Incorporating historical two-arm data in clinical trials with binary outcome: A practical approach

Manuel Feißt | Johannes Krisam | Meinhard Kieser

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Correspondence
Manuel Feißt, Institute of Medical Biometry and Informatics, Im Neuenheimer Feld 130.3, D-69120 Heidelberg, Germany.
Email: feisst@imbi.uni-heidelberg.de

SUMMARY
The feasibility of a new clinical trial may be increased by incorporating historical data of previous trials. In the particular case where only data from a single historical trial are available, there exists no clear recommendation in the literature regarding the most favorable approach. A main problem of the incorporation of historical data is the possible inflation of the type I error rate. A way to control this type of error is the so-called power prior approach. This Bayesian method does not “borrow” the full historical information but uses a parameter $0 \leq \delta \leq 1$ to determine the amount of borrowed data. Based on the methodology of the power prior, we propose a frequentist framework that allows incorporation of historical data from both arms of two-armed trials with binary outcome, while simultaneously controlling the type I error rate. It is shown that for any specific trial scenario a value $\delta > 0$ can be determined such that the type I error rate falls below the prespecified significance level. The magnitude of this value of $\delta$ depends on the characteristics of the data observed in the historical trial. Conditionally on these characteristics, an increase in power as compared to a trial without borrowing may result. Similarly, we propose methods how the required sample size can be reduced. The results are discussed and compared to those obtained in a Bayesian framework. Application is illustrated by a clinical trial example.

KEYWORDS
clinical trials, historical data, power prior, sample size determination, type I error rate

1 | INTRODUCTION

Clinical trials often examine the efficacy of a new experimental treatment vs placebo or, if available, the current gold standard therapy. 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 patients in order to achieve a sufficiently high probability for detecting a positive treatment effect, that is, a high statistical power. 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, the 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.1-7 Viele et al gave an excellent overview over existing methods.1 In comparison to the exclusive use of historical control data, there is only sparse literature on the integration of historical data from both arms, that is, “borrowing” information from both intervention and control arm.8,9 However, there are situations where incorporation of historical data from two treatment arms may be useful, for example, 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.

Due to the potential heterogeneity between the results from the historical data and the new study population, the integration of historical data might result in a type I error rate inflation. However, control of this quantity is particularly required in the regulatory context. 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 can be handled in a straightforward way.10 However, the optimal determination of the scaling factor is currently still under discussion.11

In this article, 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 binary endpoints, the required 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. Therefore, the appropriate handling of this nuisance parameter is of particular importance.

The outline of this article is as follows. At first, the general framework and notation is presented in section 2.1. Afterward, we introduce the power prior framework in the case of borrowing historical data from both arms of two-armed clinical trials with binary outcome and explore its relationship to the frequentist analysis of the fourfold table in section 2.2. The problem of type I error rate inflation is addressed in section 2.3. In section 2.4, we introduce our proposed approach for the determination of the power prior factor. Section 2.5 describes the potential gain in power, while section 2.6 extends this approach into a sample size calculation procedure. Furthermore, a related Bayesian framework is presented in section 2.7. The proposed sample size calculation procedure is examined by various scenarios in section 3 and is illustrated in more detail by means of a clinical trial example in section 4. Finally, the proposed methods and the corresponding results are discussed in section 5.

2 | METHODS

2.1 | Notations and framework

We suppose a two-armed clinical trial with binary outcome, where higher rates indicate a preferable outcome, for example, response to treatment. Our primary analysis is a two-sided test at a significance level \( \alpha \) assessing the test problem

\[
H_0 : \pi_C = \pi_T \text{ vs } H_1 : \pi_C \neq \pi_T,
\]

where \( \pi_C \) denotes the true response rate for the control arm and \( \pi_T \) is the true response rate for the treatment arm. Furthermore, we assume the existence of historical data from a single historical two-armed trial with binary data in the following form: \( n_{CH}, c_H, \) and \( n_{CH} - c_H \) denote the number of patients, responders, and nonresponders in the historical control arm, respectively. Similarly, \( n_{TH}, t_H, \) and \( n_{TH} - t_H \) denote the number of patients, responders, and nonresponders in the historical treatment arm, respectively. Analogously, the data of the new clinical trial are denoted by \( n_C, c, \) and \( n_C - c \) as number of patients, responders and nonresponders in the control arm of the new trial and, similarly, \( n_T, t, \) and \( n_T - t \) as number of patients, responders, and nonresponders in the new treatment arm.

As in Reference 1, we limit our investigations to the case where the historical data is fixed and, therefore, the performance characteristics (eg, type I error rate and power) are conditional on fixed historical response rates. This refers to the case in which we want to incorporate the data of an already completed historical trial into a new trial. Note, that the historical data may also be handled as random thus taking into account the uncertainty of historical data.
Furthermore, there also exists an alternative framework, where one plans to use the historical data prior to the conduct of the historical study. This would, for example, be the case in an inferentially 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.¹

2.2 The power prior

Viele et al¹ presented different methods of incorporating historical control data into a new trial ranging from 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). Furthermore, Viele et al¹ considered a meta-analysis approach based on more than one historical study. However, meta-analyses based on very few studies have shown to perform rather poorly¹² and are thus not considered in this article.

One method that presents a compromise between separating and pooling is the so-called power prior approach, a Bayesian approach first introduced by Chen and Ibrahim.¹³ The main idea is to assign a “weight” to the historical data. This weight 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).

In its original form, the power prior updates an initial prior \( f_0 \) of a treatment effect \( \pi \) by the likelihood \( L \) based on the historical data \( x_H \), raised to the power of a weight \( \delta \in [0, 1] \):

\[
f_H(\pi|x_H, \delta) \propto L_H(\pi|x_H)^\delta f_0(\pi),
\]

where \( f_H \) denotes the power prior. The principle of the weighting is straightforward: when \( \delta = 0 \) (ie, separate analysis), the likelihood factor becomes 1, and thus only the initial prior is used. Similarly, when \( \delta = 1 \) (ie, pooling), the likelihood factor is fully used (not downweighted) and therefore becomes the same as the usual Bayesian updating of the initial prior, but now for the power prior as the initial prior.

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 \( L_x \):

\[
f(\pi|x_H, \delta) \propto L(\pi|x)f_H(\pi|x_H, \delta) \propto L(\pi|x)L_H(\pi|x_H)\delta f_0(\pi)
\]

In the case of two-armed 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 (while of course, also odds ratio or risk ratio might be the summary measure of interest in such situations). In a binomial framework, beta distributions are used for a conjugate analysis resulting in a Beta posterior distribution. For Beta \( (\alpha, \beta) \), the determination of the parameters \( c \) and \( \beta \) is in the case of binary data straightforward. For example, if we have \( c \) 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 likelihood Beta\( (\alpha_0 + c, \beta_0 + n_C - c) \). Similarly, if we update the initial prior with the power prior likelihood, we simply downweight the historical data with the factor \( \delta \) to the power prior of Beta\( (\alpha_0 + \delta c_H, \beta_0 + \delta (n_{CH} - c_H)) \), where \( c_H \) denotes the 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.

However, analytical computation of the posterior distribution of \( \Delta = \pi_T - \pi_C \), that is, the difference of two beta-distributions, is not trivial (ie, no conjugated model is available).¹⁴–¹⁸ Trying to achieve a conjugated model, in the two-armed framework there exists a further Bayesian approach based directly on the difference \( \Delta = \pi_T - \pi_C \) as parameter of interest and simultaneously assuming normal distribution for the risk difference. In this framework, the influence of the true control proportion \( \pi_C \) on the type I error and the power (see sections 2.3 and 2.5) is limited, since it can be achieved that the type I error rate is independent of \( \pi_C \). We refer to this framework in more detail in section 2.7.

Interestingly, the analysis of 2 \( \times \) 2 tables together with prior information was already addressed long time ago (1877) approach by Liebermeister.¹⁹ He solved this problem in a kind of Bayesian way (however, the Bayesian framework was not yet developed at that time). The interested reader is referred to the chapter “Bayesian tail probabilities for decision making” (by Leonard Held) in the book of Lesaffre et al.²⁰

Based on the principle of the simple form of the beta distribution, the power prior approach for binary outcomes can be transformed straightforwardly to a frequentist fourfold table with subsequent analysis.²¹ One simply adds the weighted historical data to the respective cell of the fourfold table, see Table 1. If there is initial prior information in the
higher the type I error rate inflation is. Similarly, the smaller the amount of borrowing rate inflation. For
based on the data of the historical and the new study (eg, empirical Bayes approach)11. Nevertheless, all methods have
is incorporated in the new trial,
are the number of patients, responders and nonresponders in the new treatment arm.
value chosen for the weighting parameter
is a topic, which is currently still under discussion. Methods range from
δ ∈[0, 1] determines the amount of historical data that
is incorporated in the new trial, \( n_{\text{CH}} \), \( c_H \), and \( n_H - c_H \) are the number of patients, responders and nonresponders, respectively, in the historical control arm, and \( n_{TH}, t_H, \) and \( n_{TH} - t_H \) are the number of patients, responders and nonresponders in the historical treatment arm.
form of a prior distribution, it can again be similarly added to the respective cell. Note that, in the following, we limit
our investigations to the case where there is no initial prior information, that is, a vague, uninformative prior is used. Our systematic investigations (see section 3) 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. Nevertheless, for a specific clinical trial application, the use of the latter class of tests could be more favorable.22 Note that depending on the chosen test procedure one may have to adjust the cell counts of the fourfold table for noninteger cell counts. The choice of the value chosen for the weighting parameter \( \delta \) is a topic, which is currently still under discussion. Methods range from
determination (eg, by an expert) to treating \( \delta \) as an unknown parameter (eg, fully Bayes approach11) or estimation based on the data of the historical and the new study (eg, empirical Bayes approach11). Nevertheless, all methods have in common that they aim to reduce the occurring type I error inflation, that is, when the type I error inflates over the predetermined significance level \( \alpha \).

### 2.3 Type I error rate inflation

As a result of integrating historical data, an inflation of the type I error rate may occur. In our framework (fixed historical data), the actual type I error rate (which is depending on the value of \( \pi_C \) in the case of incorporating historical data) can be determined by calculating the proportion of fourfold tables that reject the null hypothesis weighted by their probability of occurrence:

\[
P\left(\sum_{c=0}^{n_C} \sum_{t=0}^{n_T} P(c, n_C|\pi_C) P(t, n_T|\pi_T = \pi_C) I(P(c, n_C, t, n_T, \delta, c_H, n_{\text{CH}}, t_H, n_{TH}|\pi_T = \pi_C) < \alpha)\right),
\]

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

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 \( \pi_C \). For two-armed borrowing, the type I error rate is rather independent from the difference between the observed historical and the true rate \( \pi_C \) which is equal to \( \pi_T \) under the null-hypothesis. Figure 1 depicts the actual type I error rate for the scenario where we assume

- 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, and
- a significance level of \( \alpha = 0.05 \),

considering both the approach of control arm borrowing and two-armed 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).

For the control arm borrowing approach (Figure 1 left), we see that the more the observed historical control rate (ie, number of responses divided by number of patients in the control arm) differs from the true control rate \( \pi_C \), the higher the type I error rate inflation is. Similarly, the smaller the amount of borrowing \( \delta \), the smaller is the type I error rate inflation. For \( \delta = 0 \), the type I error rate is about 0.05, which corresponds to the nominal significance level \( \alpha \). For
the two-arm borrowing approach (Figure 1 right), the type I error rate is mainly influenced by $\delta$ (and by the observed historical rate difference, which we can see later on in Figure 2 and section 3) but not by the true control response rate $\pi_C$. For $\delta = 0.2$ and 0.4, the type I error rate is actually completely below the nominal significance level. Thus, 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 (left), the type I error rate for increasing $\delta$ (0-1, 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 \( \pi_C = 0.65 \), and
• a significance level of \( \alpha = .05 \).

It can be observed that for every scenario, the type I error functions (ie, the actual type I error rate depending on the amount \( \delta \) of historical data that is included) are nearly convex and show some values below the significance level. Thus, for every scenario, there exists a \( \delta > 0 \) such that the significance level is controlled at \( \alpha = .05 \). The dots identify \( \delta^* \) the maximum value of \( \delta \) controlling the type I error rate at the minimal significance level of \( \alpha = .05 \). For small observed historical differences, even full borrowing is possible while at the same time controlling the type I error rate. The maximal value of \( \delta \) 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 \( \delta \).

### 2.4 Determination of \( \delta^* \)

If we aim to determine the maximal \( \delta \) which still guarantees type I error rate control (\( \delta^* \)) we also have to take into account the true control response rate \( \pi_C \), which can be regarded as a so-called “nuisance” parameter. Nuisance parameters are parameters that we primarily do not intend to estimate but need to be considered nonetheless.\(^23\) The unknown value of \( \pi_C \) has a rather small impact on the type I error rate. However, type I error rate control needs to be ensured for all possible values of \( \pi_C \) because \( \delta^* \) depends on \( \pi_C \), and since \( \delta^* \) 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 \( \pi_C \). Therefore, \( \delta^* \) can be determined by calculating the maximal \( \delta \) for all values of \( \pi_C \) and then taking the minimum of all maximal \( \delta \) (this approach is valid due to the convexity of the type I error functions, see Figure 2, left6:)

\[
\delta^* = \min_{\pi_C \in [0,1]} \max \{0 < \delta \leq 1 | \alpha(\delta|c_H, n_{CH}, t_H, n_{TH}, n_C, n_{T}, \pi_C) < \alpha \}. \tag{2}
\]

Since the actual type I error rate is sensitive to very small or large values of \( \pi_C \), it is sensible that control of the type I error rate does not need to be ensured for the whole range of \( \pi_C \). One possible way to achieve this is the Berger and Boos procedure.\(^24\) Berger and Boos propose to control the type I error rate only in a \( 1 - \gamma \) confidence interval of a nuisance parameter, here \( \pi_C \), referring to the historical control data. However, to guarantee global control of the significance level by \( \alpha \), we have to locally adjust it to \( \alpha - \gamma \). Then, \( \delta^* \) can be determined by calculating the maximum \( \delta \) in the \( 1 - \gamma \) confidence interval of \( \pi_C \) based on a local significance level of \( \alpha - \gamma \):

\[
\delta^* = \min_{\pi_C \in [a,b]} \max \{0 < \delta \leq 1 | \alpha(\delta|c_H, n_{CH}, t_H, n_{TH}, n_C, n_{T}, \pi_C) < \alpha - \gamma \}, \tag{3}
\]

where \([a, b]\) is the respective \( 1 - \gamma \) confidence interval and \( \gamma \) has to be prespecified. Since the type I error functions for \( \pi_C \) are nearly flat (Figure 1), a quite small value can be chosen for \( \gamma \), 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.\(^25\) Therefore, in our framework, we recommend to determine \( \gamma = 0.0001 \) as well (for details, see Supplementary Appendix A). In the following, we refer to this procedure as the “local approach” and the procedure without restriction to a confidence interval as the “global approach” for the determination of \( \delta^* \).

Note that for the practical calculation, we propose to calculate the maximal \( \delta \) for \( \pi_C \) by the nested intervals procedure (which works due to the convexity of the type I error functions, Figure 2) in order to reduce the computational effort. Thereby, the convexity of the type I error functions in terms of \( \delta \) 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,\(^23\) see Supplementary Appendix B), it is possible to prove the convexity of the type I error functions. The proof can be found in Supplementary Appendix B. Also, we suggest the use of the Clopper-Pearson confidence interval since it always fulfills the coverage criterion and thus guarantees the maintenance of the confidence level.\(^23\)
2.5 | Power calculation

The aim of incorporating historical data is that the use of additional information leads to a benefit, that is, an increase in power over, vice versa, a reduction of the required sample size for the new trial.

At first, we investigate the potential increase in power. Similarly to the case of calculating the type I error rate (1), 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 \( \pi_T - \pi_C > 0 \), which we aim to detect.

Accordingly, using the same fourfold table approach, the power amounts to

\[
\sum_{c=0}^{n_C} \sum_{t=0}^{n_T} P(c, n_C | \pi_C) P(t, n_T | \pi_T) I(P(c, n_C, t, n_T, \delta, c_H, n_{CH}, t_H, n_{TH} | \pi_T = \pi_C) < \alpha)
\]  (4)

where again, \( P(x, n | \pi) \) denotes the probability for \( x \) success in \( n \) trials by a binomial success probability of \( \pi \) 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 (right) shows the power for increasing \( \delta \) (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 above (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 \( \delta^* \), the maximum value of \( \delta \) by controlling the type I error rate at the significance level of \( \alpha = .05 \) (for a fixed \( \pi_C = 0.65 \)). Therefore, we can quantify the “price” that has to be paid when borrowing full information: there is no or merely a slight increase in power for the scenarios where the type I error 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), \( \delta^* \) gets smaller, that is, fewer information can be borrowed and, thus, the gain in power gets 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/treatment group differences (0.05-0.2).

2.6 | Sample size calculation

When planning a new trial, one usually aims to achieve a pre-specified power \( 1 - \beta \), for example, 0.8 or 0.9. As we previously showed how our proposed method can be employed to yield an increased power, it can, vice versa, also be used to reduce the sample size required to achieve a pre-specified 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 section 2.5) whether there is an advantage or not. If in a respective scenario an increase of power can be achieved, there also exist combinations of \( \delta^* \), \( n_C \), and \( n_T \) that fulfill predefined conditions, for example, \( \alpha = .05 \), \( 1 - \beta = 0.8 \) for prespecified response rates \( \pi_C \) and \( \pi_T \). To identify the optimal combination, that is, those resulting in the largest reduction in sample size, the combination with the smallest \( n_C \) and \( n_T \) has to be determined. Finding the solution via a grid approach requires a rather large computational effort, since for every possible combination of \( n_C \), \( n_T \), and \( \delta^* \), the type I error rate and the power have to be calculated and compared. Therefore, we suggest the following algorithm to find this combination in a less time-consuming way, based on a predetermined significance level \( \alpha \), power \( 1 - \beta \), and \( \pi_C \) and \( \pi_T \):

1. Calculate \( n_C \) and \( n_T \) based on the predetermined parameters (standard sample size calculation) and opt for the local or global approach (see above)
2. Calculate \( \delta^*_0 \) as in Equation (2) or (3) (depending on the determined approach) based on \( c_{H}, n_{CH}, t_{H}, n_{TH} \) and \( \pi_C \). By integration of the historical data, we should generally obtain an increase in power. If no increase in power is obtained, then stop, as the integration of historical data does yield any benefit.
3. Decrease \( n_C \) and \( n_T \) until the power still lies above the predefined power level \( 1 - \beta \)
4. Calculate \( \delta^* \) as in Equation (2) or (3) based on \( c_{H}, n_{CH}, t_{H}, n_{TH} \) and \( \pi_C \), and on the new \( n_C \) and \( n_T \) from 3.
5. a. 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 \( n_C \) and \( n_T \) from 5.b. If \( \delta^*_1 \leq \delta^*_0 \), stop. \( n_C, n_T \) (from 4) and \( \delta^*_0 \) is 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 and is due to the fact that for increasing $n_c$ and $n_t$ and fixed $\delta'$, the “weight” of the historical data decreases and, therefore, the type I error rate approaches the nominal significance level (see Supplementary Appendix C).

Note that if in the algorithm $\delta_1^* > \delta_0^*$ (5a) and we therefore have to go back to 3, this is regarded as an additional algorithm step in the following section.

## 2.7 Bayesian framework

In this section, we address the Bayesian framework which we already mentioned in section 2.2. 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 $\Delta$ 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 are given as follows (using the same notation as in section 2.2):

$$f(\Delta| x, x_H, \delta) \propto L(\Delta| x) L_H(\Delta| x_H)^\delta f_0(\Delta)$$

Based on the assumption of normality, we get

$$L(\Delta| x) \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_T} + \frac{p_2(1-p_2)}{n_C}\right)$$

where $p_1 = t/n_T$ and $p_2 = c/n_C$. Similarly:

$$L_H(\Delta| x_H) \sim N\left(p_{1H} - p_{2H}, \frac{p_{1H}(1-p_{1H})}{n_{TH}} + \frac{p_{2H}(1-p_{2H})}{n_{CH}}\right)$$

where $p_{1H} = t_H/n_{TH}$ and $p_{2H} = c_H/n_{CH}$. Since we are working now with continuous distributions (instead of discrete count data, as in the previous sections), we have to change further calculations which are now based on integrals instead of sums. Thus, the type I error for the assessment of the test problem:

$$H_0 : \Delta = 0 \text{ vs } H_1 : \Delta \neq 0,$$

can be calculated as:

$$\text{Type I Error} = \int \phi(\Delta) \left[ I(P(\Delta > 0| n_{CH}, n_{TH}, n_C, n_T) > 0.975) + I(P(\Delta < 0| n_{CH}, n_{TH}, n_C, n_T) > 0.975) \right] d\Delta$$

where $\phi$ denotes the density function of a standard normal distribution and

$$P(\Delta > 0| n_{CH}, n_{TH}, n_C, n_T) = \int_0^\infty n(\Delta, \frac{n_C + n_T}{2n_C n_T}) \phi(\Delta_H, \frac{n_{CH} + n_{TH}}{2n_{CH} n_{TH}}) d\Delta$$

with $\Delta_H = p_{1H} - p_{2H}$ and $n(\mu, \sigma^2)$ denoting the density function of the normal distribution with mean $\mu$ and variance $\sigma^2$. Note that compared to formula (1), in this type I error definition, we have to split the rejection region into two parts.

Additionally, we applied a variance-stabilizing transformation (arcsine transformation) to achieve an expression which is independent of the parameter $\pi_C$.

Similarly, the formula for the calculation of the power can be derived (see section 2.5).

Based on this definition of the type I error rate, $\delta^*$ can be determined (independent on $\pi_C$) by
Sample size calculation can be performed analogously to section 2.6 by implementing a slight modification of the proposed algorithm.

3 | RESULTS

In the previous sections, we saw that an increase in power or, vice versa, a sample size reduction by incorporation of historical data is not possible for all scenarios, since this depends on the observed historical data. To detect factors that favor a successful incorporation (ie, type I error rate control and sample size reduction), we examined various scenarios for both the global and the local approach. Our main outcome of interest was the proportion of sample size saved. Therefore, we investigated situations with the following parameters:

\[ \pi_C = 0.1, 0.2, 0.3, 0.4, \]

\[ \Delta = \pi_T - \pi_C = 0.1, 0.15, 0.2, 0.25, \]

\[ c_H = 20; t_H = 21, 22, \ldots, 50, \]

\[ n_{CH} = n_{TH} = 100 \]

\[ \alpha = .05, 1 - \beta = .8, \text{ and } \gamma = 0.0001. \]

Thus, a total of 480 scenarios (4 different values for \( \pi_C \), 4 different values for \( \Delta \), 30 different values for \( t_H \)) were included in the investigation. Note that due to the symmetry of the binomial distribution, we only investigated response rates smaller than 0.5. Variation of further parameters was not examined due to the following reasons:

- \( c_H \): the difference between \( c_H \) and \( \pi_C \) is reflected by the variation of \( \pi_C \).
- \( n_C, n_T \): the sample size is calculated in the initial step of the algorithm.
- \( n_{CH}, n_{TH} \): the amount of historical data that is borrowed remains always the same; for example, if \( n_{CH} = n_{TH} = 100 \) and \( \delta^* = 0.25 \) then, if \( n_{CH} = n_{TH} = 50 \), we get \( \delta^* = 0.5 \).
- \( n_{CH} \neq n_{TH} \) 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 the scenarios are unbalanced, the larger the inflation of the type I error rate is (since it becomes more and more similar to the case of one-armed [control arm] borrowing) and thus, less sample size can be reduced. Nevertheless, there are situations in clinical trial practice where unbalanced designs are used or unbalanced data of historical trials are received, which our approach is also able to deal with (see section 4).
- \( \gamma \): as mentioned earlier, we identified \( \gamma = 0.0001 \) is an appropriate value for our framework.
- \( \alpha \) and \( 1 - \beta \): our proposed algorithms work in the same way for (realistic) values other than \( \alpha = .05 \) and \( 1 - \beta = .8 \) with very similar results.

The results are depicted in Figure 3. Up to 22% of the sample size could be saved by integration of historical data. However, in 10.5% 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 11% (top left). In more than 75% of the scenarios, the local approach lead to an equal or even further reduced sample size, while in more than 20% of the scenarios, the global approach yielded a larger sample size reduction than the global approach (top right). When the observed historical control rate differs from the true value \( \pi_C \), we saw a slightly smaller amount of reduced sample size (middle left). Regarding computational effort, in some cases, the algorithm needed up to five steps (see section 2.6) to converge. In most cases, however, only one or two steps were required (middle right). The largest benefit was found for moderate
FIGURE 3 Results of the sample size calculation procedure for various scenarios (see section 3). Top left: boxplots of the proportion of saved samples size for the global and local approach, respectively. Top right: boxplot of the difference in the proportion of saved sample size between the global and the local approach. Middle left: proportion of saved sample size depending on the difference between observed historical and true control rate. Middle right: relative frequencies of the number of steps until convergence of the algorithm described in section 2.6. Bottom left: mean (± SD) proportion of saved sample size depending on the observed historical rate difference. Bottom right: mean (±SD) δ*(the maximum value of δ controlling the type I error rate) depending on the observed historical rate difference.
(0.5-0.15) observed historical rate differences (bottom left) as $\delta^*$ decreased for larger differences (bottom right, cf. Figure 2). It should be noted that the results in proportion sample size saved for varying $\pi_C$ and $\Delta$ did not much differ between the respective values, which is illustrated in Supplementary Appendix D.

At the end of this chapter, we compare the results from this section to the respective results obtained with the above described Bayesian approach based on the difference parameter $\Delta$. In total, it can be stated that for most scenarios, the results behave very similar compared to the local and global approach. Figure 4 (left) depicts the proportion of saved sample size compared between the Bayes approach and the local procedure. Thereby, the same scenarios are considered. It can be seen that the proportion of sample size saved is slightly higher in the Bayesian approach (up to 28% of sample size can be saved). Figure 4 (middle) confirms this fact since in most of the scenarios the Bayesian procedure shows slightly better results in terms of saved sample size than the local approach. Figure 4 (right) shows that the for the majority of scenarios a similar value of $\delta^*$ results. However, there are also few scenarios, which show a large difference in these values.

4 | CLINICAL TRIAL EXAMPLE

To demonstrate our proposed methods for the incorporation of historical data in two-armed trials with binary outcome, we consider a clinical trial example, the “Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis” (FaSScinate) trial. In this trial, the efficacy and safety of tocilizumab in patients with systemic sclerosis (SSc), a rare connective tissue disorder, were 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

---

**Figure 4** Top left: boxplots of the proportion of saved sample size for the Bayesian and the local approach, respectively. Top right: scatterplot of the proportion of saved sample size for the Bayesian and the local approach for each scenario, respectively. Bottom left: boxplot of the difference in the proportion of saved sample size between the Bayesian and the local approach. Bottom right: boxplot of the difference in $\delta^*$ between the Bayesian and the local approach.
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 (0.227) and 16 responders within 43 tocilizumab patients (0.372) were observed.

Let us assume that we plan a subsequent new trial in SSC investigating the former secondary endpoint as the new primary endpoint. With our developed framework, we can integrate the (historical) FaSScinate study data into our new trial in order to potentially achieve a gain in power or, vice versa, a reduced sample size.

First, we are interested in the gain of power while simultaneously controlling the type I error rate. Therefore, based on the results of the FaSScinate trial, we assume the following parameters for the sample size calculation: \( \pi_C = 0.23 \) and \( \pi_T - \pi_C = 0.14 \). To achieve a power of \( 1 - \beta = 0.8 \) with a two-sided significance level of \( \alpha = 0.05 \), a sample size of \( n_C = n_T = 167 \) is needed in a trial without borrowing. The historical data observed in the FaSScinate are \( c_H = 10, t_H = 16, n_{CH} = 44, \) and \( n_{TH} = 43 \). We choose \( \gamma = 0.0001 \) and thus, the respective 0.9999 confidence interval (Clopper-Pearson, based on the observed historical control rate) for \( \pi_C \) is \([0.050, 0.526]\).

The results for \( \delta^* \) based on Equation (2) and (3) are shown in Table 2. The value for \( \delta^* \) is 0.35 for local and 0.37 for global type I error control, respectively. Since the global (minimum) \( \delta^* \) was found to lie in the confidence interval (at \( \pi_C = 0.39 \)), the global approach leads to a higher value \( \delta^* \) and therefore, the maximum gain in power is achieved for the global approach. Nearly for every \( \pi_C \) a gain in power can be observed; only for the unrealistic values near \( \pi_C = 0.8 \) (for a observed historical rate of 0.23) a decrease in power occurs. Therefore, since the local approach reduces the range of \( \pi_C \) on the more stable and realistic scenarios in the confidence interval, there are better results for minimum and mean gain in power for in this approach. Furthermore, the local approach guarantees a positive minimum gain in power (0.01). Note that these considerations do not apply to the Bayesian approach, since it is independent of \( \pi_C \).

Considering the rejection regions for this clinical trial example may help to illustrate the idea and strategy of our proposed procedure. Therefore, we construct the rejection region based on the same control proportion in the observed data (ie, 0.23; we fix the control proportion due to its impact on the rejection region). Rejection of the null-hypothesis involves the test procedure, but in a right side rejection region that can be classified to lie between the rejection regions are illustrated in Figure 5.

The results for \( \delta^* \) based on Equation (2) and (3) are shown in Table 2. The value for \( \delta^* \) is 0.35 for local and 0.37 for global type I error control, respectively. Since the global (minimum) \( \delta^* \) was found to lie in the confidence interval (at \( \pi_C = 0.39 \)), the global approach leads to a higher value \( \delta^* \) and therefore, the maximum gain in power is achieved for the global approach. Nearly for every \( \pi_C \) a gain in power can be observed; only for the unrealistic values near \( \pi_C = 0.8 \) (for an observed historical rate of 0.23) a decrease in power occurs. Therefore, since the local approach reduces the range of \( \pi_C \) on the more stable and realistic scenarios in the confidence interval, there are better results for minimum and mean gain in power for in this approach. Furthermore, the local approach guarantees a positive minimum gain in power (0.01). Note that these considerations do not apply to the Bayesian approach, since it is independent of \( \pi_C \).

Considering the rejection regions for this clinical trial example may help to illustrate the idea and strategy of our proposed procedure. Therefore, we construct the rejection region based on the same control proportion in the observed data (ie, 0.23; we fix the control proportion due to its impact on the rejection region). Rejection of the null-hypothesis with a test without historical data in such a situation would occur if 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 \( n = 38 \)). With historical data based on an optimal \( \delta^* \) (global approach) of 0.37, the rejection region changes to \( R = [0;21] \cup [54;167] \), and based on a \( \delta^* \) 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, we also compare these rejection regions of a two-sided test with \( \alpha = 1 \) without any historical data, resulting in a rejection region \( R = [0;26] \cup [52;167] \). Thus, our procedure results in an even smaller left side rejection region than the \( \alpha = 0.05 \) test procedure, but in a right side rejection region that can be classified to lie between the \( \alpha = 0.05 \) and \( \alpha = 1 \) test procedure. The rejection regions are illustrated in Figure 5.

As we will explain in further detail in section 5, the inclusion of historical data (by increasing delta) increases the power in favor of an effect observed in the historical data and thus expands the rejection region for an effect in favor of the observed effect in the historical but shrinks the rejection region for an effect in the opposite direction.

Next, we consider the benefit in reduced sample size that can be achieved by incorporating the historical information of the FaSScinate trial into a new trial. With the same parameter values as above (\( \pi_C = 0.23, \pi_T - \pi_C = 0.14, 1 - \beta = 0.8, \alpha = 0.05, c_H = 10, t_H = 16, n_{CH} = 44, n_{TH} = 43, \gamma = 0.001 \)), the sample size can be reduced by 24 patients.

| Parameter                  | Local approach | Global approach | Bayesian approach |
|----------------------------|----------------|-----------------|------------------|
| Minimum gain in power      | 0.01 (at \( \pi_C = 0.05 \)) | -0.002 (at \( \pi_C = 0.82 \)) | Not applicable |
| Maximum gain in power      | 0.058 (at \( \pi_C = 0.33 \)) | 0.060 (at \( \pi_C = 0.31 \)) | Not applicable |
| Mean gain in power         | 0.039           | 0.031           | Not applicable   |
| Power at \( \pi_C = 0.25 \)| 0.848           | 0.851           | 0.861            |

Note: \( \delta^* \) determines the maximum amount of incorporated historical data that still guarantees type I error rate control and \( \pi_C \) denotes the true control rate.
with the local approach, by 26 patients (15.6%) with the global approach and by 28 patients (16.8%) using the Bayesian approach, see Table 3. Thus, in this specific scenario, the Bayesian approach leads to a higher benefit as compared to the global and local approach.

## 5 | DISCUSSION

In this article, we presented a framework which allows to integrate historical two-arm data of a previous trial in the planning and analysis of a two-armed clinical trial with binary outcome while simultaneously controlling the type I error rate. Our approach is based on the Bayesian power prior approach that can be transferred straightforwardly into a frequentist fourfold table approach with analysis by the common chi-square test. Thereby, the amount of historical data is simply controlled by a factor $\delta$ that ranges from 0 (no borrowing) to 1 (full borrowing). The maximum amount of borrowing which still ensures control of the type I error rate depends substantially on the characteristics of the historical data. In our examples, we showed that in the frequentist framework of the global and local approach 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 (eg, an observed historical rate differences $\leq 0.02$, see Figure 3E).

Furthermore, we presented a Bayesian approach, which performs slightly better for most scenarios in terms of increased power or, vice versa, in terms of saved sample size. This is due to the fact that the Bayesian approach is independent of the nuisance parameter $\pi_C$ and does not to have to deal with the discreteness of the chi-square distribution. However, the concepts of type I error rate and power were initially created regarding for frequentist testing and then translated to the Bayesian setting and their correct interpretation is still under discussion in the literature. Therefore, we decided to work mainly on the fourfold table approach. Nevertheless, the presented Bayesian setting provides a direct way into an, in our opinion, mathematically elegant setting and shows benefit in terms of increased power and saved sample size compared to the frequentist local and global approaches. In addition, in this setting, it is easily possible to prove of the convexity of the type I error function for increasing $\delta$ (see section 2.4). We give the proof in Supplementary Appendix B which only approximately fits for the approaches based on the chi-square test statistics. Furthermore, due to the independence of $\pi_C$, the calculations based on the Bayesian approach are substantially faster, thus making this approach more appealing compared to the local and global approach for the use in simulation studies. Further discussion of the Bayesian inference in binary data can be found in the study by Lesaffre et al.20

In a recently published paper, Kopp-Schneider et al28 stated that “strict control of type I error implies that no power gain is possible under any mechanism of incorporation of prior information, including dynamic borrowing.” Thus, the question arises, how this fits with our results. To clarify this fact, we try to give an intuitive and illustrated access to our approach:

### FIGURE 5
Rejection regions for different test procedures with fixed number of control responses $n = 38 (\pi_C = 0.23)$

### TABLE 3
Results of the sample size calculation procedure for the FaSScinate trial for local and global approach

|                      | Local approach | Global approach | Bayesian approach |
|----------------------|----------------|-----------------|------------------|
| Initial sample size  | 167            | 167             | 167              |
| New sample size      | 143            | 141             | 139              |
| Saved sample size (%)| 24 (14.4%)     | 26 (15.6%)      | 28 (16.8%)       |
| $\delta$             | 0.37           | 0.44            | 0.46             |
| Steps                | 1              | 2               | 2                |

Note: $\delta$ determines the maximum amount of incorporated historical data that still guarantees type I error rate control.
To achieve an increase in power while simultaneously controlling the type I error rate in our approach, it is crucial that the type I error rate function (Figure 2 left) is partly below the prespecified significance level \( \alpha \). To add some intuition to this fact, one can consider the distribution of a standard normal test statistic (which is closely related to the chi-square test statistic on which our calculations are based). 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 \(-\infty\) to the \(a/2\)-quantile and the integral from the \(1-a/2\) to \(\infty\). For an increasing amount of historical data incorporated (controlled by \(\delta\)), 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 \(\delta\), 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.

Based on this illustration, we can see 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 predetermined significance level and it would not be possible to control the type I error rate by the predetermined significance level when increasing \(\delta\). Incorporation of historical data would then always result in a type I error inflation. It should, however, be noted that our proposed test procedure could be easily adapted to a one-sided setting under the specification of an upper limit for the one-sided significance level, for example, \(2\alpha\).

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 5). Thus, in summary, our results are in line with the results of Kopp-Schneider et al\(^{28}\) who argue that “borrowing of information cannot lead to an increased power while strictly controlling type I error.” Our proposed procedure achieves this goal 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 at the prespecified significance level \(\alpha\). 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 in our proposed method translated into a frequentist setting while at the same time assuring control of the type I error rate.

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 for binary outcome. Here, the power to reveal a given effect size is depending on the true control proportion and thus, there is no uniform power for revealing an effect into both directions of the alternative hypothesis. Furthermore, this shift of the rejection regions embodies and translates the idea of the Bayesian learning into the frequentist framework, as we already stated in section 2.2. Based on these perspectives, our proposed procedure delivers on its promises, that is, an increase in power (for true treatment effects favoring the treatment group over the control group) by simultaneously controlling the type I error rate.

Since the calculation of \(\delta\) depends on the true control rate \(\pi_C\) and \(\delta\) mainly influences the inflation of the type I error rate, \(\pi_C\) has to be handled as a nuisance parameter in the frequentist framework and thus one has to guarantee control of the type I error over the whole range of \(\pi_C\). To handle this problem, we suggest two approaches to determine the maximum value of \(\delta\) controlling the type I error rate (\(\delta^*\)). We denote them as the local approach and the global approach, respectively, where the type I error rate is either controlled by a slightly more stricter significance level \(\alpha - \gamma\) for values of \(\pi_C\) in a \(1 - \gamma\) confidence interval (local approach) or by the nominal significance level \(\alpha\) for all values of \(\pi_C\) (global approach). In the scenarios, we considered, the local approach leads for more situations to a bigger benefit, but there are also scenarios where the global approach is advantageous. If the value of \(\delta^*\) 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 \(\pi_C\) while simultaneously reducing the local significance level. However, if one is mainly interested in ensuring a substantial gain in power, it is usually better to reduce the range of \(\pi_C\) to the more realistic and therefore more beneficial values in the confidence interval of the local approach. Nevertheless, this issue is negligible regarding the sample size calculation since it is only based on one particular value of \(\pi_C\). Since the determination of \(\delta^*\) requires a large computational effort, especially for sample size calculation, we suggested several practical recommendations. In detail, we developed an algorithm to find the most beneficial (in terms
of saved sample size) combination of \( n_C, n_T \) (sample sizes of the new trial) and \( \delta^* \), which usually converges in only one or two steps. Nevertheless, the approach remains 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 exponentially with increasing \( n \) (the computation time ranged from about 20 minutes in a scenario where \( n_C = n_T = 50 \) to about 8 hours in a scenario of \( n_C = n_T = 400 \)). Therefore, for very large sample sizes (\( n > 300 \)) we recommend to determine by simulations type I error rate and power. We gladly share the R-code for the presented sample size calculation procedure, which is available on GitHub (https://github.com/manuelfeisst/manuel). Note that due to practical reasons (the function chisq.test from R produces NA for zeros in the fourfold tables), the sums in the code which are based on Equations (1) and (4) 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 paper (the differences in power and type I error are smaller than 0.0000001 for all scenarios).

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 (ie, 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 (ie, 0.05-0.2) historical rate differences. At first sight, it seems arguable that our proposed method penalizes 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. 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, our proposed method results 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 is certainly 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 our proposed approach.

Since the extent of the type I error inflation mainly depends on the historical difference, this causes a problem in the case of two-armed 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. Thus, for two-armed 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, we highly recommend to verify the choice of a historical trial by Pocock's criteria\(^29\) (see chapter “Acceptable Historical Control” of his 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.”

by extending them to the case of two-armed clinical trials. By extending this claim and the other criteria to the case of two-armed borrowing, they should be verified equally for both arms of the studies.

In comparison, other adaptive weighting approaches considered in the literature\(^1^1\) do not work directly with the aim to control the type I error at a prespecified significance level but merely assess the agreement between the current and the historical data. Contrarily, in our procedure, the observed data of the new trial do not play a role, but only the data of the historical trial does. Most of the discussed procedures can be extended to the case of two-armed borrowing...
of historical data. However, the resulting calculations may differ considerably from each other. It should be noted that the empirical Bayes estimator of Gravestock and Held\(^ {11} \) shows very high similarity to our proposed estimator regarding the value of \( \delta^* \) in relation to the observed historical effect size (see Figure 3 bottom right).

With our work, we propose a framework to integrate existing data on control and experimental treatment gained in a previously conducted clinical trial for the planning and analysis of a subsequent two-armed trial with binary outcome. By doing so, a gain in power can be achieved or the required sample size can be reduced and thus, resources can be saved. In future work, we plan to extend the framework to other outcomes, for example, continuous or survival endpoints. We hope that our work supports the further streamlining of the development of drugs and medical devices, especially in the field of rare diseases, and that it will find its way into clinical research.

ACKNOWLEDGMENT
The authors thank the reviewers for their very helpful comments, detailed evaluation and constructive criticism of our work. Their extensive and comprehensive recommendations helped us to substantially improve the manuscript. Especially, the authors would like to thank Professor Leonard Held whose challenging and stimulating comments were extremely valuable for us.

DATA AVAILABILITY STATEMENT
We gladly share the R-code for the presented sample size calculation procedure and the scenario data generated in section 3, which is available on GitHub (https://github.com/manuelfeisst/manuel).

ORCID
Manuel Feißt https://orcid.org/0000-0002-5016-9731
Meinhard Kieser https://orcid.org/0000-0003-2402-4333

REFERENCES
1. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13(1):41-54.
2. Chen MH, Ibrahim JG, Lam P, Yu A, Zhang Y. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics*. 2011;67(3):1163-1170.
3. Chen MH, Ibrahim JG, Shao QM. Power prior distributions for generalized linear models. *J Stat Plan Infer*. 2000;84(1–2):121-137.
4. Duan YY, Smith EP, Ye KY. Using power priors to improve the binomial test of water quality. *J Agricul Biol Environ Stat*. 2006;11(2):151-168.
5. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67(3):1047-1056.
6. Rietbergen C, Klugkist I, Janssen KJ, Moons KG, Hoijtink HJ. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemp Clin Trials*. 2011;32(6):848-855.
7. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70(4):1023-1032.
8. Gamalo-Siebers M, Savic J, Basu C, et al. Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharm Stat*. 2017;16(4):232-249.
9. Weber S, Gelman A, Lee D, Betancourt M, Vehtari A, Racine-Poon A. Bayesian aggregation of average data: an application in drug development. *Ann Appl Stat*. 2018;12(3):1583-1604.
10. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724-3749.
11. Gravestock I, Held L, C.O.-N. consortium. Adaptive power priors with empirical Bayes for clinical trials. *Pharm Stat*. 2017;16(5):349-360.
12. Seide SE, Röver C, Friede T. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC Med Res Methodol*. 2019;19(1):16.
13. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60.
14. Kawasaki Y, Miyaoka E. A Bayesian inference of \( P(\pi_1 > \pi_2) \) for two proportions. *J Biopharm Stat*. 2012;22(3):425-437.
15. Lee PM. *Bayesian Statistics: An Introduction*. Arnold, London: John Wiley; 2012.
16. Howard J. The 2 x 2 table: a discussion from a Bayesian viewpoint. *Stat Sci*. 1998;13(4):351-367.
17. Altham PM. Exact Bayesian analysis of a 2 times 2 contingency table, and Fisher’ exact” significance test. *J R Stat Soc B Methodol*. 1969;31(2):261-269.
18. Nemtin M, Mutanen P. Exact Bayesian analysis of two proportions. *Scand J Stat*. 1987;14(1):67-77.
19. Liebermeister K. *Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik*. Sammlung Klinischer Vorträge (Innere Medicin No. 31–64)110, 935–962 Breitkopf & Härtel; 1877.
20. Lesaffre E, Biao G, Boulanger B. *Bayesian Methods in Pharmaceutical Research*. Boca Raton: CRC Press; 2020.
21. Zaslavsky BG. Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. Biometrics. 2013;69(1):157-163.
22. Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2×2 tables. Stat Med. 2009;28(7):1159-1175.
23. Fagerland MW, Lydersen S, Laake P. Statistical Analysis of Contingency Tables. Boca Raton: CRC Press/Taylor & Francis Group; 2017:634.
24. Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. J Am Stat Assoc. 1994;89(427):1012-1016.
25. Lydersen S, Langaas M, Bakke Ø. The exact unconditional z-pooled test for equality of two binomial probabilities: optimal choice of the Berger and Boos confidence coefficient. J Stat Comput Simul. 2012;82(9):1311-1316.
26. Yu G. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. Stat Probab Lett. 2009;79(14):1621-1629.
27. Khanna D, Denton CP, Jahreis A, et al. Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis (faSScinate): a phase 2, randomised, controlled trial. Lancet. 2016;387(10038):2630-2640.
28. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biom J. 2020;62(2):361-374.
29. Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis. 1976;29(3):175-188.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Feißt M, Krisam J, Kieser M. Incorporating historical two-arm data in clinical trials with binary outcome: A practical approach. Pharmaceutical Statistics. 2020;19:662–678. https://doi.org/10.1002/pst.2023