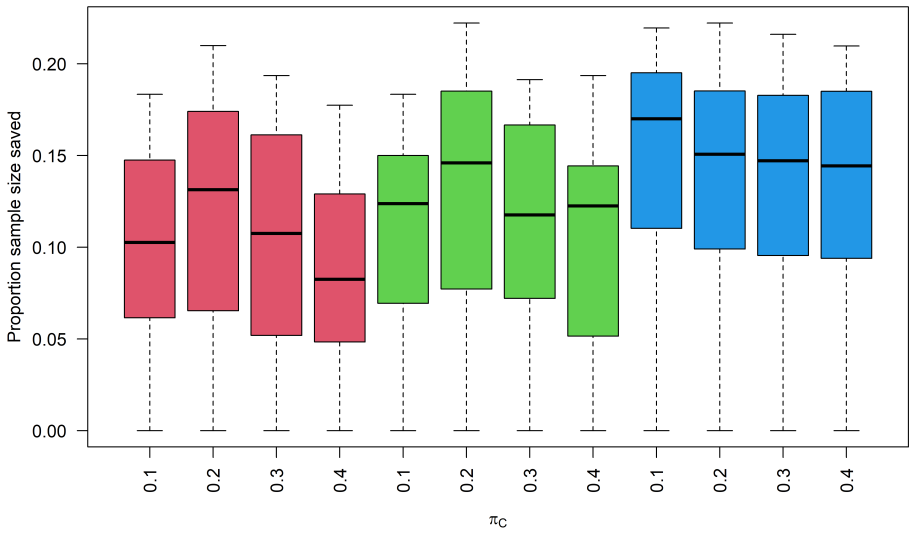

Figure 3.18: Proportion of sample size saved for the three different approaches (red=global, green=local, blue=independent) for varying true control proportion π_C.

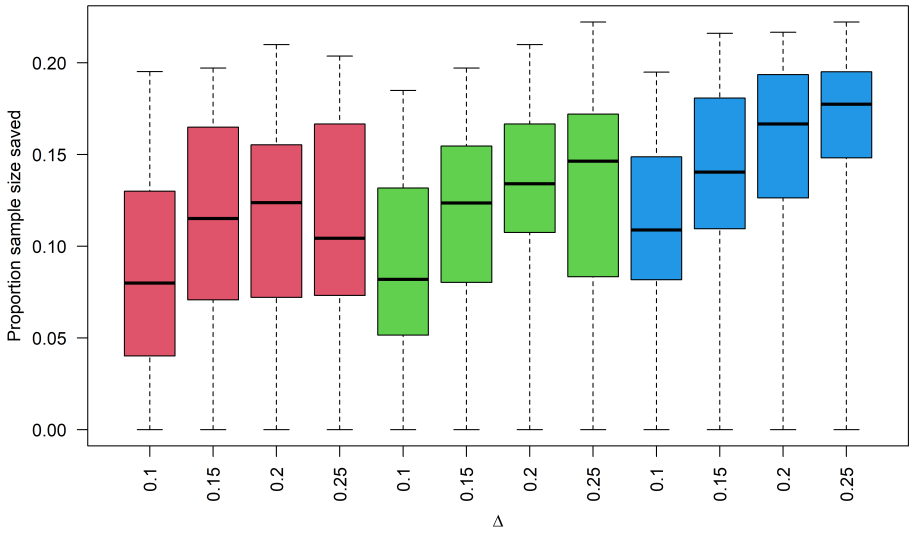

Figure 3.19: Proportion of samples size saved for the three different approaches (red=global, green=local, blue=independent) for varying true effect size Δ.

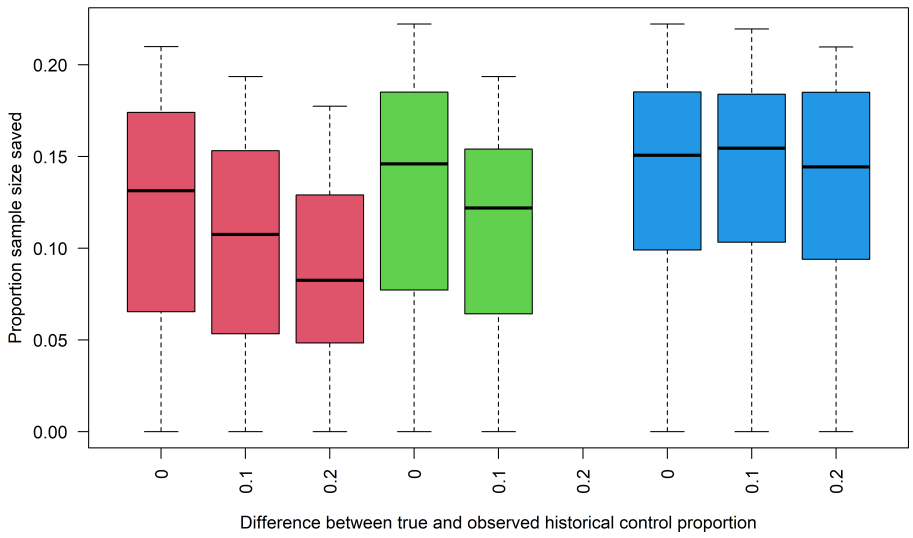

Figure 3.20: Proportion of samples size saved for the three different approaches (red=global, green=local, blue=independent) for varying difference between true and observed historical control proportion. The boxplot for 0.2 in the local approach (green) is missing, since for these scenarios the respective value of π_C was not located in the respective $1-\gamma$ confidence interval of π_C (see Subsection 3.1.1).

Note that for the local approach, there is no boxplot for a difference of 0.2 since the respective value for π_C was located outside the corresponding confidence interval (see Subsection 3.1.1).

3.1.5 Summary

In summary, up to 22.2% of the initial sample size can be saved by incorporating historical data. Thereby, the algorithm proposed in Subsection 2.9.2 can highly reduce the computational effort to find the most beneficial combination of δ^* and the samples sizes of the new trial n_C and n_T since, in the majority of the scenarios, it converges after one additional step. As already forecasted in Chapter 2, the main influencing factor on a beneficial incorporation of historical data is the observed difference in sample size saved, as can be seen from Figure 3.3. The further possible influencing parameters show only slight differences when varied. In contrast to a setting in which only historical control arm data is incorporated, the two-arm case is merely slightly impacted by a difference between true and assumed response proportions. Comparing the three different approaches to estimate δ^*, i.e. the global, local, and independent approach, the

independent approach shows the slightly best performance in terms of sample size saved, while the local approach performs slightly better than the global approach. No scenarios occurred in which the different approaches produced widely dissimilar results.

3.2 Clinical Trial Example - The FaSScinate Trial

To demonstrate the proposed methods for the incorporation of historical data in two-arm trials with binary outcome, a clinical trial example is considered, the 'Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis' (FaSScinate) trial (Khanna et al., 2016). In this trial, the efficacy and safety of tocilizumab in patients with Systemic sclerosis (SSc), a rare connective tissue disorder, was investigated. It is characterized by tightening and thickening of the skin, whereby multiple internal organs are involved including heart, lung, kidneys, and gastrointestinal tract. The FaSScinate trial was a randomized, double-blind, placebo-controlled phase II trial. An important secondary binary endpoint was the proportion of patients achieving an improvement in the so-called modified Rodnan skin score (mRSS) by at least 4.7 points from baseline to week 24. A change in the mRSS of 4.7 points or more was deemed clinically important and can be regarded as a treatment response. In the FaSScinate trial, 10 responders within 44 placebo patients and 16 responders within 43 tocilizumab patients were observed.

It is assumed that a subsequent new trial in SSc is planned which investigates the former secondary endpoint as the new primary endpoint. With the developed framework, the (historical) FaSScinate study data can be integrated into a new trial in order to potentially achieve a gain in power or, *vice versa*, a reduced sample size.

First, the gain in power while simultaneously controlling the type I error rate is considered. For this, based on the results of the FaSScinate trial, the following parameters for the sample size calculation are assumed: $\pi_C = 0.23$ and $\pi_T - \pi_C = 0.14$. To achieve a power of $1 - \beta_0 = 0.8$ with a two-sided significance level of $\alpha_0 = 0.05$, a sample size of $n_C = n_T = 167$ is needed in a trial without borrowing using the chi-square test for the analysis. The historical data observed in the FaSScinate are $c_H = 10$, $t_H = 16$, $n_{CH} = 44$, and $n_{TH} = 43$. Furthermore, $\gamma = 0.0001$ is chosen (see Subsection 2.7.2) and the respective 0.9999 confidence interval (Clopper-Pearson, based on the observed historical control rate) for π_C is $[0.050; 0.526]$.

The results for δ^* are shown in Table 3.2. The value for δ^* amounts to 0.35 for local, and 0.37 for global type I error rate control, respectively. Since the global (minimum) δ^* was contained in the confidence interval (at $\pi_C = 0.39$),

Table 3.2: δ^* and gain in power by incorporating historical data while simultaneously controlling the type I error rate for the FaSScinate trial for local, global, and independent approach. Thereby, δ^* denotes the maximum amount of incorporated historical data that still guarantees type I error rate control, π_C denotes the true control rate.

	Global approach	Local approach	Independent approach
δ^*	0.37 (minimum at $\pi_C = 0.39$)	0.35 (minimum at $\pi_C = 0.37$)	0.46
Minimum gain in power	-0.002 (at $\pi_C = 0.82$)	0.01 (at $\pi_C = 0.05$)	Not applicable
Maximum gain in power	0.060 (at $\pi_C = 0.31$)	0.058 (at $\pi_C = 0.33$)	Not applicable
Mean gain in power	0.031	0.039	Not applicable
Power at $\pi_C = 0.25$	0.851	0.848	0.861

the global approach leads to a higher δ^*, and, therefore, the maximum gain in power is achieved for the global approach. For almost every value of π_C, a gain in power can be observed; only for the unrealistic values near $\pi_C = 0.8$ (differing greatly from the observed historical rate of 0.23), a decrease in power occurs. Therefore, since the local approach restricts the range of π_C to the more stable and realistic scenarios in the confidence interval, there are better results for minimum and mean gain in power for this approach. Furthermore, the local approach guarantees a positive minimum gain in power (0.01). Note that these considerations do not apply to the independent approach, since it is, as its name suggests, independent of π_C. With this approach a value of $\delta^* = 0.46$ was obtained.

Considering the rejection regions for this clinical trial example may help to illustrate the idea and strategy of the proposed procedures. Therefore, the rejection regions are built based on the same control proportion in the observed data (i.e. 0.23; the control proportion is fixed due to its impact on the rejection region). Rejection of the null hypothesis using a test without incorporation of historical data in such a situation occurs when the number of responders in the intervention group either lies in the interval $[0; 24]$ or in the interval $[55; 167]$; thus the rejection region would be $R = [0; 24] \cup [55; 167]$ (with fixed number of control responses $c = 38$). With historical data based on an optimal δ^* (global approach) of 0.37, the rejection region changes to $R = [0; 21] \cup [54; 167]$, and based on a δ^* of 1 (which does not control the type I error rate) the rejection region is $R = [0; 16] \cup [50; 167]$. To provide a more comprehensive evaluation of these results, they are also compared to rejection regions of a two-sided test with

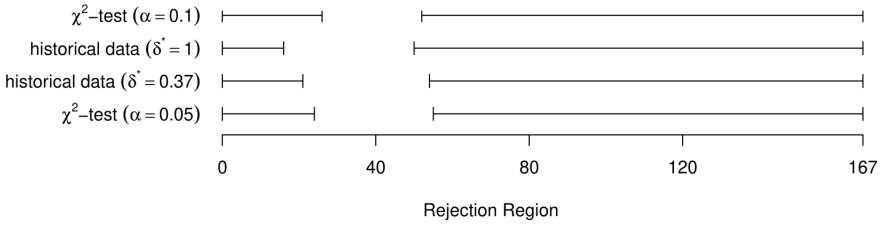

Figure 3.21: Rejection regions in terms of number of responders in the treatment group t for different test procedures with fixed number of control responses c=38 ($\pi_C = 0.23$).

$\alpha_0 = 0.1$ without any historical data, resulting in a rejection region $R = [0; 26] \cup [52; 167]$. Thus, the proposed procedure results in an even smaller lower part of the rejection region than the $\alpha_0 = 0.05$ test procedure, but in an upper part of the rejection region that can be classified to lie between the $\alpha_0 = 0.05$ and $\alpha_0 = 0.1$ test procedure. The rejection regions are illustrated in Figure 3.21.

As it was illustrated already in Subsection 2.6.3 and will be considered in further detail in Chapter 4, the inclusion of historical data (by increasing δ) increases the power only in favor of an effect observed in the historical data. Thus, the rejection region is expanded for an effect in favor of the observed effect in the historical, but at the same time the rejection region for an effect in the opposite direction is shrunken.

Additionally, the benefit in reduced sample size that can be achieved by incorporating the historical information of the FaSScinate trial into a new trial is considered. With the same parameter values as above ($\pi_C = 0.23$, $\pi_T - \pi_C = 0.14$, $1 - \beta_0 = 0.8$, $\alpha_0 = 0.05$, $c_H = 10$, $t_H = 16$, $n_{CH} = 44$, $n_{TH} = 43$, $\gamma = 0.001$), the sample size can be reduced by 24 patients (14.4%) with the local approach, by 26 patients (15.6%) with the global approach and by 28 patients (16.8%) using the independent approach (see Table 3.3). Thus, in this specific scenario the independent approach leads to a higher benefit as compared to the global and local approach.

Furthermore, the global approach leads to a higher benefit than the local approach. On the first sight this is surprising, since the local approach just shrinks the range of values for π_C ensuring type I error rate control to a $1 - \gamma$ confidence interval (by simultaneously decreasing the significance level to $\alpha_0 - \gamma$). If the maximum of the actual type I error rate lies outside the corresponding confidence interval, this reduction would lead to a benefit in terms of sample size saved compared to the global approach. Simultaneously decreasing the

Table 3.3: Results of the sample size calculation procedure for the FaSScinate trial for local and global approach. δ^* denotes the maximum amount of incorporated historical data that still guarantees type I error rate control.

	Global approach	Local approach	Independent approach
Initial sample size	167	167	167
New sample size	141	143	139
Saved sample size (%)	26 (15.6 %)	24 (14.4 %)	28 (16.8%)
δ^*	0.44	0.37	0.46
Steps	2	1	2

Table 3.4: Results of the sample size calculation procedure for the FaSScinate trial for local and global approach.The use of the global grid indicates that the values of π_C are corrected to two decimal places. The local grid divides the respective $1-\gamma$ confidence interval in equidistant parts (number= number grid steps; without rounding to a specific number of decimal places).

Number grid steps	δ^*	Sample size saved	π_C with max α
49 (global grid)	0.44	26 (15.6%)	0.35
50 (local grid)	0.37	24 (14.4%)	0.469
100 (local grid)	0.37	23 (13.8%)	0.469
200 (local grid)	0.37	23 (13.8%)	0.501
400 (local grid)	0.37	22 (13.2%)	0.431

significance level by a small value of $\gamma = 0.0001$ should not noticeably influence the results. Therefore, one would expect that for nearly every scenario, the local approach is at least as beneficial as the global approach. The reason why there are scenarios where the global approach is still more beneficial than the local approach is due to the respective values of π_C for which the actual type I error rate is calculated. There are two reasons why, in this specific scenario, the global approach performs better than the local approach. At first, for a fixed true control proportion π_C, a fixed δ, and for increasing sample size of the new trial n_C and n_T, the actual type I error rate does not increase monotonously for all scenarios but 'jumps' upwards and downwards. This can be seen in Figures 2.4 and A.1, and is due to the character of the chi-square distribution (Fagerland et al., 2017). This 'jumping' intensifies if values of π_C are near $\pi_C = 0.5$ (see Figure A.1). The second reason is that, in the local approach, the range of values for π_C is shrunken but the number of values where the type I error rate is controlled remains identical to the global approach. This results in a grid of these values that is more dense which may result in finding a higher maximum of the actual type I error rate. This fact is depicted in Table 3.4. It can be seen that changing the grid from a global to a local one (49 to 50 grid steps) and thus controlling the type I error rate for other values of π_C changes the benefit

in δ^* and sample size saved. Increasing the number of grid steps up to 400 steps further slightly decreases this benefit.

In summary, the benefit in terms of power increase or, *vice versa*, in terms of sample size saved in specific scenarios is not only depending on the parameters of the specific scenario but also on the distribution of the test statistic which is used (in this thesis: the chi-square distribution) as well as on the grid on which the calculations are based. Note that these latter consideration of the global and local approach do not refer to the independent approach, since this approach was built independently from π_C and is based on the normal distribution. However, as already stated above, there are other aspects of this approach which can be criticized and which will be dealt with in more detail in Chapter 4.

Chapter 4

Discussion

4.1 Contributions to Research and Discussion

In this thesis, a framework is presented which allows to integrate historical two-arm data of a previous trial in the planning and analysis of a two-arm clinical trial with binary outcome while simultaneously controlling the type I error rate. The resulting approaches, i.e. the global, local, and independent approach, are based on the Bayesian power prior method. It is shown that the idea of the power prior method can be transferred straightforwardly into a frequentist fourfold table setting. For all three approaches, the amount of historical data that may be incorporated into a new clinical trial is controlled by a factor δ that ranges from 0 (no borrowing) to 1 (full borrowing). The maximum amount of borrowed data which still ensures control of the type I error rate depends substantially on the characteristics of the historical data. The conducted systematic investigations show that up to 22% of the initial sample size can be saved by incorporation of historical data. However, the integration of the historical data is not always accompanied by a benefit in terms of an increased power or a reduced sample size (e.g. no benefit occures in case of an observed historical rate difference $\leq$0.02).

Comparing the three approaches presented in this thesis with regard to benefit in terms of sample size saved, systematic investigations showed that the independent approach performs slightly better than the local approach, whereas the local approach performs in summary slightly better than the global approach. Since the type I error rate depends on the true control proportion π_C, the local and global approach control this rate over the whole range $[0; 1]$ at the nominal significance level of α_0 (globally) or within a $1 - \gamma$ confidence interval to a reduced significance level of $\alpha_0 - \gamma$, resulting in a total type I error rate of α_0 (locally), respectively. As the name suggests, the independent approach

works independently of π_C, directly targets the rate difference Δ. Therefore, the computation effort for this approach is rather smaller in comparison to the two other approaches. However, some aspects of this approach can be seen critically. At first, working with a variance stabilizing transformation is arguable since, especially for the very high and the very low values of π_C, the transformation used in this thesis is inexact (Warton and Hui, 2011). Despite this, alternatively proposed transformation methods are subject to other weaknesses as well (Feder et al., 2020). Furthermore, working in a Bayesian framework and subsequently considering and comparing operating characteristics like type I error rate and power may be seen critically since the concepts of type I error rate and power were initially created for frequentist testing and then translated to the Bayesian setting, and the correct interpretation of Bayesian hypothesis test results is still under discussion in the literature (Lesaffre et al., 2020). Nevertheless, the presented Bayesian setting provides a direct way and represents a mathematically elegant solution.

In the scenarios that were considered in this thesis, the local approach leads for most situations to a larger benefit compared to the global approach; however, there are also scenarios where the latter one is advantageous. On the first sight, this is surprising since the local approach simply shrinks the range of admissible values for π_C where the type I error rate has to be controlled to a $1-\gamma$ confidence interval (by simultaneously decreasing the significance level to $\alpha_0-\gamma$). However, if the value of δ^* determined with the global method was found in the respective confidence interval of the local approach, the global approach is the better choice because there is no benefit for reducing the range of values for π_C while simultaneously reducing the local significance level. Furthermore, the calculations on which the results for the global and the local approach are based on only ensure control of the type error on a grid of values of π_C. Therefore, changing the grid (by changing the approach) may lead to a change in the results of the calculations. In addition, the discrete character of the chi-square distribution may lead to non-monotonous behavior of the model parameters for increasing or decreasing values. Nevertheless, if one is mainly interested in ensuring a substantial gain in power, it is usually better to reduce the range of π_C to the more realistic and therefore more beneficial values in the confidence interval of the local approach. However, this issue is negligible regarding the sample size calculation since it is only based on one particular value of π_C.

The main influencing factor determining the amount of historical data that is allowed to be incorporated, while at the same time controlling the type I error rate, is the observed difference of the historical response rates. This is due to the fact that large differences are in contrast to the assumption of equal response rates assumed under the null hypothesis, which leads to an inflation

of the type I error rate. Although full borrowing (i.e. use of the total historical data) for small observed historical differences is possible without inflation of the type I error rate, these scenarios do not lead to a gain in power if the observed historical difference is small in relation to the true difference. Therefore, the largest benefit is achieved for moderate (i.e. from 0.05 to 0.2) historical rate differences. At first sight, it seems arguable that the proposed methods penalize a large observed treatment effect in the historical data by limiting the amount of borrowable historical information. However, the probability for rejecting the null hypothesis in a new trial becomes shifted by the historical data and, therefore, influences the type I error rate. Thus, it has to be taken into account that a larger observed treatment effect increases the shift in the decision of a rejection of the null hypothesis. Furthermore, the proposed methods result in a benefit for a wide range of values for the observed historical rate differences. On the one hand, if there is a large observed historical rate difference, there certainly is a lower need for a benefit in saved sample size since the required sample size in a new trial would be considerably smaller (if the sample size calculation is based on the historical data). On the other hand, a small observed difference indicates a small true effect resulting in a larger sample size needed to achieve a sufficiently high power to detect the effect in a new trial. In this case, borrowing a large amount of historical data is desirable and simultaneously favored by the proposed approaches.

Since the extent of the type I error rate inflation mainly depends on the historical rate difference, this causes a problem in the case of two-arm borrowing: Here, a heterogeneity between observed historical data and true underlying effects of the new trial is not necessarily penalized in terms of an increased type I error rate as it is in the case of one-armed borrowing (i.e. borrowing solely in the control arm). Thus, for two-arm borrowing, it is even more important that the choice of a possible historical trial does not solely depend on a data-driven justification but is additionally based on further external criteria. Therefore, it may be useful to verify the choice of a historical trial by Pocock's criteria (Pocock, 1976) (see chapter 'Acceptable Historical Control' of this paper) for integration of historical data:

'The acceptability of a historical control group requires that it meets the following conditions:

1. Such a group must have received a precisely defined standard treatment which must be the same as the treatment for the randomized controls.

2. The group must have been part of a recent clinical study which contained the same requirements for patient eligibility.

3. The methods of treatment evaluation must be the same.

4. The distributions of important patient characteristics in the group should

be comparable with those in the new trial.

5. The previous study must have been performed in the same organization with largely the same clinical investigators.

6. There must be no other indications leading one to expect differing results between the randomized and historical controls. For instance, more rapid accrual on the new study might lead one to suspect less enthusiastic participation of investigators in the previous study so that the process of patient selection may have been different.'

It appears reasonable to extend these criteria to the case of two-arm clinical trials. By doing, they should be verified to an equally rigorous extent for both arms of the studies.

To achieve an increase in power while simultaneously controlling the type I error rate, it is crucial that the type I error rate function depending on δ (the parameter controlling the amount of historical data which will be incorporated) is at least partly below the prespecified significance level. To add some intuition to this fact, one can consider the distribution of a standard normal test statistic. In the two-sided test procedure, the type I error rate can be illustrated as the sum of the integral of the density function of the test statistic from -∞ to the $\alpha/2$-quantile and the integral from the 1- $\alpha/2$-quantile to ∞. For an increasing amount of historical data incorporated (controlled by δ), this distribution gets more peaked and shifted in direction of the observed rate difference in the historical data. As a result, the integral in direction of the observed effect increases while the other integral simultaneously decreases. However, for small δ the first integral's increase is not as substantial as the second integral's decrease. Thus, the type I error rate first slightly decreases before it increases. This is depicted in Figure A.2 in the Appendix. Based on this illustration, it follows that the procedure does not achieve the same beneficial result in the case of a one-sided testing procedure since in this case, only the first integral increases while there is no part of the function that decreases. Thus, the type I error function would not fall below the nominal significance level and it would not be possible to control the type I error rate by the nominal significance level when increasing δ. Incorporation of historical data would then always result in a type I error rate inflation. It should, however, be noted that the proposed test procedure could easily be adapted to a one-sided setting under the specification of an upper limit for the one-sided significance level, e.g. $2\alpha_0$. Furthermore, it follows from this illustration that the possible gain in power only occurs when the true effect indicates a favorable treatment effect for the treatment group as compared to the control group. Simultaneously, the power to reveal a favorable effect for the control group as compared to the treatment group decreases. This can also be seen by considering the rejection regions presented for the clinical trial example

(Figure 3.21) and the rejection regions in Figure 2.3. Thus, the proposed procedures achieve their benefit only in the two-arm borrowing and two-sided testing case by reducing the power to reveal a favorable treatment effect for the control group as compared to the treatment group, while still controlling the type I error rate at the prespecified significance level α_0. Thus, the incorporation of historical data 'shifts' the rejection regions under the null hypothesis in favor of the effect observed in the historical data. In summary, the increase in power is mainly based on the shift of the rejection regions in favor of the effect observed in the historical data. This follows the Bayesian idea of using prior knowledge for decision-making which is translated into a frequentist setting in the proposed methods while at the same time assuring control of the type I error rate.

Achieving an increase in power while simultaneously controlling the type I error rate seems at first glance to contradict the theory of the most powerful test (Kopp-Schneider et al., 2020): Since the likelihood ratio test (which is equivalent to the chi-square test in the case of a fourfold table) is the uniformly most powerful test according to the Neyman-Pearson lemma, it is not possible to find a 'better' test (in terms of higher power by simultaneously controlling the type I error rate). Thus, in this thesis the benefit in terms of an increase in power or sample size saved is defined only on the basis of revealing a treatment benefit which corresponds with the direction of the treatment effect observed in the historical data. From a theoretical point of view, a two-sided test procedure should achieve a sufficiently high power in both directions of the alternative hypothesis. However, in practice this is hardly ever considered, especially in the case of a binary outcome, as in this case the power is not symmetric for an assumed effect. Here, the power to reveal a given effect size is depending on the true control proportion and thus there is not the same power for revealing an effect into both directions of the alternative hypothesis, respectively (e.g., for a fixed sample size n the power for revealing an effect between 0.3 and 0.4 is higher than the power for revealing an effect between 0.4 and 0.5). Based on these perspectives, the proposed procedure delivers on its promises, i.e. an increase in power (for true treatment effects favoring the treatment group over the control group) by simultaneously controlling the type I error rate.

The determination of δ^* requires a large computational effort, especially for sample size calculation, since the most beneficial combinations of a wide range of parameters has to be found. Therefore, several practical recommendations are suggested in this thesis. In detail, an algorithm is developed (see Subsection 2.9.2) to find this most beneficial (in terms of saved sample size) combination of n_C, n_T (sample sizes of the new trial) and δ^*, which was found to usually converge in only one or two steps. Nevertheless, the approaches remain computationally intensive, especially in case of large sample sizes, since

the computation time for the repeated calculation of the actual type I error rate and the power increases quadratically with increasing n. Therefore, for very large sample sizes (e.g. $n >> 300$) it is recommended to determine type I error rate and power by simulations.

In comparison, other adaptive weighting approaches considered in the literature (e.g. Gravestock & Held 2017) do not work directly with the aim to control the type I error rate at a prespecified significance level, but merely assess the agreement between the current and the historical data. Contrarily, in the procedures presented in this thesis, the observed data of the new study are not included in the calculations, but only the data of the historical trial are.

4.2 Limitations and Directions for Future Research

In this work, the evaluations were limited to the case where the historical data is fixed. However, there are further frameworks for merging historical data and data of a current or planned trial. These include, for example, that the historical data may also be handled as random thus taking into account the uncertainty of historical data. Furthermore, there also exists a framework where one plans to use the historical data prior to the conduct of the historical study. This would, e.g., be the case in a seamless phase II/III trial, where it is prospectively decided that the phase II data will be combined and analyzed together with the phase III data. Nevertheless, an extensive examination and comparison of these approaches would go beyond the scope of this thesis and therefore, is not further considered.

Furthermore, for the results presented in this thesis extensive calculation had to be performed. Therefore, the resulting fourfold tables were analyzed using the commonly used chi-square test for the global and local approach. This choice can be seen critically, since this test is an approximate test and therefore type I error rate control may not be met in general since the true type I error rate occasionally exceeds the nominal significance level. Furthermore, its discreteness may lead to inhomogeneous results (e.g. the type I error function is not homogeneously convex, see Figure 2.2). However, calculations based on the chi-square test are considerably faster compared to unconditional tests, such as the Fisher Boschloo test. For a specific clinical trial application, the use of this class of tests could be more favorable.

Furthermore, the systematic investigations could have been done more extensively by extending them to a lot of more scenarios and parameters to be evaluated. However, to handle the computational effort, the investigations per-

formed in this thesis were selected to represent a plausible range of scenarios common in praxis.

In future work, the framework could be extended to other outcomes, e.g. continuous or survival endpoints. Furthermore, the presented framework and its methods could be applied or compared to alternative frameworks allowing to incorporate the historical data in a current study.

4.3 Conclusion

Within this thesis, a framework is proposed to integrate existing data for the planning and analysis of a subsequent clinical trial. The focus was on the application to two-arm trials with binary outcome. It was shown that for specific scenarios a gain in power can be achieved or the required sample size can be reduced and thus resources can be saved. These methods were developed with the vision that they will support the further streamlining of the development of drugs and medical devices, especially in the field of rare diseases.

Chapter 5

Summary

5.1 Summary (English)

The aim of this thesis was to examine whether and how two-arm data of a historical clinical trial can be incorporated into a newly performed clinical trial. It was investigated whether this incorporation can be accompanied with a benefit in terms of increasing the power or, *vice versa*, reducing the required sample size of a new clinical trial as compared to a trial without borrowing. Reducing the required sample size generally also reduces the time effort and the cost of the new clinical trial and can thus be regarded as highly desirable from an operational perspective. Furthermore, reducing the sample size and duration of a clinical trial can also be considered as beneficial from a patient perspective, as efficacious treatments will find their way into clinical practice more rapidly.

In a regulatory context, a necessary condition for the successful incorporation of historical data into a new study is the control of the type I error rate by predefined significance level. In general, the type I error rate inflates with increasing amount of historical data. Thus, in this thesis approaches were developed which are based on a method that allows controlling the amount of historical data incorporated into the new trial – the so-called 'power prior' approach. This Bayesian method was transferred into a frequentist framework, since the statistical concepts of type I error rate and power were originally developed within the inference theory of a frequentist setting.

In the course of this thesis, it was shown that for a two-sided statistical test procedure incorporating an increasing amount of historical two-arm data leads to a type I error rate that initially decreases before increasing. Thus, it was possible to determine an amount of historical data which could be incorporated under a simultaneous control of the type I error rate at the predefined significance level. It was demonstrated that the extent of this amount depends

on various parameters. In order to reduce and control the impact of these so-called nuisance parameters, three different approaches were developed, that determine the amount of historical data to be incorporated. In the further course of this thesis, these three approaches were examined and compared with focus on benefit in terms of proportion of sample size saved. It was shown that, by incorporation of historical two-arm data, the power to reveal the effect in favor of the same effect as observed in the historical data can be increased in numerous scenarios. Consequently, the required sample size for a new trial can be reduced for many practically relevant situations. However, some scenarios were identified in which the incorporation of historical data is not accompanied with a benefit.

The approaches proposed in this thesis were particularly computationally intensive. Therefore, general recommendations were given in order to diminish the computational effort. In addition, an algorithm was developed that substantially reduces the amount of calculations that have to be performed by using the proposed procedures.

In summary, in this thesis it could be shown that incorporation of historical two-arm data into a new clinical trial can be beneficial in terms of an increase in power or, *vice versa*, a reduction in required sample size, while the type I error rate can simultaneously be retained under the nominal significance level. However, the existence and magnitude of this benefit largely depends on the underlying historical data. Thus, scenarios were identified that are accompanied with a high benefit or with no benefit all.

5.2 Zusammenfassung (Deutsch)

Das Ziel dieser Arbeit war es, zu untersuchen, ob und wie Daten einer bereits durchgeführten (historischen) zweiarmigen klinischen Studie in eine neue klinische Studie eingebunden werden können. Es wurde überprüft, ob diese Einbindung mit einem Mehrwert im Sinne einer Erhöhung der Power beziehungsweise einer Reduzierung des erforderlichen Stichprobenumfangs einer neuen klinischen Studie im Vergleich zu einer konventionellen Studie ohne Einbindung historischer Daten einhergehen kann. Eine Reduzierung des erforderlichen Stichprobenumfangs reduziert in der Regel auch den zeitlichen Aufwand und die Kosten einer neuen klinischen Studie. Dies kann daher aus operativer Sicht als sehr wünschenswert angesehen werden. Darüber hinaus kann eine Reduzierung des Stichprobenumfangs und der Dauer einer klinischen Studie auch aus Sicht der Patienten als vorteilhaft angesehen werden, da wirksame Behandlungen schneller ihren Weg in die klinische Praxis finden können.

In einem regulatorischen Kontext ist eine notwendige Bedingung für die

erfolgreiche Einbindung historischer Daten in eine neue Studie die Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art unterhalb eines vorgegebenen Signifikanzniveaus. Im Allgemeinen vergrößert sich jedoch die Wahrscheinlichkeit des Fehlers 1.Art mit steigendem Anteil an eingebundenen historischen Daten. Daher wurden in dieser Arbeit Ansätze entwickelt, die auf einer der sogenannten Power-Prior-Methode beruhen, welche es erlaubt, den Anteil der in die neue Studie einfließenden historischen Daten zu kontrollieren. Diese Bayes'sche Methode wurde in einen frequentistischen Rahmen überführt, da die statistischen Konzepte des Fehlers 1. Art und der Power ursprünglich innerhalb der Inferenztheorie eines frequentistischen Settings entwickelt wurden.

Im Rahmen dieser Arbeit wurde gezeigt, dass für ein zweiseitiges statistisches Testproblem mit steigendem Anteil an historischen Daten aus zwei Studienarmen die Wahrscheinlichkeit eines Fehlers 1. Art zunächst abnimmt, bevor er zunimmt. Dadurch war es möglich, bei gleichzeitiger Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art zum vorgegebenen Signifikanzniveau, einen entsprechenden Anteil an historischen Daten in eine neue Studie einzubinden. Es wurde gezeigt, dass das Ausmaß dieses Anteils von verschiedenen Parametern abhängt. Unter der Berücksichtigung dieser sogenannten Störparameter, wurden drei verschiedene Ansätze entwickelt um den Anteil der einzubeziehenden historischen Daten zu bestimmen. Im weiteren Verlauf dieser Arbeit wurden diese drei Ansätze insbesondere bezüglich der Möglichkeit Stichprobenumfang einzusparen untersucht und miteinander verglichen. Es konnte gezeigt werden, dass durch die Einbeziehung historischer Daten in vielen Szenarien die Power zur Aufdeckung des gleichen Effekts, wie er in den historischen Daten beobachtet wurde, erhöht werden kann. Folglich kann der erforderliche Stichprobenumfang für eine neue Studie für viele praktisch relevante Situationen reduziert werden. Es wurden jedoch auch einige Szenarien identifiziert, in denen die Einbeziehung historischer Daten nicht mit einem Mehrwert verbunden ist.

Die in dieser Arbeit entwickelten Ansätze sind mit einem hohen Rechenaufwand verbunden. Es wurden daher praktische Empfehlungen gegeben, um diesen zu verringern. Darüber hinaus wurde ein Algorithmus für die Bestimmung des optimalen Stichprobenumfangs entwickelt, der den Rechenaufwand bei den entwickelten Verfahren deutlich reduziert.

Zusammenfassend wurde in dieser Arbeit gezeigt, dass die Einbeziehung historischer Daten aus zwei Studienarmen in eine neue Studie mit einem Mehrwert verbunden sein kann. Dieser Mehrwert spiegelt sich im Sinne eine Erhöhung der Power zugunsten des Effekts, wie er in den historischen Daten beobachtet wurde beziehungsweise in einer Reduzierung des erforderlichen Stichprobenumfangs wider. Gleichzeitig wird dabei die Wahrscheinlichkeit eines Fehlers 1. Art durch das vorgegebene Signifikanzniveau eingehalten. Die Existenz und das

Ausmaß dieses Mehrwerts hängt jedoch maßgeblich von den zugrundeliegenden historischen Daten ab. Es wurden sowohl Szenarien identifiziert, die mit einem hohen als auch solche, die mit gar keinem Mehrwert einhergehen.

Bibliography

Altham, P. M. (1969). Exact bayesian analysis of a 2 times 2 contingency table, and fisher's "exact" significance test. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):261–269.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical science*, pages 101–117.

Chen, M.-H., Ibrahim, J. G., Lam, P., Yu, A., and Zhang, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*, 67(3):1163–1170.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84(1-2):121–137.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Duan, Y., Smith, E. P., and Ye, K. (2006). Using power priors to improve the binomial test of water quality. *Journal of agricultural, biological, and environmental statistics*, 11(2):151.

Fagerland, M., Lydersen, S., and Laake, P. (2017). *Statistical analysis of contingency tables*. CRC press.

Feder, P. I., Aume, L. L., Triplett, C. A., Simmons, J. E., and Narotsky, M. G. (2020). Analysis of proportional data in reproductive and developmental toxicity studies: Comparison of sensitivities of logit transformation, arcsine square root transformation, and nonparametric analysis. *Birth Defects Research*, 112(16):1260–1272.

Feißt, M., Krisam, J., and Kieser, M. (2020). Incorporating historical two-arm data in clinical trials with binary outcome: A practical approach. *Pharmaceutical statistics*, 19(5):662–678.

Gamalo-Siebers, M., Savic, J., Basu, C., Zhao, X., Gopalakrishnan, M., Gao, A., Song, G., Baygani, S., Thompson, L., Xia, H. A., et al. (2017). Statistical modeling for bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharmaceutical statistics*, 16(4):232–249.

Gravestock, I., Held, L., and consortium, C.-N. (2017). Adaptive power priors with empirical bayes for clinical trials. *Pharmaceutical statistics*, 16(5):349–360.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056.

Howard, J. (1998). The 2× 2 table: A discussion from a bayesian viewpoint. *Statistical Science*, pages 351–367.

Ibrahim, J. G., Chen, M.-H., et al. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749.

ICH E9 expert working group (1999). Statistical principles for clinical trials. international conference on harmonisation e9 expert working group. *Stat. Med.*, 18:1905–1942.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314.

Kawasaki, Y. and Miyaoka, E. (2012). A bayesian inference of p ($\pi 1$¿ $\pi 2$) for two proportions. *Journal of biopharmaceutical statistics*, 22(3):425–437.

Khanna, D., Denton, C. P., Jahreis, A., van Laar, J. M., Frech, T. M., Anderson, M. E., Baron, M., Chung, L., Fierlbeck, G., Lakshminarayanan, S., et al. (2016). Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis (fasscinate): a phase 2, randomised, controlled trial. *The Lancet*, 387(10038):2630–2640.

Kopp-Schneider, A., Calderazzo, S., and Wiesenfarth, M. (2020). Power gains by using external information in clinical trials are typically not possible when requiring strict type i error control. *Biometrical Journal*, 62(2):361–374.

Lee, P. (2004). Bayesian statistics: An introduction. 2004. *London: Arnold*.

Lesaffre, E., Baio, G., and Boulanger, B. (2020). *Bayesian Methods in Pharmaceutical Research*. CRC Press.

Lydersen, S., Fagerland, M. W., and Laake, P. (2009). Recommended tests for association in 2× 2 tables. *Statistics in medicine*, 28(7):1159–1175.

Lydersen, S., Langaas, M., and Bakke, Ø. (2012). The exact unconditional z-pooled test for equality of two binomial probabilities: optimal choice of the berger and boos confidence coefficient. *Journal of Statistical Computation and Simulation*, 82(9):1311–1316.

Nurminen, M. and Mutanen, P. (1987). Exact bayesian analysis of two proportions. *Scandinavian Journal of Statistics*, pages 67–77.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, 29(3):175–188.

Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G., and Hoijtink, H. J. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, 32(6):848–855.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

Seide, S. E., Röver, C., and Friede, T. (2019). Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC medical research methodology*, 19(1):16.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54.

Warton, D. I. and Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10.

Weber, S., Gelman, A., Lee, D., Betancourt, M., Vehtari, A., Racine-Poon, A., et al. (2018). Bayesian aggregation of average data: An application in drug development. *The Annals of Applied Statistics*, 12(3):1583–1604.

Witting, H. (1985). Mathematische statistik i [mathematical statistics, vol. i]. *Stuttgart, Germany: Teubner. Measurement of Short-Term Changes of Ability*.

Yu, G. (2009). Variance stabilizing transformations of poisson, binomial and negative binomial distributions. *Statistics & Probability Letters*, 79(14):1621–1629.

Zaslavsky, B. G. (2013). Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. *Biometrics*, 69(1):157–163.

Appendix A

Additional Tables and Figures

A.1 Systematic investigations concerning the choice of the parameter γ

In the following, the impact of various values of γ on the resulting value of δ^*, is investigated (see Subsection 2.7.2). Hereby the following scenario are examined:

- 50, 60 and 70 responders within 100 patients in the historical control arm (c_H),
- 5, 10, 15, 20 and 25 more responders within 100 patients in the historical treatment arm than in the historical control arm ($t_H - c_H$)
- 200 patients per arm in the new trial (n_C and n_T),
- $\gamma = 0, 0.00001, 0.0001, 0.0005, 0.001, 0.002$ and 0.01

Table A.1: δ^*(the maximum value of δ controlling the type I error rate) for various scenarios and values of γ.

c_H	t_H	$\gamma = 0$	$\gamma = 0.00001$	$\gamma = 0.0001$	$\gamma = 0.0005$	$\gamma = 0.001$	$\gamma = 0.002$	$\gamma = 0.01$
50	55	1	1	1	1	1	1	1
50	60	0.46	0.46	0.46	0.45	0.44	0.43	0
50	65	0.2	0.2	0.19	0.19	0.14	0.12	0
50	70	0.09	0.09	0.09	0.09	0	0	0
50	75	0.03	0.03	0.03	0	0	0	0
60	65	1	1	1	1	1	1	1
60	70	0.43	0.43	0.43	0.42	0.4	0.38	0
60	75	0.18	0.18	0.18	0.16	0.14	0	0
60	80	0.09	0.09	0.09	0.06	0	0	0
60	85	0.03	0.03	0.03	0	0	0	0
70	75	1	1	1	1	1	1	1
70	80	0.38	0.37	0.37	0.36	0.34	0.32	0
70	85	0.15	0.15	0.14	0.13	0.11	0	0
70	90	0.06	0.06	0.06	0.03	0	0	0
70	95	0.03	0.02	0	0	0	0	0

Table A.2: $1 - \gamma$ confidence interval for the true control proportion for various scenarios and values of γ.

c_H	$\gamma = 0$	$\gamma = 0.00001$	$\gamma = 0.0001$	$\gamma = 0.0005$	$\gamma = 0.001$	$\gamma = 0.002$	$\gamma = 0.01$
50	[0;1]	[0.29;0.71]	[0.31;0.69]	[0.33;0.67]	[0.34;0.66]	[0.34;0.66]	[0.37;0.63]
60	[0;1]	[0.38;0.79]	[0.40;0.78]	[0.42;0.76]	[0.43;0.75]	[0.44;0.75]	[0.47;0.72]
70	[0;1]	[0.48;0.87]	[0.50;0.86]	[0.52;0.84]	[0.53;0.84]	[0.54;0.83]	[0.57;0.81]

A.2 Actual type I error rate of the chi-square test

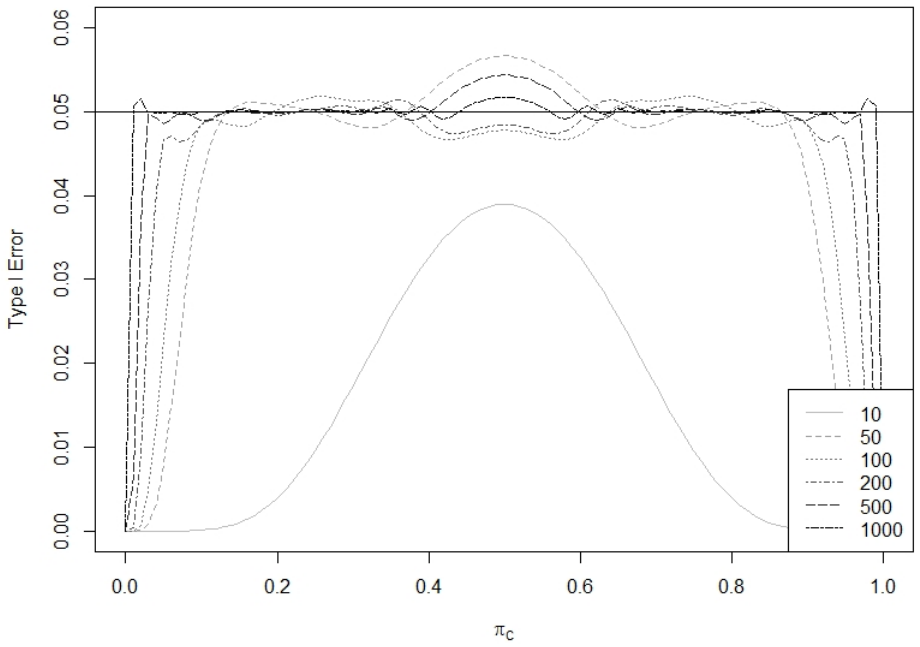

Figure A.1: Actual type I error rate of the chi-square test for different sample sizes over the range of the true control proportion (π_C). The darker the colour the larger the sample size (ranging from 10 to 1000).

A.3 Actual type I error rate of the chi-square test

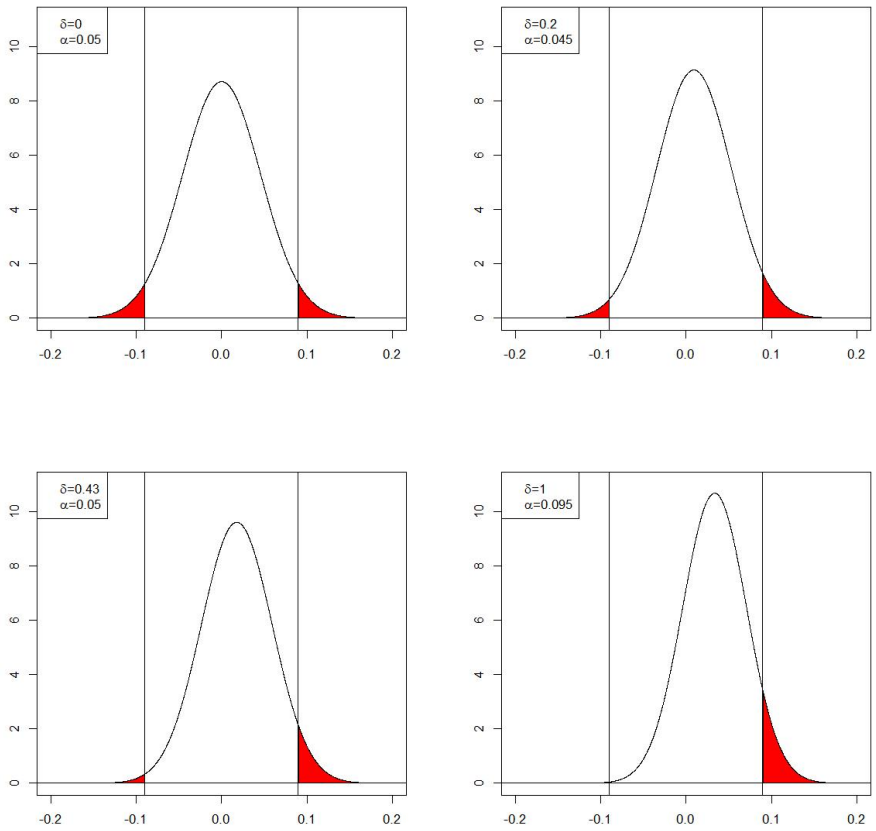

Figure A.2: Test statistics and actual type I error rates α of a normal distributed test statistics for various values of δ. The red area represents the true type I error rate α. The underlying scenario is: $c_H = 65$ responders out of $n_{CH} = 100$ patients in the historical control arm, $c_H = 75$ responders out of $n_{TH} = 100$ patients in the historical treatment arm, $n = 200$ patients per arm in the new trial and a true control proportion $\pi_C = 0.7$.

Appendix B

R Code

```
#####Input:

#c_old:   number of responders in the historical control arm
#t_old:   number of responders in the historical treatment arm
#nc_old:  number of patients in the historical control arm
#nt_old:  number of patients in the historical treatment arm
#nc:      number of patients in the new control arm
#nt:      number of patients in the new tretment arm
#delta:   factor that determines the amount of borrowed historical
    information
#pi:      true control proportion
#ES:      effect size pi_T-pi_C
#parts:   number of parts in which the area of the true control
    proportion is divided (rec.:100)
#alpha:   significance level
#power:   power to detect the effect
#gamma:   parameter of the Berger and Boos procedure


######### General functions
###################################

require(compiler)


#####Indicator function
###################################
Indicator<-function(x,min,max){
if(min<=x && x<=max){
y<-1

}
else{
y<-0
```

```
}
return(y)
}

##### Function for calculating the true type I error
###################################
truealpha<-function(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi){
p<-matrix(c(rep(0,nc*nt)),nrow=nc)
for (i in 1:(nc-1)){
for(j in 1:(nt-1)){
tab<-matrix(c(c_old*delta+i,nc_old*delta+nc-c_old*delta-i,t_old*
    delta+j,nt_old*delta+nt-t_old*delta-j),nrow=2)
p[i,j]<-dbinom(i,nc,pi)*dbinom(j,nt,pi)*(1-Indicator(chisq.test(tab
    ,correct=FALSE)$statistic,-1,qchisq(0.95,1)))
}
}

g3<-sum(p)
return(g3)
}
truealpha<-Vectorize(truealpha)
truealpha<-cmpfun(truealpha)



##### Function for calculating the true power
###################################
truepower<-function(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES){
p<-matrix(c(rep(0,nc*nt)),nrow=nc)
for (i in 1:(nc-1)){
for(j in 1:(nt-1)){
tab<-matrix(c(c_old*delta+i,nc_old*delta+nc-c_old*delta-i,t_old*
    delta+j,nt_old*delta+nt-t_old*delta-j),nrow=2)
p[i,j]<-dbinom(i,nc,pi)*dbinom(j,nt,pi+ES)*(1-Indicator(chisq.test(
    tab,correct=FALSE)$statistic,-1,qchisq(0.95,1)))
}
}

g3<-sum(p)
return(g3)
}
truepower<-Vectorize(truepower)
truepower<-cmpfun(truepower)









###################################
```

```
#### Local Procedure
##################################


#### Function that calculates delta^* for a fixed true control
    proportion
##################################
sc<-function(c_old,t_old,nc_old,nt_old,nc,nt,pi,ES,gamma,alpha){

thealpha<-alpha-gamma


#Nested intervals procedure
delta_opt<-0
delta<-0:2/2
alpha1<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[1],pi)
alpha2<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[2],pi)
alpha3<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[3],pi)

#1. Step
if(thealpha>alpha3)
{delta_opt<-1}

else{

if(alpha2>thealpha)
{delta1<-0
delta2<-0.25
delta3<-0.5
alpha_l<-alpha1
alpha_r<-alpha2}

else
{delta1<-0.5
delta2<-0.75
delta3<-1
alpha_l<-alpha2
alpha_r<-alpha3}


alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)



#2.Step
if(alpha_m<thealpha){
delta_l<-delta2
delta_m<-round((delta2+delta3)/2,2)
delta_r<-delta3
}
```

```
128
129 else{
130 delta_l<-delta1
131 delta_m<-round((delta1+delta2)/2,2)
132 delta_r<-delta2
133 }
134
135 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)
136
137
138 #3.Step
139 if(alpha_m<thealpha){
140 delta1<-delta_m
141 delta2<-round((delta_m+delta_r)/2,2)
142 delta3<-delta_r
143 }
144
145 else{
146 delta1<-delta_l
147 delta2<-round((delta_l+delta_m)/2,2)
148 delta3<-delta_m
149 }
150
151 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)
152
153 #4.Step
154 if(alpha_m<thealpha){
155 delta_l<-delta2
156 delta_m<-round((delta2+delta3)/2,2)
157 delta_r<-delta3
158 }
159
160 else{
161 delta_l<-delta1
162 delta_m<-round((delta1+delta2)/2,2)
163 delta_r<-delta2
164 }
165
166 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)
167
168 #5.Step
169 if(alpha_m<thealpha){
170 delta1<-delta_m
171 delta2<-round((delta_m+delta_r)/2,2)
172 delta3<-delta_r
173 }
174
175 else{
176 delta1<-delta_l
177 delta2<-round((delta_l+delta_m)/2,2)
```

```
delta3<-delta_m
}

alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)

#6.Step
if(alpha_m<thealpha){
delta_l<-delta2
delta_m<-round((delta2+delta3)/2,2)
delta_r<-delta3
}

else{
delta_l<-delta1
delta_m<-round((delta1+delta2)/2,2)
delta_r<-delta2
}

alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)

#7.Step
if(alpha_m<thealpha){
delta1<-delta_m
delta2<-round((delta_m+delta_r)/2,2)
delta3<-delta_r
}

else{
delta1<-delta_l
delta2<-round((delta_l+delta_m)/2,2)
delta3<-delta_m
}

alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)

#8.Step
if(alpha_m<thealpha){
delta_l<-delta2
delta_m<-round((delta2+delta3)/2,2)
delta_r<-delta3
}

else{
delta_l<-delta1
delta_m<-(floor(100*(delta1+delta2)/2))/100
delta_r<-delta2
}

delta_opt<-max(0,delta_m)
}
```

```
228
229 return(list(delta_opt=delta_opt))
230
231
232 }
233
234
235
236 ### Function that calculates delta^*
237 ###################################
238 localpro<-function(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,gamma,
        alpha){
239 cpu<-0
240 cpo<-0
241 delta<-0
242 power0<-0
243 power1<-0
244 power2<-0
245 alpha0<-0
246 alpha1<-0
247 alpha2<-0
248 power<-0
249 delta_min<-0
250
251
252 #Calculate the Pearson Clopper Confidence Interval
253 cpu<-qbeta(gamma/2,c_old,nc_old-c_old+1) #  pu = BETAINV(x/2;k;n-k
        +1)
254 cpo<-qbeta(1-gamma/2,c_old+1,nc_old-c_old) #po = BETAINV(1-x/2;k+1;
        n-k)
255
256 #Seperate the interval in parts
257 steps<-0
258 step<-(cpo-cpu)/parts
259 for(i in 1:parts){
260 steps[i]<-cpu+(i)*step
261 }
262 steps<-c(cpu,steps)
263
264 #Calculate delta for every pi in the Confidence interval
265 for(i in 1:length(steps)){
266 fit<-sc(c_old,t_old,nc_old,nt_old,nc,nt,steps[i],ES,gamma,alpha)
267 delta[i]<-fit$delta_opt
268 }
269
270 delta_max<-min(delta)
271
272 alpha<-0
273 power<-0
274 for(i in 1:length(steps)){
```

```
alpha1[i]<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_max,
    steps[i])
power1[i]<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta_max,
    steps[i],ES)
alpha0[i]<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,0,steps[i])
power0[i]<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,0,steps[i],ES)
}

return(list(alpha1=alpha1,power1=power1,delta_max=delta_max,delta=
    delta,power0=power0,
alpha0=alpha0,steps=steps))


}


##### Function for the sample size calculation for the local
    procedure
###################################
samplesizerl<-function(c_old,t_old,nc_old,nt_old,pi,ES,parts,alpha,
    power,gamma){

#1.SampleSizeCalculation initial
library(pwr)
w<-ES.h(pi+ES,pi)
n<-ceiling(pwr.2p.test(h=w,power=power,sig.level=alpha)$n)
nc<-n
nt<-n

#2.Algorithm
#Calculate Confidence interval
cpu<-qbeta(gamma/2,c_old,nc_old-c_old+1) #  pu = BETAINV(x/2;k;n-k
    +1)
cpo<-qbeta(1-gamma/2,c_old+1,nc_old-c_old) #po = BETAINV(1-x/2;k+1;
    n-k)

# Steps in the algorithm
if(pi<cpu || pi>cpo){print('Error: pi is not in the confidence
    interval')}
else{

steps<-1
fit<-localpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,gamma,alpha)
delta<-fit$delta_max
pi<-max(fit$steps[which(round(fit$steps,digits=2)==pi)])
power1<-fit$power1[which(fit$steps==pi)]

while(power1>power){
power1<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES)
nc<-nc-1
```

```
nt<-nt-1
}
nc<-nc+1
nt<-nt+1

fit<-localpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,gamma,alpha)
delta1<-fit$delta_max
power1<-fit$power1[which(fit$steps==pi)]

while(delta1>delta){
steps<-steps+1
while(power1>power){
power1<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta1,pi,ES)
nc<-nc-1
nt<-nt-1
}
delta<-delta1
nc<-nc+1
nt<-nt+1
fit<-localpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,gamma,alpha)
delta1<-fit$delta_max

}
nc+1
nt+1
n_won1<-n-nc
n_won2<-n-nt

return(list(fit=fit,n_won1=n_won1,n_won2=n_won2,n=n,nc=nc,nt=nt,
    delta=delta,delta1=delta1,steps=steps))}
}

###Output
#fit:            alpha and power values at each pi
#n_won1;n_won2: saved sample size in the control arm and the
    treatment arm, respectively
#n:              initial sample size
#nc;nt:          new reduced sample size
#delta:          delta^*
#delta1:         last delta in the algorithm
#steps:          values of pi where delta^* was calculated

###################################
##### Global procedure
###################################

### Function that calculates delta^* for a fixed true control
    proportion pi
###################################
sc1<-function(c_old,t_old,nc_old,nt_old,nc,nt,pi,ES,alpha){
```

```
364
365 thealpha<-alpha
366 delta_opt<-0
367
368 #Nested intervals procedure
369 delta<-0:2/2
370 alpha1<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[1],pi)
371 alpha2<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[2],pi)
372 alpha3<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[3],pi)
373
374
375 #1. Step
376 if(thealpha>alpha3)
377 {delta_opt<-1}
378
379 else{
380
381 if(alpha2>thealpha)
382 {delta1<-0
383 delta2<-0.25
384 delta3<-0.5
385 alpha_l<-alpha1
386 alpha_r<-alpha2}
387
388 else
389 {delta1<-0.5
390 delta2<-0.75
391 delta3<-1
392 alpha_l<-alpha2
393 alpha_r<-alpha3}
394
395
396 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)
397
398
399 #2.Step
400 if(alpha_m<thealpha){
401 delta_l<-delta2
402 delta_m<-round((delta2+delta3)/2,2)
403 delta_r<-delta3
404 }
405
406 else{
407 delta_l<-delta1
408 delta_m<-round((delta1+delta2)/2,2)
409 delta_r<-delta2
410 }
411
412 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)
413
```

```
414
415 #3.Step
416 if(alpha_m<thealpha){
417 delta1<-delta_m
418 delta2<-round((delta_m+delta_r)/2,2)
419 delta3<-delta_r
420 }
421
422 else{
423 delta1<-delta_l
424 delta2<-round((delta_l+delta_m)/2,2)
425 delta3<-delta_m
426 }
427
428 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)
429
430
431 #4.Step
432 if(alpha_m<thealpha){
433 delta_l<-delta2
434 delta_m<-round((delta2+delta3)/2,2)
435 delta_r<-delta3
436 }
437
438 else{
439 delta_l<-delta1
440 delta_m<-round((delta1+delta2)/2,2)
441 delta_r<-delta2
442 }
443
444 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)
445
446 #5.Step
447 if(alpha_m<thealpha){
448 delta1<-delta_m
449 delta2<-round((delta_m+delta_r)/2,2)
450 delta3<-delta_r
451 }
452
453 else{
454 delta1<-delta_l
455 delta2<-round((delta_l+delta_m)/2,2)
456 delta3<-delta_m
457 }
458
459 alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)
460
461
462 #6.Step
463 if(alpha_m<thealpha){
```

```
delta_l<-delta2
delta_m<-round((delta2+delta3)/2,2)
delta_r<-delta3
}

else{
delta_l<-delta1
delta_m<-round((delta1+delta2)/2,2)
delta_r<-delta2
}

alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_m,pi)


#7.Step
if(alpha_m<thealpha){
delta1<-delta_m
delta2<-round((delta_m+delta_r)/2,2)
delta3<-delta_r
}

else{
delta1<-delta_l
delta2<-round((delta_l+delta_m)/2,2)
delta3<-delta_m
}

alpha_m<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta2,pi)

#8.Step
if(alpha_m<thealpha){
delta_l<-delta2
delta_m<-round((delta2+delta3)/2,2)
delta_r<-delta3
}

else{
delta_l<-delta1
delta_m<-(floor(100*(delta1+delta2)/2))/100
delta_r<-delta2
}

delta_opt<-max(0,delta_m)
}

return(list(delta_opt=delta_opt))
}
```

```
514 ### Function that calculates delta^*
515 ###################################
516 globalpro<-function(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,alpha)
        {
517 cpu<-0
518 cpo<-0
519 delta<-0
520 power0<-0
521 power1<-0
522 power2<-0
523 alpha0<-0
524 alpha1<-0
525 alpha2<-0
526 power<-0
527 delta_min<-0
528
529
530 #seperate the values of pi in parts
531 steps<-(0:parts/parts)[2:parts]
532
533 #Calculate delta for each pi
534 for(i in 1:length(steps)){
535 fit<-sc1(c_old,t_old,nc_old,nt_old,nc,nt,steps[i],ES,alpha)
536 delta[i]<-fit$delta_opt
537 }
538
539 delta_max<-min(delta)
540
541 alpha<-0
542 power<-0
543 for(i in 1:length(steps)){
544 alpha1[i]<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,delta_max,
        steps[i])
545 power1[i]<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta_max,
        steps[i],ES)
546 alpha0[i]<-truealpha(c_old,t_old,nc_old,nt_old,nc,nt,0,steps[i])
547 power0[i]<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,0,steps[i],ES)
548 }
549
550 return(list(alpha1=alpha1,power1=power1,delta_max=delta_max,delta=
        delta,power0=power0,
551 alpha0=alpha0,steps=steps))
552
553
554 }
555
556
557 ##### Function for the sample size calculation for the global
        procedure
558 ###################################
```

```
samplesizeri<-function(c_old,t_old,nc_old,nt_old,pi,ES,parts,alpha,
    power){

#1.SampleSizeCalculation inital
library(pwr)
w<-ES.h(pi+ES,pi)
n<-ceiling(pwr.2p.test(h=w,power=power,sig.level=alpha)$n)
nc<-n
nt<-n

#2.Algorithm
steps<-1
fit<-globalpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,alpha)
delta<-fit$delta_max
power1<-fit$power1[pi*100]
while(power1>power){
power1<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES)
nc<-nc-1
nt<-nt-1
}
nc<-nc+1
nt<-nt+1

fit<-globalpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,alpha)
delta1<-fit$delta_max
power1<-fit$power1[pi*100]

while(delta1>delta){
steps<-steps+1
while(power1>power){
power1<-truepower(c_old,t_old,nc_old,nt_old,nc,nt,delta1,pi,ES)
nc<-nc-1
nt<-nt-1
}
delta<-delta1
nc<-nc+1
nt<-nt+1
fit<-globalpro(c_old,t_old,nc_old,nt_old,nc,nt,ES,parts,alpha)
delta1<-fit$delta_max

}
nc+1
nt+1
n_won1<-n-nc
n_won2<-n-nt

return(list(fit=fit,n_won1=n_won1,n_won2=n_won2,n=n,nc=nc,nt=nt,
    delta=delta,delta1=delta1,steps=steps))
}

```

```
607 ##################################
608 #### Independent Procedure
609 ##################################
610
611
612 #### Function that calculates the true typ I error
613 ##################################
614 normalalpha<-function(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES){
615 p1<-c_old/nc_old
616 p2<-t_old/nt_old
617 n_neu<-nc
618 mu<-p2-p1
619 pi1<-pi
620 pi2<-pi+ES
621 sigma<-sqrt(p1*(1-p1)/nc_old+p2*(1-p2)/nt_old)
622 sigma_neu<-sqrt(pi1*(1-pi1)/nc+pi2*(1-pi2)/nt)
623
624 #function that simulates the indicator function
625 big0<-function(y){
626 big00<-0
627 for(i in 1:length(y)){
628 wkt<-function(x){(dnorm(x,y[i],sigma_neu)*((dnorm(x,mu,sigma))^
    delta))}
629 int<-integrate(wkt,lower=-Inf,upper=Inf)
630 wkt1<-function(x){(dnorm(x,y[i],sigma_neu)*((dnorm(x,mu,sigma))^
    delta))/int$value}
631
632 big<-integrate(wkt1,lower=-Inf,upper=0)$value
633
634
635 if(big>0.975|big<0.025){
636 big00[i]<-1
637 }
638 else big00[i]<-0
639 }
640 return(big00)
641 }
642 w<-function(x){dnorm(x,(pi2-pi1),sigma_neu)}
643 ww<-function(x){w(x)*big0(x)}
644 prob<-integrate(ww,lower=-1.5,upper=1.5)$value
645 return(prob)
646 }
647
648
649 ### Function that calculates delta^*
650 ##################################
651
652 optdelta<-function(c_old,t_old,nc_old,nt_old,nc,nt,pi,ES){
653 delta0<-0
654 delta<-0:100/100
```

```
655 for(i in 1:101){
656 delta0[i]<-normalalpha(c_old,t_old,nc_old,nt_old,nc,nt,delta[i],pi,
        ES)
657 }
658 delta_opt<-delta[max(which(delta0<0.05))]
659 return(delta_opt)
660 }
661
662 ##### Function for the sample size calculation for the independent
        procedure
663 ###################################
664 samplesizer<-function(c_old,t_old,nc_old,nt_old,pi,ES){
665
666 #1.SampleSizeCalculation
667 power<-0.8
668 library(pwr)
669 w<-ES.h(pi+ES,pi)
670 n<-ceiling(pwr.2p.test(h=w,power=0.8,sig.level=0.05)$n)
671 nc<-n
672 nt<-n
673
674 #2.Calculate delta^* and power
675 delta<-optdelta(c_old,t_old,nc_old,nt_old,nc,nt,pi,0)
676 power1<-normalalpha(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES)
677
678
679 while(power1>power){
680 power1<-normalalpha(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES)
681 nc<-nc-1
682 nt<-nt-1
683 }
684 nc<-nc+1
685 nt<-nt+1
686
687 delta1<-optdelta(c_old,t_old,nc_old,nt_old,nc,nt,pi,0)
688
689 steps<-1
690 while(delta1>delta){
691 steps<-steps+1
692 while(power1>power){
693 power1<-normalalpha(c_old,t_old,nc_old,nt_old,nc,nt,delta,pi,ES)
694 nc<-nc-1
695 nt<-nt-1
696 }
697 delta<-delta1
698 nc<-nc+1
699 nt<-nt+1
700 fit<-optdelta(c_old,t_old,nc_old,nt_old,nc,nt,pi,0)
701
702 }
```

```
703
704
705 n_won1<-n-nc
706 n_won2<-n-nt
707
708 return(list(fit=fit,n_won1=n_won1,n_won2=n_won2,n=n,nc=nc,nt=nt,
        delta=delta,delta1=delta1,steps=steps))
709 }
710
711
712 ###################################
713 ###################################
714 ###Input:
715 #c_old:   number of responders in the historical control arm
716 #t_old:   number of responders in the historical treatment arm
717 #nc_old:  number of patients in the historical control arm
718 #nt_old:  number of patients in the historical treatment arm
719 #nc:      number of patients in the new control arm
720 #nt:      number of patients in the new tretment arm
721 #delta:   factor that determines the amount of borrowed historical
        information
722 #pi:      true control proportion
723 #ES:      effect size pi_T-pi_C
724 #parts:    number of parts in which the area of the true control
        proportion is divided (rec.:100)
725 #alpha:   the significance level
726 #power:   the power to detect the effect
727 #gamma:   parameter of the Berger and Boos procedure
728
729 ###Output
730 #fit:                    # alpha1 and power1: alpha and power values
         after borrowing for each pi
731 # delta_max, delta: delta^* and the max delta for each pi
732 # power0, alph0: alpha and power values without borrowing for each
        pi
733 #n_won1;n_won2: saved sample size in the control arm and the
        treatment arm, respectively
734 #n:              initial sample size
735 #nc;nt:          new reduced sample size
736 #delta:          delta^*
737 #delta1:         last delta of the algorithm
738 #steps:          values of pi where delta^* was calculated
739
740 ###################################
741 ###### Example from the Manuscript (Clincial trial example from
        Chapter 3)
742 ###################################
743
744 ##########Local procedure:
745 ###################################
```

```
746 fit1<-samplesizerl(c_old = 10,t_old = 16,nc_old = 44,nt_old = 43,pi
        = 0.23,ES = 0.14,parts = 100,alpha = 0.05,power = 0.8,gamma =
    0.0001)
747
748 ##########Global procedure:
749 ###################################
750 fit2<-samplesizer(c_old = 10,t_old = 16,nc_old = 44,nt_old = 43,pi
    = 0.23,ES = 0.14,parts = 100,alpha = 0.05,power = 0.8)
751
752 ########## Independent procedure:
753 ###################################
754 fit2<-samplesizeri(c_old = 10,t_old = 16,nc_old = 44,nt_old = 43,pi
        = 0.23,ES = 0.14,parts = 100,alpha = 0.05,power = 0.8)
```

Curriculum Vitae

Manuel Feißt
born 10 December 1989 in Herbolzheim, Germany

Education

Ruprecht-Karls-University of Heidelberg	Since 05/2016
Doctoral student (Dr. sc. hum.)	
Baden-Württemberg Zertifikat für Hochschuldidaktik	14/10/2019
Albert-Ludwigs-University of Freiburg	10/2010 – 04/2016
Master of Science Mathematics (M.Sc.)	04/2016
Bachelor of Science Mathematics (B.Sc.)	09/2013
Städtisches Gymnasium Ettenheim	08/2000 – 06/2009
A-Level (Abitur)	06/2009
Ferdinand-Ruska-Schule Grafenhausen	09/1996 – 06/2000

Professional experience

Ruprecht-Karls-University of Heidelberg Head of the working group 'Statistical Modelling' at the Institute of Medical Biometry and Informatics	Since 01/2020
Ruprecht-Karls-University of Heidelberg Head of the working group 'Lehre' at the Institute of Medical Biometry and Informatics	05/2017 – 12/2019
Ruprecht-Karls-University of Heidelberg Research fellow at the Institute of Medical Biometry and Informatics	Since 05/2016
Albert-Ludwigs-University of Freiburg Student assistant at the Institute of Medical Biometry and Statistics	01/2015 - 12/2015
Albert-Ludwigs-University of Freiburg Teaching assistant at the Institute of Mathematics	10/2011 - 03/2016

Publications

Methodological publications

Feißt M., Hennigs A.,Heil J., Moosbrugger H., Kelava A., Stolpner I., Kieser M., Rauch G. (2019): Refining scores based on patient reported outcomes statistical and medical perspectives. *BMC Medical Research Methodology,19:167,2019*

Feißt M., Krisam J., Kieser M. (2020): Incorporating historical two-arm data in clinical trials with binary outcome: A practical approach. *Pharmaceutical Statistics,19/5:662-678,2020*

Project-based publications

Wallwiener M, Matthies L, Simoes E, Keilmann L, Hartkopf A.D, Sokolov A.N, Walter C.B, Sickenberger N, Wallwiener S, **Feisst M,** Gass P, Fasching P.A, Lux M.P, Wallwiener D, Taran F.A, Rom J, Schneeweiss A, Graf J, Brucker S.Y. (2017) Reliability of an e-PRO Tool of EORTC QLQ-C30 for Measurement of Health-Related Quality of Life in Patients With Breast Cancer: Prospective Randomized Trial. *Journal of Medical Internet Research* ;**19**/9:e322.

Hennigs, A., Heil, J., Wagner, A., Rath, M., Moosbrugger, H., Kelava, A., ... & **Feißt, M**. (2018). Development and psychometric validation of a shorter version of the Breast Cancer Treatment Outcome Scale (BCTOS-12). *The Breast*, *38, 58-65, 2020.*

Fink C. A., Friedrich M., Frey P. E., Raedeker L., Leuk A., Bruckner T., **Feisst M**., Tenckhoff S., Klose C., Doerr-Harim C., Neudecker J., Mihaljevic A. L. (2018): Prospective multicentre cohort study of patient-reported outcomes and complications following major abdominal neoplastic surgery (PATRONUS) - study protocol for a CHIR-Net student-initiated German medical audit study (CHIR-Net SIGMA study). *BMC Surgery,18:90,2018*

Heil J., Hug S., Martiny H., Golatta M., **Feisst M**., Madjar H., Bader W., Hahn M. (2018): Standards of hygiene for ultrasound-guided core cut biopsies of the breast. Hygienestandards bei der ultraschallgesteuerten Stanzbiopsie an der Mamma. *Ultraschall in der Medizin / European Journal of Ultrasound,39/6:636-641,2018*

Smetanay K., Junio P., **Feißt M.**, Seitz J., Hassel J. C., Mayer L., Matthies L.M., Schumann A., Hennigs A., Heil J., Sohn C., Jaeger D., Schneeweiss A., Marmë F. (2019): COOLHAIR: a prospective randomized trial to investigate the efficacy and tolerability of scalp cooling in patients undergoing (neo)adjuvant chemotherapy for early breast cancer. *Breast Cancer Research and Treatment, 172/3:135-145,2019*

Hennigs A., Riedel F., **Feißt M.**, Köpke M., Rezai M., Nitz U., Moderow M., Golatta M., Sohn C., Heil J. (2019): Evolution of the Use of Completion Axillary Lymph Node Dissection in Patients with T1/2N0M0 Breast Cancer and Tumour-Involved Sentinel Lymph Nodes Undergoing Mastectomy: A Cohort Study. *Annals of Surgical Oncology, 26/8:2435-2443, 2019*

Ullrich P., Werner C., Eckert T., Bongartz M., Kiss R., **Feißt M.**, Delbaere K.,Bauer J. M., Hauer K. (2019): Cut-off for the Life-Space Assessment in persons with cognitive impairment. *Aging Clinical and Experimental Research, 31/9:1331-1335, 2019*

Hennigs A., Köpke M., **Feißt M.**, Riedel F., Rezai M., Nitz U., Moderow M., Golatta M., Sohn C., Schneeweiss A., Heil J. (2019): Which patients with sentinel node-positive breast cancer after breast conservation still receive completion axillary lymph node dissection in routine clinical practice? *Breast Cancer Research and Treatment,173/3:429-438,2019*

Wallwiener S., Goetz M., Lanfer A., Gillessen A., Suling M., **Feisst M.**, Sohn C., Wallwiener M. (2019): Epidemiology of mental disorders during pregnancy and link to birth outcome: a large-scale retrospective observational database study including 38,000 pregnancies. *Archives of Gynecology and Obstetrics,299/3:755-763,2019*

Matthies L., Taran F. A., Keilmann L., Schneeweiss A., Simoes E., Hartkopf A. D., Sokolov A. N., Walter C. B., Sickenberger N., Wallwiener S., **Feisst M.**, Gass P., Lux M. P., Schütz F., Fasching P. A., Sohn C., Brucker S. Y., Graf J., Wallwiener M. (2019): An Electronic Patient-Reported Outcome Tool for the FACT-B (Functional Assessment of Cancer Therapy-Breast) Questionnaire for Measuring the Health-Related Quality of Life in Patients With Breast Cancer:

Reliability Study. *Journal of Medical Internet Research,21/1:e10004,2019*

Wallwiener M., Nabieva N., **Feisst M.**, Fehm T., de Waal J., Rezai M., Baier B., Baake G., Kolberg H. C., Guggenberger M., Warm M., Harbeck N., Wuerstlein R., Deuker J. U., Dall P., Richter B., Wachsmann G., Brucker C., Siebers J. W., Popovic M., Kuhn T., Wolf C., Vollert H. W., Breitbach G. P., Janni W., Landthaler R., Kohls A., Rezek D., Noesselt T., Fischer G., Henschen S., Praetz T., Heyl V., Kühn T., Krauss T., Thomssen C., Hohn A., Tesch H., Mundhenke C., Hein A., Rauh C., Bayer C. M., Schmidt K., Belleville E., Brucker S. Y., Hadji P., Beckmann M. W., Wallwiener D., Kümmel S., Hartkopf A., Fasching P. A. (2019): Influence of patient and tumor characteristics on therapy persistence with letrozole in postmenopausal women with advanced breast cancer: results of the prospective observational EvAluate-TM study. *BMC Cancer,19:611,2019*

Riedel F., Heil J., **Feißt M.**, Rezai M., Moderow M., Sohn C., Schütz F., Golatta M., Hennigs A. (2019): Non-sentinel axillary tumor burden applying the ACOSOG Z0011 eligibility criteria to a large routine cohort. *Breast Cancer Research and Treatment,177/2:457-467,2019*

Stolpner I., Heil J., **Feißt M.**, Karsten M. M., Weber W. P., Blohmer J. U., Forster T., Golatta M., Schütz F., Sohn C., Hennigs A. (2019): Clinical Validation of the BREAST-Q Breast-Conserving Therapy Module. *Annals of Surgical Oncology,26/9:2759-2767,2019*

Rädecker L., Schwab M., Frey P., Friedrich M., Sliwinski S., Steinle J., Fink C., Leuk A., Ganschow P., Ottawa G., Klose C., **Feisst M.**, Dörr-Harim C., Tenckhoff S., Mihaljevic A. (2019): Design und Evaluation eines Prüf-Studierenden-Kurses für studentische prospektive Multizenterstudien – ein CHIR-Net SIGMA-Projekt zum forschenden Lernen. *Zentralblatt für Chirurgie,DOI 10.1055/a-1007-1995,2019(elektronischer Sonderdruck)*

Bongartz M., Kiss R., Lacroix A., Eckert T., Ullrich P., Jansen C. P., **Feißt M.**, Mellone S., Chiari L., Becker C., Hauer K. (2019): Validity, reliability, and feasibility of the uSense activity monitor to register physical activity and gait performance in habitual settings of geriatric patients. *Physiological Measurement,40/9,2019*

Feißt M., Heil J., Stolpner I., von Au A., Domschke C., Sohn C., Kieser M., Rauch G., Hennigs A. (2019): Psychometric Validation of the Breast Cancer Treatment Outcome Scale (BCTOS-12): a Prospective Chohort Study. *Archives of Gynecology and Obstetrics (ARCH),300/6:1679-1686,2019*

Heger P., **Feißt M.**, Krisam J., Klose C., Dörr-Harim C., Tenckhoff S., Büchler M. W., Diener M. K., Mihaljevic A. L. (2019): Hernia reduction following laparotomy using small stitch abdominal wall closure with and without mesh augmentation (the HULC trial): study protocol for a randomized controlled trial. *Trials,20/1:738,2019*

Glaeser A., Sinn H.-P., Garcia-Etienne C., Riedel F., Hug S., Schaefgen B., Golatta M., Hennigs A., **Feisst M.**, Sohn C., Heil J. (2019): Heterogeneous Responses of Axillary Lymph Node Metastases to Neoadjuvant Chemotherapy are Common and Depend on Breast Cancer Subtype. *Annals of Surgical Oncology,26/13:4381-4389,2019*

Stefanovic S., Deutsch T. M., Riethdorf S., Fischer C., Hartkopf A., Sinn P., **Feisst M.**, Pantel K., Golatta M., Brucker S. Y. Sütterlin M., Schneeweiss A., Wallwiener M. (2020):The Lack of Evidence for an Association between Cancer Biomarker Conversion Patterns and CTC-Status in Patients with Metatstasic Breast Cancer *International Journal of Molecular Science,21/6:2161,2020*

Janke F., Bozorgmehr F., Wrenger S., Dietz S., Heussel C. P., Heussel G., Silva C. F., Rheinheimer S., **Feisst M.**, Thomas M., Golpon H., Günther A., Sültmann H., Muley T., Janciauskiene S., Meister M., Schneider M. A. (2020): Novel Liquid Biomarker Panels for A Very Early Response Capturing of NSCLC Therapies in Advanced Stages. *Cancers,12/4:954,2020*

Schwab M., Brindl N., Studier-Fischer A., Tu T., Gsenger J., Pilgrim M., Friedrich M., Frey P.-E., Achilles C., Leuck A., Bürgel T., **Feisst M.**, Klose C., Tenckhoff S., Dörr-Harim C., Mihaljevic A. L. (2020): Postoperative Complications and Mobilisation Following Major Abdominal Surgery With vs. Without Fitness Tracker-Based Feedback (EXPELLIARMUS): Study Protocol for a Student-Led Multicentre Randomised Controlled Trial (CHIR-Net SIGMA Study Group). *Trials,21/1:293,2020*

Pfob A., Koelbel V., Schuetz F., **Feißt M.**, Blumenstein M., Hennings A., Golatta M., Heil J. (2020): Surgeon's preference of subcutaneous tissue resection: most important factor for short-term complications in subcutaneous implant placement after mastectomy-results of a cohort study. *Archives of Gynecology and Obstetrics,301/4:1037-1045,2020*

Deutsch T. M., Stefanovic S., **Feisst M.**, Fischer C., Riedel F., Fremd C., Domschke C., Pantel K., Hartkopf A. D., Sutterlin M., Brucker S. Y., Schneeweiss A., Wallwiener M. (2020): Cut-Off Analysis of CTC Change under Systemic Therapy for Defining Early Therapy Response in Metastatic Breast Cancer. *Cancers, 12/4:1055,2020*

Haßdenteufel, K., **Feißt M.**, Brusniak K., Lingenfelder K., Matthies L. M., Wallwiener M., Wallwiener S. (2020): Reduction in physical activity significantly increases depression and anxiety in the perinatal period: a longitudinal study based on a self-report digital assessment tool. *Archives of Gynecology and Obstetrics,302/1:53-64,2020*

Deutsch T. M., Riethdorf S., Fremd C., **Feisst M.**, Nees J., Fischer C., Hartkopf A. D., Pantel K., Trumpp A., Schütz F., Schneeweis A., Wallwiener M. (2020): HER2-targeted Therapy Influences CTC Status in Metastatic Breast Cancer. *Breast Cancer Research and Treatmant,182/1:127-136,2020*

Sulieman I., Strobel O., Scharenberg C., Mihaljevic A. M., Müller B. M., Diener M. K., Mehrabi A., Schneider M., Berchtold C., Tjaden C., Hinz U., **Feisst M.**, Büchler M. W., Hackert T., Loos M. (2020): Symptomatic marginal ulcer after pancreatoduodenectomy. *Surgery,168/1:67-71,2020*

Czerny M., Siepe M., Beyersdorf F., **Feisst M.**, Gabel M., Pilz M., Pöling J., Dohle D.-S., Sarvanakis K., Luehr M., Hagl C., Rawa A., Schneider W., Detter C., Holubec T., Borger M., Böning A., Rylski B. (2020): Prediction of mortality rate in acute type A dissection: the German Registry for Acute Type A Aortic Dissection score. *European Journal of Cardio-Thoracic Surgery,58/4:700-706,2020*

Lossnitzer N., **Feisst M.** *(shared co-autorship)*, Wild B., Katus H. A., Schultz J.-H., Frankenstein L., Stock C.(2020): Cross-lagged analyses of the bidirectional relationship between depression and markers of chronic heart failure Short running title: Depression and chronic heart failure. *Depression and Anxiety,37/9:898-907,2020*

Brusniak K., Arndt H. M., **Feisst M.**, Haßdenteufel K., Matthies L. M., Deutsch T. M., Hudalla H., Abele H., Wallwiener M., Wallwiener S. (2020): Challenges in Acceptance and Compliance in Digital Health Assessments During Pregnancy: Prospective Cohort Study. *JMIR Mhealth Uhealth,8/10:e17377,2020*

Riedel F., Heil J., **Feisst M.**, Moderow M., von Au A., Domschke C., Michel L., Schaefgen B., Golatta M., Hennings A. (2020): Analyzing non-sentinel axillary metastases in patients with T3-T4 cN0 early breast cancer and tumor-involved sentinel lymph nodes undergoing breast-conserving therapy or mastectomy. *Breast Cancer Research and Treatment,184/2:627-636,2020*

Rädeker L., Schwab M., Frey P. E., Friedrich M., Sliwinski S., Steinle J., Fink C. A., Leuk A., Ganschow P., Ottawa G. B., Klose C., **Feißt M.**, Dörr-Harim C., Tenckhoff S., Mihaljevic A. L. (2020): Design and Evaluation of a Clinical Investigator Training for Student-lead Prospective Multicentre Clinical Trials: a CHIR-Net SIGMA Research-based Learning Project. *Zentralblatt für Chirurgie,145/6:521-530,2020*

Loos M., Strobel O., Dietrich M., Mehrabi A., Ramouz A., Al-Saeedi M., Müller-Stich B. P., Diener M. K., Schneider M., Berchtold C., **Feisst M.**, Hinz U., Mayer P., Giannakis A., Schneider D., Weigand M. A., Büchler M. W., Hackert T. (2021): Hyperamylasemia and acute pancreatitis after pancreatoduodenectomy: Two different entities. *Surgery,169/2:369-376,2021*

Conference contributions

Feißt M. & Kieser, M. (2018). Incorporating historical data in two-armed clinical trials with binary outcome. *64. Biometrisches Kolloquium*, 25.-28.03.2018, Frankfurt, Germany.

Feißt M. & Kieser, M. (2018). Incorporating historical data in two-armed clinical trials with binary outcome. *Herbstworkshop der Arbeitsgruppen Statistische Methoden in der Medizin (IBS-DR), Statistische Methoden in der Epidemiologie (IBS-DR, DGEpi), Statistische Methoden in der klinischen Forschung (GMDS) und Epidemiologische Methoden (DGEpi, GMDS, DGSMP)*, 22.-23.11.2018, München, Germany.

Feißt M. & Kieser, M. (2019). Incorporating historical data in two-armed clinical trials with binary outcome. *5. Konferenz der Deutschen Arbeitsgemeinschaft Statistik (DAGStat)*, 18.-22.03.2019, München, Germany.

Acknowledgement

First of all, I would like to thank my supervisor Prof. Dr. Meinhard Kieser for giving me the opportunity to write this thesis at the Institute of Medical Biometry and Informatics. I am very grateful for his guidance, constructive suggestions and continued support that contributed greatly to the development of this work, and moreover, to the maturing of my professional profile.

Furthermore, I would like to thank my colleague Johannes Krisam. The joint in-depth discussions and consultations have contributed significantly to the success of this work.

In addition, I would like to thank all other people who accompanied and supported me directly or indirectly in the last years: I thank my colleagues at the Institute of Medical Biometry and Informatics for creating such a friendly and inspiring atmosphere and for their constant support and encouragement.

A final and special thank you goes to my wife Kathrin, who has always encouraged me to persevere and continue even in difficult times.

EIDESSTATTLICHE VERSICHERUNG

1. Bei der eingereichten Dissertation zu dem Thema "Incorporation of historical two-arm data in clincal trials with binary outcome" handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Heidelberg, den 07.04.2021

_____
(Manuel Feißt)