Improving the efficiency of clinical trial designs by using historical control data or adding a treatment arm to an ongoing trial

Maxine Sarah Bennett

St Catharine's College
University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy.

January, 2018

Improving the efficiency of clinical trial designs by using historical control data or adding a treatment arm to an ongoing trial

Maxine Sarah Bennett

Abstract

The most common type of confirmatory trial is a randomised trial comparing the experimental treatment of interest to a control treatment. Confirmatory trials are expensive and take a lot of time in the planning, set up and recruitment of patients. Efficient methodology in clinical trial design is critical to save both time and money and allow treatments to become available to patients quickly.

Often there are data available on the control treatment from a previous trial. These historical data are often used to design new trials, forming the basis of sample size calculations, but are not used in the analysis of the new trial. Incorporating historical control data into the design and analysis could potentially lead to more efficient trials. When the historical and current control data agree, incorporating historical control data could reduce the number of control patients required in the current trial and therefore the duration of the trial, or increase the precision of parameter estimates. However, when the historical and current data are inconsistent, there is a potential for biased treatment effect estimates, inflated type I error and reduced power.

We propose two novel weights to assess agreement between the current and historical control data: a probability weight based on tail area probabilities; and a weight based on the equivalence of the historical and current control data parameters. For binary outcome data, agreement is assessed using the posterior distributions of the response probability in the historical and current control data. For normally distributed outcome data, agreement is assessed using the marginal posterior distributions of the difference in means and the ratio of the variances of the current and historical control data. We consider an adaptive design with an interim analysis. At the interim, the agreement between the historical and current control data is assessed using the probability or equivalence probability weight approach. The allocation ratio is adapted to randomise fewer patients to control when there is agreement and revert back to a standard trial design when there is disagreement. The final analysis is Bayesian utilising the analysis approach of the power prior with a fixed weight. The operating characteristics of the proposed design are explored and we show how the equivalence bounds can be chosen at the design stage of the current study

to control the maximum inflation in type I error.

We then consider a design where a treatment arm is added to an ongoing clinical trial. For many disease areas, there are often treatments in different stages of the development process. We consider the design of a two-arm parallel group trial where it is planned to add a new treatment arm during the trial. This could potentially save money, patients, time and resources. The addition of a treatment arm creates a multiple comparison problem. Dunnett [39] proposed a design that controls the family-wise error rate when comparing multiple experimental treatments to control and determined the optimal allocation ratio. We have calculated the correlation between test statistics for the method proposed by Dunnett when a treatment arm is added during the trial and only concurrent controls are used for each treatment comparison. We propose an adaptive design where the sample size of all treatment arms are increased to control the family-wise error rate. We explore adapting the allocation ratio once the new treatment arm is added to maximise the overall power of the trial.

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

Signed . . . .
Maxine Sarah Bennett

Acknowledgements

Firstly, I would like to thank Adrian Mander for his supervision, for giving me the opportunity to undertake this project and for the patient guidance, encouragement and advice he has provided throughout my PhD. I would also like to thank him for giving me independence in my PhD to explore my own research interests.

Secondly, I would like to thank Simon White who became my secondary supervisor during the third year of my PhD. His attention to detail improved my understanding of my research area and his support and advice improved the work in this dissertation. I would also like to thank GSK who partly funded my PhD, in particular William Powley, Jeffrey Wetherington and Nicky Best who have provided great advice throughout.

I am grateful to the MRC Biostatistics Unit which is a friendly and supportive environment for conducting research, with people always on hand to offer help and advice.

I would like to give special thanks to my mother and all my family and friends who have always encouraged me in my ambitions and provided support through the ups and downs of my research. Last, but not least, I would like to thank my partner Tom who has provided endless support throughout my PhD and without his help and patience I would not have been able to complete this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Clinical trial design

A confirmatory trial is necessary to provide firm evidence of efficacy or safety of an experimental treatment [1]. The most common type of confirmatory trial is a randomised controlled trial (RCT) comparing the experimental treatment of interest to a standard of care or placebo. Confirmatory trials are expensive and take a lot of time in the planning, set up and recruitment of patients [2]. The failure rate for phase III trials is approximately 50 percent [3] and therefore the big effort in terms of cost, time and resources of running a confirmatory trial may not lead to a beneficial treatment. Efficient methodological design in all phases of clinical trials is critical in saving both time and money, and allowing treatments to become available to patients sooner. Also, regulatory bodies recognise the need for more efficient and innovative trial designs [4], and guidelines have been produced for adaptive clinical trial designs and designs using Bayesian methodology [5, 6].

This thesis considers a Bayesian approach to trial design as a way to design efficient clinical trials and allow exploration of flexible design options. In a Bayesian framework, the design or analysis of the trial can be adapted while the trial is ongoing, different sources of information can be combined or expert opinion can be used to inform both the design and the analysis of a trial. However, regulatory approval of a treatment is based on the frequentist operating characteristics of a trial design, therefore throughout the thesis we aim to find a Bayesian design that has good frequentist operating characteristics. Frequentist operating characteristics, such as the type I and type II error rate, can be determined for any Bayesian design through simulation and can be calculated exactly for some specific Bayesian designs.

Where possible, conjugate Bayesian analysis is used throughout this thesis. A conjugate prior leads to a posterior distribution that is available in closed form and is a member of the same distributional family as the prior. Often a conjugate prior is flexible enough to represent the prior belief about the parameter and simplifies the calculation of the posterior distribution. The availability of an analytical expression for the posterior distribution is especially important in design where simulation is often required to determine the operating characteristics of a design and an analytical expression for the posterior distribution will reduce the computation time.

The Food and Drug Administration (FDA) acknowledge that their guidance for industry document on statistical principles for clinical trials focuses on frequentist methods for the design and analysis of clinical trials but suggest that "the use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust" [1]. Further, the FDA have provided a guidance document for the use of Bayesian statistics in medical device clinical trials [5]. The advantages of the Bayesian approach to trial design suggests that for trials where recruitment is difficult and for certain therapeutic areas, this criterion will be met, even in later phase trials.

In a standard two-arm confirmatory randomised controlled trial comparing an experimental treatment to a standard of care or control treatment, the sample size is fixed at the design stage to achieve the desired error rates for a specified treatment effect. Patients are randomised equally between the treatment groups and the analysis is on a defined endpoint, using only data from patients randomised within the current trial. Often the sample size required is large.

The FDA's guidance for industry document on adaptive design clinical trials for drugs [6] defines an adaptive design clinical study as "A study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study". The motivation for using adaptive designs is to increase the efficiency of a trial compared to using a traditional design approach. One or more of the following gains in efficiency could be achieved depending on the design chosen: reducing the number of patients and therefore the duration of the current trial; increasing the power of the study; or randomising more patients to the better performing treatment.

The use of adaptive designs has increased over the past 10 years [7]. The limitations of adaptive designs seem to mainly lie in logistical issues. In an adaptive design the endpoints used need to be observed relatively quickly to allow the trial to be adapted based on data collected in the current trial. Additional resources are required in an adaptive design compared to a standard trial design in order to plan and carry out interim analyses and the subsequent adaptations. An adaptive design that may lead to a trial ending before the planned duration also has implications on trial staffing.

In this thesis we specifically consider the following Bayesian ideas and adaptive methods: incorporating historical control data into both the design and analysis of a trial; allowing interim looks at the current trial data; incorporating expert opinion into the design of the trial, which will also affect inference from the trial; adding treatment arms to an ongoing trial; and adapting the treatment allocation ratio during the trial. Confirmatory trials will be the main focus of this thesis. Confirmatory trials are designed to have a low type I error rate, typically 2.5% or 5% and a low type II error, typically 10% to 20%. However, the methods considered here could be applicable to early phase trials which are more exploratory and typically allow a higher type I error rate and a similar type II error rate compared to confirmatory trials.

1.2 Historical data methods

Historical data can arise in many situations. There may be outcome information available on both the experimental treatment and the standard of care from earlier phases of treatment development. For the control treatment or standard of care, there are usually data available from a previous trial. An expert's opinion or experience could also be considered historical data, where a standard of care treatment has been used in clinical practice for a long duration. Historical data are often used informally to design new trials. Historical data may be used to choose endpoints and the treatment effect to be detected in the new study. Historical data therefore inform the design of a new study by forming the basis of the sample size calculations but the historical data are not used in the analysis of the new trial.

Designing a trial with the aim to incorporate historical data into the analysis has the potential to increase the efficiency of the trial. Historical data can be used as prior information in a fully Bayesian framework, where the historical data are updated with the current trial likelihood using Bayes theorem [8]. The posterior distribution is then a weighted average of the historical and current data. More recently historical data methods have been proposed that dynamically borrow historical data when the historical data are in agreement with the data observed in the current trial. Throughout this thesis we only consider incorporating historical information on the control treatment into the current trial. When the current and historical control data are in agreement, one or more of the following advantages could be achieved depending on the design chosen: more patients can be randomised to the experimental treatment and fewer to the control treatment; the total sample size of the trial can be reduced; the trial duration can be decreased or the precision of parameter estimates can be increased. However, when the historical and current control data are inconsistent, there is a potential for biased treatment effects, inflated type I error rate and reduced power.

A limitation of incorporating historical control data is that although the current study is a randomised comparison of the experimental treatment to control, the inclusion of historical controls in the final analysis means that the final analysis is not a fully randomised comparison. However, due to the advantages of the Bayesian approach to trial design and utilising prior information, there is an argument for using designs without a fully randomised comparison in certain settings. The benefits of adaptive designs, historical data and non-randomised comparisons are discussed in detail by Berry in a comment paper for Statistical Science [9] where he comments that "good statistical methods for using historical control data seem not to be available". A recent systematic review by Wadsworth et al. provides details of methods relevant for extrapolations which includes many historical data methods [10] and the following sections review some of the methods that have been proposed in the literature for incorporating historical data into the design and analysis of a current trial.

One of the earliest papers to discuss incorporating historical data into both the design and analysis of a new study is a seminal paper by Pocock [11]. This paper considers a design where patients are randomised to both treatment and control in the current study, even when "acceptable" historical control data are available. The historical control data are then incorporated into the final analysis of the current trial.

An important question is which historical data are relevant to the current study, especially when there are a large amount of historical control data available spanning many years and from multiple historical studies. In this thesis we focus on methods for incorporating control data from a single historical study into both the design and analysis of a current trial. However, Pocock's acceptability criteria [11] for choosing relevant historical studies to incorporate into a current trial design and analysis are applicable whether there are multiple historical studies or a single historical study available. These criteria state that the acceptability of a historical control group requires:

- the group must have received a precisely defined standard treatment which must be the same as the treatment for the current trial controls;
- the group must have been part of a recent clinical study which contained the same requirements for patient eligibility;
- the method of treatment evaluation must be the same;
- the distributions of important patient characteristics in the historical group should be comparable with those in the new trial;
- the previous study must have been performed in the same organisation with largely the same clinical investigators; and

- there must be no other indications leading one to expect different results between the randomised and historical controls (e.g. differing enrolment rates).

Some of the aforementioned criteria are strict. To satisfy the criterion that the historical study must have been performed in the same organisation as the current trial is likely to rule out "literature controls" and a large amount of potentially relevant information. Also, requiring the enrolment rates to be the same seems unnecessary unless there are major differences between the trials, which would only be known after the current trial had finished recruiting. However, these criteria provide a good starting point for selecting relevant historical studies. Furthermore, it has been shown by Neuenschwander et al. [12] that selecting a small number of historical studies that are very similar to each other and are thought to be similar to the current study, leads to a larger prior effective sample size than incorporating all historical data where there is lots of heterogeneity between studies. The prior effective sample size when there are multiple historical studies is defined as how many patients the historical studies represent when the historical studies are combined, taking into account the heterogeneity between the historical studies.

In this chapter the historical data methods are described in their general form, the specific details are kept for the main chapters. The main chapters include both previously published and new methods. This structure is chosen since both published and new methods are explored and new methods are often extensions to the published methods. Where a new method is proposed or a published method has been extended should be clear from the text. Historical data methods for binary outcome data are considered in Chapter 2 and methods for normally distributed outcome data are considered in Chapter 3.

To introduce the historical data approaches, we consider a standard trial design comparing one experimental treatment to control, assuming historical data are available for the control arm only. Let D_t, D_c and D_h denote data from the current treatment group, current control group and historical data, respectively. Let θ_c denote the parameter of interest in the controls and θ_t the parameter of interest in the treatment group. Where a method assumes that the true underlying parameters are different in the historical and current controls, θ_h denotes the parameter of interest in the historical controls.

1.2.1 Pooling

The simplest approach to incorporate historical data into the analysis of the current trial is to assume that the historical control data are exchangeable with the control data in the current trial. The historical data forms the prior for the current study control arm which is then updated with the current study control data using Bayes theorem [8]. We assume an initial prior, $\pi_0(\theta_c)$, for the control parameter of interest before the historical data are observed, $\pi_0(\theta_c)$ is updated to form a posterior distribution that incorporates the historical data, which forms the prior for the current study control parameter, given by,

$$\pi(\theta_c \mid D_h) \propto L(\theta_c \mid D_h)\pi_0(\theta_c), \tag{1.1}$$

where $L(\theta_c \mid D_h)$ is the likelihood of the historical data. $\pi(\theta_c \mid D_h)$ is then updated with the control data from the current study using Bayes theorem. The posterior distribution for the control parameter is then,

$$\pi(\theta_c \mid D_c, D_h) \propto L(\theta_c \mid D_c)L(\theta_c \mid D_h)\pi_0(\theta_c) = L(\theta_c \mid D_c, D_h)\pi_0(\theta_c),$$

where $L(\theta_c \mid D_c)$ is the likelihood of the current control data. This is the same as pooling the current and historical control data as if they were from the same study.

1.2.2 Pocock's approach to modelling bias

Pocock [11] proposed an approach to incorporating historical control data that acknowledges that the underlying parameters of interest in the historical and current controls may not be the same, even with careful selection of the historical data. The possibility of unknown bias is modelled as $\eta = \theta_c - \theta_h$ where η is treated as a random variable with mean zero and fixed variance σ^2_η. In practice, it is recommended to conduct a sensitivity analysis considering a range of values for σ^2_η. An initial vague prior is assumed for θ_h before the historical data are observed, denoted $\pi_0(\theta_h)$. This vague prior is updated with the historical data to obtain the posterior distribution, $\pi(\theta_h \mid D_h) = \pi_0(\theta_h)L(\theta_h \mid D_h)$. The sum of the two random variables $\pi(\theta_h \mid D_h)$ and η forms the prior for the control arm in the current study. This approach was extended to give the commensurate prior approach described in Section 1.2.5. A bias variance approach to discounting historical data was also proposed in a recent paper by Galwey [13].

1.2.3 Power prior

The power prior, originally proposed by Ibrahim and Chen [14] assumes that the historical data and current control data are estimating the same underlying parameter of interest, θ_c. An initial non-informative prior $\pi_0(\theta_c)$ is assumed for θ_c before the historical data are observed. This is then updated with the likelihood of the historical data raised to a power α_0, where the power quantifies the uncertainty in the similarity between the historical and current studies. The prior for the current study control arm is then,

$$\pi(\theta_c \mid D_h) \propto \pi_0(\theta_c) L(\theta_c \mid D_h)^{\alpha_0},$$

where α_0 is a fixed value and usually lies between zero and one. When α_0 is zero, no historical data are used in the final analysis and the prior reduces to the initial non-informative prior,

$$\pi(\theta_c \mid D_h) = \pi_0(\theta_c),$$

and when α_0 is one, all of the historical data are used in the final analysis, pooling the current and historical controls, and the prior becomes Equation 1.1. The power can be given a value above one, however, in the area of historical data, we consider the most reliable information to be the data from the current randomised controlled trial and are therefore unlikely to want to give the historical data more weight in the final analysis than the current control data. The power α_0 can be interpreted as a relative precision parameter for the historical data.

1.2.4 Modified power prior

As we are usually unsure about the agreement between the current and historical data, α_0 can be treated as a random variable [14, 15, 16]. Ibrahim and Chen proposed a joint prior for the parameter of interest and the power of the form [14],

$$\pi(\theta_c, \alpha_0 \mid D_h) \propto \pi_0(\theta_c) L(\theta_c \mid D_h)^{\alpha_0} \pi(\alpha_0). \tag{1.2}$$

A natural choice for $\pi(\alpha_0)$ is a beta distribution, given the desirability of a power between zero and one.

A few problems arise with the formulation of the joint prior given in Equation 1.2. Firstly, it violates the likelihood principle [17], since multiplying the likelihood by a constant would change the joint prior and therefore the posterior distribution. Secondly, the historical data has little influence in the analysis even when there is complete agreement between the historical and current data [15].

This joint prior is missing a factor dependent on α_0 and therefore the power parameter always has a tendency to be close to zero, making no use of the historical data. Duan [15] proposed a correction to the original joint power prior formulation, defined as the modified power prior, given by,

$$\pi(\theta_c, \alpha_0 \mid D_h) \propto C(\alpha_0)\pi_0(\theta_c)L(\theta_c \mid D_h)^{\alpha_0}\pi(\alpha_0),$$

where,

$$C(\alpha_0) = \frac{1}{\int_{\theta_c} L(\theta_c \mid D_h)^{\alpha_0}\pi_0(\theta_c)d\theta_c},$$

in the region of α_0 such that $\int_{\theta_c} L(\theta_c \mid D_h)^{\alpha_0}\pi_0(\theta_c)d\theta_c$ is finite. The modified power prior does satisfy the likelihood principle.

1.2.5 Commensurate prior

The commensurate prior approach [18, 19, 20] assumes different underlying parameters for the current and historical controls. The location commensurate prior for θ_c is a conditional prior distribution, centred at the historical parameter θ_h with a fixed value τ that controls the cross-study borrowing. The joint distribution of θ_c and θ_h before the current trial is then given by,

$$\pi(\theta_c, \theta_h \mid D_h, \tau) \propto \pi(\theta_c \mid \theta_h, 1/\tau)\pi_0(\theta_h)L(\theta_h \mid D_h),$$

where $L(\theta_h \mid D_h)$ is the likelihood of the historical data and $\pi_0(\theta_h)$ is an initial prior for the historical parameter before the historical data are observed. Larger values of τ indicate increased commensurability (reduced variability) between the current and historical data parameters and induce increased borrowing from the historical data to inform inference on θ_c.

Similar to the modified power prior, there is a single parameter that governs how much historical data are borrowed and incorporated into the final inference on the current study control parameter τ, τ can also be treated as a random variable rather than a fixed value. The choice of prior for τ is similar to the prior for the between study variance parameter in a meta-analysis. Priors for variance parameters are discussed comprehensively in [21]. It is generally recommended that an informative prior should be used on the between study variance parameter in a meta-analysis since the parameter is not well estimated from the data when there are few studies [22]. An informative prior is recommended for the commensurate prior for similar reasons. An informative prior is required to induce sufficient borrowing from the historical data. For one historical study, there is a direct relationship between a meta-analysis and the commensurate prior approach to incorporating historical data [19].

An extension to the location commensurate prior also places a commensurate prior on the parameter that represents the true underlying variance in the control group [18]. To induce borrowing from the historical data to the current trial parameters, the mean and variance in the historical and current data both have to agree.

1.2.6 Robust mixture prior

When there are historical data available from only one historical study, the robust mixture prior [23] is a two-component mixture distribution of conjugate priors. The first component of the mixture distribution is an informative component based on the historical data and the second component is a weakly-informative component. The form of the weakly-informative component is dependent on the type of outcome data. The weights given to each component of the mixture distribution in the prior are chosen by the study designer based on how relevant the historical data are thought to be to the current study control data. The prior weight given to the informative component of the robust mixture distribution based on the historical data can be interpreted in a similar way to the power chosen in the power prior, when α_0 is a fixed value. The weakly-informative component of the mixture distribution gives a heavy tailed prior distribution compared to using only the historical data as a prior and adds robustness against prior-data conflict [23, 24]. Using a mixture prior allows added flexibility while maintaining the convenience of using a conjugate prior. Here we will review properties of mixture distributions that are used throughout this thesis.

Let $\pi_1(\theta_c), \ldots, \pi_J(\theta_c)$ be proper probability density functions. Then given weights, $w_1, \ldots, w_J$, where $w_j > 0$ and $\sum\limits_{j=1}^{J} w_j = 1$, the mixture distribution,

$$\pi(\theta_c) = \sum_{j=1}^{J} w_j \pi_j(\theta_c),$$

is also a proper probability density.

If the individual mixture components are conjugate prior distributions, then the posterior distribution is also a mixture of conjugate distributions with updated parameter values and weights. Assuming the prior density for θ_c of [25],

$$\pi^{(0)}(\theta_c) = \sum_{j=1}^{J} w_j^{(0)} \pi_j^{(0)}(\theta_c),$$

where the superscript (0) denotes a prior distribution or weight, $w_j^{(0)}$ are the prior weights and $\pi_j^{(0)}(\theta_c)$ are the individual conjugate prior mixture distribution components (for the robust mixture prior, $\pi_1^{(0)}(\theta_c \mid D_h)$ would be the informative component of the mixture distribution based on the historical data and $\pi_2^{(0)}(\theta_c)$ would be a weakly-informative mixture component), the posterior distribution is given by [25],

$$\begin{aligned}\pi^{(1)}(\theta_c \mid D_c) &= \frac{\sum\limits_{j=1}^{J} w_j^{(0)} \pi_j^{(0)}(\theta_c) L(\theta_c \mid D_c)}{C} \\ &= \sum_{j=1}^{J} w_j^{(1)} \pi_j^{(1)}(\theta_c \mid D_c),\end{aligned}$$

where the superscript (1) denotes a posterior distribution or weight, $L(\theta_c \mid D_c)$ is the likelihood of the current trial control data and $C = \sum\limits_{j=1}^{J} w_j^{(0)} c_j$,

$$\pi_j^{(1)}(\theta_c \mid D_c) = \frac{\pi_j^{(0)}(\theta_c) L(\theta_c \mid D_c)}{c_j},\ w_j^{(1)} = \frac{w_j^{(0)} c_j}{\sum\limits_{j=1}^{J} w_j^{(0)} c_j},\ c_j = \int\limits_{-\infty}^{\infty} \pi_j^{(0)}(\theta_c) L(\theta_c \mid D_c) d\theta_c,$$

where θ_c is either a single parameter or a vector of parameters. The posterior mixture distribution components $\pi_j^{(1)}(\theta_c \mid D_c)$ are then obtained from standard conjugate Bayesian prior to posterior updates. The updated posterior weights sum to one and are calculated using the marginal likelihood of the data for each component of the mixture prior distribution.

Morita et al. [26] present a definition for the effective sample size of a parametric distribution and their method is used to calculate the effective sample size of the posterior mixture distribution. The method proposed by Morita et al. [26] approximates the information at the mode of the mixture distribution using a quadratic approximation. A vague prior is constructed which is updated with a sequence of known sample sizes to get a sequence of posterior distributions. The information of these posterior distributions with known sample sizes are then compared to the information of the mixture posterior distribution and the closest in terms of information at the mode gives the effective sample size of the mixture distribution.

1.3 Equivalence and choice of equivalence bounds

Throughout this thesis, the concept of equivalence is used. Equivalence testing is typically used to compare an experimental treatment with a standard of care treatment, where the experimental treatment may offer benefits such as lower cost or fewer side effects compared to the standard of care. In a frequentist framework, to test the equivalence of two treatments, the null hypothesis is that the treatments differ and the alternative hypothesis is that the treatments are equivalent. When comparing an experimental treatment to a standard of care, equivalence represents the belief that the two treatments are close enough that neither treatment is considered more efficacious or less futile than the other. How close two treatments are required to be for them to be considered equivalent is governed by the equivalence interval (δ_l, δ_u). The hypotheses for testing equivalence of two treatments in terms of the absolute difference of the parameter of interest in each treatment group are given by [27],

$$\begin{aligned} &H_0 : \theta_t - \theta_c \leq \delta_l \quad \text{or} \quad \theta_t - \theta_c \geq \delta_u \\ &H_1 : \delta_l < \theta_t - \theta_c < \delta_u \end{aligned}$$

The hypotheses H_0 and H_1 are tested as two separate one-sided hypothesis tests [27],

$$\begin{aligned} &H_{01} : \theta_t - \theta_c \leq \delta_l \\ &H_{11} : \theta_t - \theta_c > \delta_l \end{aligned}$$

and,

$$\begin{aligned} &H_{02} : \theta_t - \theta_c \geq \delta_u \\ &H_{12} : \theta_t - \theta_c < \delta_u \end{aligned}$$

In order to conclude equivalence, both null hypotheses H_{01} and H_{02} must be rejected at the chosen significance level.

Typically, when testing equivalence in terms of the absolute difference between two parameters, the equivalence interval is symmetric and there is one equivalence margin parameter given by $\delta = \delta_u = -\delta_l$. The choice of equivalence margin is subjective, however approaches have been proposed for choosing this parameter when comparing an experimental treatment to a standard of care treatment. The equivalence margin can be chosen based on expert opinion and interpreted as the largest difference that can be judged as clinically acceptable for the experimental treatment to be used in practice as an equivalent to the standard of care. Ng [28] discusses some interpretations of the equivalence margin in the literature, most of which relate to clinical judgement. An alternative approach proposed in the literature for choosing δ [28] is based on the comparison of the active standard of care treatment to a placebo using prior data. The equivalence margin is chosen to be a small fraction (e.g. 0.2) of the treatment effect comparing the active standard of care treatment to a placebo or a fraction of the lower limit of a confidence interval of the difference between the active standard of care and the placebo obtained from a meta-analysis of historical studies.

Throughout this thesis, equivalence is used for the comparison of historical control data and the control data from the current trial, where the current trial is assumed to be a superiority trial. The more formal methods proposed for determining the equivalence bounds can not be used in this setting. The control treatment may not be an active control, as is likely with a standard equivalence design for an experimental treatment. Therefore, the equivalence bounds are initially considered to be chosen based on expert opinion and alternative approaches for choosing the equivalence bounds are considered.

1.4 Adding a treatment arm to an ongoing trial

There are circumstances that may arise where it would be beneficial to add a treatment arm to an ongoing clinical trial, such as a treatment about to complete phase II development in the same disease area or a treatment currently awaiting regulatory approval.

The advantages of adding a treatment arm to an ongoing trial include: there is only one protocol, with new treatments incorporated as an amendment; utilising the existing trial infrastructure (e.g. staff, protocols, recruitment, randomisation); increasing the chance of allocation to a research arm, which may boost accrual; it potentially requires fewer patients; less money and a shorter trial duration [29].

Only a few papers have discussed methods for adding treatment arms to ongoing trials [29, 30, 31, 32]. Elm et al. [30] consider a design where a new treatment arm is added during an ongoing trial and randomisation continues to all treatment arms until the end of the study. Elm et al. [30] consider before and after the new treatment arm is added as two cohorts and assume a random cohort effect is present. They compare four different analysis methods and their operating characteristics in this setting. The four methods are: a linear model adjusting for a cohort effect; pooling the data from both stages; an inverse chi-square combination test and the weighted inverse normal combination test. Wason et al. [31] briefly discuss adding treatment arms in multi-arm multi-stage trials, Sydes et al. [29] discuss the methodology of the STAMPEDE trial, which added multiple treatment arms throughout the study and the most recent paper by Cohen et al. [32] gives a review of the methodology and practice of adding a treatment arm to a study.

One of the main design considerations when adding a treatment arm during an ongoing trial is controlling the family-wise error rate (FWER) for multiple comparisons of treatment to control. The FWER is defined as the probability of at least one type I error. The definition of the family of hypotheses in multi-arm trials has been discussed in many papers [29, 33, 34]. The main argument against correcting for multiple testing when comparing multiple experimental treatments to a single control treatment is that no adjustment is made for multiple testing when multiple treatments are compared to the same control treatment in separate trials. Therefore an adjustment is not necessary when the same comparisons are made in a single trial. The main difference when conducting a single trial rather than a separate trial for each experimental treatment is that in a single trial, the same control group is used for all treatment comparisons, which induces a correlation between the test statistics. The European Medicines Agency states that for a trial where there are more than two treatment arms "control of the FWER in the strong sense is a minimal prerequisite for confirmatory claims" [35]. A multiple testing procedure controls the FWER in the strong sense if the FWER control at level α is guaranteed regardless of which or how many null hypotheses are true. A multiple testing procedure controls the FWER in the weak sense if the FWER control at level α is guaranteed only when all null hypotheses are true [36]. The aim of the methods proposed in Chapter 4 when designing a trial where a treatment arm is added is to control the FWER of the design at a specified level. The next section gives an introduction to standard methods that correct for multiple testing.

1.4.1 Multiple testing procedures

Consider a multi-arm trial comparing J experimental treatments to a control treatment. There is a family of hypotheses which requires J statistical tests. A type I error in this setting is defined as rejecting any true null hypothesis (false positive) and interest is in controlling the risk of any false positives.

If no adjustment is made for multiplicity, then the overall sample size is reduced by a factor of $(J-1)/2J$ for a single trial of J experimental treatments with a single control group compared to running J separate trials, each with its own control arm, however the FWER will be inflated above the desired level.

Many procedures have been developed to control the FWER when performing multiple tests. The choice of method depends on: the aim of the study; the relationship among the null hypotheses; whether there is a logical relationship and hypotheses have a pre-specified ordering or an ordering that is data driven; and finally, whether there is a distributional relationship between the hypotheses and whether a non-parametric, semi-parametric or parametric method should be used.

Multiple testing procedures fall into two main categories. Single step methods, where individual test statistics are compared to critical values simultaneously, and secondly, sequential procedures, where adaptive adjustments are made to the critical values for the remaining hypotheses dependent on the previous hypotheses tested. A comprehensive review of multiple testing procedures is given in [37] and a nice summary of which multiple testing procedures are best in different scenarios is given in the book Multiple testing problems in pharmaceutical statistics [36]. In this thesis only single step methods are considered. We give a summary of some of the simple single step non-parametric procedures below [38], but focus on the single step Dunnett procedure which is a parametric procedure that accounts for the correlation between test statistics when comparing multiple experimental treatments to a single control treatment [39].

1.4.2 Single step multiple comparison adjustments

The two simplest corrections for multiple testing are given by Bonferroni and Dunnett. Both methods maintain strong control of the FWER.

Bonferroni

For J hypotheses being tested and FWER specified at level α. The Bonferroni correction simply tests each hypothesis at significance level α/J. The weighted Bonferroni method tests each hypothesis at level $w_j\alpha/J$ where $\sum_{j=1}^{J} w_j = 1$. Similarly, the Sidak method [38] tests each hypothesis at significance level $1-(1-\alpha)^{1/J}$. However, all of these methods are conservative when the test statistics are correlated.

Dunnett

The method proposed by Dunnett [39] to correct for multiple testing is designed for comparing multiple experimental treatments to a single control treatment, as it accounts for the correlation between test statistics. Throughout Chapter 4 we assume a design where multiple experimental treatments are compared to a single control arm. Consider a trial comparing J experimental treatments to a control treatment, with test statistics for each experimental treatment given by,

$$Z_j = \frac{\bar{X}_j - \bar{X}_0}{\sqrt{\dfrac{1}{n_j} + \dfrac{1}{n_0}}} \sim N(0, \sigma^2),$$

with $j = 1, 2, ..., J$ hypotheses being tested. Where $\bar{X}_j$ and $\bar{X}_0$ are the sample means of the experimental treatment j and the control group respectively, which are assumed to be independently and normally distributed. μ_j and μ_0 are the true underlying means in the treatment and control groups. σ is the assumed known common standard deviation across all treatment groups and n_j and n_0 are the sample sizes of the experimental and control treatment arms, respectively.

The joint distribution of the J test statistics is multivariate normal with Z_j having mean 0, variance σ^2 and correlation between any two test statistics, Z_1 and Z_2 for example, given by [39],

$$\rho_{Z_1 Z_2} = 1 \Bigg/ \sqrt{\left(\frac{n_0}{n_1}+1\right)\left(\frac{n_0}{n_2}+1\right)} .$$

The critical values are determined that control the FWER at the desired level and these are used to calculate the sample size needed per treatment group for a given marginal power for each of the experimental treatment to control comparisons. Dunnett shows that the optimal allocation ratio for this design is approximately $n_0 = n\sqrt{J}$, where n_0 is the number of control patients and n the number of patients in each experimental treatment group, further details of this design are given in Chapter 4 where the design is extended to account for adding a treatment arm during the trial. The methods described in this section are frequentist and in this thesis we mainly aim to use Bayesian methodology for the reasons discussed in Section 1.1. Whitehead et al. [40] proposed a Bayesian design comparable to the approach proposed by Dunnett [39] for comparing multiple experimental treatments to a control treatment. The design proposed by Whitehead et al. was for phase II trials with the aim of identifying treatments that are worth further investigation in a phase III trial. However, the methodology can be used for confirmatory trials. The next section describes the design proposed by Whitehead et al. [40], this design is described in detail here since it is only briefly considered in Chapter 4.

1.4.3 Bayesian design for comparing multiple experimental treatments to a control treatment

The sample sizes derived using the Bayesian approach proposed by Whitehead et al. [40] match the sample sizes obtained from a frequentist design where the aim for the multi-arm trial is to control the probability of detecting one or more truly beneficial experimental treatment (family-wise power) and the probability of continuing with a single experimental treatment that has no benefit over control (marginal type I error rate). For a confirmatory trial the aim would usually be to control the probability of detecting a single experimental treatment that is better than control (marginal power) and the probability of continuing with one or more experimental treatment that has no benefit over the control treatment (FWER). The Bayesian approach described in this section incorporates prior information into both the design and analysis of the trial.

Following a similar notation to Whitehead et al. [40]. Let Δ^* denote the treatment effect to be detected, the clinically important treatment difference between a single experimental treatment and control and let Δ denote the treatment effect of a single experimental treatment compared to control. For a single experimental treatment, the sample size is chosen to satisfy the Bayesian criterion that the posterior belief that $\Delta > 0$ is large enough to conclude that the experimental treatment is promising and research proceeds to phase III, or the posterior belief that $\Delta < \Delta^*$ is large enough that the treatment is not considered further. Let $j = 0, \ldots, J$ represent the treatment group, where $j = 0$ represents the control treatment and there are J experimental treatments. Let i represent the ith patient in treatment group j, $i = 1, \ldots, n_j$. n_j denotes the number of patients in treatment group j. Let x_{ij} denote the observed value for patient i on treatment j. In this section, let the superscript 0 represent prior information and the superscript 1 represent posterior information. Let $\bar{x}_j$ denote the sample mean for treatment j, μ_j the true underlying mean for treatment j and σ^2 the true underlying variance, assumed to be common across treatment groups.

The observed value x_{ij} of the ith patient on treatment j is distributed normally with mean μ_j and precision $v = 1/\sigma^2$ $(v^{-1} = \sigma^2)$ for $i = 1, ..., n_j, j = 0, 1, ..., J$, assuming v is known. The treatment effects comparing each experimental treatment to control are denoted by $\Delta_j = \mu_j - \mu_0$ for $j = 1, \ldots, J$.

We assume independent normal prior distributions for each treatment mean μ_j,

$$\pi(\mu_j) \sim N(\mu_j^0, (n_j^0 v)^{-1}),$$

where n_j^0 represents the prior effective sample size for treatment group j.

The posterior distribution for each treatment mean is then normally distributed and given by,

$$\pi(\mu_j) \sim N(\mu_j^1, (n_j^1 v)^{-1}),$$

where $\mu_j^1 = (\mu_j^0 n_j^0 + n_j \bar{x}_j)/(n_j^0 + n_j)$ and $n_j^1 = (n_j^0 + n_j)$, which represents the total amount of information (both prior and observed) for treatment j.

The joint posterior distribution for the vector of treatment effects $(\Delta_1, \ldots, \Delta_J)$ is multivariate normal with Δ_j having mean $\mu_j^1 - \mu_0^1$ and variance $(v n_j^1 n_0^1/(n_j^1 + n_0^1))^{-1}$. The covariance between any two treatment effects is given by $(n_0^1 v)^{-1}$.

It is assumed that the trial is designed to obtain equal information for each experimental treatment (prior effective sample size plus current data sample size), therefore $n_1^1 = \ldots n_j^1 = n^1$. Then the variances for each treatment effect are equal, denoted by $(vn^1n_0^1/(n^1+n_0^1))^{-1}$.

Sample Size Criteria

Let $\mathbf{x}$ denote data collected in the current trial, the sample size should be chosen to ensure that either one treatment is developed further ($\Pr(\Delta_j > 0 \mid \mathbf{x}) \geq 1-\alpha$) or all treatments are abandoned ($\Pr(\Delta_j < \Delta^*$ for all $j = 1, \ldots J \mid \mathbf{x}) > 1-\beta$), for any possible outcome dataset $\mathbf{x}$.

To determine a suitable sample size before the study is conducted, let Δ' denote the value of the posterior mean of Δ_j, $\mu_j^1 - \mu_0^1 = \Delta', j = 1, \ldots, J$ for which $\Pr(\Delta_1 > 0 \mid \mathbf{x}) = \ldots \Pr(\Delta_J > 0 \mid \mathbf{x}) = 1-\alpha$ and $\Pr(\Delta_j < \Delta^*$ for all $j = 1, \ldots J \mid \mathbf{x}) = 1-\beta$.

Then,

$$\Pr(\Delta_j > 0 \mid \mathbf{x}) = \Phi\left(\Delta'\sqrt{\frac{n^1n_0^1v}{(n^1+n_0^1)}}\right) \quad \Longrightarrow \quad \Delta' = \Phi^{-1}(1-\alpha)\left(\frac{n_j^1n_0^1v}{(n_j^1+n_0^1)}\right)^{-1/2}.$$

$$\begin{aligned}
&\Pr(\Delta_j < \Delta^* \text{ for all } j = 1, \ldots J \mid \mathbf{x}) = \Pr(\max(\Delta_1, \ldots, \Delta_J) < \Delta^* \mid \mathbf{x}) = \\
&\Pr\left(\max\left((\Delta_1 - \Delta')\sqrt{\frac{n^1n_0^1v}{(n^1+n_0^1)}},\right.\right. \\
&\left.\left.\ldots, (\Delta_J - \Delta')\sqrt{\frac{n^1n_0^1v}{(n^1+n_0^1)}} < (\Delta^* - \Delta')\sqrt{\frac{n^1n_0^1v}{(n^1+n_0^1)}}\middle|\mathbf{x}\right)\right) = 1-\beta \\
&\Longrightarrow \left(\Delta^* - \Phi^{-1}(1-\alpha)\left(\frac{n_j^1n_0^1v}{(n_j^1+n_0^1)}\right)^{-1/2}\right)\sqrt{\frac{n^1n_0^1v}{(n^1+n_0^1)}} = c_{1-\beta,\rho,J},
\end{aligned}$$

where $c_{1-\beta,\rho,J}$ is the value such that $\Pr(C_1, \ldots, C_J < c_{1-\beta,\rho,J}) = 1-\beta$, where the vector $(C_1, \ldots, C_J)$ is multivariate normal with C_j having mean 0, variance 1 and correlation between any two C_j, $\rho = \left(1 + \frac{n_0^1}{n^1}\right)^{-1}$.

Sample sizes are chosen such that,

$$\frac{(n^1+n_0^1)}{n^1n_0^1} = \frac{1}{n_0^0+n_0} + \frac{1}{n_j^0+n_j} = \frac{v}{\{\Phi^{-1}(1-\alpha) + c_{1-\beta,\rho,J}/\Delta^*\}^2} \text{ for } j = 1, \ldots, J.$$

The minimum total sample size required with respect to the constraint above is obtained from,

$$\begin{aligned}
n_j &= \left(\sigma^2 + \frac{\sigma^2}{\sqrt{J}}\right)\left(\frac{\Phi^{-1}(1-\alpha) - c_{1-\beta,\rho,J}}{\Delta^*}\right)^2 - n_j^0 \\
n_0 &= \left(\sigma^2 + \sqrt{J}\sigma^2\right)\left(\frac{\Phi^{-1}(1-\alpha) - c_{1-\beta,\rho,J}}{\Delta^*}\right)^2 - n_0^0,
\end{aligned}$$

and the optimal allocation ratio is $(n_0 + n_0^0) = (n_j + n_j^0)\sqrt{J}$, consistent with the frequentist approximation for the optimal allocation. Note that, when $n_j^0 = 0$ and $n_0^0 = 0$ these formulae reduce to the sample size formulae for a standard frequentist multi-arm trial.

If informative priors are used, the prior means will be used in the Bayesian analysis of the trial, although they play no part in determining the sample size. The design set-up here considers controlling the marginal type I error rate and the family-wise power, rather than the FWER and the marginal power which we consider in Chapter 4, however, the same theory applies.

1.5 Outline of thesis

In this thesis, three methods published in the literature for incorporating historical data into the design and the analysis of a current trial are assessed. It is assumed throughout that the historical data available are for the control arm only and from a single study. The methods reviewed are: the modified power prior [15]; the commensurate prior; [18] and the robust mixture prior [23]. In Chapter 2 these methods are explored for binary outcome data, response and non-response to a treatment. In Chapter 3, the aforementioned methods are explored for normally distributed outcome data. The limitations of the published methods are discussed. A novel equivalence weight approach for assessing the agreement between historical and current controls is then proposed for binary data in Chapter 2 and normally distributed outcome data in Chapter 3, the equivalence weight is used to down-weight the historical data. Two designs are explored, a design where the down-weighted historical data are incorporated as additional information to increase the size of the current trial control arm, and an adaptive design where the down-weighted historical data replace controls yet to be randomised in the current trial. The equivalence weight is used alongside the analysis approach of the power prior [14] using the equivalence weight as a fixed power. Methods are discussed for choosing the equivalence bounds to control the maximum inflation in the type I error rate when there is disagreement between the historical and current control data.

In Chapter 4, the focus of the thesis then turns to trials where a treatment arm is added during the trial. A design is proposed that adapts the sample size of all treatment arms when a new treatment arm is added to control the FWER for a specific marginal power for each treatment to control comparison. Only concurrent controls are used for each treatment to control comparison. Optimal allocation to each treatment group is then explored when a trial starts with a single experimental treatment and control arm, and an additional experimental treatment is added to the trial.

Finally, the work presented in this thesis is summarised in Chapter 5 and ideas for future research are outlined.

The statistical package Stata is used to generate results throughout the thesis [41].

Chapter 2

Historical data methods for the design and analysis of a trial with a binary outcome

2.1 Introduction

This chapter is structured to address the five questions about the use of historical data posed by Pocock in his 1976 seminal paper [11] when the outcome data are binary. These five questions are: what is relevant historical data; how many additional patients does the historical data provide; how do we assess agreement between historical and current data; what sample size is required for the new study; and how can historical data be incorporated into the analysis. The first question was addressed comprehensively in the original paper where Pocock defined six acceptability criteria for using historical data in both the design and analysis of a new study, these criteria were discussed in Chapter 1. It is assumed throughout this chapter that the historical data chosen are relevant to the current study taking into account the six acceptability criteria defined by Pocock.

In this chapter, it is assumed that only one historical study is available and therefore the maximum number of additional patients that the historical study may provide is the sample size of the historical study. Depending on the agreement between the historical and current control data, the historical data may be down-weighted and therefore the additional number of patients that the historical data provides will be reduced. To address the remaining questions, this chapter is structured as follows: how to assess agreement between the current and historical data is discussed in Sections 2.2 and 2.3; how to incorporate historical data into the design of a current study in Section 2.4; and finally how to incorporate the historical data into the analysis of a current study is addressed in Section 2.5.

Given the assumption that only one relevant historical study is available and that this historical study provides information on the control arm only, the aim in this chapter is to assess the conflict between the historical and current control data and to determine how much weight to give the historical data in the final analysis of the current trial. When outcome data are binary, response or non-response, there is only one unknown parameter, the response probability. To assess agreement between the historical and current trial controls, the response probabilities are compared.

The advantages and disadvantages of several historical data methods published in the literature are explored for how they assess agreement between historical and current controls, and how they incorporate historical data into the design and analysis of the current study when the outcome data are binary. Two novel methods are then proposed for assessing the agreement between historical and current control data. These two approaches specify a weight which is used to down-weight the historical data. A design that incorporates the historical data as additional information and an adaptive design are both explored for incorporating the historical data into the design and analysis of a new study. The analysis approach of the power prior is utilised, where the calculated weight is used as a fixed power to down-weight the historical data [14]. The operating characteristics of these two new approaches are compared to the methods proposed in the literature for binary outcome data.

For all historical data methods, the main disadvantage is that incorporating historical control data affects the operating characteristics of the current trial. Consider a two-arm trial for a binary outcome, with historical data available for the control arm only. Testing the null hypothesis of no treatment difference against the alternative hypothesis that the new treatment has a higher response probability than the controls. In general, if the current trial control estimate of the response probability is lower than the historical, incorporating historical data into the final analysis will result in reduced power and a deflated type I error rate for estimating the treatment effect, compared to a standard design not incorporating any historical data. If the current trial estimate of the response probability in the control arm is higher than the historical data estimate, the power and type I error rate will be inflated compared to a standard design. The effect on the operating characteristics from incorporating the historical data is due to the historical data causing bias in the control response probability estimate and therefore bias in the treatment effect estimate.

2.1.1 Notation

Let x_h and y_h be the number of responses and non-responses in the historical data, respectively. Let x_c and y_c be the number of responses and non-responses in the current control group and x_t and y_t be the number of responses and non-responses in the current treatment group. n_h, n_c and n_t are the historical, current control and experimental treatment sample sizes, respectively. Let p_c and p_t denote the true underlying response probabilities in the control and treatment arms, respectively. For models where it is assumed that the true underlying response probabilities in the current and historical studies are not the same, let p_h denote the underlying true response probability in the historical controls.

2.1.2 Illustrative example

Throughout this chapter, we consider an example from Viele et al. [42], which is representative of a confirmatory two-arm randomised controlled trial. The primary analysis of interest is a hypothesis test of $H_0 : p_c = p_t$ against $H_1 : p_c < p_t$. For a standard design, incorporating no historical data, assuming a control response probability of 65%, 200 patients are required per treatment arm to detect a treatment difference of 12% with approximately 76% power and a one-sided type I error rate of 2.5%. In addition, for the historical data designs considered in this chapter, it is assumed there are 100 (n_h) historical control patients available with a response probability of 65%.

Throughout sections 2.2 and 2.3 where methods are explored that assess agreement between historical and current control data, the example used assumes that the historical data are fixed at 65/100 responses and that there are 100 control patients in the current trial. A range of response proportions in the current control data are explored.

2.2 Published methods for assessing agreement between historical and current control data

For the historical data methods discussed in this section, each method assesses the agreement between the historical and current control data and calculates either a weight w or a prior effective sample size (ESS). For methods that calculate a weight, this weight is used to down-weight the historical data by raising the likelihood of the historical data to the power of the weight and we define the effective historical sample size ($EHSS$) to be the weight times the historical sample size wn_h. For binary data, $EHSS = wn_h$, since the posterior distribution based on the weighted historical data likelihood is identical to a posterior distribution based on the likelihood of wn_h observations of which there were wx_h responses and wy_h non-responses. A distinction is made between the $EHSS$ which is based only on the historical data and the ESS which also incorporates the information contained in the prior before the historical data are observed.

2.2.1 Modified power prior

The modified power prior assumes that the historical data and current control data are estimating the same underlying parameter of interest ($p_h = p_c$). For binary outcome data, the modified power prior has the general form [15, 43],

$$\pi(p_c, \alpha_0 \mid x_h, y_h) \propto \frac{L(p_c \mid x_h, y_h)^{\alpha_0} \pi_0(p_c)}{\int L(p_c \mid x_h, y_h)^{\alpha_0} \pi_0(p_c) dp_c} \pi(\alpha_0),$$

where $\pi_0(p_c)$ is the initial prior for p_c before the historical data are observed, $L(p_c \mid x_h, y_h)^{\alpha_0}$ is the likelihood of the historical data raised to a power α_0, with prior $\pi(\alpha_0)$. $\pi(\alpha_0)$ is assumed to be a Beta(a, b) distribution. A beta distribution is an intuitive choice for $\pi(\alpha_0)$ since it lies between zero and one and covers a wide variety of shapes depending on the parameter values chosen. In the historical data setting, it is unlikely we would want to give more weight to the historical data than the current control data, therefore a beta prior that does not allow a weight above one is chosen.

For binary data and only one historical study, the likelihood of the historical data follows a binomial distribution and beta prior distributions are assumed for p_c and α_0. The modified power prior is then given by [15, 16],

$$\pi(p_c, \alpha_0 \mid x_h, y_h) = \frac{p_c^{\alpha_0 x_h + c - 1}(1 - p_c)^{\alpha_0 y_h + d - 1}}{\mathrm{B}(\alpha_0 x_h + c, \alpha_0 y_h + d)} \frac{\alpha_0^{a-1}(1 - \alpha_0)^{b-1}}{\mathrm{B}(a, b)},$$

where $\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and an initial Beta(c,d) prior is assumed for p_c.

The joint posterior distribution is then [15],

$$\pi(p_c, \alpha_0 \mid x_h, y_h, x_c, y_c) \propto \frac{p_c^{\alpha_0 x_h + x_c + c - 1}(1 - p_c)^{\alpha_0 y_h + y_c + d - 1} \alpha_0^{a-1}(1 - \alpha_0)^{b-1}}{\mathrm{B}(\alpha_0 x_h + c, \alpha_0 y_h + d)}.$$

The marginal posterior distribution for the power is given by [16],

$$\pi(\alpha_0 \mid x_h, y_h, x_c, y_c) \propto \frac{\Gamma(\alpha_0 n_h + c + d)\Gamma(\alpha_0 x_h + x_c + c)\Gamma(\alpha_0 y_h + y_c + d)}{\Gamma(\alpha_0 x_h + c)\Gamma(\alpha_0 y_h + d)\Gamma(\alpha_0 n_h + n_c + c + d)} \alpha_0^{a-1}(1 - \alpha_0)^{b-1}. \tag{2.1}$$

Throughout the thesis, for the modified power prior approach we explore using a summary measure of the marginal posterior distribution of α_0 as a measure of the agreement between the current and historical controls and as a fixed value to down-weight the historical data. A fixed summary value allows exact knowledge of how much of the historical data are incorporated into the final analysis. We explore using the mean, median or mode of the posterior distribution as a fixed value for α_0 (substitution method). We denote this approach the modified power prior approach because we are placing a prior on α_0, we then obtain a single value for the power from the marginal distribution of α_0. An alternative approach is to incorporate the full uncertainty around α_0 into the analysis which we denote the fully Bayesian modified power prior approach.

The marginal distribution of p_c can be used to explore the behaviour of p_c by integrating α_0 out of the joint posterior distribution $\pi(p_c, \alpha_0 \mid x_h, y_h, x_c, y_c)$, we denote this the fully Bayesian modified power prior approach. However, the marginal distribution of p_c does not have a closed form distribution and requires numerical integration. The marginal distribution of p_c is given by,

$$\pi(p_c \mid x_h, y_h, x_c, y_c) = \int_0^1 \pi(p_c, \alpha_0 \mid x_h, y_h, x_c, y_c) d\alpha_0,$$

where,

$$\pi(p_c, \alpha_0 \mid x_h, y_h, x_c, y_c) = \frac{\dfrac{p_c^{\alpha_0 x_h + x_c + c - 1}(1-p_c)^{\alpha_0 y_h + y_c + d - 1}\alpha_0^{a-1}(1-\alpha_0)^{b-1}}{\mathrm{B}(\alpha_0 x_h + c, \alpha_0 y_h + d)}}{\int_0^1\int_0^1 \dfrac{p_c^{\alpha_0 x_h + x_c + c - 1}(1-p_c)^{\alpha_0 y_h + y_c + d - 1}\alpha_0^{a-1}(1-\alpha_0)^{b-1}}{\mathrm{B}(\alpha_0 x_h + c, \alpha_0 y_h + d)} d\alpha_0 dp_c} \tag{2.2}$$

Choice of prior for the power

The modified power prior allows a prior to be placed on α_0, with the aim of learning about what the discounting parameter should be from the data. If there is no prior opinion on the agreement between the historical and current control then a minimally-informative prior for α_0 may be chosen. Various minimally-informative priors have been discussed for Bernoulli random variables [17, 44]. An intuitive choice of minimally-informative prior for α_0 would be a uniform distribution on [0,1] (or equivalently a Beta(1,1) distribution), giving equal probability to all values of the power between zero and one. The Jeffrey's minimally-informative prior [17] for a Bernoulli outcome that is invariant to transformation is a Beta(0.5,0.5). Another minimally-informative prior that has been suggested is a Beta(c,c) prior where c < 1 [45]. Here, these priors are considered as $c \to 0$, these priors are denoted quasi-dichotomous priors. The intuition behind using these priors for the power prior is that they are heavy-tailed, which may help the choice of whether the historical data should be borrowed or not depending on the observed current control data.

These priors are considered minimally-informative because the ESS, given by the sum of the parameter values, is small. However, they can be quite informative for the power prior, as illustrated in Figure 2.1.

A final minimally-informative prior for a beta distribution is an improper Beta(0,0) prior. The intuition behind this prior is to choose a prior which produces a Bayesian posterior estimate that coincides with the maximum likelihood estimate, since the maximum likelihood estimate does not consider prior information. This is an intuitive prior for a binomial likelihood and results in a proper posterior distribution, except when there are zero responses or non-responses in the data. However, using a Beta(0,0) prior on the power α_0 results in an improper marginal posterior distribution for the α_0. A Beta(0,0) distribution can therefore not be used as a prior for the power.

Let's consider the Viele example [42]. Figure 2.1 shows three different priors for α_0, a Beta(1,1), Beta(0.5,0.5) and a Beta(0.3,0.3) prior and the posterior distributions of α_0 for each prior given different levels of agreement between the historical and current controls. It is assumed that there are 100 current control patients and 100 historical patients and the response proportion in the historical controls is fixed at 65%. A Beta(1,1) prior is assumed for p_c before the historical data are observed.

A prior that gives a posterior distribution with mass close to one when there is complete agreement between the historical and current control data is desirable. As has been previously noted [16, 18] and is illustrated in Figure 2.1, under a Beta(1,1) for α_0, the marginal posterior distribution of α_0 is almost flat at complete agreement between the historical and current data (65% response proportion). The quasi-dichotomous priors result in a posterior distribution for α_0 with most of its mass at zero and one and seems to differentiate well between agreement and disagreement in the historical and current data.

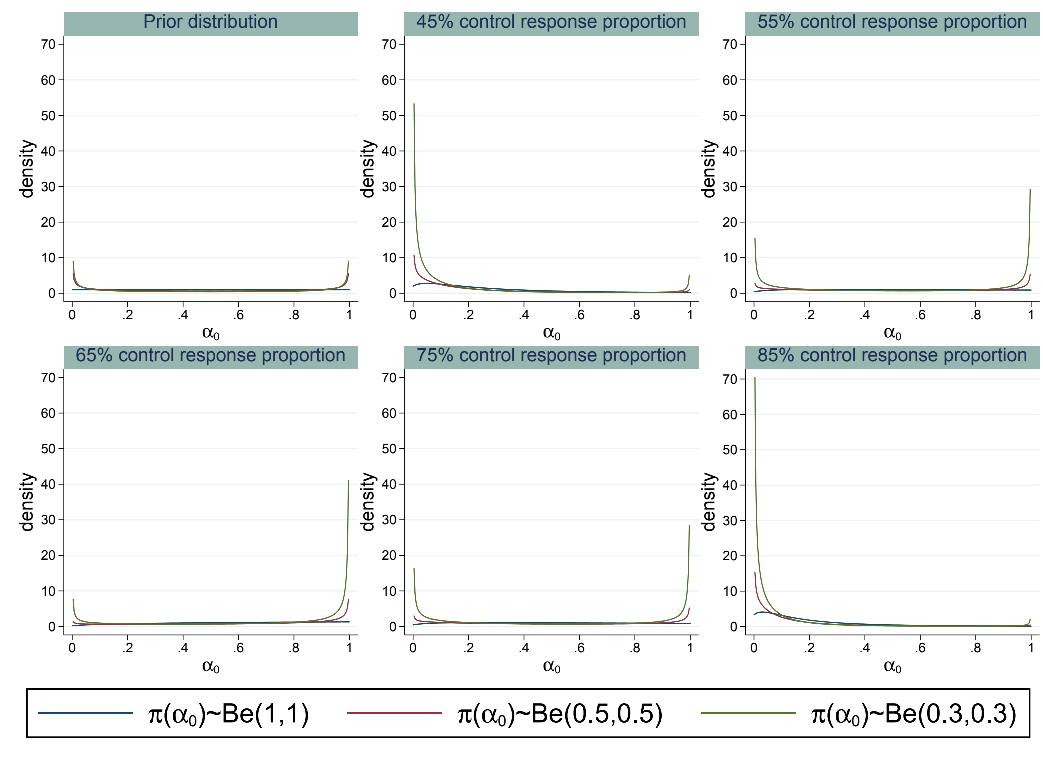
From the marginal posterior distribution of α_0, given in Equation 2.1, the mean is calculated using numerical integration,

$$\mathbf{E}(\alpha_0) = \int_0^1 \alpha_0 \, \frac{\pi(\alpha_0 \mid x_h, y_h, x_c, y_c)}{\int_0^1 \pi(\alpha_0 \mid x_h, y_h, x_c, y_c) d\alpha_0} d\alpha_0. \tag{2.3}$$

To calculate the median of the distribution, numerical integration and optimization can be used. The function to be minimised is given by,

$$\left\{ \int_0^{q(0.5)} \frac{\pi(\alpha_0 \mid x_h, y_h, x_c, y_c) d\alpha_0}{\int_0^1 \pi(\alpha_0 \mid x_h, y_h, x_c, y_c) d\alpha_0} - 0.5 \right\}^2, \tag{2.4}$$

Figure 2.1: Marginal distributions of α_0 for different observed current control response proportions and different priors on the power. Viele example, historical data 65/100 responses, 100 current controls.



with respect to q(0.5). The same approach can be used to obtain the 95% credible interval for α_0. Minimising the functions,

$$\left\{\int_0^L \frac{\pi(\alpha_0 \mid x_h, y_h, x_c, y_c)d\alpha_0}{\int_0^1 \pi(\alpha_0 \mid x_h, y_h, x_c, y_c)d\alpha_0} - 0.025\right\}^2 \text{ and } \left\{\int_0^U \frac{\pi(\alpha_0 \mid x_h, y_h, x_c, y_c)d\alpha_0}{\int_0^1 \pi(\alpha_0 \mid x_h, y_h, x_c, y_c)d\alpha_0} - 0.975\right\}^2, \tag{2.5}$$

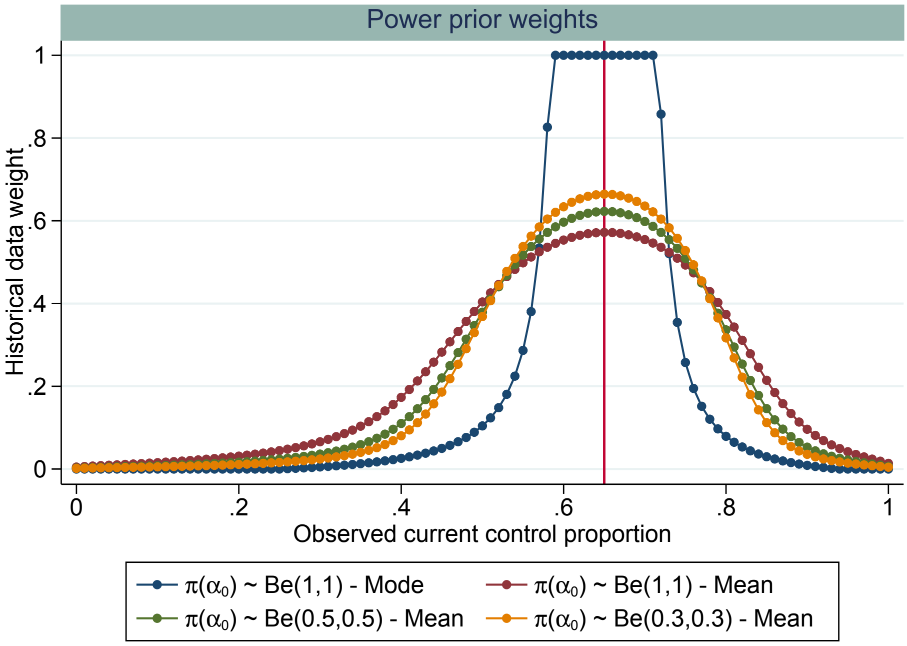
respectively, with respect to L and U.

The mode M is the value of α_0 such that $\pi(M) \geq \pi(\alpha_0), \forall\ \alpha_0 \in [0, 1]$. The mode can be calculated by differentiating $\pi(\alpha_0 \mid x_h, y_h, x_c, y_c)$ and equating to zero. Or the mode may need to be determined through searching over the function values or by using graphical methods where the mode lies at the bounds of the distribution.

It can be difficult to evaluate Equations 2.3 to 2.5 when a quasi-dichotomous prior is chosen for α_0. When the parameter values of the quasi-dichotomous prior are small, the marginal distribution of α_0 is heavily peaked at the limits of the distribution (near zero and one) and can be difficult to integrate using numerical integration due to the discontinuity at zero and/or one. Furthermore, because numerical optimisation is required to calculate the median, this approach could be computationally intensive depending on the current control sample size. Since the power needs to be calculated for each observed current control response. The terms power and weight are used interchangeably to denote the summary measure of the posterior distribution of α_0 that is used to down-weight the historical data.

Figure 2.2 and Table 2.1 show the mean of the posterior distribution of α_0 for the three beta prior distributions discussed above: Beta(1,1); Beta(0.5,0.5); and Beta(0.3,0.3), and the mode of the posterior distribution of α_0 for a Beta(1,1) prior for a range of observed current control proportions and fixed historical data. The mode of the quasi-dichotomous prior was not considered here, as this would always give either zero and one, similar to the test and pool approach discussed by Viele et al. [42]. Table 2.2 gives the median and 95% credible interval of the power α_0 for five observed current control response proportions. The optimisation technique used to find the median value and 95% credible interval by minimising the functions in Equations 2.4 and 2.5 was the modified Newton-Raphson technique implemented in Mata [46]. For a single observed response proportion, calculating the median took 2.88 seconds, therefore calculating the median for all possible responses in a dataset or for all possible combinations of control responses in a two-stage adaptive design may be time consuming.

Figure 2.2: Mean and mode of the marginal distribution of α_0 for different observed control response proportions and different priors on α_0. Viele example, historical data 65/100 responses, 100 current controls. The vertical red line represents complete agreement between the historical and current control proportions.



Using the mean of the posterior distribution of α_0 to down-weight the historical data

Table 2.1: Mean and mode of the marginal distribution of α_0 for a range of observed current control proportions and different priors on α_0. Viele example, historical data 65/100 responses, 100 current controls.

Current control proportion	Mean (α_0)			Mode (α_0)
	Beta(1,1)	Beta(0.5,0.5)	Beta(0.3,0.3)	Beta(1,1)
0.45	0.307	0.250	0.218	0.057
0.55	0.512	0.538	0.563	0.380
0.65	0.571	0.622	0.664	1
0.75	0.474	0.481	0.494	0.195
0.85	0.185	0.119	0.088	0.024

Table 2.2: Median and 95% credible interval of the marginal distribution of α_0 for different observed current control proportions. Viele example, historical data 65/100 responses, 100 current controls.

	Median(95% credible interval)		
Current control proportion	Beta(1,1)	Beta(0.5,0.5)	Beta(0.3,0.3)
0.45	0.210 (0.011,0.889)	0.115 (0,0.956)	0.062 (0,0.986)
0.55	0.490 (0.040,0.972)	0.514 (0.006,0.998)	0.557 (0,1)
0.65	0.594 (0.065,0.981)	0.692 (0.020,0.999)	0.793 (0.003,1)
0.75	0.483 (0.038,0.972)	0.500 (0.006,0.998)	0.535 (0,1)
0.85	0.145 (0.007,0.806)	0.067 (0,0.826)	0.030 (0,0.813)

gives a low weight to the historical data even when the historical and current observed response proportions are the same. The weight discounts slowly to zero as the difference increases. The quasi-dichotomous priors give a higher weight at agreement and discount more quickly to zero than the Beta(1,1) prior. The mode gives a weight of one to the historical data (pooling of the historical and current controls) for a range of response proportions around complete agreement and discounts quickly to zero as the difference between the current and historical control response proportions increases. However, for a Beta(1,1) prior, at complete agreement in the historical and current control response proportions, the posterior distribution of α_0 is almost flat across all values of α_0, therefore a weight of one does not seem representative of the posterior belief about the power. Using the median of the posterior distribution of α_0 to down-weight the historical data gives a higher weight at complete agreement in the historical and current control response proportions and discounts more quickly to zero as the difference increases compared to using the mean of the posterior distribution of α_0. However, there is a computation cost to using the median, each calculation of the power requires integration and optimisation and therefore using the median would be too computationally intensive in the trial design setting. The credible intervals of α_0 are wide for all priors and different levels of agreement between the historical and current data, indicating there is a lot of uncertainty in estimating α_0. Power prior and modified power prior are used interchangeably throughout the thesis, where the power is treated as a random variable, the modified power prior is always used.

2.2.2 Commensurate prior

The commensurate prior [18, 19, 20] assumes different parameters for the true underlying response probabilities in the current and historical controls. The outcome data are modelled on the log odds scale. The model fitted is given by,

$$\begin{aligned}
X_c &\sim \text{Bin}\left(n_c, p_c\right), \\
X_h &\sim \text{Bin}(n_c, p_h), \\
X_t &\sim \text{Bin}(n_t, p_t), \\
\text{logit}(p_c) &= \psi_c, \\
\text{logit}(p_h) &= \psi_h, \\
\text{logit}(p_t) &= \psi_t, \\
\psi_c &\sim \text{N}(\psi_h, 1/\tau), \\
\psi_h &\sim \text{N}(0, 1000), \\
\psi_t &\sim \text{N}(0, 1000),
\end{aligned}$$

where $\text{logit}(p) = \log(p/(1-p))$. Three different priors for τ (the commensurability parameter) are considered in the original paper by Hobbs et al. [19]. The first two priors are conjugate gamma priors of the form, $\pi(\tau) \sim \text{Gamma}(c(\tilde{\tau}), c)$ where $\tilde{\tau}$ is a prior guess of τ and c represents the belief in the prior estimate of τ, with a smaller value of c corresponding to weaker prior belief, $\tilde{\tau} > 0$ and $c > 0$. Hobbs et al. [19] consider two gamma priors, a Gamma(1,0.01) and a Gamma(1,0.05). The gamma prior here is parameterised in terms of the shape and rate, therefore, a Gamma(α, β) prior has mean α/β and variance α/β^2. The Gamma(1,0.01) prior has mean 100 and variance 10000 and the Gamma(1,0.05) has mean 20 and variance 400. The third prior considered by Hobbs et al. [19] is a spike and slab prior. This prior distribution is a two-component mixture distribution made up of a uniform distribution between two limits (slab) and a degenerate distribution at a selected large value (spike).

Let S_l and S_u be the limits between which the spike and slab prior is uniformally distributed and let K be a point mass that denotes commensurability between the historical and current controls. Then the prior distribution is specified as [19],

$$\begin{aligned}
&\Pr(\tau < S_l) = 0, \\
&\Pr(\tau < u) = p_0\{(u - S_l)/(S_u - S_l)\}, S_l \leq u \leq S_u, \\
\text{and } &\Pr(\tau > S_u) = P(\tau = K) = 1 - p_0,
\end{aligned}$$

where p_0 denotes the prior probability of existence within the slab, $0 \leq S_l < S_u$ and $K > S_u$.

The argument for this type of prior is that for small differences between the current and historical data response proportions, the marginal likelihood of the data is nearly flat with a gradually decreasing slope as a function of τ, when τ is sufficiently large. It is therefore suggested that it may be sensible to choose a large value of τ that represents commensurability [19]. Hobbs et al. [19] warn that the results of the commensurate prior model are dependent on the calibration of the spike and slab prior, i.e. on the choice of the parameters p_0, S_l, S_u and K. We therefore do not consider the spike and slab prior in this thesis. In this thesis, to illustrate the commensurate prior method we consider the Gamma(1,0.01) prior for τ.

Priors used for the between study variance parameter in a meta-analysis are also appropriate priors for the commensurability parameter. An informative prior is required to induce borrowing across studies in a meta-analysis when there is only a small amount of data available to inform on the between study variance parameter. Hobbs et al. [19] discuss the connection between the commensurate prior and meta-analysis. Priors of the form Gamma$(1, \beta)$ and Gamma(ϵ, ϵ) were also considered by Viele et al. [42] for the between study variance parameter in hierarchical modelling to incorporate historical data.

In WinBUGS [47], the commensurate prior, as it was originally proposed, would be fitted using the code in Figure 2.3(a). In this specification of the model, the parameters p_h and p_c are jointly estimated. This joint specification means that the current control data informs inference on p_h. In this thesis, the current trial data is considered the most reliable information and the aim is to assess the agreement between the historical and current control data and to utilise the historical data in estimating p_c when the current and historical control data agree. A model where the historical data inform inference on p_c but does not allow feedback from the current data model back to the historical data model may be more appropriate. We explore using the cut function in WinBUGS which was designed for this purpose [17]. The model code for the commensurate prior using the cut function is shown in Figure 2.3(b). For the example model given in Figure 2.3(b), the cut function makes a copy of ψ_h^*, but otherwise severs the link between ψ_h^* and ψ_h. Hence, ψ_h always has the same value as ψ_h^* but ψ_h^* is isolated from X_c and cannot be influenced by it [17]. Possible alternative approaches to cutting the feedback from the current control data to the historical model were explored, such as a two step procedure, where the historical data model is estimated and the mean of the prior for ψ_c is fixed at the mean of the posterior distribution of ψ_h with precision τ. A fully Bayesian version was also considered, taking the posterior samples from the distribution of ψ_h as the mean of the prior distribution for ψ_c. However, only the cut function approach stopped the feedback from the current data model to the historical data model.

Figure 2.3: Commensurate prior model WinBUGS code (a) Standard approach and (b) using the cut function

```
model {
    Xc ~ dbin(pc, nc)
    Xh ~ dbin(ph, nh)
    Xt ~ dbin(pt, nt)
    logit(pc) <- psic
    logit(ph) <- psih
    logit(pt) <- psit

    psic ~ dnorm(psih, tau)
    psih ~ dnorm(0,0.001)
    psit ~ dnorm(0,0.001)
    tau ~ dgamma(1,0.1)

    prob <- step(pt-pc)
    trt  <- pt-pc

}
```

(a)

```
model {
    Xc ~ dbin(pc, nc)
    Xh ~ dbin(ph, nh)
    Xt ~ dbin(pt, nt)
    logit(pc) <- psic
    logit(ph) <- psihstar
    logit(pt) <- psit

    psic ~ dnorm(psih, tau)
    psihstar ~ dnorm(0,0.001)
    psit ~ dnorm(0,0.001)
    tau ~ dgamma(1,0.1)

    psih <- cut(psihstar)
    prob <- step(pt-pc)
    trt  <- pt-pc
}
```

(b)

Once the commensurate prior model has been fitted in WinBUGS, the ESS of the control response probability posterior distribution needs to be determined to assess how much information from the historical data has been incorporated into the posterior distribution of p_c. Hobbs et al. [20] propose using the method of Morita et al. [26] to determine the $EHSS$, which assumes a normal approximation with known variance for the posterior distribution of the control response probability. The effective historical sample size is given by [20],

$$EHSS \approx n_c \left\{ \frac{prec(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))}{prec(\pi(p_c \mid x_c, y_c))} - 1 \right\}, \tag{2.6}$$

where n_c is the number of current controls randomised, $prec(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))$ denotes the posterior precision of the control parameter under the model incorporating historical data and $prec(\pi(p_c \mid x_c, y_c))$ denotes the posterior precision of the control parameter under the model not incorporating any historical data. Equation 2.6 is derived assuming a normal approximation with known variance for the distribution of the response probability. The Morita algorithm [26] compares the information of the posterior distribution that we want to approximate the ESS of, to the information of a distribution with known sample size. Assuming a normal approximation, the information of a normal distribution with known variance is $1/\sigma^2$ and the information of a normal likelihood with known variance is n/σ^2. Minimising the distance between the information of these two distributions gives the ESS to be the ratio of the known variance in the likelihood to the variance of the posterior distribution. The posterior distribution of the current control parameter obtained from fitting the commensurate prior model not incorporating any historical data is used as the known distribution, the ESS of the control posterior distribution incorporating the historical data is then derived as follows,

$$\begin{aligned} n \approx \frac{\sigma^2}{\mathrm{Var}(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))} &= \frac{n_c/prec(\pi(p_c \mid x_c, y_c))}{\mathrm{Var}(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))} \\ &= \frac{n_c prec(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))}{prec(\pi(p_c \mid x_c, y_c))} \end{aligned}$$

$$\implies EHSS = \frac{n_c prec(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))}{prec(\pi(p_c \mid x_c, y_c))} - n_c = n_c \left\{ \frac{prec(\pi(p_c \mid x_h, y_h, x_c, y_c, \tau))}{prec(\pi(p_c \mid x_c, y_c))} - 1 \right\} \tag{2.7}$$

For binary outcome data, both the mean and variance are dependent on the response probability. For low and high observed response proportions the variance of the posterior distribution of the response probability is small and therefore the precision is large. The normal approximation of the posterior distribution works well for the range of response proportions where the precision under the reference model is reasonably constant, but not at the extremes, since the normal approximation is using the relative difference in precisions. Depending on the disagreement between the historical and current controls and the prior on τ, incorporating historical data will bias the current control response probability estimate, which will also affect the variance. The change in variance from the reference model to the model incorporating historical data for binary data is dependent on both the bias and the additional patients from the historical data.

The second approach to approximating the ESS of the posterior distribution of p_c is to use the method proposed by Schmidli et al. [23] for the robust meta-analytic prior [23], where the posterior distribution of the control response probability is approximated by a mixture of beta distributions and the Morita algorithm is used to approximate the ESS of this mixture distribution. We explore whether this approach can be used for the commensurate prior.

Approximating the posterior distribution of the commensurate prior model using a single beta/mixture of beta distributions

For this approach, the commensurate prior model is fitted in WinBUGS, the Markov chain Monte Carlo (MCMC) samples of the posterior distribution of the control parameter are then used to optimise the parameters of a mixture of beta distributions. We explore whether a one-component or two-component mixture distribution provides the best approximation to the kernel density estimate of the posterior distribution of the control response probability, constructed from the MCMC samples. We use the Kullback-Liebler divergence [48], which is a measure of how one probability distribution diverges from another, expected probability distribution, to assess whether a one-component or two-component mixture distribution is a better approximation to the kernel density estimate. The density estimate is constructed on the logit scale and transformed to the probability scale to ensure the estimated density values lie between zero and one. Here, the density estimate of the control posterior distribution is constructed as follows.

Given the M posterior sample estimates for the current control response probability on the logit scale, which are denoted by $\hat{\psi}_c^{(1)}, ..., \hat{\psi}_c^{(M)}$. The density estimate, $\pi(\psi_c)$, is a mixture of normal distributions,

$$\pi(\psi_c) = \frac{1}{Mh} \sum_{i=1}^{M} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{\left(\frac{\psi_c - \hat{\psi}_c^{(i)}}{h}\right)^2}{2}\right)},$$

where
$h = \frac{0.8r}{M^{1/5}}$ and $r = min\left(\sqrt{\mathrm{Var}(\psi_c)}, \frac{IQR(\psi_c)}{1.349}\right)$.

This is the standard approach for calculating univariate density values in Stata [49, 50].

The density is estimated on the logit scale and transformed to the probability scale using the change of variables method to ensure that the estimated density values lie between zero and one. The Kullback-Liebler divergence is then used to compare the density estimate as the true distribution with the mixture of beta distributions approximation obtained from optimisation using the Broyden-Fletcher-Goldfarb-Shanno technique implemented in Mata [46]. Letting, $\psi_c = g(p_c) = log\left(\frac{p_c}{1-p_c}\right)$ and $p_c = g^{-1}(\psi_c)$, then the density values on the probability scale are given by,

$$\pi(p_c) = \pi(g(p_c)) \left| \frac{dg(p_c)}{dp_c} \right|,$$

where

$$\begin{aligned} g(p_c) &= log(p_c) - log(1 - p_c) \\ \implies \frac{dg(p_c)}{dp_c} &= \frac{1}{p_c} + \frac{1}{(1 - p_c)}. \end{aligned}$$

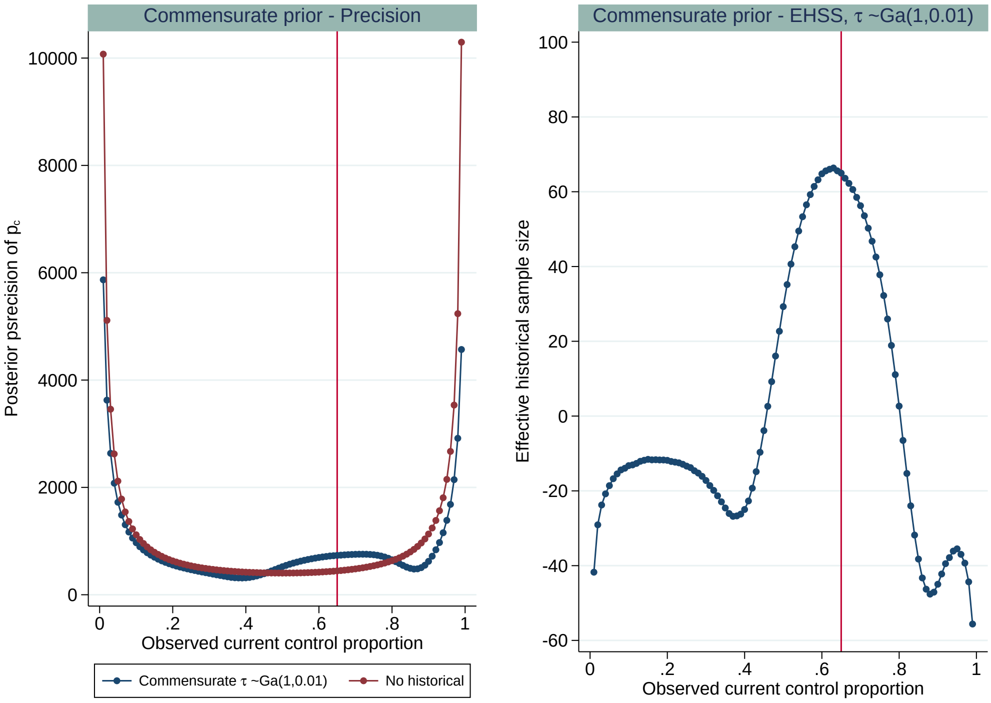
The accuracy of the mixture of beta distribution approximation to the posterior distribution will vary depending on the observed response proportion. Depending on the number of control patients in the current trial, it may be computationally intensive to determine the $EHSS$ for all possible responses in the current trial, especially if a two-component mixture of beta distributions approximation is used and the Morita algorithm is required to calculate the $EHSS$.

For the Viele example, a single beta distribution provides a reasonable approximation to the posterior distribution of the current control proportion and the ESS of a beta distribution is the sum of the parameter values, which simplifies determining the $EHSS$. A two-component mixture of beta distributions may provide a better approximation of the posterior distribution but adds complexity for the optimisation, and the Morita algorithm described in Section 2.2.3 is required to calculate the ESS of a mixture of beta distributions.

Example of implementing the commensurate prior for the Viele example

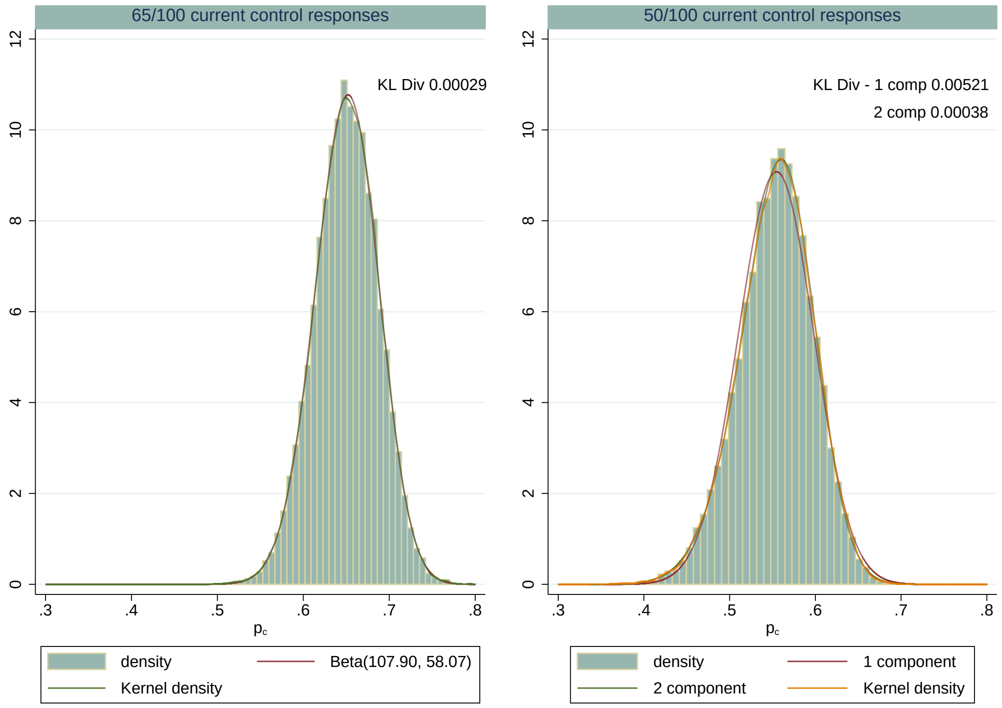
Figure 2.4 shows the precision of the posterior distribution of p_c under a model assuming no historical data and under the commensurate prior model. For the model assuming no historical data are available a normal prior distribution with mean 0 and precision 0.001 was assumed for p_c. Figure 2.4 also shows the $EHSS$ for the commensurate prior model for different observed response proportions in the current controls, assuming 65 responses out of 100 patients in the historical data and 100 current controls. A Gamma(1,0.01) prior is assumed for τ and the standard commensurate prior model is fitted, not using the cut function to stop the feedback from the current data model to the historical data model. At low and high observed response proportions in the current controls, small changes in the estimated response proportion result in large changes in the precision. The normal approximation of the $EHSS$ given in Equation 2.6 breaks down at the extremes because it is based on the relative difference in precision of p_c under the model with and without historical data. Furthermore, since the standard commensurate prior approach draws the estimated current and historical parameter estimates towards each other, the maximum $EHSS$ does not occur at complete agreement, but when the observed current control response proportion is slightly lower than the historical response proportion in this example. The $EHSS$ is estimated to be below zero for some observed current control response proportions because the variance of the posterior distribution of p_c is increased from incorporating historical data into the model that is different from the current control data. The increased variance is largest for an observed current control response proportion of approximately 0.4 and 0.85 where there are substantial differences between the current and historical response proportions but some of the historical data are still incorporated into inference on p_c.

Figure 2.4: Left graph shows the posterior precision of p_c for different observed current control proportions when using the commensurate prior model and a model incorporating no historical data. Right graph shows the $EHSS$ (calculated using Equation 2.6) using the commensurate prior model for different observed current control proportions. Viele example, historical data 65/100 responses, 100 current controls. The vertical red lines represent complete agreement between the historical and current control proportions.



Approximating the posterior distribution of p_c by a single beta distribution or a two-component mixture of beta distributions is explored in Figure 2.5 using the original commensurate prior model and in Figure 2.6 using the commensurate prior model with the cut function. The commensurate prior approach here is illustrated having observed 65 responses out of 100 in the current controls (complete agreement with the historical data) and also having observed 50 responses out of 100 in the current controls (substantial disagreement). A Gamma(1,0.01) prior is assumed for τ.

Figure 2.5: MCMC samples, kernel density and beta distribution approximations of the control posterior distribution for the standard commensurate prior model for two different observed current control response proportions, assuming a Gamma(1,0.01) prior for τ. Historical data 65/100 responses, 100 current controls.



one-component - Beta(71.10, 57.30), two-component - $0.68\text{Beta}(99.45, 62.70) + 0.32\text{Beta}(76.34, 55.90)$

Figure 2.5 shows the posterior MCMC samples, kernel density estimate and the beta distribution approximation of the control posterior distribution for 65/100 and 50/100 current control responses. At complete agreement between the current and historical data, a single beta distribution approximates the control posterior distribution well. The ESS of the control posterior distribution at complete agreement is approximately 166. For 50 current control responses out of 100, a two-component mixture of beta distributions provides a better approximation to the control posterior distribution than a single beta distribution. The Morita algorithm is then required to determine the ESS of the two-component mixture distribution. The ESS for the two-component mixture distribution is 146 compared to the ESS of 128 from using the single beta distribution approximation.

The control posterior distributions obtained in Figure 2.5 are using the original commensurate prior model which allows feedback from the current control data to inform about the historical parameter values. This can be seen from the summary statistics of the control posterior distribution when there is disagreement between the historical and current controls, given in Table 2.3.

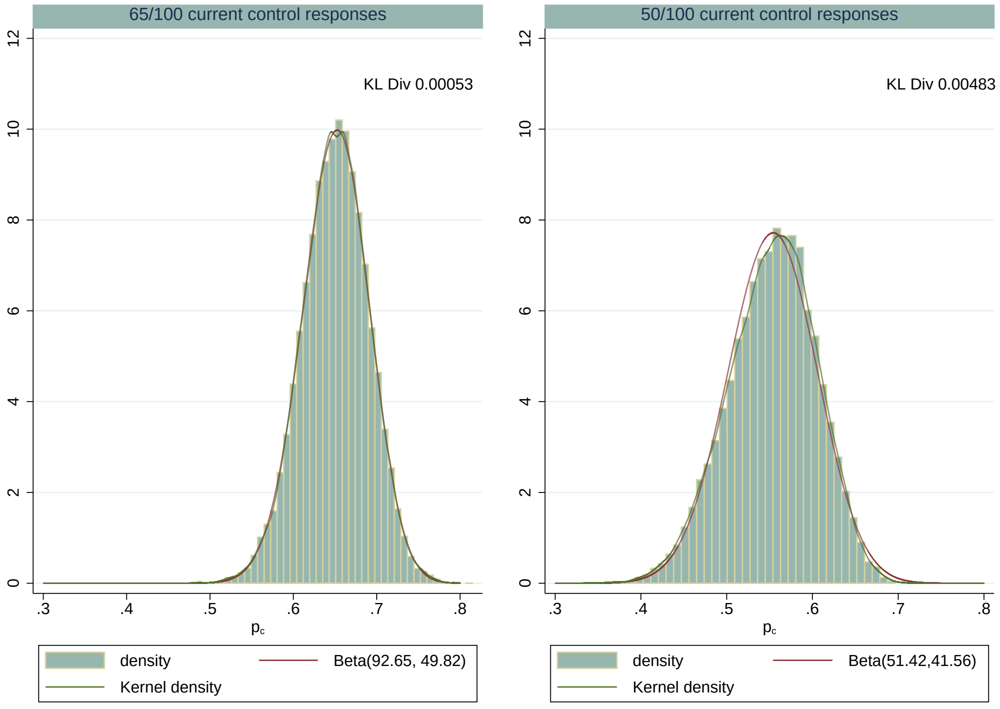
Table 2.3: Posterior summaries for the standard commensurate prior with 50/100 current control responses. Historical data 65/100 responses.

Parameter	mean	sd	MC error	2.5%	median	97.5%	burn in	sample
p_h	0.596	0.043	0.0003	0.516	0.595	0.685	1000	50000
p_c	0.554	0.044	0.0003	0.462	0.556	0.634	1000	50000
τ	80.180	91.829	0.5639	1.842	47.760	334.397	1000	50000

From the posterior summaries in Table 2.3, the historical posterior mean is drawn down from 0.65 towards the observed response proportion in the current control data of 0.5 and similarly the current control posterior mean is drawn slightly towards the historical response proportion.

Using the cut function, we obtain the control posterior distributions and posterior summaries given in Figure 2.6 and Table 2.4.

Figure 2.6: MCMC samples, kernel density and beta distribution approximations of the control posterior distribution for the commensurate prior model using the cut function for two different observed current control response proportions, assuming a Gamma(1,0.01) prior for τ. Historical data 65/100 responses, 100 current controls.



From the posterior summaries given in Table 2.4, the mean of the historical data parameter is now the estimated response proportion in the historical data. Similar to the original specification of the commensurate prior, the mean of the current control response

Table 2.4: Posterior summaries for the commensurate prior model using the cut function with 50/100 current control responses. Historical data 65/100 responses.

Parameter	mean	sd	MC error	2.5%	median	97.5%	burn in	sample
p_h	0.650	0.047	0.0002	0.555	0.651	0.739	1000	50000
p_c	0.553	0.051	0.0003	0.447	0.556	0.646	1000	50000
τ	34.278	58.427	0.3018	0.690	12.680	204.598	1000	50000

probability parameter has been drawn closer to the historical parameter estimate. Comparing Table 2.4 to Table 2.3, when using the cut function, the standard deviation of the posterior distribution of p_c increases. The standard deviation increases because there is no feedback from the current control data to the historical parameter estimate and therefore the posterior estimate of p_h is not drawn closer to the posterior estimate of p_c when using the cut function, the posterior estimate of τ is then smaller, and less historical information is used to inform inference on p_c. Figure 2.6 shows that the single beta distribution approximation of the control posterior distribution when there were 50 current control responses is not very close to the kernel density estimate, with a Kullback-Liebler divergence measure of 0.0048. However, for this example, numerical optimisation of the two-component mixture distribution converged to a mixture distribution that was not close to the kernel density estimate and therefore did not provide a better approximation. The effective sample sizes at complete agreement and a 15% difference in the historical and current control response proportions are 142 and 93 respectively.

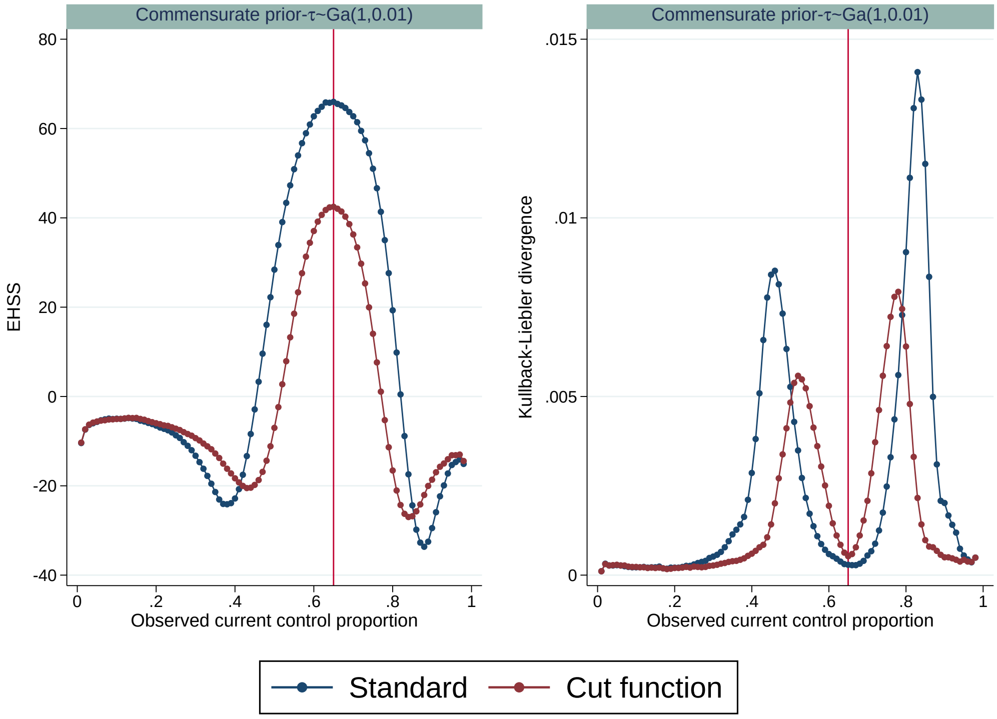
The two-component mixture distribution is difficult to optimise for some observed current control response proportions and sometimes converges to a distribution that is not close to the kernel density estimate. Figure 2.7 illustrates the $EHSS$ obtained using a single beta distribution approximation for all possible current control response proportions for the Viele example. The $EHSS$ is given assuming a Gamma(1,0.01) prior for τ using the standard commensurate prior model and the commensurate prior model using the cut function. The Kullback-Liebler divergence is also given to show the error in each beta distribution approximation to the posterior distribution.

From Figure 2.7, the commensurate prior model using the cut function results in a smaller $EHSS$ at complete agreement between the current and historical data compared to the standard commensurate prior model. The historical response probability is estimated less precisely without the current control data feeding back into inference on the historical data parameter.

2.2.3 Robust mixture prior

For one historical study, the robust mixture prior [23] is a two-component mixture prior, with a mixture component based on the historical data and a weakly-informative mixture

Figure 2.7: Commensurate prior model $EHSS$ (left) and Kullback-Liebler divergence (right) for different observed current control response proportions using the standard commensurate prior model and the commensurate prior model using the cut function, $\tau \sim \text{Gamma}(1, 0.01)$. Viele example, historical data 65/100 responses, 100 current controls. The vertical red lines represent complete agreement between the historical and current control proportions.



component,

$$\pi(p_c \mid x_h, y_h, w) = w\text{Beta}(x_h, y_h) + (1 - w)\text{Beta}(1, 1),$$

where the weight, $w \in [0, 1]$, is pre-specified.

The posterior distribution for the response probability in the control arm is then a mixture of beta distributions with updated parameter values and weights [23],

$$\pi(p_c \mid x_h, y_h, x_c, y_c, w) = \tilde{w}\text{Beta}(x_h + x_c, y_h + y_c) + (1 - \tilde{w})\text{Beta}(1 + x_c, 1 + y_c),$$

where,

$$\tilde{w} \propto \frac{w \, \frac{\text{B}(x_h + x_c, y_h + y_c)}{\text{B}(x_h, y_h)}}{w \frac{\text{B}(x_h + x_c, y_h + y_c)}{\text{B}(x_h, y_h)} + (1 - w) \, \frac{\text{B}(1 + x_c, 1 + y_c)}{\text{B}(1, 1)}}.$$

No recommendation is given for choosing the weight parameter and therefore a sensitivity analysis is required to determine how much weight is given to the historical data in the final analysis for different initial weights of the prior mixture distribution [23].

The agreement between the historical and current controls is summarised using the ESS. The ESS is determined by calculating the ESS of the control posterior distribution incorporating both the historical and randomised controls in the current trial and subtracting the number of randomised controls [23]. The ESS of the posterior distribution is calculated using the method of Morita et al. [26], which is outlined below.

For binary data, the Morita algorithm searches for a distribution with known sample size that has the same information at the mean as the posterior distribution for the current control response probability distribution. The sample size of the posterior distribution is then estimated to be the sample size of the created distribution. Here we use the algorithm by Morita but the information is compared at the mode of the distribution as proposed by Schmidli et al. [23]. Comparing the information at the mode is likely to be the reason the ESS from the one-component and two-component mixture distributions differ when the densities look similar in Figure 2.5. The original paper by Morita et al. [26] shows that when comparing the information at the mean, the ESS matches the commonly used ESS value for the beta distribution of the sum of the parameter values [26]. However, the mode is used here since this seems a more appropriate measure when the posterior mixture distribution may be bimodal.

The control response probability posterior distribution is a mixture of beta distributions and the mode is found by searching over the function values to find the maximum. The observed information at the mode is then approximated using a quadratic approximation. Let $\tilde{p}_c$ be the mode of the current control posterior distribution. The observed information is defined as,

$$I = -\left.\frac{d^2 log\pi(p_c)}{dp_c^2}\right|_{p_c=\tilde{p}_c}$$

The second derivative is calculated using a quadratic approximation. This is the quadratic function whose first and second derivatives are the same as those of the distribution of p_c at the mode. The formula is given by,

$$\pi(p_c) \approx \pi(\tilde{p}_c) + \dot{\pi}(\tilde{p}_c)(p_c - \tilde{p}_c) + \frac{1}{2}\ddot{\pi}(\tilde{p}_c)(p_c - \tilde{p}_c)^2.$$

This quadratic approximation is the second-order Taylor polynomial for the posterior mixture distribution of p_c at $p_c = \tilde{p}_c$.

The observed information of the control response probability posterior distribution is compared to a posterior distribution constructed from a non-informative prior and has a known sample size. The expected information for the posterior distribution with sample size m under a weakly-informative prior is,

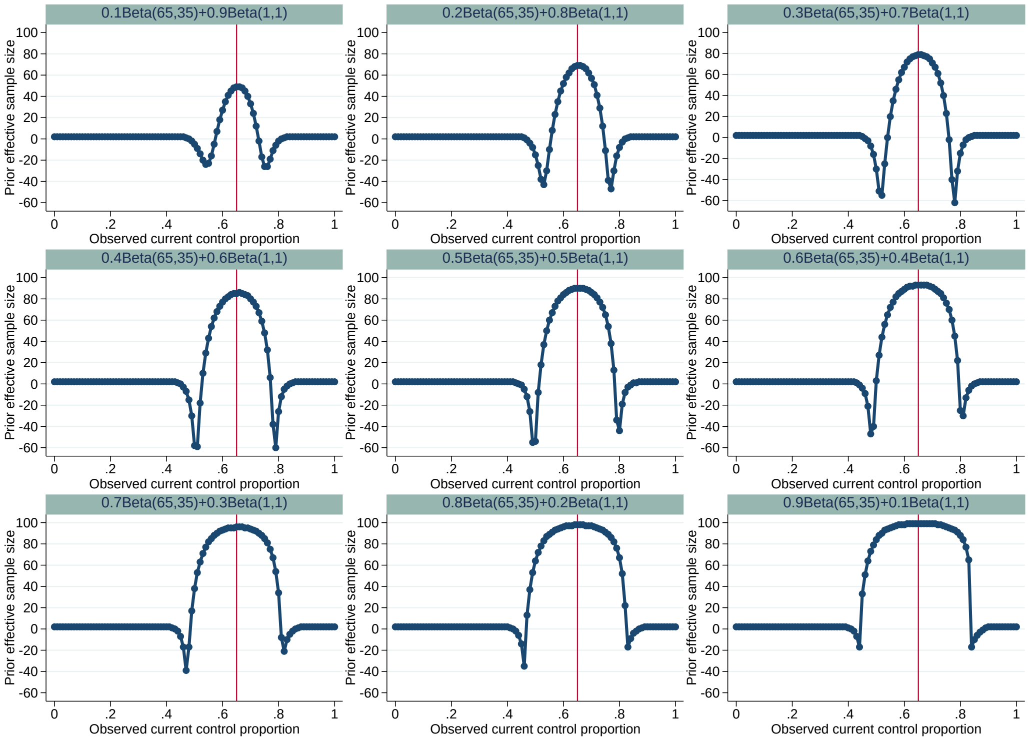
$$\mathbf{E}(I_0(m)) = -\sum_{x_c=0}^{x_c=m} \left\{ \left. \frac{d^2 log(\pi_0(p_c \mid x_c))}{dp_c^2} \right|_{p_c=\tilde{p}_c} \right\} \pi(x_c).$$

where $\pi_0(p_c \mid x_c)$ is a beta distribution with sample size m and $\pi(x_c)$ is the prior predictive distribution with respect to the informative prior $\pi(p_c)$, which is a mixture of beta-binomial distributions. As an initial non-informative prior we choose a Beta$(\tilde{p}_c/c, (1 - \tilde{p}_c)/c)$, where c is a large constant, here it is chosen to be 100. We then loop over all sample sizes m up to a reasonable maximum, which is chosen to be the sum of the beta component of the mixture distribution with the largest parameter values plus 10. The ESS is the largest m such that $\mathbf{E}(I_0(m)) < I$ [23].

Figure 2.8 shows the prior ESS for different observed response proportions in the current controls and different initial mixture prior weights. Here we have assumed there are 100 historical controls with an observed response proportion of 65% and 100 current controls. The prior ESS is n_h when $\tilde{w}$ is one and the prior ESS is two when $\tilde{w}$ is zero. Ideally, a Beta(0,0) non-informative component would be used in the mixture prior, giving a prior ESS of zero when zero weight is given to the informative component. The historical data weight would then be $EHSS/n_h$, however a Beta(0,0) is an improper distribution and cannot be used in the mixture prior.

A prior that gives an initial low weight to the informative historical data component of the mixture prior, for example a weight of 0.1, gives a low prior ESS even when there is complete agreement between the current and historical controls. A low initial prior weight also discounts quickly as the difference in the current and historical response proportions increases. When an initial weight greater than 0.8 is given to the informative component, at complete agreement between the current and historical controls all of the historical data are incorporated into the final analysis but the prior ESS remains high even when there are substantial differences between the current and historical controls. Similar to the commensurate prior approach, incorporating historical data can increase the variance of the posterior distribution of the control parameter compared to using no historical data, resulting in a negative prior ESS. For the example in Figure 2.8, calculating the ESS for all observed current control proportions for a single initial mixture prior weight took approximately one hour.

Figure 2.8: Robust mixture prior effective sample size for different observed current control response proportions. Viele example, historical data 65/100 responses, 100 current controls. The vertical red lines represent complete agreement between the historical and current control proportions.



2.2.4 Limitations of published historical data methods

The three methods proposed in the literature for incorporating historical control data into the analysis of a current trial: power priors; commensurate priors; and robust mixture priors all have their disadvantages. These are split here into the disadvantages in a non-adaptive design setting and an adaptive design setting, focusing on when there is only one historical study. For one historical study, it is particularly difficult to assess whether there is agreement between the historical and current control data. Similar to the meta-analysis problem of assessing the between study variance with only two studies, strong priors are required on these variance parameters as they are not well estimated from the data. All of the historical data methods proposed have this problem, strong priors are required on the commensurability parameter, τ, and the power, α_0, to induce borrowing. The initial weights chosen for the robust mixture prior have a strong influence on how the historical data are utilised. With only one historical study, it is required that prior knowledge is used to inform the parameters that govern how much historical data are borrowed and how quickly the historical data are discounted when there is disagreement. There are differences between all the methods in how intuitive these parameters are to choose and how easy it is to implement the method. These differences are discussed in the next sections.

Non-adaptive design

The main difference in using historical data methods for an adaptive design versus a non-adaptive design is that for an adaptive design calculation of the prior ESS given the current trial data is required at an interim analysis of the trial. The prior ESS indicates how much of the historical data will be incorporated into the final analysis given the agreement between the historical and current controls at that interim analysis.

For the commensurate prior and the robust mixture prior it can be difficult to determine the $EHSS$, this requires calculating the ESS of the control response probability posterior distribution which does not have a known distribution. However, for a non-adaptive design, the analysis approach for the commensurate prior and robust mixture prior which calculates the probability that the response in treatment is greater than control, does not require the explicit calculation of the $EHSS$, as illustrated in Figure 2.3 and Section 2.5.7. However, it would be useful to know the $EHSS$ and how much of the historical data are incorporated into the final analysis. Calculating the ESS of the fully Bayesian modified power prior was not considered in this thesis and calculating the probability that the treatment response probability is greater than the control response probability for the fully Bayesian modified power prior is computationally intensive due to the number of integrations required.

As mentioned previously, choosing a flat prior for the commensurability parameter and the power prior does not induce sufficient borrowing of the historical data when there is complete agreement between the historical and current controls, since with only one historical study there is not enough information to learn about these parameters from the data. Quasi-dichotomous priors for the power in the modified power prior approach can lead to a marginal posterior distribution that is difficult to integrate. For the commensurability parameter, for a certain range of disagreement between the historical and current controls, the variability of the posterior distribution for the control response probability can be increased compared to the posterior distribution obtained from a design not using any historical data.

The commensurate prior requires MCMC, each model was fitted in WinBUGS, this increases the computation time for calculating the operating characteristics compared to a design where a conjugate analysis is possible. For the commensurate prior with only one historical study, the cut function stops the feedback from the current control data to the historical parameter, but it is unclear how cutting this feedback impacts the estimation of τ.

Finally, the choice of prior for the commensurability parameter or the power prior is not intuitive and the operating characteristics for each prior need to be considered which can be computationally intensive. For the robust mixture prior, the operating characteristics can be determined quickly as the posterior distribution is a mixture of beta distributions and a quick method to calculate the operating characteristics is explored in Section 2.5.1.

Adaptive design

For the power prior, given an observed number of responses in the current controls, the marginal distribution of the power has a known distribution and therefore a summary measure of the marginal posterior distribution can be used to calculate the $EHSS$. Using the mode of the marginal distribution of α_0 from a Beta(1,1) prior, at complete agreement between the current and historical data the marginal posterior distribution of the power is nearly flat but gradually increasing with α_0, the mode takes the maximum to give the historical data a weight of one, which is not an intuitive weight to give to the historical data. This approach however does give desirable weights, which are high at agreement between the historical and current controls and discount to zero quickly as the difference increases. There is no prior input from the designer of the discounting of the historical data using the mode. Taking the mean or median of the posterior distribution requires integration for the mean and integration and optimisation for the median and therefore can be computationally intensive. To induce increased borrowing of historical data when there is agreement a quasi-dichotomous prior was explored. As the prior parameters decrease to zero, these priors have an increased amount of mass at the tails of the distribution, which can cause problems with numerical integration. The choice of prior for the power prior and the rate at which the historical data are discounted are not intuitive.

Calculating the $EHSS$ for the commensurate prior is particularly difficult. The control response probability posterior distribution is not a known distribution and the distribution can not always be approximated well by a mixture distribution. Further this posterior distribution has to be approximated for each observed number of responses in the current controls. If an appropriate approximation can be found then the Morita algorithm can be used to calculate the ESS. This approach is computationally intensive. The choice of prior on the commensurability parameter is not intuitive and because for some observed current control responses, incorporating historical data can result in a posterior distribution with a larger variance than when not using any historical data, it is possible that the $EHSS$ calculated may be negative. In an adaptive design setting, this would require randomising more patients to control than a design that did not consider the historical data. The commensurate prior approach using the cut function and without the cut function give different $EHSS$ values.

Finally, for the robust mixture prior and only one historical study, the control response probability posterior distribution is a two-component mixture of beta distributions. The Morita algorithm is required to calculate the ESS of this posterior distribution which can be computationally intensive, the disadvantages of this algorithm, beyond computationally intensive, are given below. Similar to the commensurate prior, the variance of the control posterior distribution can be increased when there is disagreement between the historical and current controls and can result in a negative prior ESS. Exploring the prior ESS for a given initial robust mixture prior weight and observed current control response probability is the most intuitive way to see the effect of the choice of the initial prior weight. However, this can be computationally intensive, depending on the number of initial weights explored.

The commensurate prior and the modified power prior using the mean of the marginal distribution of the power as a weight, do not give a weight of one to the historical data or an $EHSS$ equal to n_h, even when the historical and current controls completely agree. It is possible that the robust mixture prior can give an $EHSS$ equal to n_h if a high initial weight is given to the informative component of the robust mixture prior when there is only one historical study.

The Morita algorithm is used to determine the ESS of a parametric prior, this algorithm approximates the information of the robust mixture prior at the mode and matches this to the information of a distribution with known sample size, which then becomes the ESS of the prior. The Morita algorithm does not compare the whole distribution to a known distribution and it is unclear how accurate the ESS approximation is from only comparing the information at the mode when the distribution is a mixture distribution and the sample size is unknown. The original paper [26] justifies their approach by deriving the correct sample size estimates using their approach on distributions with a known sample size. The Morita algorithm also requires looping over all sample sizes and possible responses in a sample size up to a chosen maximum which is computationally intensive.

Similar to the robust mixture prior, using the fully Bayesian version of the commensurate prior with one interim analysis only requires the $EHSS$ to be calculated at the interim analysis. The power and type I error of these designs can be calculated at the end of the study, without calculating the ESS of the posterior distribution at the end of the study. However, the amount of historical information utilised in the final analysis is then not known.

Given the limitations of the historical data methods proposed in the literature, new methods for incorporating historical data are required that are intuitive and simple. These approaches need to allow control over the maximum possible type I error rate and the reduction in power when there is conflict between the historical and current control data.

Two methods are proposed in Section 2.3 for assessing the agreement between historical and current control data, an equivalence probability weight and a weight based on tail area probabilities.

2.3 Assessing agreement between historical and current control data – probability and equivalence probability weight

We now propose two novel methods that calculate a weight, between zero and one, where zero represents no historical data borrowing and one represents pooling of the historical and current data. The aim of these approaches is to obtain a high weight when there is agreement between the historical and current control data and also to recognise conflict quickly and discount the historical data, obtaining a low weight when there is disagreement between the historical and current controls.

2.3.1 Probability weight

Assuming beta distributions for the historical and current control response proportions, $p_h \sim \text{Beta}(x_h, y_h)$ and $p_c \sim \text{Beta}(x_c, y_c)$.

The probability weight is given by,

$$w = 2 \times min\{\Pr(p_c > p_h), 1 - \Pr(p_c > p_h)\}, \tag{2.8}$$

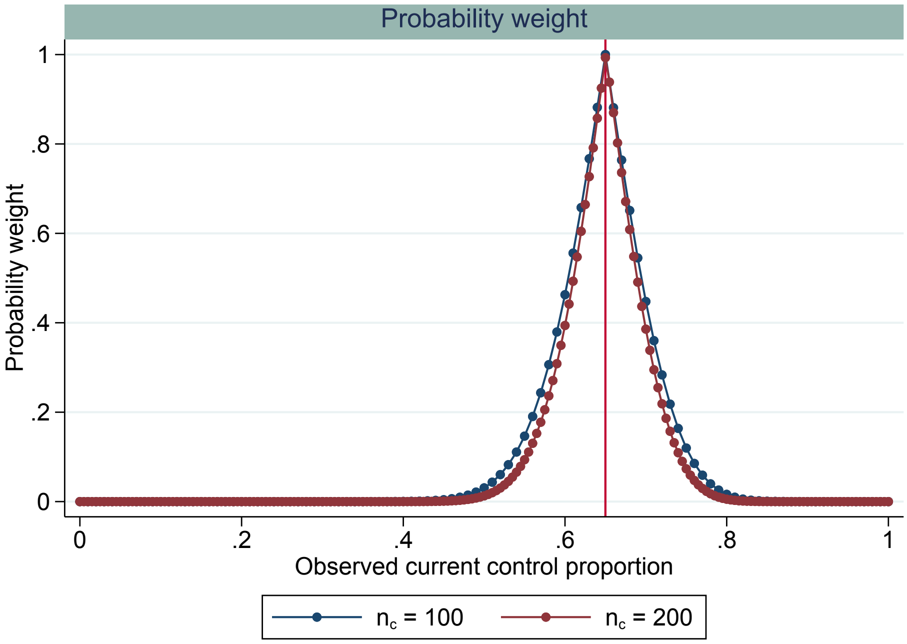
as proposed by Thompson [51] where,

$$\Pr(p_c > p_h) = \int_0^1 \int_{p_h}^1 \frac{p_h^{x_h-1}(1-p_h)^{y_h-1}}{\mathrm{B}(x_h, y_h)} \frac{p_c^{x_c-1}(1-p_c)^{y_c-1}}{\mathrm{B}(x_c, y_c)} dp_c dp_h. \tag{2.9}$$

A quick method to calculate the probability weight is described in Section 2.5.1. Note that doubling the $\Pr(p_c > p_h)$ is an approximation since the beta distribution is not always symmetric.

The probability weight is illustrated in Figure 2.9. Using the probability weight approach, when the historical and current data completely agree, the historical data are given a high weight in the final analysis. The probability weight decreases quickly to zero as the difference in response proportions increases. At a 15% difference, the historical data are almost completely discounted. When the current control sample size is larger,

Figure 2.9: Probability weight for different observed current control response proportions and current control sample sizes of 100 and 200. Viele example, historical data 65/100 responses. The vertical red line represents complete agreement between the historical and current control proportions.



the historical data are discounted at a quicker rate as the difference between the current and historical response proportions increases. The disadvantage of this approach is that there is no flexibility in how quickly the historical data are discounted.

2.3.2 Equivalence probability weight

The second approach is the equivalence probability weight and we describe two equivalence weights. The one-sample approach assumes the historical data observed response proportion is the fixed truth and assesses the equivalence of the current controls to the fixed historical response proportion. The two-sample approach acknowledges that the historical data are a sample and incorporates this additional uncertainty, by assuming a distribution for both the historical and current control response proportions.

One-sample

Assuming the historical response probability is fixed at the historical sample estimate ($\hat{p}_h$). We choose an equivalence interval around the historical response probability ($\hat{p}_h - \delta, \hat{p}_h + \delta$), where δ is the equivalence bound. The weight is the probability that the current control response distribution lies within this interval. We assume a normal approximation to the beta distribution for the response probability in the current controls. The accuracy of the approximation of the normal distribution to the beta distribution in this setting is explored in Section 2.3.4.

The normal approximation for the beta posterior distribution is derived using Laplace's technique [52]. The Laplace technique is a simple 2-term expansion on the log probability density function (pdf). Let $q(p_c)$ denote the log pdf, $q(p_c) = log_e(\pi(p_c))$. Let $\hat{p}_c$ denote the maximum likelihood estimate of p_c, which is also the maxima of $q(p_c)$, then,

$$\begin{aligned} q(p_c) &\approx q(\hat{p}_c) + (p_c - \hat{p}_c)\dot{q}(\hat{p}_c) + \frac{1}{2}(p_c - \hat{p}_c)^2\ddot{q}(\hat{p}_c) \\ &= q(\hat{p}_c) + 0 + \frac{1}{2}(p_c - \hat{p}_c)^2\ddot{q}(\hat{p}_c) \\ &= constant + \frac{1}{2}(p_c - \hat{p}_c)^2\ddot{q}(\hat{p}_c) \\ &= constant - \frac{(p_c - \mu_c)^2}{2\sigma^2}. \end{aligned}$$

This matches the log pdf of a normal distribution with mean μ_c and variance σ^2.

Assuming a flat initial Beta(1,1) prior for the control response probability, the posterior distribution given x_c responses and $n_c - x_c$ non-responses in the current trial is given by [53],

$$\pi(p_c \mid x_c, n_c) \propto p_c^{x_c}(1 - p_c)^{n_c - x_c},$$

taking logs,

$$q(p_c) = x_c log_e(p_c) + (n_c - x_c)log_e(1 - p_c),$$

then,

$$\begin{aligned} \dot{q}(p_c) &= \frac{x_c}{p_c} - \frac{n_c - x_c}{1 - p_c} = 0 \implies \hat{p}_c = \frac{x_c}{n_c}, \\ \ddot{q}(p_c) \mid_{p_c=\hat{p}_c} &= -\frac{x_c}{p_c^2} - \frac{n_c - x_c}{(1 - p_c)^2} = -\frac{n_c}{\hat{p}_c(1 - \hat{p}_c)}, \end{aligned}$$

$$\hat{\sigma}^2 = \{-\ddot{q}(p_c) \mid_{p_c=\hat{p}_c}\}^{-1} = \frac{\hat{p}_c(1 - \hat{p}_c)}{n_c}.$$

A symmetric equivalence interval around the historical response probability is assumed throughout. The historical data weight is then given by,

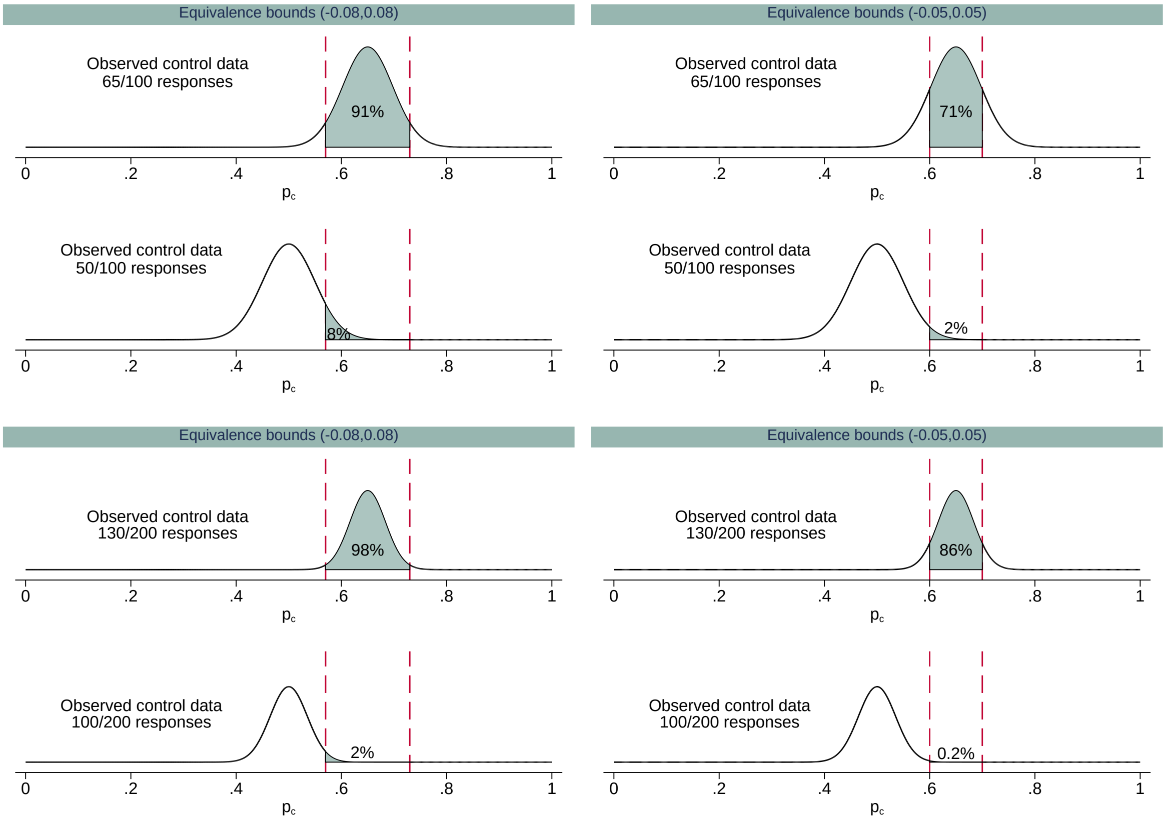
$$w = \Pr(\hat{p}_h - \delta \leq p_c \leq \hat{p}_h + \delta) \tag{2.10}$$

where,

$$\begin{aligned}\Pr(\hat{p}_h-\delta \le p_c \le \hat{p}_h+\delta) &= 1-\int_{-\infty}^{\hat{p}_h-\delta} \mathrm{N}\left(\hat{p}_c, \frac{\hat{p}_c(1-\hat{p}_c)}{n_c}\right) dp_c - \int_{\hat{p}_h+\delta}^{\infty} \mathrm{N}\left(\hat{p}_c, \frac{\hat{p}_c(1-\hat{p}_c)}{n_c}\right) dp_c \\ &= \Phi\left(\frac{\hat{p}_h+\delta-\hat{p}_c}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}}\right) - \Phi\left(\frac{\hat{p}_h-\delta-\hat{p}_c}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}}\right)\end{aligned} \tag{2.11}$$

Figure 2.10 illustrates the main features of the one-sample equivalence probability weight for different equivalence bounds and current control sample sizes at agreement between the historical and current controls (65% response proportion) and substantial disagreement (50% response proportion in the current controls).

Figure 2.10: One-sample equivalence probability weight when the historical and current controls completely agree and when they differ for 8% and 5% equivalence bounds, 100 and 200 current control patients, historical data 65/100 responses. The vertical dashed red lines represent the equivalence limits.



From Figure 2.10, comparing the 8% (i.e. $\delta = 0.08$) and 5% equivalence bounds. The larger equivalence bounds give a higher weight to the historical data, both when the historical and current controls agree and disagree. Comparing the historical data weight obtained when the current control sample size is 100 patients and when it is 200 patients.

When the current control sample size is larger, the normal approximation to the current control response probability has a smaller variance and is more peaked around the sample estimate and therefore for the same equivalence bounds a higher weight is obtained, since more of the current control response probability distribution lies within the equivalence bounds. A high weight is obtained when there is agreement and a low weight when there is disagreement between the historical and current controls. The two-sample approach has similar features, but is also dependent on the historical data sample size. The choice of equivalence bounds is key to controlling how much weight is given to the historical data when the current and historical control response proportions are in complete agreement.

The historical data estimate of the response probability is a sample estimate and the one-sample approach does not incorporate the variability around the estimated historical response probability. Therefore, we now explore a two-sample equivalence weight approach, where the variability around the historical data estimate is incorporated into calculating the historical data weight.

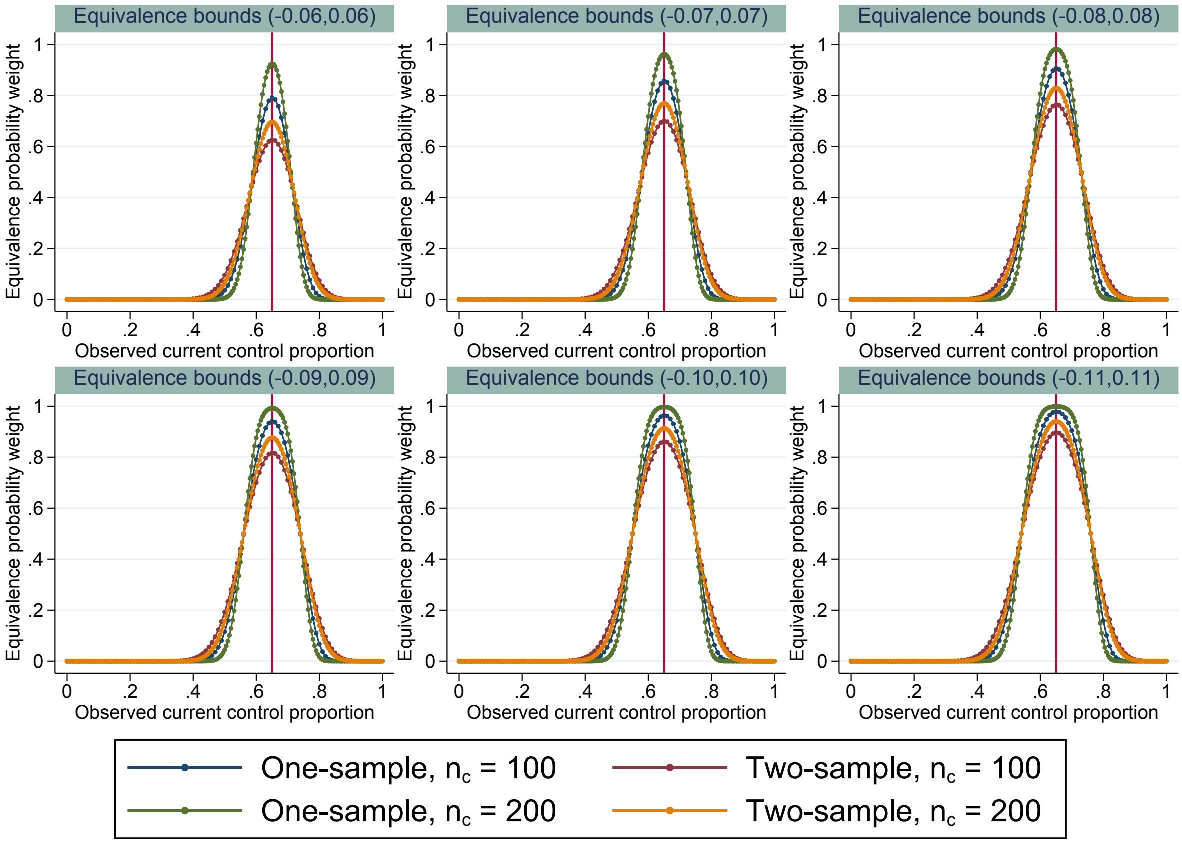
Two-sample

We assume a normal distribution approximation for both the posterior distribution of the response probability in the historical data and the current controls and calculate the probability that the difference distribution of the response proportions lies within the chosen equivalence bounds. The posterior distributions for p_c and p_h are derived assuming Beta(1,1) priors for both parameters. The two-sample equivalence probability weight is,

$$\begin{aligned} w &= 1 - \Pr(p_c - p_h > \delta) - \Pr(p_c - p_h < \delta) \\ &= \Phi\left(\frac{\delta - (\hat{p}_c - \hat{p}_h)}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c} + \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}}}\right) - \Phi\left(\frac{-\delta - (\hat{p}_c - \hat{p}_h)}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c} + \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}}}\right) \end{aligned} \tag{2.12}$$

Figure 2.11 shows the distribution of the one-sample and two-sample equivalence probability weights obtained for the Viele example assuming different observed control proportions in the current study. The historical data observed 65 responses from 100 patients and it is assumed there are either 100 or 200 controls available from the current study. For the one-sample equivalence probability weight, when there are 100 current control patients, 6% equivalence bounds gives a weight of 0.79 at complete agreement between the historical and current controls and at approximately a 15% difference, the historical data are given a weight of zero. Assuming larger, 11% equivalence bounds, when there are 100 current control patients, a weight of 0.98 is obtained at complete agreement which decreases to zero at approximately a 20% difference. When there are 200 current control patients available, the equivalence probability weight is larger at complete agreement and

Figure 2.11: One-sample and two-sample equivalence probability weights for different observed current control response proportions, different equivalence bounds and different current control sample sizes. Viele example, historical data 65/100 responses. The vertical red lines represent complete agreement between the historical and current control proportions.



discounts more quickly to zero as the difference between the current and historical control proportions increases compared to when there are 100 current control patients.

The main factors that affect the weight obtained for the one-sample equivalence approach are: the difference in response proportions between the current and historical data; the equivalence bounds chosen; and the sample size of the current control group.

The weights obtained from the two-sample equivalence approach are lower than the one-sample approach at complete agreement between the current and historical controls due to the variability incorporated from the historical data. The weight also decreases to zero at a slower rate than the one-sample approach as the difference between the historical and current controls increases. For the two-sample approach, the factors that affect the weight are: the difference in response proportions between the current and historical data;

the equivalence bounds chosen; the historical data sample size; and the current control sample size. The choice of equivalence bounds is explored further in Section 2.5.6.

The maximum equivalence probability weight does not always occur when there is complete agreement in the historical and current control response proportions. In all simulations explored, since the number of responses can only be integer values, if complete agreement in the response proportions is possible, given the sample sizes of the historical and current control data, the maximum weight was obtained at complete agreement. When the sample sizes in the current and historical controls differ and complete agreement in the response proportions between the current and historical controls is not possible, the maximum equivalence weight will occur when the current and historical control response proportions are similar, but the maximum equivalence weight may not be when the observed historical and current control proportions are closest. Where the maximum equivalence weight occurs depends on what the observed historical and current control proportions are, and the sample sizes of the historical and current control data, since the variance of the response probability distribution is dependent on the observed proportion and the sample size. In all simulation explored, the maximum equivalence weight is always close to agreement in the historical and current control proportions and the maximum weight is close to the weight obtained at complete agreement in the historical and current control proportions. Therefore, the equivalence probability weight has the desired property of giving the historical data a high weight when it is similar to the current control data.

2.3.3 Using a weight to discount the historical data

The probability weight and equivalence probability weight are used as a fixed power to down-weight the historical data. A rationale for using these probabilities to down-weight the historical data comes from the power prior with a fixed weight [14]. The power prior with a fixed weight is given by,

$$\pi(p_c, \alpha_0 \mid x_h, y_h) \propto L(p_c \mid x_h, y_h)^{\alpha_0} \pi_0(p_c).$$

If we consider a power prior of the form,

$$\pi(p_c, \alpha_0 \mid x_h, y_h) \propto L(p_c \mid x_h, y_h)^{I_{\{p_h \equiv p_c\}}} \pi_0(p_c),$$

where $I_{\{\hat{p}_h \equiv \hat{p}_c\}}$ is an indicator function of whether the historical and current control data are in agreement or not. This prior results in a pooled analysis of the current and historical control data when they “agree”, and when the historical and current data “disagree” the prior becomes the initial prior for p_c before the historical data are observed $\pi_0(p_c)$. In expectation, $\mathbf{E}(I_{\{p_h \equiv p_c\}}) = \Pr(p_h \equiv p_c)$, we use the expectation of the agreement as a plug-in value approximation of the agreement between the historical and current controls. The probability weight and equivalence probability weight are defined to be $w = \Pr(p_h \equiv p_c)$, to give the prior,

$$\pi(p_c, \alpha_0 \mid x_h, y_h) \propto L(p_c \mid x_h, y_h)^w \pi_0(p_c).$$

Then, for one historical study, the $EHSS$ for the probability and equivalence weight approaches is $w \times n_h$, as with the power prior approach.

2.3.4 Normal approximation to the beta distribution

For the equivalence probability weight approach, it is possible to directly use the beta distribution for the response probability in the historical and current controls rather than using a normal approximation. However, assuming a normal distribution gives a simple equation (Equation 2.12) for the two-sample equivalence probability weight whereas using the beta distribution would require numerical integration. Furthermore, choosing symmetric equivalence bounds seems intuitive when using the normal approximation of the difference in response proportions since when there is no difference between the observed current and historical control proportions the difference distribution will be symmetric and centred at zero. For small sample sizes the beta distribution may not be symmetric. The choice of equivalence bounds is discussed further in Section 2.5.6.

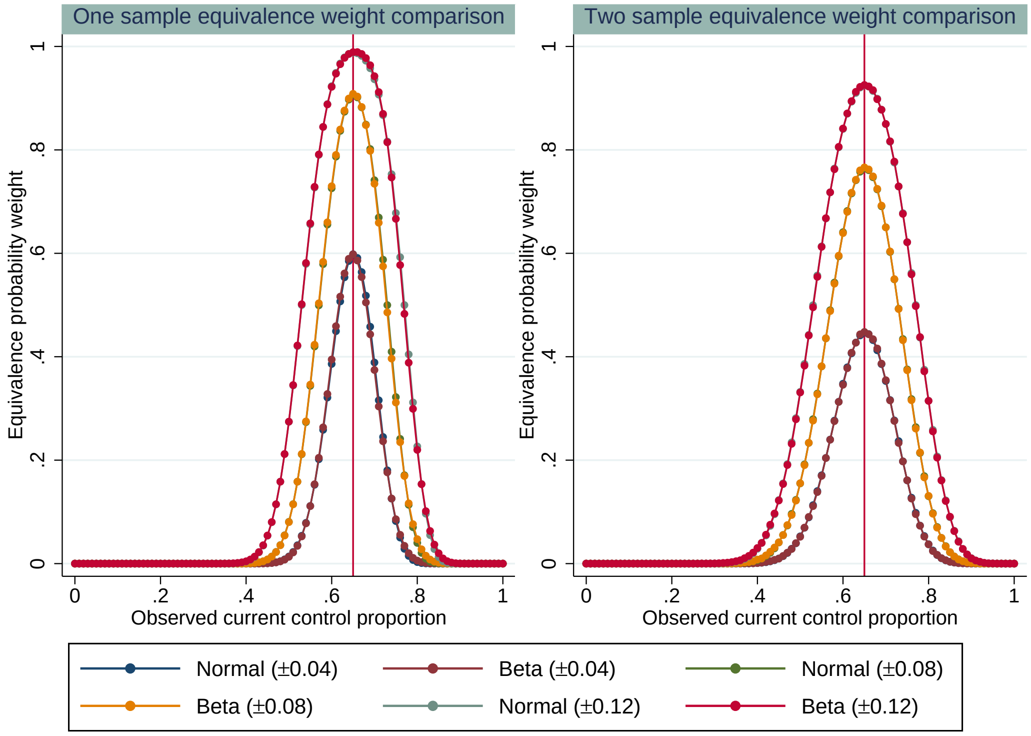
The skewness of the beta distribution with x_c responses and y_c non-responses is given by,

$$skewness = \frac{2(y_c - x_c)\sqrt{x_c + y_c + 1}}{(x_c + y_c + 2)\sqrt{x_c y_c}}.$$

When $x_c = y_c$ the skewness is zero and the beta distribution is symmetric. For $x_c < y_c$ the beta distribution is positively skewed and for $x_c > y_c$ the beta distribution is negatively skewed. As $x_c \to \infty$ and $y_c \to \infty$, the skewness tends to 0. The normal approximation to the beta distribution is best when the parameters of the beta distribution are equal or when both of the beta distribution parameter values are large.

For the Viele example, the equivalence approach using the beta distribution directly or using the normal approximation, both with symmetric equivalence bounds, give similar weights. Figure 2.12 illustrates this.

Figure 2.12: Equivalence probability weights for different observed current control proportions and different equivalence bounds when using a normal approximation to the beta distribution or the beta distribution directly. Viele example, historical data 65/100 responses, 100 current controls. The vertical red lines represent complete agreement between the historical and current control proportions.



The error in the normal approximation to each beta distribution can be explored using the difference in the cumulative distribution functions as suggested by Cook [54].

The error is then given by,

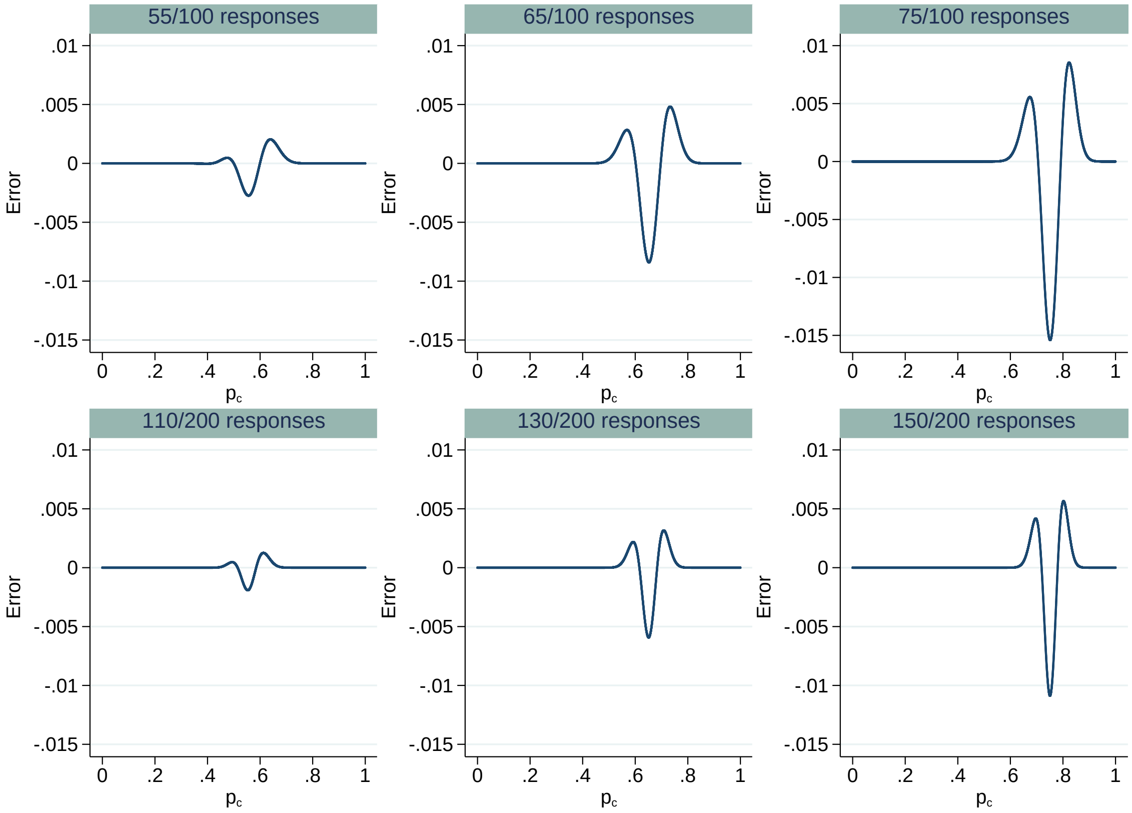
$$\text{Error} = I_x(x_c, y_c) - \Phi(z),$$

where $I_x(x_c, y_c)$ is the regularized incomplete beta function, Φ is the cumulative distribution function (CDF) of the normal distribution and $z = \frac{p_c - (\hat{p}_c)}{(\hat{p}_c(1-\hat{p}_c))/(x_c+y_c)}$.

Figure 2.13 shows the difference in the CDF of the beta distribution and the CDF of its normal approximation for response proportions of 55%, 65% and 75% and samples sizes of 100 and 200. The error is dependent on the observed response proportion as well as the sample size.

In summary, large equivalence bounds give more weight to the historical data at agreement and for a wider range of disagreement in the current and historical controls. The two-sample equivalence weight approach incorporates the additional uncertainty around the historical data sample estimate and therefore gives a lower weight to the historical data at complete agreement than the one-sample approach and discounts to zero more quickly as the difference increases. Symmetric equivalence bounds around the historical data are used since we want to down-weight the historical data for a difference in response proportions in either direction. If the historical response proportion is close to zero or one, the equivalence approach may not be plausible due to the need for symmetric equivalence bounds. The weights obtained from assuming a normal approximation to the beta distribution are similar to those obtained from directly using the beta distribution and the normal approximation is much simpler to use. However, if the response proportion is at the extremes of the probability scale and the sample size is small, the accuracy of the normal approximation will need to be explored further.

Figure 2.13: Difference in the CDF of the beta distribution and the CDF of its normal approximation for response proportions of 55%, 65% and 75% and samples sizes of 100 and 200.



2.4 Design

2.4.1 Additional information design

The primary analysis of interest in the current study is a hypothesis test of $H_0 : p_c = p_t$ against $H_1 : p_c < p_t$. The sample size of the current trial is fixed to detect a given treatment difference at a specified power and type I error rate. A Beta(1,1) prior is assumed for p_c before the historical data are observed and a Beta(1,1) prior is assumed for p_t, for which there are no historical data available. At the end of the study, the historical and current control data are compared. Weights are calculated according to the power prior, probability or equivalence probability weight approaches. The prior effective sample size for the control arm is then $ESS = n_h w + 2$, the down-weighted historical data plus the effective sample size of the prior on p_c before the historical data are observed. For the robust mixture prior, the prior ESS is calculated directly, incorporating both the down-weighted historical data and the effective sample size of the prior on p_c before the historical data are observed. The control arm sample size is then $n_c + ESS$, where n_c is the current control sample size. The prior effective sample size for the treatment arm is two from the Beta(1,1) prior on p_t. The treatment arm sample size is then $n_t + 2$.

The aim of the additional information design is to increase the power of the current study by increasing the sample size of the control arm when there is agreement between the historical and current controls, this is the design considered by Viele et al. [42].

2.4.2 Adaptive design with a single interim analysis

The primary analysis of interest is a hypothesis test of $H_0 : p_c = p_t$ against $H_1 : p_c < p_t$. We use a two-stage adaptive design proposed by Schmidli et al. [23] where the allocation ratio is adapted after the first stage. A Beta(1,1) prior is assumed for p_c before the historical data are observed and a Beta(1,1) prior is assumed for p_t, for which there are no historical data available. Therefore the prior effective sample size for the treatment arm $ESS_t = 2$. For the methods of assessing agreement between the historical and current controls that calculate a weight: the modified power prior; probability weight; and equivalence probability weight approaches, w_1 is the weight calculated at the interim analysis when comparing the first stage current control data and the historical data and w_2 is the weight re-calculated at the end of the study comparing the historical data to all of the current trial control data. For the robust mixture prior, the prior ESS is calculated directly, incorporating both the down-weighted historical data and the effective sample size of the prior on p_c before the historical data are observed. ESS_{c1} denotes the prior effective sample size at the interim analysis using only the first stage current control data and the historical data. The number of control and treatment patients randomised in stage one (n_{c1} and n_{t1} respectively) and the total number of patients required per treatment group

(n_t and n_c) are fixed, n_t and n_c are chosen as the sample sizes required to detect a given treatment difference at a specified power and type I error rate in a standard design not incorporating historical data. The adaptive design proceeds as follows:

Stage one: Randomise n_{t1} to the experimental treatment and n_{c1} to control.

Interim analysis: Calculate the prior effective sample size for the control arm, $ESS_{c1} = w_1 n_h + 2$ or ESS_{c1} directly using the first stage controls and the historical data.

Stage two: Randomise $(n_t - n_{t1} - ESS_t)$ to the experimental treatment and $\max(n_c - n_{c1} - ESS_{c1}; nmin)$ to control.

where $nmin$ is a pre-specified fixed minimum number of control patients to be randomised in stage two. A minimum number of control patients are randomised in stage two to allow a randomised comparison in the second stage even if no extra controls are required because the total sample size is achieved through the incorporation of the historical control data. The agreement between the historical and current controls is re-assessed at the end of the study using all current control data and the historical data to determine the amount of historical data to incorporate into the final analysis. We denote the prior effective sample size at the end of the study, $ESS_{c2} = w_2 n_h + 2$ for the weighting approaches. For the robust mixture prior ESS_{c2} is calculated directly. For the robust mixture prior approach, ESS_{c1} is the effective sample size of the posterior distribution at the interim minus n_{c1}. The robust mixture prior approach can give a negative prior ESS, where ESS_{c1} is calculated to be negative it is set to zero.

The adaptive design replaces current controls yet to be randomised with historical controls when the historical and current controls are in agreement. The aim of the adaptive design is to reduce the duration of the current study and the number of control patients to be randomised in the current study.

An alternative adaptive design, not considered here, would be to fix the total sample size of the current study at the sample size from a standard sample size calculation, incorporating no historical data and at the interim analysis adapt the allocation ratio to randomise more patients to treatment and fewer to control if there is agreement between the historical and current controls. The sample of the control group including historical controls will then be the same as the sample size of the treatment group at the end of the study [20].

2.5 Analysis

2.5.1 Analysis approach and operating characteristics for the additional information design using the power prior, probability and equivalence probability weight

Our primary analysis of interest is a hypothesis test of $H_0 : p_c = p_t$ against $H_1 : p_c < p_t$. We assume an initial vague Beta(1,1) prior on the control response probability before the historical data are observed. At the end of the study, using all the current trial control data and the historical data, the weight, w, to be given to the historical data is calculated using the power prior (a summary measure of Equation 2.1), probability (Equation 2.9), one-sample (Equation 2.11) or two-sample (Equation 2.12) equivalence weight approach. The initial vague Beta(1,1) prior is updated with the weighted historical data and updated again with the current trial control data. A vague Beta(1,1) prior is assumed for the treatment response probability which is updated at the end of the trial.

The posterior distributions at the end of the study for the control and treatment groups are then given by,

$$\begin{aligned}\pi(p_c \mid x_h, y_h, x_c, y_c, w) &\sim \text{Beta}(1 + x_h w + x_c, 1 + y_h w + y_c),\\ \pi(p_t \mid x_t, y_t) &\sim \text{Beta}(1 + x_t, 1 + y_t).\end{aligned}$$

The weight, w, is dependent on x_c, y_c, x_h and y_h. The historical data are fixed and therefore w is deterministic given x_c and y_c. The total control sample size also varies but is deterministic given the number of current control responses observed. The primary analysis declares trial success if $\Pr(p_c < p_t \mid Data) > 0.975$. We propose using the iterative procedure described by Cook to calculate $\Pr(p_c < p_t \mid Data)$ exactly [55]. The $\Pr(p_c < p_t \mid Data)$ can only be calculated exactly using the iterative procedure described by Cook [55] when one of the beta parameters of the treatment or control posterior distributions: $1 + x_h w + x_c$; $1 + y_h w + y_c$; $1 + x_t$ or $1 + y_t$ are integer. Given that here we have assumed an initial Beta(1,1) prior for the treatment response probability, $1 + x_t$ and $1 + y_t$ will both be integer, therefore the following iterative procedure can be used to calculate $\Pr(p_c < p_t \mid Data)$.

Let,

$$\begin{aligned}\Pr(p_c < p_t \mid Data) &= g(1 + x_t, 1 + y_t, 1 + x_h w + x_c, 1 + y_h w + y_c)\\ &= \int_0^1 \frac{p^{x_t}(1-p)^{y_t}}{\text{B}(1 + x_t, 1 + y_t)} I_p(1 + x_h w + x_c, 1 + y_h w + y_c)dp,\end{aligned}$$

where $I_p(1 + x_h w + x_c, 1 + y_h w + y_c)$ is the incomplete beta function, the CDF of a Beta$(1 + x_h w + x_c, 1 + y_h w + y_c)$ random variable.

We then use the following symmetries to get the smallest integer parameter value in the last position of the function g.

$$\begin{aligned}
&g(1 + x_t, 1 + y_t, 1 + x_h w + x_c, 1 + y_h w + y_c)\\
=&g(1 + y_h w + y_c, 1 + x_h w + x_c, 1 + y_t, 1 + x_t)\\
=&1 - g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t)\\
=&1 - g(1 + y_t, 1 + x_t, 1 + y_h w + y_c, 1 + x_h w + x_c).
\end{aligned} \tag{2.13}$$

Assuming the smallest integer parameter is $1 + y_t$. The probability that the treatment response probability is greater than the control response probability is calculated using the following recurrence relation [55],

$$\begin{aligned}
&g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t + 1)\\
=&g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t)\\
+&h(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t)/(1 + y_t),
\end{aligned}$$

where,

$$\begin{aligned}
&h(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t)\\
=&\frac{\mathrm{B}(2 + x_h w + x_c + x_t, 2 + y_h w + y_c + y_t)}{\mathrm{B}(1 + x_h w + x_c, 1 + y_h w + y_c)\mathrm{B}(1 + x_t, 1 + y_t)}\\
=&\frac{\Gamma(2 + x_h w + x_c + x_t)\Gamma(2 + y_h w + y_c + y_t)}{\Gamma(1 + x_h w + x_c)\Gamma(1 + y_h w + y_c)\Gamma(1 + x_t)\Gamma(1 + y_t)}\\
\times&\frac{\Gamma(2 + x_h w + x_c + y_h w + y_c)\Gamma(2 + x_t + y_t)}{\Gamma(4 + x_h w + x_c + y_h w + y_c + x_t + y_t)},
\end{aligned}$$

and,

$$\begin{aligned}
&g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1)\\
=&\frac{\Gamma(2 + x_h w + x_c + y_h w + y_c)\Gamma(2 + x_h w + x_c + x_t)}{\Gamma(3 + x_h w + x_c + y_h w + y_c + x_t)\Gamma(1 + x_h w + x_c)}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t) \\
=&g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1) \\
+&\sum_{i=0}^{y_t} h(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + i)/(1 + i) \\
\implies &\Pr(p_t > p_c) = 1 - g(1 + x_h w + x_c, 1 + y_h w + y_c, 1 + x_t, 1 + y_t), \quad \text{from Equation 2.13.}
\end{aligned}
$$

For binary data there are a finite number of possible responses. The number of responses in the current control group and the number of responses in the treatment group follow a binomial distribution, $X_c \sim \text{Bin}(n_c, p_c)$, under the alternative hypothesis $X_t \sim \text{Bin}(n_t, p_t)$ and under the null hypothesis $X_t \sim \text{Bin}(n_t, p_c)$. The weight is deterministic given the observed number of control responses. Therefore, considering all possible combinations of control and treatment response, the operating characteristics of the study design can be calculated exactly.

Let $x_c = 0, \ldots, n_c$ be the number of control responses, $x_t = 0, \ldots, n_t$ the number of treatment responses, then the probability of observing each combination of control and treatment response, due to independence, is given by,

$$
\Pr(X_c = x_c \cap X_t = x_t) = \Pr(X_c = x_c \mid p_c, n_c) \Pr(X_t = x_t \mid p_t, n_t),
$$

where $\Pr(X_c = x_c \mid p_c, n_c) = \binom{n_c}{x_c} {p_c}^{x_c} (1 - p_c)^{n_c - x_c}$ and $\Pr(X_t = x_t \mid p_t, n_t) = \binom{n_t}{x_t} {p_t}^{x_t} (1 - p_t)^{n_t - x_t}$. $p_t = p_c + \Delta$ under the alternative and $p_t = p_c$ under the null hypothesis, where Δ denotes the treatment effect.

The power is given by,

$$
\begin{aligned}
1 - \beta = &\sum_{x_c=0}^{x_c=n_c} \sum_{x_t=0}^{x_t=n_t} \Pr(X_c = x_c \mid p_c, n_c) \Pr(X_t = x_t \mid p_t, n_t) \\
&\times \mathbb{1}(\Pr(p_t > p_c) > 0.975 \mid x_h, x_c, x_t, n_h, n_c, n_t, w),
\end{aligned}
\tag{2.14}
$$

where $\mathbb{1}$ is an indicator function.

The type I error rate is calculated using Equation 2.14, assuming the true underlying treatment response probability p_t is equal to p_c when calculating $\Pr(X_t = x_t \mid p_t, n_t)$.

The expected *EHSS* for a given true underlying control probability can be calculated using,

$$n_h\mathbf{E}(w(x_h,n_h,p_c,n_c)) = n_h \sum_{x_c=0}^{x_c=n_c} \Pr(X_c = x_c \mid p_c, n_c) w(x_c,n_c,x_h,n_h). \tag{2.15}$$

The expected sample size in the control group for a given control proportion is then $n_c + n_h\mathbf{E}(w(x_h,n_h,p_c,n_c))$ plus the prior effective sample size for p_c before the historical data are observed and the mean squared error (MSE) is given by,

$$\mathbf{E}(\hat{p}_c - p_c)^2 = \sum_{x_c=0}^{x_c=n_c} \Pr(X_c = x_c \mid p_c, n_c) \left(\left(\frac{x_h w(x_c,n_c,x_h,n_h) + x_c}{n_h w(x_c,n_c,x_h,n_h) + n_c} \right) - p_c \right)^2. \tag{2.16}$$

No historical data design

All of the historical data methods above are compared to a design not incorporating any historical data. The operating characteristics for this design are calculated as with the weighting approaches above assuming the weight is zero.

2.5.2 Additional information design – frequentist operating characteristics for the Viele example using the probability and equivalence probability weight

For the Viele example, we assume $n_c = n_t = 198$, along with the Beta(1,1) prior assumed for the treatment response probabilities before any data are observed, gives an effective sample size of 200 patients per treatment group. The true underlying response probability in the current control arm is varied and the treatment response probability is assumed to be 12% higher than the control response probability. The historical data are fixed at $x_h = 65, n_h = 100$. For a standard design, incorporating no historical data, assuming a control response probability of 65%, 200 patients per treatment arm would give 76% power and a one-sided type I error rate of 2.5% to detect a treatment difference of 12%.

The operating characteristics of all designs incorporating historical data depend on how quickly the historical data are discounted and the direction of the difference between the historical and current control response proportions. All historical data methods perform similarly when the true control proportion is close to 0.65 (the historical response probability), the power is increased compared to a design not incorporating the historical data and the type I error rate is lower than the desired 2.5% level. Since this design incorporates the historical data as additional information, at agreement between the current and historical controls, utilising the historical data results in an overpowered study. When the current control response probability is higher than the historical, the historical data draws the estimated control response probability down, increasing the treatment effect estimate and inflating the type I error rate. When the current control response probability is less than the historical, the estimated treatment effect is reduced and the power is reduced when compared to a design not incorporating any historical data. When the difference between the historical and current controls is large, on average the historical data are given zero weight in the analysis and the operating characteristics revert back to those of a standard trial design.

Probability weight

Figure 2.14: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true current control proportions for the additional information design using the probability weight approach and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.

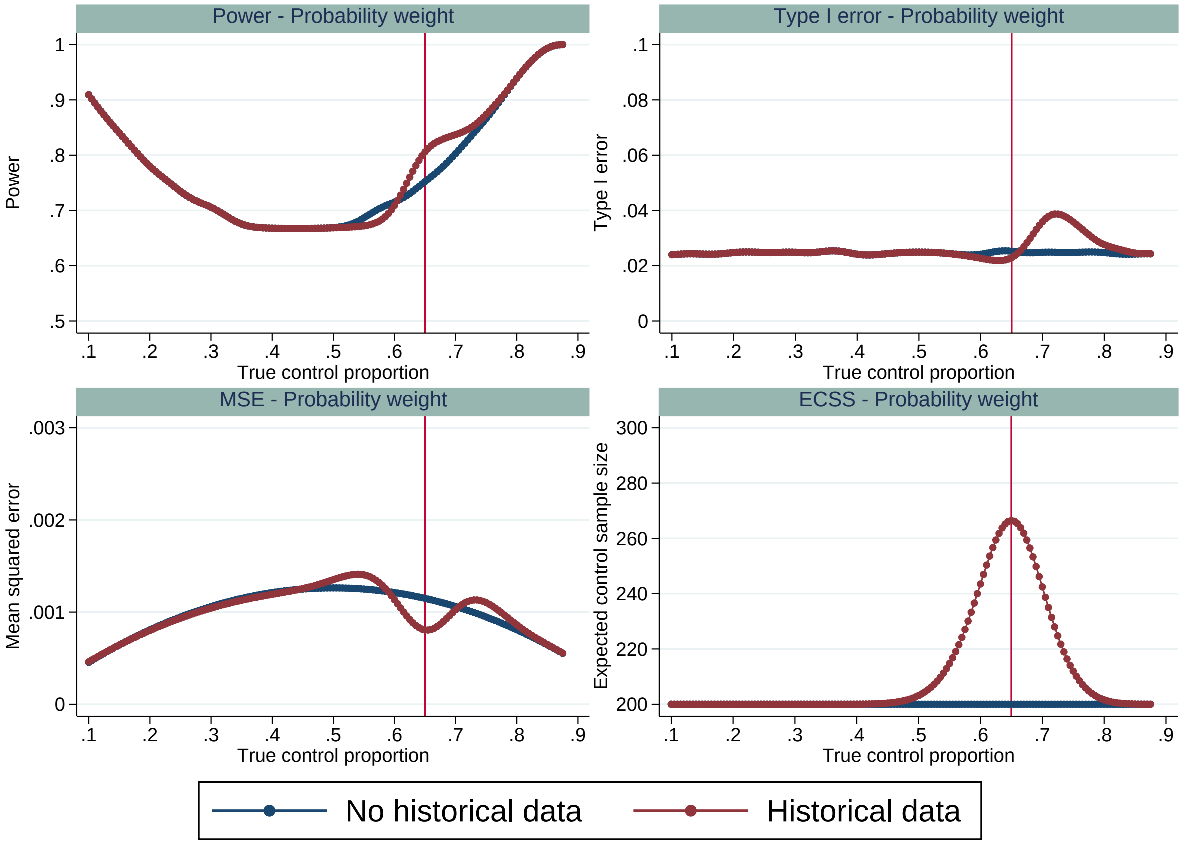


Figure 2.14 shows that the probability weight approach quickly reverts back to the operating characteristics of a standard design as the difference between the historical and current controls increases and the maximum possible type I error rate is 3.9%. On average an additional 66.5 patients are incorporated into the analysis at complete agreement,

giving a power of 81%. The probability weight is fixed and therefore does not allow control over the rate at which the historical data are discounted. The range of true control proportions around the historical data response probability that give a mean squared error lower than the standard design indicates where the historical data design provides some advantage over the standard trial design. For the probability weight approach this range is 0.59 to 0.70. The range of true control proportions for which the mean squared error is lower using the historical data design compared to the mean squared error for a standard trial design are compared in table 2.5 on page 83 for all of the historical data methods explored.

Equivalence probability weight

Figure 2.15: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true current control proportions for the additional information design using the one-sample and two-sample equivalence probability weight approaches with 8% equivalence bounds and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.

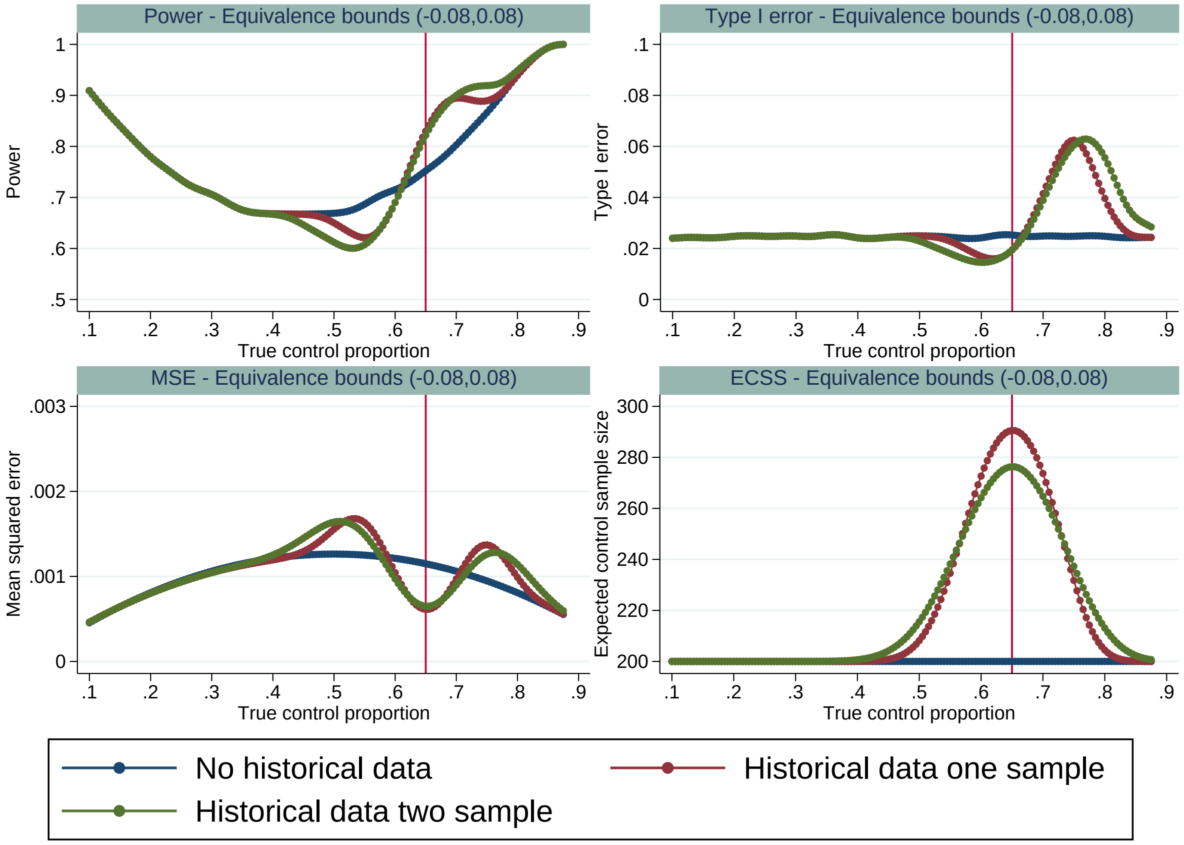


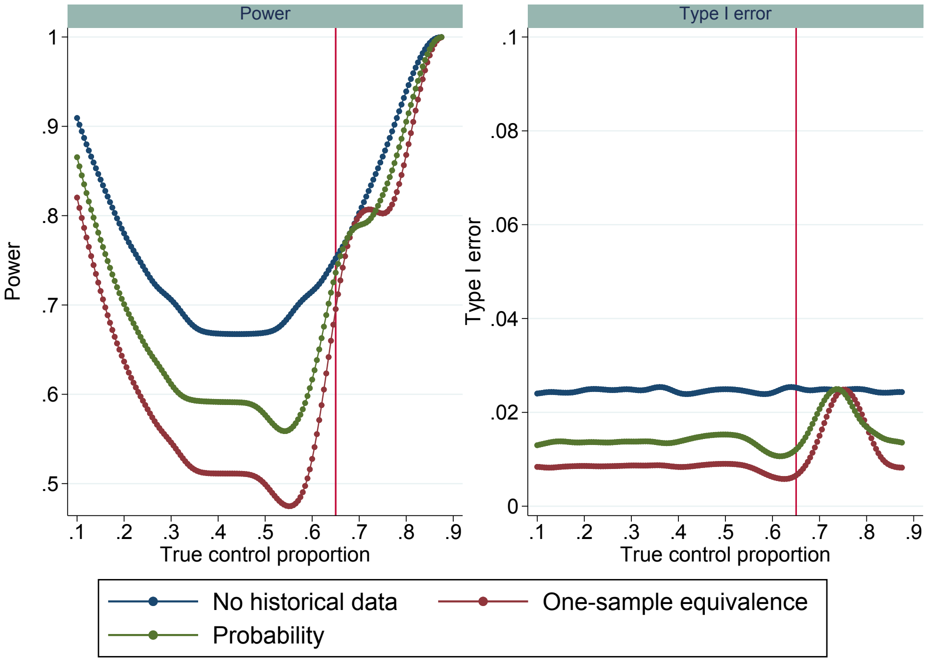
Figure 2.15 shows the design characteristics of the one-sample and two-sample equiv-

alence approaches with 8% equivalence bounds. 8% equivalence bounds were chosen as sensible bounds for a 12% treatment effect to illustrate the equivalence method. However, as discussed in Section 2.5.6 the equivalence bounds should be chosen based on prior knowledge or to control the operating characteristics of the study design. Narrower equivalence bounds borrow less historical data both when the historical and current controls agree and disagree. This reduces the maximum possible type I error rate but also reduces the power gained at compete agreement. The choice of bounds allows control over the maximum inflation in the type I error rate. This is explored further in Section 2.5.6. The two-sample equivalence approach incorporates the additional uncertainty of the historical data being a sample and therefore on average borrows less information at complete agreement compared to the one-sample equivalence approach. On average an additional 90.5 patients are incorporated into the analysis at complete agreement between the current and historical controls for the one-sample equivalence approach and 76.3 patients for the two-sample approach.

2.5.3 Controlling the maximum type I error rate of the additional information design incorporating historical data at the same level as a standard trial design

When incorporating historical data into the design and analysis of a current trial there is always a risk of inflating the type I error rate above the desired level when the historical and current control data do not agree. The historical data are carefully selected using the six acceptability criteria defined by Pocock [11] in the hope that the current and historical controls are similar and incorporating the historical data improves the operating characteristics of the current study. If we wish to incorporate historical data and control the maximum type I error rate at the level required under a standard trial design for any possible observed level of disagreement between the historical and current controls, then there is little to no benefit in terms of power of incorporating the historical data into the design and analysis of the current study, as illustrated in Figure 2.16 for the one-sample equivalence weight approach with 8% equivalence bounds and the probability weight approach. We searched for the probability used to declare trial success that gave a maximum type I error rate across all current control proportions to be less than 2.5%. Figure 2.16 shows that when the probability of declaring trial success is chosen to control the maximum type I error rate across all true control proportions there is little to no gain in power compared to a standard trial design when using the one-sample equivalence probability weight approach or the probability weight approach. Therefore, for a design where historical data are used, it has to be accepted that to gain power when there is agreement between the current and historical controls, there is a risk of inflating the type I error rate above the desired level if there is disagreement between the current and historical controls. A similar result was observed by Cuffe [56] who showed that for outcomes that are normally distributed with known variance, to control the type I error rate at the level of the current study when incorporating no historical data for all values of the unknown expected outcome on control, then a conservative critical value will be needed which will eliminate most (if not all) of the gains in power made by incorporating the historical data.

Figure 2.16: Comparison of the power and type I error rate across different true current control proportions for the additional information design using the one-sample equivalence probability weight approach with 8% equivalence bounds and the probability weight approach controlling the maximum possible type I error rate at 2.5% and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



2.5.4 Analysis approach and operating characteristics for the adaptive design using the power prior, probability and equivalence probability weight

We assume an initial vague Beta(1,1) prior on the control response probability before the historical data are observed. This prior is updated with the first stage control data at the interim analysis. At the interim analysis, using the first stage control data and the historical data, the weight (w_1) to be given to the historical data is calculated using the power prior (a summary measure of Equation 2.1), probability (Equation 2.9), one-sample (Equation 2.11) or two-sample (Equation 2.12) equivalence weight approach. The ESS_{c1} ($w_1 n_h + 2$) is calculated and the number of control patients to be randomised in stage two is determined. At the end of the study the weight is re-calculated using all of the current study control data, this is denoted w_2. Similarly, a vague Beta(1,1) prior is assumed for the treatment response probability which is updated after stage two of the trial. At the interim analysis, only the response proportion in the control arm data is required to adapt the trial, the difference in response proportions between the treatment arm and the control arm can remain un-blinded.

The posterior distributions at the end of the study for the control and treatment groups are then given by,

$$\pi(p_c \mid x_h, y_h, x_{c1}, x_{c2}, y_{c1}, y_{c2}, w_1, w_2) \sim \text{Beta}(1 + x_h w_2 + x_{c1} + x_{c2}, 1 + y_h w_2 + y_{c1} + y_{c2}),$$
$$\pi(p_t \mid x_{t1}, x_{t2}, y_{t1}, y_{t2}) \sim \text{Beta}(1 + x_{t1} + x_{t2}, 1 + y_{t1} + y_{t2}).$$

where x_{c1} and x_{c2} are the number of responses in the control group in stage one and two of the trial respectively, y_{c1} and y_{c2} the non-responses and similarly with subscript t for the treatment group. The total number of controls randomised in stage two $(x_{c2} + y_{c2})$ is dependent on w_1, the weight given to the historical data at the interim analysis and w_1 is deterministic given the historical data and the first stage current control data. w_2 is the weight given to the historical data, calculated at the end of the study, using the historical data and all of the control data from the current trial.

The primary analysis declares trial success if $\Pr(p_c < p_t \mid Data) > 0.975$. The $\Pr(p_c < p_t \mid Data)$ can be calculated exactly when one of the beta parameters of the treatment or control posterior distributions are integer using the iterative procedure proposed by Cook [55], described in Section 2.5.1. We know that the number of responses in the control and treatment group follow a binomial distribution. Further, the weight given to the historical data is deterministic given the observed number of control responses and therefore the number of second stage controls required is also deterministic given the number of control responses in stage one. Considering all possible combinations of first and second stage control responses and all possible treatment responses, we can calculate the operating characteristics of this adaptive design exactly. The interim analysis is at a fixed time in the trial, when n_{c1} patients have been randomised to control.

Let $x_{c1} = 0, \dots, n_{c1}$ be the number of first stage control responses, $x_{c2} = 0, \dots, n_{c2|x_{c1}}$, the number of control responses in stage two given the number of controls randomised in stage two. Where $n_{c2|x_{c1}}$ denotes the number of controls in stage two given the number of control responses in stage one. The total number of controls randomised in stage two $(n_{c2|x_{c1}})$ is dependent on $nmin$ and w_1, where w_1 is the weight given to the historical data at the interim analysis and w_1 is deterministic given the historical data and the first stage current control data. Since x_h, y_h, n_{c1} and $nmin$ are all fixed by design, the number of second stage controls is denoted $n_{c2|x_{c1}}$, dependent on the number of observed control responses in stage one x_{c1}. $x_t = 0, \dots, n_t$ denotes all possible treatment responses. The response distributions are given by,

$$\begin{aligned}
\Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) &= \binom{n_{c1}}{x_{c1}} {p_c}^{x_{c1}} (1 - p_c)^{n_{c1} - x_{c1}}, \\
\Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) &= \binom{n_{c2|x_{c1}}}{x_{c2}} {p_c}^{x_{c2}} (1 - p_c)^{n_{c2|x_{c1}} - x_{c2}}, \\
\Pr(X_t = x_t \mid p_t, n_t) &= \binom{n_t}{x_t} {p_t}^{x_t} (1 - p_t)^{n_t - x_t}.
\end{aligned} \tag{2.17}$$

The power is given by,

$$\begin{aligned}
1 - \beta = \sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} \sum_{x_t=0}^{x_t=n_t} &\Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) \\
&\times \Pr(X_t = x_t \mid p_t, n_t) \mathbb{1}(\Pr(p_t > p_c) > 0.975 \mid x_h, x_{c1}, x_{c2}, x_t, n_{c1}, n_{c2|x_{c1}}, n_h, n_t, w_2).
\end{aligned}$$

The type I error rate is calculated using the formula for the power assuming the true underlying treatment response probability p_t is equal to p_c when calculating $\Pr(X_t = x_t \mid p_t, n_t)$.

The expected sample size of the current trial control group ($ECCSS$) is given by,

$$\begin{aligned}
ECCSS = n_{c1} + \sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} &\Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \\
&\times \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}})(\max(n_c - n_{c1} - ESS_{c1}, nmin)),
\end{aligned}$$

and the expected effective historical sample size at the end of the current study for a true underlying control response probability is given by,

$$\begin{aligned}
n_h \mathbf{E}(w_2(x_h, n_h, p_c, n_{c1}, n_{c2|x_{c1}})) = n_h \sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} &\Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \\
&\times \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) w_2(x_{c1}, x_{c2}, n_{c1}, n_{c2|x_{c1}}, x_h, n_h).
\end{aligned}$$

The expected total sample size of the control group ($ECSS$) for a given control response probability is then $ECSS = ECCSS + n_h \mathbf{E}(w_2(x_h, n_h, p_c, n_{c1}, n_{c2/x_{c1}})) + 2$. Note that this could be greater or less than the total number of controls required n_c. The weight given to the historical data is re-calculated at the end of the study, where the level of agreement between the historical and current controls may have changed from the agreement at the interim analysis. Further, even when no extra controls are required in stage two of the trial, *nmin* controls are randomised to ensure a randomised comparison can be made in both stages of the trial.

The mean squared error is given by,

$$\mathbf{E}(\hat{p}_c - p_c)^2 = \sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} \Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) \\ \left(\left(\frac{x_h w_2(x_{c1}, x_{c2}, n_{c1}, n_{c2|x_{c1}}, x_h, n_h) + x_{c1} + x_{c2}}{n_h w_2(x_{c1}, x_{c2}, n_{c1}, n_{c2|x_{c1}}, x_h, n_h) + n_{c1} + n_{c2|x_{c1}}} \right) - p_c \right)^2 . \tag{2.18}$$

No historical data design

All of the historical data methods above are compared to a design not incorporating any historical data. The operating characteristics for this design are calculated as with the weighting approaches above assuming the weight is zero.

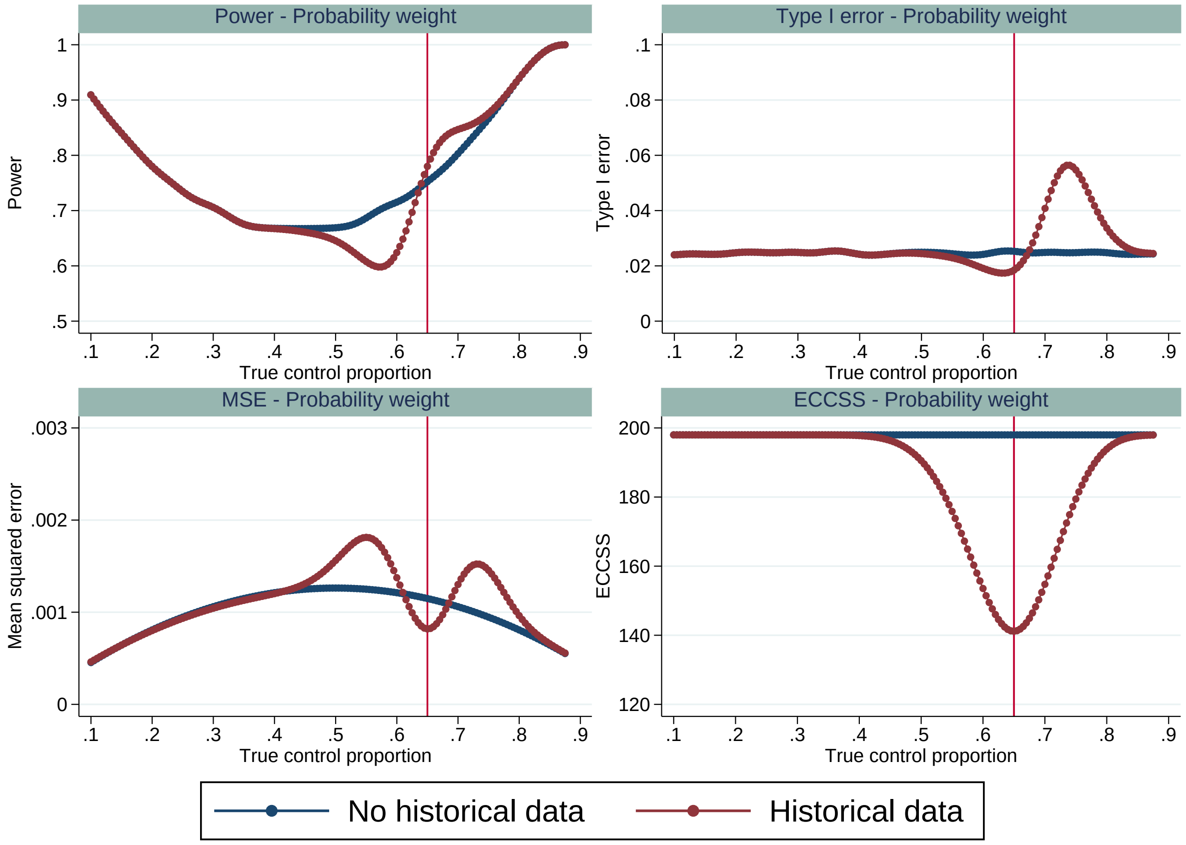
2.5.5 Adaptive design – frequentist operating characteristics for the Viele example using the probability and equivalence probability weight

We explore the operating characteristics of the adaptive design proposed in Section 2.4.2 for a range of true control response probabilities in the current study. The treatment response is always assumed to be 12% higher than the control and there are 100 (n_h) historical control patients available with a response probability of 65%. The interim analysis is conducted after 100 patients have been randomised to both the control and treatment group ($n_{c1} = n_{t1} = 100$) and $nmin = 20$. In Appendix A the expected total sample size of the control group, incorporating both current and historical control data is illustrated for a range of true response probabilities using the probability and the equivalence probability weight approaches. In Appendix B the expected historical data weights at the interim analysis and the final analysis are compared for a range of true control response probabilities using the probability and the equivalence probability weight approaches.

Probability weight

Figure 2.17 shows the operating characteristics of the adaptive design using the probability weight approach. Comparing the operating characteristics of the adaptive design to

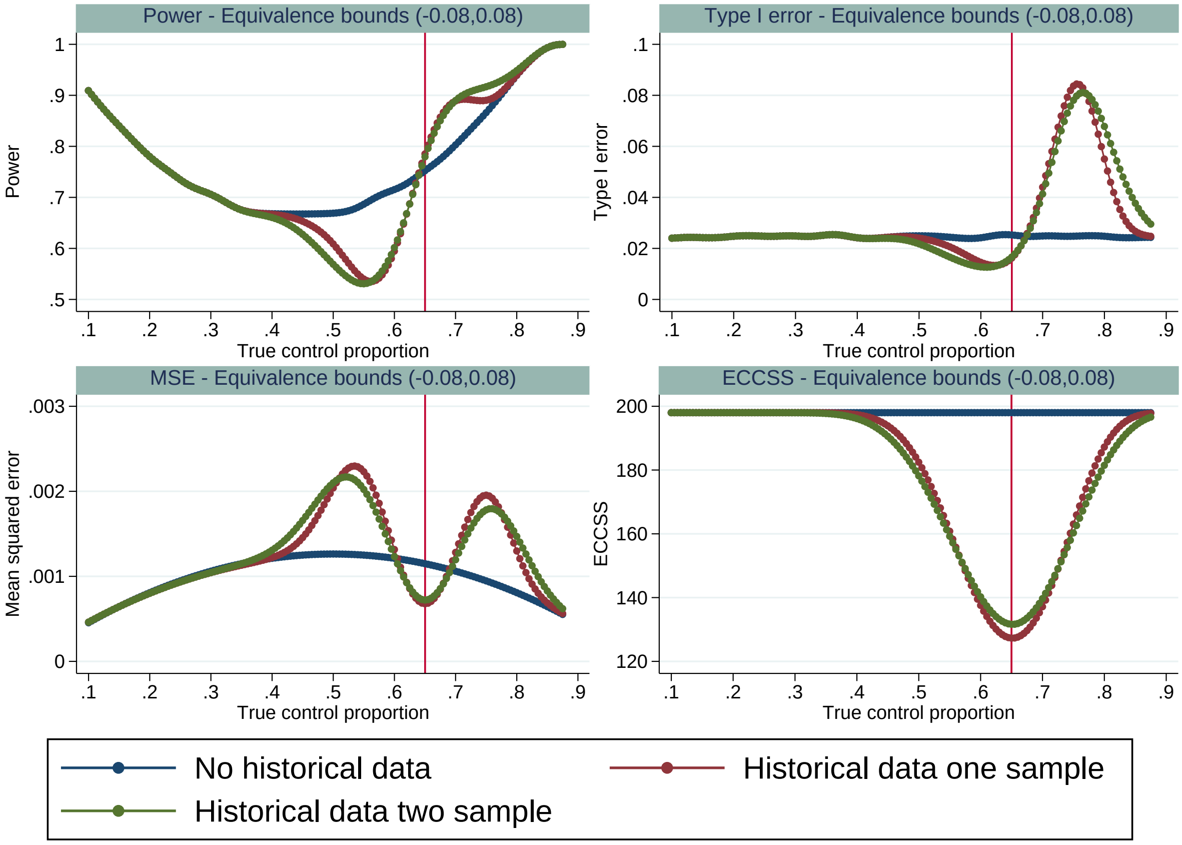
Figure 2.17: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true current control proportions for the adaptive design using the probability weight approach and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



the additional information design given in Figure 2.14, the adaptive design has a lower power at complete agreement and the maximum possible type I error rate is slightly higher. This is because the adaptive design is replacing the current control data yet to be randomised with historical control data when there is agreement between the current and historical controls, therefore at agreement, the additional information design will have a larger control sample size than the adaptive design. However, there is a substantial saving in the current control sample size required in the adaptive design compared to the additional information design. For the adaptive design, at complete agreement between the current and historical control data, the expected sample size of the current control group is 141.17, 56 patients fewer than the standard trial design. The comparisons of the operating characteristics of all the historical data designs considered are compared in Table 2.5 on page 83.

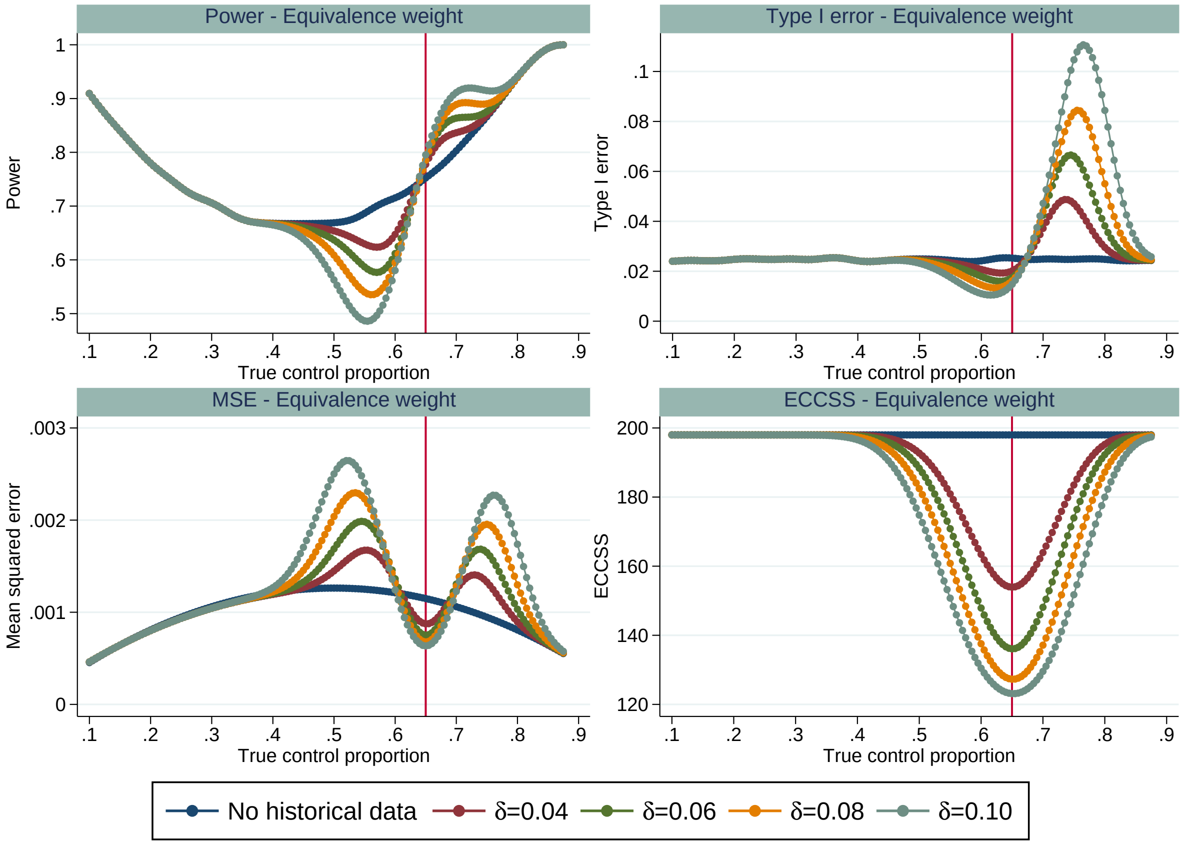
Equivalence probability weight

Figure 2.18: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true current control proportions for the adaptive design using the one-sample and two-sample equivalence probability weight approaches with 8% equivalence bounds and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



Similar comparisons can be made between the additional information and adaptive design for the equivalence approach as with the probability weight approach. The additional variance incorporated into the two-sample equivalence approach results in less information being borrowed at complete agreement and this design requires a larger difference between the historical and current controls for the historical data to be completely discounted and the operating characteristics to revert back to a standard design. At complete agreement between the current and historical control data, the expected sample size of the current control group using the one-sample equivalence probability weight approach is 127.30 and using the two-sample equivalence probability weight approach is 131.63, as shown in Figure 2.18. Figure 2.19 illustrates the operating characteristics for the one-sample

Figure 2.19: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true current control proportions for the adaptive design using the one-sample equivalence probability weight approach with 4%, 6%, 8% and 10% equivalence bounds and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



equivalence weight approach for different equivalence bounds (4%,6%,8% and 10%). As the equivalence bounds increase, at complete agreement between the historical and current control data, the amount of historical data borrowed and the power both increase, however this is at the expense of a higher maximum possible type I error rate for larger equivalence bounds when the estimated current control response probability is larger than the historical response probability.

2.5.6 Choosing the equivalence bounds

The equivalence bounds chosen have a large effect on how much historical data are borrowed and how quickly the historical data are discounted when there is disagreement between the historical and current controls. The equivalence bounds therefore have a large effect on the operating characteristics of the current study. How to choose the equivalence bounds is therefore an important question. Similar to a test of equivalence of an experimental and control treatment, the equivalence bounds are chosen by the study designer based on prior knowledge. The bounds are chosen to represent a clinically relevant equivalence distance, which forms a region of acceptable deviation of the current control data from the historical. In the absence of knowledge about acceptable equivalence bounds, the equivalence bounds can be chosen to minimise trial risk based on statistical properties of the study design. For example controlling the maximum type I error rate across all possible current control response probabilities. The main concern when using historical data is the risk of inflating the type I error rate. A design which aims to incorporate historical data at the design stage when the current trial data has not yet been observed risks inflating the type I error rate. If the design parameters of the current study are chosen to strictly control the type I error rate when incorporating historical data, there is little to no benefit in using the historical data design over a standard trial design, as discussed in Section 2.5.3. However, the equivalence bounds can be chosen so that the maximum type I error rate across all possible true current control probabilities is capped at a chosen value, minimising the risk if the current and historical control response probabilities do in fact differ.

Further considerations in choosing the equivalence bounds are:

- The equivalence bounds should be less than the treatment effect to be detected.
- Narrow equivalence bounds will require large amounts of data to achieve a high weight even when the historical and current controls are in complete agreement. Using the one-sample equivalence approach the weight depends on the current control sample size and using the two-sample equivalence approach the weight depends on both the current control and historical sample size. To avoid the historical data having a larger influence on the control parameter estimate than the current control data, one may wish to only consider equivalence bounds that give an effective historical sample size less than the current control study sample size at complete agreement between the historical and current controls.
- The equivalence bounds should be chosen to give a large weight to the historical data around agreement between the current and historical control data and a small weight when there is substantial disagreement between the current and historical controls.
- The equivalence bounds can be chosen to govern how quickly the discounting of the historical data occurs as the difference between the historical and current controls increases.

- A criterion other than the maximum type I error rate could be used for choosing the equivalence bounds, such as controlling the maximum mean squared error or the maximum $EHSS$.

Equivalence bounds that control the maximum type I error rate

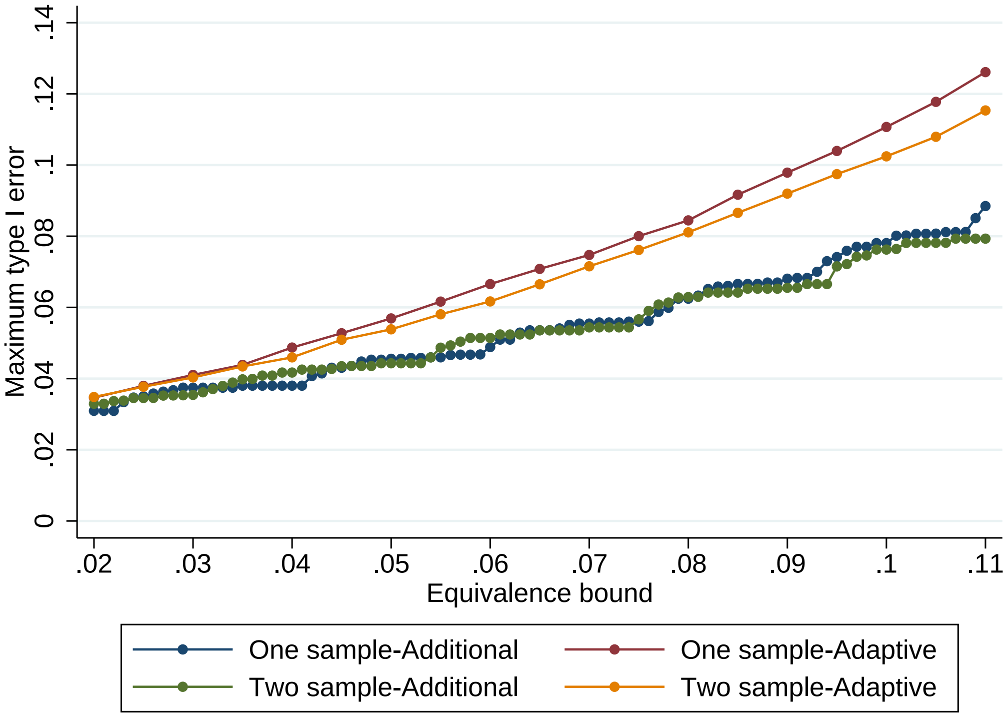
For a design where the weighted historical data are incorporated as additional information at the end of the current study, numerical optimisation can be used to determine the equivalence bounds that control the maximum type I error rate at a chosen value. The bounds are determined by minimising the squared error of the maximum type I error rate obtained for a given equivalence bound and an assumed underlying current control response probability to the desired maximum type I error rate.

Optimisation is possible for the additional information design because the type I error rate can be calculated quickly since there are few possible combinations of control and treatment responses. For the two-stage adaptive design the number of possible combinations of first and second stage control responses and treatment responses is much higher. The weight given to the historical data also has to be calculated at the interim analysis and re-calculated at the final analysis. The adaptive design is therefore more computationally expensive. Depending on the number of patients in each treatment group, a quicker approximation of the bounds can be determined by plotting the maximum error distribution for a range of equivalence bounds and using interpolation to choose the bounds that control the maximum type I error rate across all possible true control response probabilities.

Figure 2.20 illustrates the maximum type I error rate across all true control proportions in the current study for different equivalence bounds. Both the additional information and adaptive design for the one-sample and two-sample equivalence approaches are shown. Note that controlling the maximum possible type I error rate at 2.5% (the type I error rate for a design not incorporating historical data) is only possible when incorporating no historical data, as shown in Figure 2.16. Controlling the maximum type I error rate at 2.5% while incorporating historical data would require a larger sample size than a standard trial design. However, the equivalence bounds can be chosen to cap the maximum type I error rate at a chosen value to minimise risk in a study design where the use of historical data is required. The maximum type I error rate is higher for the adaptive design since bias is introduced from the historical data but the sample size is not increased. The power at complete agreement between the current and historical controls for different equivalence bounds is illustrated in Appendix C.

Using the modified Newton-Raphson optimisation technique implemented in Mata [46], the equivalence bounds that control the maximum type I error rate at 5% for the

Figure 2.20: Distribution of the maximum possible type I error rate across a range of equivalence bounds using the one-sample and two-sample equivalence probability weight approaches for the additional information and adaptive design.



one-sample equivalence probability weight approach are ±0.060466 and for the two-sample equivalence probability weight approach are ±0.056281 for the additional information design. For the adaptive design the optimisation did not converge within a week of the code running and therefore plotting the distribution of the maximum type I error rate is the most efficient way of determining the optimal equivalence bounds.

2.5.7 Analysis approach and operating characteristics for the additional information design using the robust mixture prior

For one historical study, the robust mixture prior [23] is a two-component mixture prior, with the historical data component and a weakly-informative component,

$$\pi(p_c \mid x_h, y_h, w) = w\text{Beta}(x_h, y_h) + (1-w)\text{Beta}(1,1),$$

where the prior mixture weight w is pre-specified. A Beta(1,1) prior is assumed for the treatment response probability.

The posterior distributions at the end of the study for the control and treatment groups are [23],

$$\pi(p_c \mid x_c, y_c, x_h, y_h, w) = \tilde{w}\text{Beta}(x_h + x_c, y_h + y_c) + (1 - \tilde{w})\text{Beta}(1 + x_c, 1 + y_c),$$
$$\pi(p_t \mid x_t, y_t) = \text{Beta}(1 + x_t, 1 + y_t),$$

where,

$$\tilde{w} \propto \frac{w\,\dfrac{\mathrm{B}(x_h + x_c, y_h + y_c)}{\mathrm{B}(x_h, y_h)}}{w\,\dfrac{\mathrm{B}(x_h + x_c, y_h + y_c)}{\mathrm{B}(x_h, y_h)} + (1 - w)\dfrac{\mathrm{B}(1 + x_c, 1 + y_c)}{\mathrm{B}(1, 1)}}. \tag{2.19}$$

For the additional information design we do not need to calculate the prior ESS at the final analysis to calculate the power and type I error rate of the design. However, if we wanted to know how much weight was given to the historical data in the final analysis for a given true control response probability, the expected prior ESS can be calculated using the method of Morita et al. [26], described in Section 2.2.3. Given that the number of possible control responses for the additional information design is not too large, the prior ESS should not be too computationally intensive to calculate using the Morita algorithm.

The power and type I error rate for the robust mixture approach can also be calculated directly. The final analysis calculates whether $\Pr(p_t > p_c)$ is greater than 0.975. However, now the posterior distribution for the response probability in the controls is a mixture of beta distributions. Cook's method [55] can be applied comparing the treatment response distribution to each mixture component of the control posterior distribution separately. These probabilities are weighted by the control posterior mixture weights to obtain the overall $\Pr(p_t > p_c \mid Data)$.

For the robust mixture prior,

$$\begin{aligned}&\Pr(p_t > p_c \mid x_h, y_h, x_c, y_c, x_t, y_t, w) = \\ &\qquad \tilde{w}\Pr(p_t > p_c \mid 1 + x_t, 1 + y_t, x_h + x_c, y_h + y_c) + \\ &\qquad (1 - \tilde{w})\Pr(p_t > p_c \mid 1 + x_t, 1 + y_t, 1 + x_c, 1 + y_c),\end{aligned}$$

where $\tilde{w}$ is given in Equation 2.19, the power is given by,

$$\begin{aligned}1 - \beta = &\sum_{x_c=0}^{x_c=n_c}\sum_{x_t=0}^{x_t=n_t} \Pr(X_c = x_c \mid p_c, n_c)\Pr(X_t = x_t \mid p_t, n_t) \\ &\mathbb{1}(\Pr(p_t > p_c) > 0.975 \mid x_c, x_t, y_c, y_t, x_h, y_h, w),\end{aligned} \tag{2.20}$$

and the type I error rate is calculated using Equation 2.20 assuming the true underlying treatment response probability p_t is equal to p_c when calculating $\Pr(X_t = x_t \mid p_t, n_t)$.

The expected control sample size can be calculated using,

$$ECSS = \sum_{x_c=0}^{x_c=n_c} \Pr(X_c = x_c \mid p_c, n_c)(ESS \mid x_c, n_c, x_h, n_h, w),$$

where ESS is the effective sample size of the control posterior mixture distribution calculated using the Morita algorithm.

The mean squared error is calculated using,

$$\mathbf{E}(\hat{p}_c - p_c)^2 = \sum_{x_c=0}^{x_c=n_c} \Pr(X_c = x_c \mid p_c, n_c)\left(\left(\tilde{w}\left(\frac{x_h + x_c}{n_h + n_c}\right) + (1-\tilde{w})\left(\frac{1+x_c}{2+n_c}\right)\right) - p_c\right)^2,$$

where $\hat{p}_c$ is derived from the sum of the individual component means of the mixture distribution weighted by the posterior weights.

2.5.8 Analysis approach and operating characteristics for the adaptive design using the robust mixture prior

A vague Beta(1,1) prior is assumed for the treatment response probability. The robust mixture prior is assumed for the control response probability, the prior parameters and weights are updated with the first stage control data at the interim analysis. The ESS of the updated mixture distribution is calculated at the interim analysis using the Morita algorithm, to determine the number of controls to be randomised in stage two. Where $ESS < n_{c1}$, the ESS was set to n_{c1}. At the end of the study, the mixture parameters and weights are updated with the second stage control data.

The posterior distributions at the end of the study for the control and treatment groups are,

$$\begin{aligned}\pi(p_c \mid x_h, y_h, x_{c1}, x_{c2}, y_{c1}, y_{c2}, w) &= \tilde{w}_2\text{Beta}(x_h + x_{c1} + x_{c2}, y_h + y_{c1} + y_{c2}) \\ &\quad + (1-\tilde{w}_2)\text{Beta}(1 + x_{c1} + x_{c2}, 1 + y_{c1} + y_{c2}) \\ \pi(p_t \mid x_{t1}, x_{t2}, y_{t1}, y_{t2}) &= \text{Beta}(1 + x_{t1} + x_{t2}, 1 + y_{t1} + y_{t2}),\end{aligned}$$

where,

$$\tilde{w}_2 \propto \frac{\tilde{w}_1 \dfrac{\mathrm{B}(x_h + x_{c1} + x_{c2}, y_h + y_{c1} + y_{c2})}{\mathrm{B}(x_h + x_{c1}, y_h + y_{c1})}}{\tilde{w}_1 \dfrac{\mathrm{B}(x_h + x_{c1} + x_{c2}, y_h + y_{c1} + y_{c2})}{\mathrm{B}(x_h + x_{c1}, y_h + y_{c1})} + (1 - \tilde{w}_1) \dfrac{\mathrm{B}(1 + x_{c1} + x_{c2}, 1 + y_{c1} + y_{c2})}{\mathrm{B}(1 + x_{c1}, 1 + y_{c1})}},$$

and,

$$\tilde{w}_1 \propto \frac{w \dfrac{\mathrm{B}(x_h + x_{c1}, y_h + y_{c1})}{\mathrm{B}(x_h, y_h)}}{w \dfrac{\mathrm{B}(x_h + x_{c1}, y_h + y_{c1})}{\mathrm{B}(x_h, y_h)} + (1 - w) \dfrac{\mathrm{B}(1 + x_{c1}, 1 + y_{c1})}{\mathrm{B}(1, 1)}}.$$

The operating characteristics for the robust mixture prior approach using the adaptive design can also be calculated exactly using Cook's method [55]. The probability that treatment response is greater than control is given by,

$$\begin{aligned}
&\Pr(p_t > p_c \mid x_t, y_t, x_h, y_h, x_{c1}, x_{c2}, y_{c1}, y_{c2}, w) = \\
&\qquad \tilde{w}_2 \Pr(p_t > p_c \mid 1 + x_t, 1 + y_t, x_h + x_{c1} + x_{c2}, y_h + y_{c1} + y_{c2}) \\
&\qquad (1 - \tilde{w}_2) \Pr(p_t > p_c \mid 1 + x_t, 1 + y_t, 1 + x_{c1} + x_{c2}, 1 + y_{c1} + y_{c2}).
\end{aligned}$$

The power is given by,

$$\begin{aligned}
1 - \beta = &\sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} \sum_{x_t=0}^{x_t=n_t} \Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) \\
&\times \Pr(X_t = x_t \mid p_t, n_t) \mathbb{1}(\Pr(p_t > p_c) > 0.975 \mid x_{c1}, x_{c2}, x_t, y_{c1}, y_{c2}, y_t, x_h, y_h, w).
\end{aligned} \tag{2.21}$$

The type I error rate is calculated using Equation 2.21, assuming the true underlying treatment response probability p_t is equal to p_c when calculating $\Pr(X_t = x_t \mid p_t, n_t)$.

The mean squared error is calculated using,

$$\mathbf{E}(\hat{p}_c - p_c)^2 = \sum_{x_{c1}=0}^{x_{c1}=n_{c1}} \sum_{x_{c2}=0}^{x_{c2}=n_{c2|x_{c1}}} \Pr(X_{c1} = x_{c1} \mid p_c, n_{c1}) \Pr(X_{c2} = x_{c2} \mid p_c, n_{c2|x_{c1}}) (\hat{p}_c - p_c)^2,$$

where $\hat{p}_c = \tilde{w}_2 \left(\frac{x_h + x_{c1} + x_{c2}}{n_h + n_{c1} + n_{c2|x_{c1}}}\right) + (1 - \tilde{w}_2) \left(\frac{1 + x_{c1} + x_{c2}}{2 + n_{c1} + n_{c2|x_{c1}}}\right)$.

For the adaptive design, the *ESS* of the posterior distribution for the control parameter at the interim determines how many control patients are to be randomised in stage two of the trial. The ESS at the interim is required for the power and type I error rate calculations. Where the prior ESS was calculated to be negative, it was set to zero. Calculating the expected control sample size at the end of the study would require applying the Morita algorithm for all possible combinations of first stage control responses and second stage control responses. The number of possible combinations also increases since the sample size in the second stage controls will vary. This is too computationally intensive to calculate for most sample sizes and is not considered here. Not knowing how much historical information is incorporated into the final analysis is a disadvantage of the robust mixture prior approach in an adaptive setting.

2.5.9 Additional information design – frequentist operating characteristics for the Viele example using the robust mixture prior and the power prior

As in the robust mixture prior paper [23], we consider a two-component mixture prior for the control response probability with two different initial prior weights,

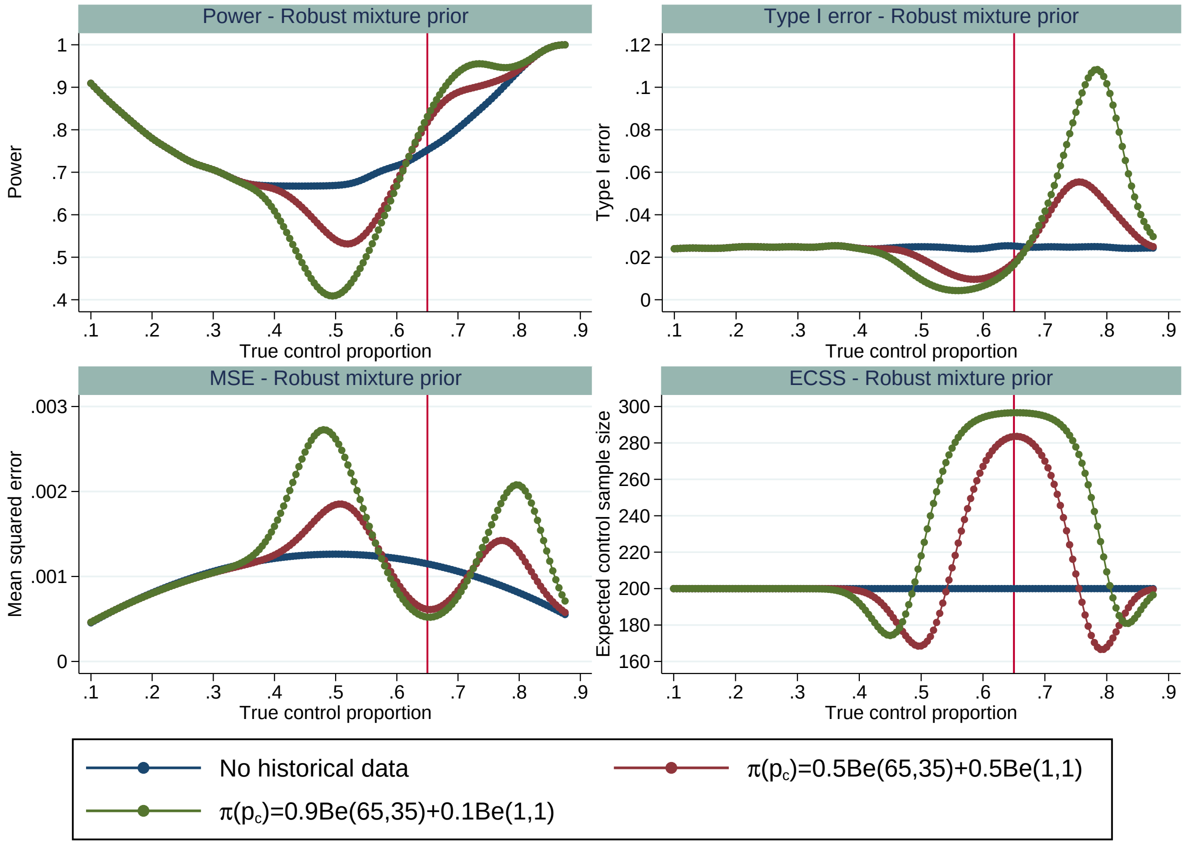
$$\begin{aligned}\text{Prior 1 (weight 0.9)} &: 0.9\text{Beta}(x_h, y_h) + 0.1\text{Beta}(1,1),\\ \text{Prior 2 (weight 0.5)} &: 0.5\text{Beta}(x_h, y_h) + 0.5\text{Beta}(1,1).\end{aligned}$$

A Beta(1,1) initial prior is assumed for p_t.

Figure 2.21 displays the operating characteristics for the additional information design using the robust mixture prior approach with weights of 0.9 and 0.5 on the informative component of the mixture prior. A lower weight on the informative component borrows less information in the final analysis even when there is complete agreement between the historical and current controls. Both weights give a negative expected control sample size for a range of differences between the historical and current controls. The initial mixture prior weights can be chosen to control the maximum possible type I error rate in the final analysis across all true current control proportions. This is done using optimisation in the same way as for the equivalence approach. Figure 2.22 shows the distribution of the maximum type I error rate across all initial mixture prior weights. Using numerical optimisation, the weight that controls the maximum type I error rate at 5% is 0.371902 on the informative component of the robust mixture prior and (1-0.371902) on the weakly-informative component.

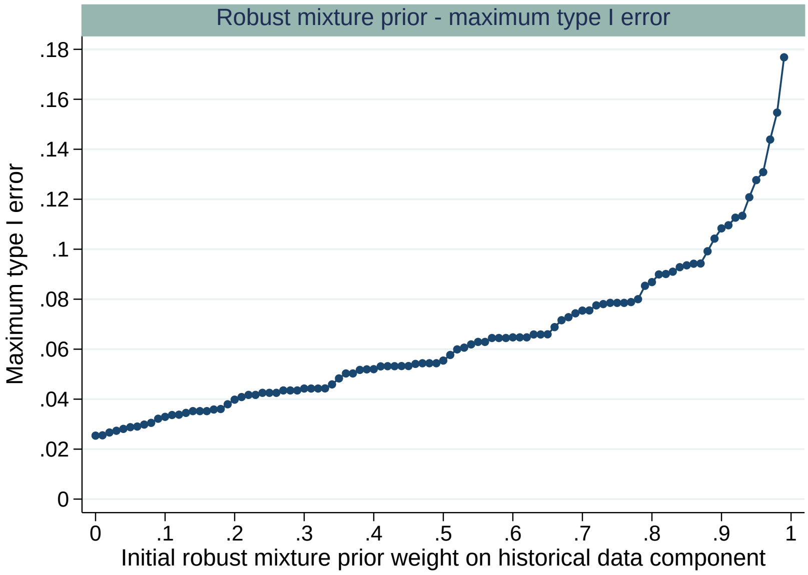
Figure 2.23 shows the operating characteristics of the additional information design for the power prior approach. The power prior with a Beta(1,1) prior on the power,

Figure 2.21: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true current control proportions for the additional information design using the robust mixture prior approach with 0.9 and 0.5 initial weight on the informative component of the mixture prior and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



taking either the posterior mode or mean as the weight are considered. Also, the power prior with a Beta(0.5,0.5) and Beta(0.3,0.3) prior on the power taking the posterior mean are explored. Using the mean of the power prior marginal distribution as a weight gives similar operating characteristics for all priors on the power. A large difference between the current and historical controls is required for the historical data to be completely discounted and the operating characteristics to revert back to those of a standard design. Using the mode of the marginal distribution of the power gives a high weight at agreement and discounts quickly. This results in a large power gain at agreement and a small maximum possible type I error rate. However, there is no flexibility in this approach to control the maximum possible type I error rate across all true control proportions at a desired level through the prior on the power and discounting the historical data. It is possible to maintain control of the maximum type I error by calibrating the success threshold c, such

Figure 2.22: Distribution of the maximum type I error rate across a range of initial weights on the informative component of the robust mixture prior for the additional information design.



that when trial success is declared using $\Pr(p_c < p_t \mid Data) >$ c, the desired maximum type I error is achieved.

Appendix D compares the power and type I error rate for the additional information design using the fully Bayesian modified power prior approach with a Beta(1,1) prior on the power and the modified power prior approach using the mean of the marginal distribution of the power as a fixed weight, assuming a Beta(1,1) prior for the power. The operating characteristics for the two approaches are only slightly different and due to the computational cost and difficulty in calculating the ESS of the historical data using the fully Bayesian approach, the fully Bayesian approach was not considered further in this chapter or Chapter 3.

The operating characteristics for the adaptive design using the robust mixture prior and power prior approaches are given in Appendix E. Also in Appendix E, the historical data weights at the interim analysis and at the end of the study are compared for the power prior approach. For the robust mixture prior approach, the prior ESS was not calculated at the end of the study because of the large number of possible combinations of first stage and second stage control responses in the adaptive design this would be too computationally intensive.

Table 2.5 compares the design characteristics of all the historical data methods explored in this chapter.

Figure 2.23: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true current control proportions for the additional information design using the power prior, assuming different priors on the power and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.

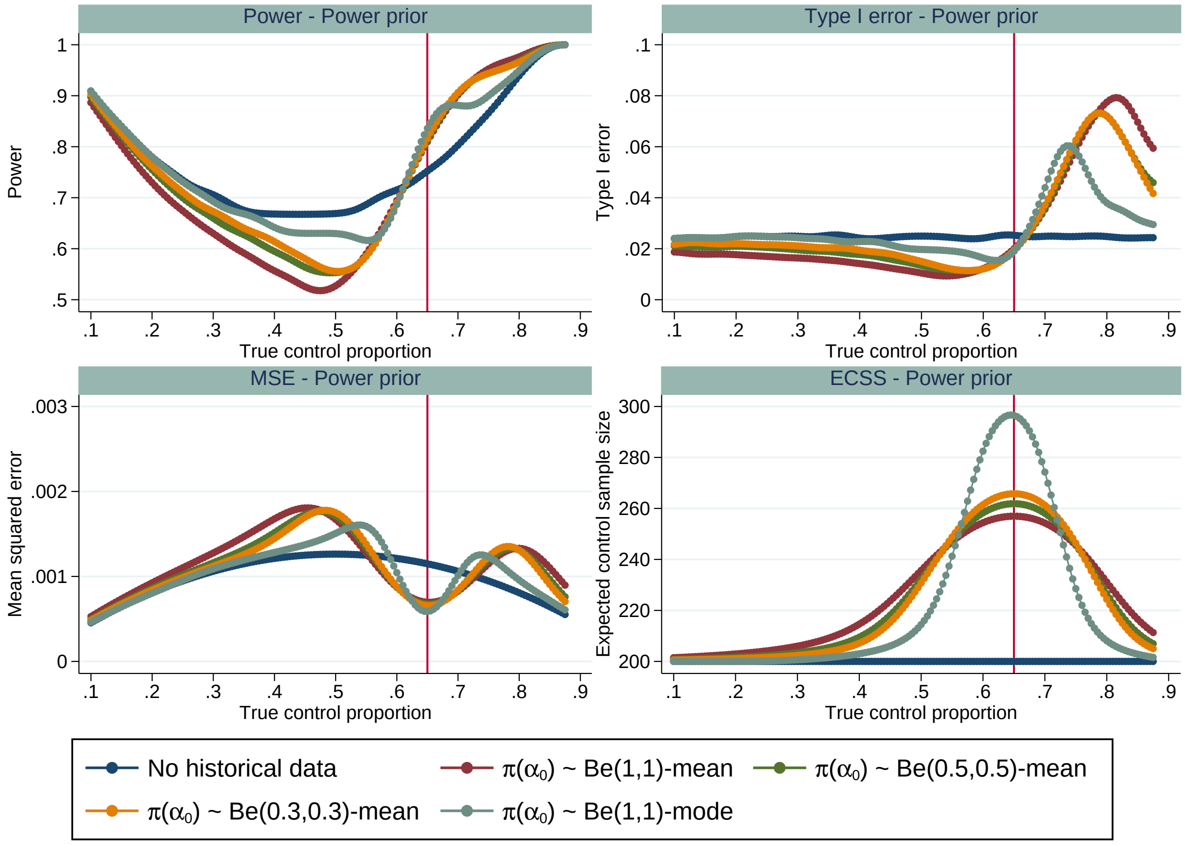


Table 2.5: Comparison of the design characteristics for all historical data methods considered – Viele example

Method	Power[†]	Type I error[†]	ECSS[†]	Max error	WI[†]	WE[†]	ECCSS [†]	MSE[††]
Additional information design								
Probability weight	0.8060	0.0229	266.46	0.0387	-	0.6646	198	(0.59,0.7)
Equivalence – one-sample (8% bounds)	0.8299	0.0195	290.53	0.0624	-	0.9053	198	(0.59,0.705)
Equivalence – two-sample (8% bounds)	0.8216	0.0195	276.34	0.0629	-	0.7634	198	(0.58,0.71)
Power prior – Be(1,1) – mean	0.8097	0.0200	256.93	0.0755	-	0.5693	198	(0.555,0.725)
Power prior – Be(0.5,0.5) – mean	0.8118	0.0194	261.84	0.0715	-	0.6184	198	(0.565,0.72)
Power prior – Be(0.3,0.3) – mean	0.8160	0.0194	265.71	0.0718	-	0.6571	198	(0.56,0.72)
Power prior – Be(1,1) – mode	0.8359	0.0190	296.50	0.0603	-	0.9650	198	(0.59,0.7)
Robust mixture prior – 0.9Be(x_h, y_h)+0.1Be(1,1)	0.8312	0.0165	296.58	0.1083	-	0.9658	198	(0.575,0.72)
Robust mixture prior – 0.5Be(x_h, y_h)+0.5Be(1,1)	0.8171	0.0178	283.53	0.0554	-	0.8353	198	(0.58,0.715)
Adaptive design								
Probability weight	0.7800	0.0185	209.75	0.0564	0.6070	0.6658	141.17	(0.615,0.68)
Equivalence – one-sample (8% bounds)	0.7852	0.0162	212.73	0.0844	0.7646	0.8364	127.30	(0.61,0.685)
Equivalence – two-sample (8% bounds)	0.7795	0.0166	205.43	0.0811	0.6674	0.7199	131.63	(0.605,0.69)
Power prior – Be(1,1) – mean	0.7714	0.0177	201.33	0.1011	0.5551	0.5635	142.98	(0.59,0.7)
Power prior – Be(0.5,0.5)– mean	0.7712	0.0170	201.59	0.0973	0.5983	0.6095	138.63	(0.59,0.695)
Power prior – Be(0.3,0.3)– mean	0.7751	0.0163	201.60	0.0986	0.6335	0.6461	134.98	(0.595,0.695)
Power prior – Be(1,1) – mode	0.8004	0.0155	222.39	0.0952	0.9175	0.9554	125.30	(0.61,0.685)
Robust mixture prior – 0.9Be(x_h, y_h)+0.1Be(1,1)	0.7831	0.0127	-	0.1529	98.47*	-	120.01	(0.6,0.695)
Robust mixture prior – 0.5Be(x_h, y_h)+0.5Be(1,1)	0.7557	0.0129	-	0.0787	84.50*	-	122.13	(0.605,0.695)

[†] At complete agreement between the current and historical control data.

[††] Range of MSE where the design incorporating historical data has a lower MSE than a design incorporating no historical data.

* Prior expected ESS at the interim analysis.

WI – historical data weight at interim analysis. WE – historical data weight at the end of the study. Max error – maximum type I error rate across all true current control proportions.

2.6 Normal approximation of the operating characteristics

When the beta posterior distributions for the control or treatment response probabilities do not contain an integer value, for example if an integer prior is not used for the response probability, then the iterative approach described by Cook [55] cannot be used to calculate the $\Pr(p_t > p_c)$ and numerical integration is required. A normal approximation can be used for the posterior distribution of the response probabilities to allow quicker calculation of the operating characteristics when the sample size of each treatment group is reasonably large (a common rule for approximating the binomial distribution by a normal distribution is that both np and $n(1-p)$ should be greater than 5, where n is the number of observations and p is the response probability, this rule can be applied here). Furthermore, if the additional information design is used and the historical data are given a fixed weight that does not vary depending on the agreement between the current and historical controls, a formula approximation can be derived for the operating characteristics of the design. We initially consider the additional information design incorporating historical data with a fixed weight in a frequentist framework.

For a standard trial design with a binary outcome, when no historical data are incorporated into the design or the final analysis of the current trial, the Bayesian and frequentist approaches to determining the operating characteristics provide similar results, as shown in the next section.

2.6.1 No historical data design

Frequentist

We are interested in testing the null hypothesis, $H_0 : p_c = p_t$ against the alternative hypothesis, $H_1 : p_c < p_t$.

Let $\hat{p}_c$ and $\hat{p}_t$ be the maximum likelihood estimates of the response probabilities in the control and treatment group, respectively. The estimated treatment effect is approximately normally distributed for reasonably large sample sizes in each treatment group [57],

$$\hat{p}_t - \hat{p}_c \approx \mathrm{N}\left(p_t - p_c, \frac{p_c(1-p_c)}{n_c} + \frac{p_t(1-p_t)}{n_t}\right).$$

Under the null hypothesis $p_c = p_t = p$ and,

$$\sigma^2_{\hat{p}_t-\hat{p}_c} = \frac{p_c(1-p_c)}{n_c} + \frac{p_t(1-p_t)}{n_t} = \frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_t}.$$

Since the value of p is not known, the pooled estimate $\hat{\bar{p}} = \frac{\hat{p}_c n_c + \hat{p}_t n_t}{n_c+n_t}$ is used to estimate $\sigma^2_{\hat{p}_t-\hat{p}_c}$,

$$\sigma^2_{\hat{p}_t-\hat{p}_c} \approx \frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{n_c} + \frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{n_t} = \hat{\bar{p}}(1-\hat{\bar{p}})\left(\frac{1}{n_c}+\frac{1}{n_t}\right).$$

A test statistic can then be constructed under the null hypothesis as,

$$Z = \frac{(\hat{p}_t - \hat{p}_c)}{\sqrt{\hat{\bar{p}}(1-\hat{\bar{p}})\left(\frac{1}{n_c}+\frac{1}{n_t}\right)}}.$$

For all possible control responses $x_c = 0, \ldots, n_c$ and all possible treatment responses $x_t = 0, \ldots, n_t$,

$\Pr(X_c = x_c \mid p_c, n_c) = \binom{n_c}{x_c} {p_c}^{x_c}(1-p_c)^{n_c-x_c}$ and $\Pr(X_t = x_t \mid p_t, n_t) = \binom{n_t}{x_t} {p_t}^{x_t}(1-p_t)^{n_t-x_t}$,

where $p_t = p_c + \Delta$ under the alternative hypothesis and $p_t = p_c$ under the null hypothesis, Δ denotes the treatment effect.

The power and type I error rate are then calculated as,

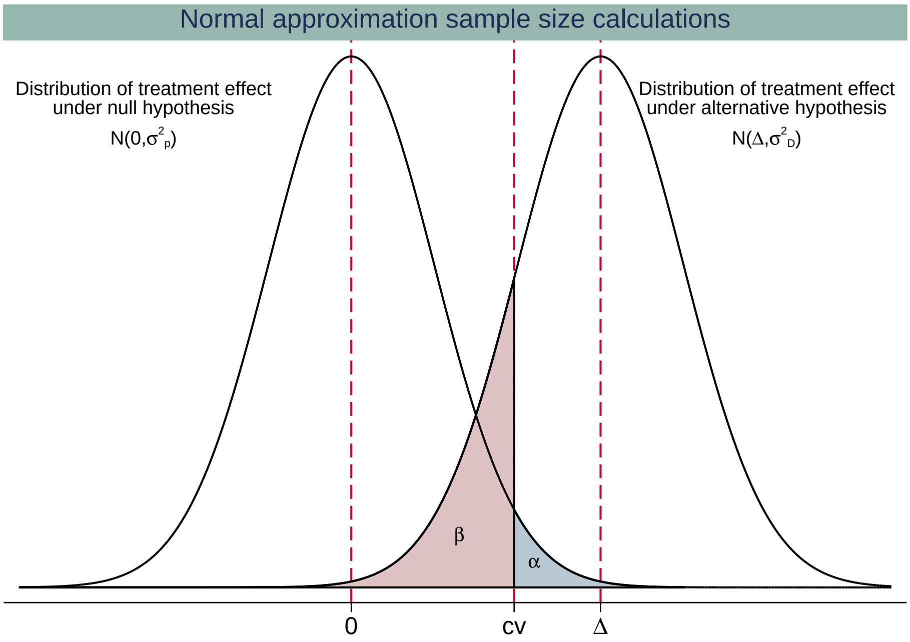
$$\begin{aligned} 1-\beta &= \sum_{x_c=0}^{x_c=n_c}\sum_{x_t=0}^{x_t=n_t} \Pr(X_c = x_c \mid p_c, n_c)\Pr(X_t = x_t \mid p_t, n_t)\mathbb{1}(Z > \Phi^{-1}(0.975) \mid x_c, x_t, n_c, n_t), \\ \alpha &= \sum_{x_c=0}^{x_c=n_c}\sum_{x_t=0}^{x_t=n_t} \Pr(X_c = x_c \mid p_c, n_c)\Pr(X_t = x_t \mid p_c, n_t)\mathbb{1}(Z > \Phi^{-1}(0.975) \mid x_c, x_t, n_c, n_t). \end{aligned} \tag{2.22}$$

where the sample sizes are large ($n_c > 200$ and $n_t > 200$), a large sample approximation may be used, given in the next section.

Approximate power - large samples

The large sample approximation is calculated from the assumed true distribution of the treatment difference under the null and alternative hypothesis using the central limit theorem. Figure 2.24 illustrates how the power and type I error rate are determined.

Figure 2.24: Sample size calculation assuming a large sample size normal approximation.


Normal approximation sample size calculations
Distribution of treatment effect under null hypothesis
N(0,σ²p)
Distribution of treatment effect under alternative hypothesis
N(Δ,σ²D)
β
α
0
cv
Δ

where cv denotes the critical value.

The power is calculated as [57],

$$\begin{aligned} 1-\beta = 1-\Phi\left(\frac{\text{cv}-\Delta}{\sigma_D}\right) &= 1-\Phi\left(\frac{\Phi^{-1}(0.975)\sigma_p-(p_t-p_c)}{\sigma_D}\right) \\ &= \Phi\left(\frac{(p_t-p_c)-\Phi^{-1}(0.975)\sigma_p}{\sigma_D}\right). \end{aligned} \tag{2.23}$$

Where, $\sigma_p = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_c}+\frac{1}{n_t}\right)}$, $\bar{p} = \frac{n_c p_c + n_t p_t}{n_c+n_t}$, $\sigma_D = \sqrt{\frac{p_t(1-p_t)}{n_t}+\frac{p_c(1-p_c)}{n_c}}$.

The approximate type I error rate is 2.5%, since,

$$\alpha = 1-\Phi\left(\frac{\text{cv}-0}{\sigma_p}\right) = 1-\Phi\left(\frac{\Phi^{-1}(0.975)\sigma_p}{\sigma_p}\right) = 1-\left(\Phi\left[\Phi^{-1}(0.975)\right]\right) = 0.025.$$

Bayesian

Assuming no prior data are available, a minimally-informative Beta(1,1) prior is used for the response probabilities in the control and treatment group and the posterior distributions are given by,

$$\pi(p_c \mid x_c, y_c) \sim \text{Beta}(1 + x_c, 1 + y_c),$$
$$\pi(p_t \mid x_t, y_t) \sim \text{Beta}(1 + x_t, 1 + y_t).$$

The power is then,

$$\begin{aligned} 1 - \beta = \sum_{x_c=0}^{x_c=n_c} \sum_{x_t=0}^{x_t=n_t} & \Pr(X_c = x_c \mid p_c, n_c) \Pr(X_t = x_t \mid p_t, n_t) \\ & \times \mathbb{1}(\Pr(p_t > p_c) > 0.975 \mid x_c, x_t, n_c, n_t). \end{aligned} \tag{2.24}$$

The type I error rate is calculated using Equation 2.24 assuming the true underlying value of p_t is p_c when calculating $\Pr(X_t = x_t \mid p_t, n_t)$. The $\Pr(p_t > p_c)$ is calculated using the iterative procedure proposed by Cook [55].

Operating characteristics comparison for the Viele example incorporating no historical data

For the example considered here, 200 patients are available per treatment group. The current control proportion varies with the treatment effect always 12% higher. Figure 2.25 compares the operating characteristics of this design using Equations 2.22 (Normal exact), 2.23 (Normal large sample approx) and 2.24 (Bayesian exact).

All approaches give similar design characteristics as expected. The operating characteristics for the exact approaches are not continuous functions since the number of responses are discrete values.

2.6.2 Incorporating historical data with a fixed weight

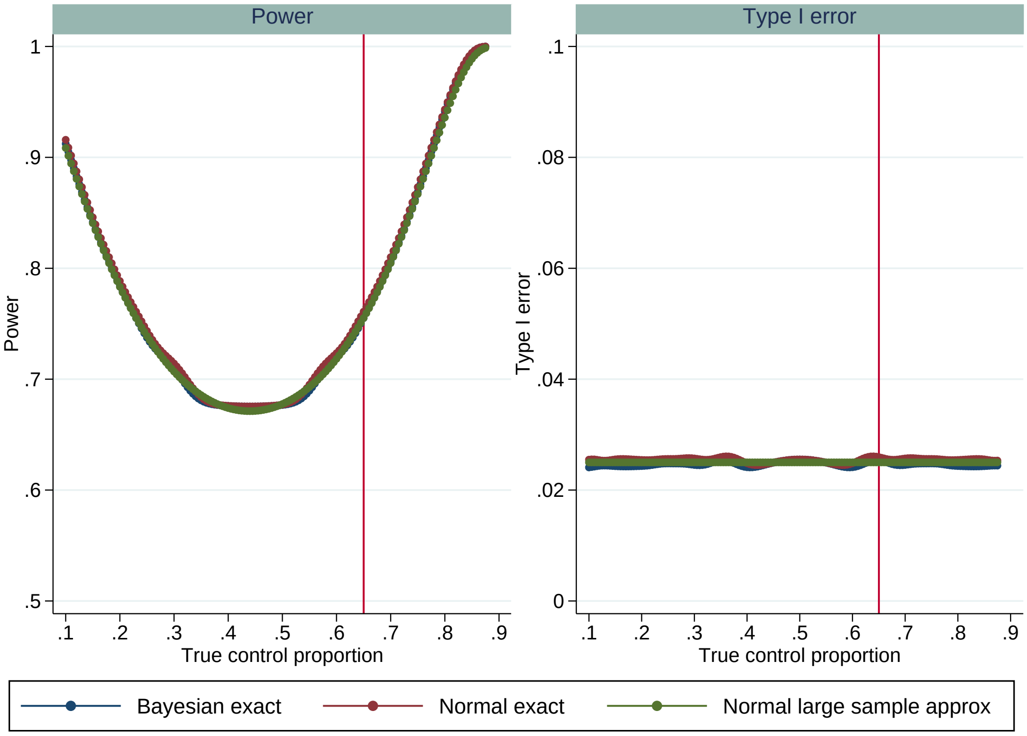
In this section it is assumed that only one historical study is available and a fixed weight is given to the historical data w, chosen based on expert opinion.

Frequentist

Let $\hat{p}_0$ denote the control response proportion incorporating both historical and current control data, to differentiate from the response proportion in the current controls only, which is denoted, $\hat{p}_c$. Then,

$$\hat{p}_0 = \frac{x_c + x_h w}{n_c + n_h w}. \tag{2.25}$$

Figure 2.25: Operating characteristics for the exact Bayesian design, the exact frequentist approach assuming a normal Z-statistic and the large sample approximation assuming the test statistic follows a normal distribution – no historical data design. Viele example, $n_c = n_t = 200$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



Since x_h and w do not vary, $x_h w$, $n_h w$ and n_c are all fixed constants in this estimator.

Then,

$$\mathbf{E}(\hat{p}_0 \mid x_h, n_h, n_c, w) = \frac{\mathbf{E}(x_c) + x_h w}{n_c + n_h w} = \frac{n_c p_c + x_h w}{n_c + n_h w},$$
$$\mathrm{Var}(\hat{p}_0 \mid n_h, n_c, w) = \mathrm{Var}\left(\frac{x_c}{n_c + n_h w}\right) = \frac{n_c p_c (1 - p_c)}{(n_c + n_h w)^2},$$

and $\hat{p}_t - \hat{p}_0$ is approximately normally distributed with mean $p_t - \dfrac{n_c p_c + x_h w}{n_c + n_h w}$ and variance $\dfrac{n_c p_c (1 - p_c)}{(n_c + n_h w)^2} + \dfrac{p_t(1 - p_t)}{n_t}$.

A test statistic is constructed under the null hypothesis of $p_c = p_t$, but the control estimate is $\hat{p}_0$, which incorporates the historical data. Note that this test statistic will only follows a normal distribution with mean zero and variance one under the assumption that $p_h = p_c = p_t$. The pooled estimate of the true response probability under the null hypothesis is then,

$$\hat{\hat{p}} = \frac{(n_c + n_h w)\hat{p}_0 + n_t\hat{p}_t}{n_c + n_h w + n_t} = \frac{n_c\hat{p}_c + n_h w\hat{p}_h + n_t\hat{p}_t}{n_c + n_h w + n_t},$$

and the test statistic, under the null hypothesis, is given by,

$$Z = \frac{(\hat{p}_t - \hat{p}_0)}{\sqrt{\hat{\hat{p}}(1-\hat{\hat{p}})\left(\frac{n_c}{(n_c+n_h w)^2} + \frac{1}{n_t}\right)}}. \tag{2.26}$$

To explore the effect of the historical data, the current control response probability is varied. The power is calculated assuming $p_t = p_c + 0.12$ and the type I error rate assuming $p_t = p_c$, using the formulas given in Equation 2.22, where Z is now also dependent on the fixed weight given to the historical data, w.

The large sample size approximation is calculated assuming the true distribution of the control response proportion, p_0, is $\mathrm{N}\left(\dfrac{n_c p_c + p_h n_h w}{n_c + n_h w}, \dfrac{n_c p_c(1-p_c)}{(n_c + n_h w)^2}\right)$. Therefore, Equation 2.23 can be used to calculate the power of this design using the assumed distribution for p_0 instead of p_c.

Bayesian

Using the Bayesian approach, at the end of a single study the posterior distributions for the response probabilities in the control and the treatment groups are given by,

$$\begin{aligned} \pi(p_c \mid x_c, y_c, x_h, y_h, w) &\sim \mathrm{Beta}(1 + x_h w + x_c, 1 + y_h w + y_c), \\ \pi(p_t \mid x_t, y_t) &\sim \mathrm{Beta}(1 + x_t, 1 + y_t), \end{aligned} \tag{2.27}$$

and the power and type I error rate are calculated as in Equation 2.14. The $\Pr(p_t > p_c)$ is calculated using the iterative procedure proposed by Cook [55].

Normal approximation of the Bayesian design - exact operating characteristics using the Z test statistic

If instead of the frequentist approach, a normal test statistic is constructed based on approximating the beta posterior distributions for the control and treatment response probabilities by a normal distribution,

$$\begin{aligned}\pi(p_c \mid x_c, n_c, x_h, n_h, w) &\sim \mathrm{N}\left(\hat{p}_0, \frac{\hat{p}_0(1-\hat{p}_0)}{wn_h + n_c}\right),\\ \pi(p_t \mid x_t, n_t) &\sim \mathrm{N}\left(\hat{p}_t, \frac{\hat{p}_t(1-\hat{p}_t)}{n_t}\right),\end{aligned} \tag{2.28}$$

where $\hat{p}_0$ is given in Equation 2.25.

The test statistic is given by,

$$Z_B = \frac{(\hat{p}_t - \hat{p}_0)}{\sqrt{\hat{\bar{p}}(1-\hat{\bar{p}})\left(\frac{1}{n_c+n_h w} + \frac{1}{n_t}\right)}}, \tag{2.29}$$

where,

$$\hat{\bar{p}} = \frac{n_c\hat{p}_c + n_h w \hat{p}_h + n_t \hat{p}_t}{n_c + n_h w + n_t},$$

and the operating characteristics can be calculated using Equation 2.22.

Large sample normal approximation of the Bayesian design operating characteristics

A formula approximation of the type I error rate of the Bayesian design is given by,

$$\alpha = \Phi\left(\frac{(p_c - p_0) - \Phi^{-1}(0.975) \times \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_c+n_h w} + \frac{1}{n_t}\right)}}{\sqrt{\left(\frac{p_c(1-p_c)}{n_t} + \frac{n_c p_c(1-p_c)}{(n_c+n_h w)^2}\right)}}\right), \tag{2.30}$$

and the power is given by,

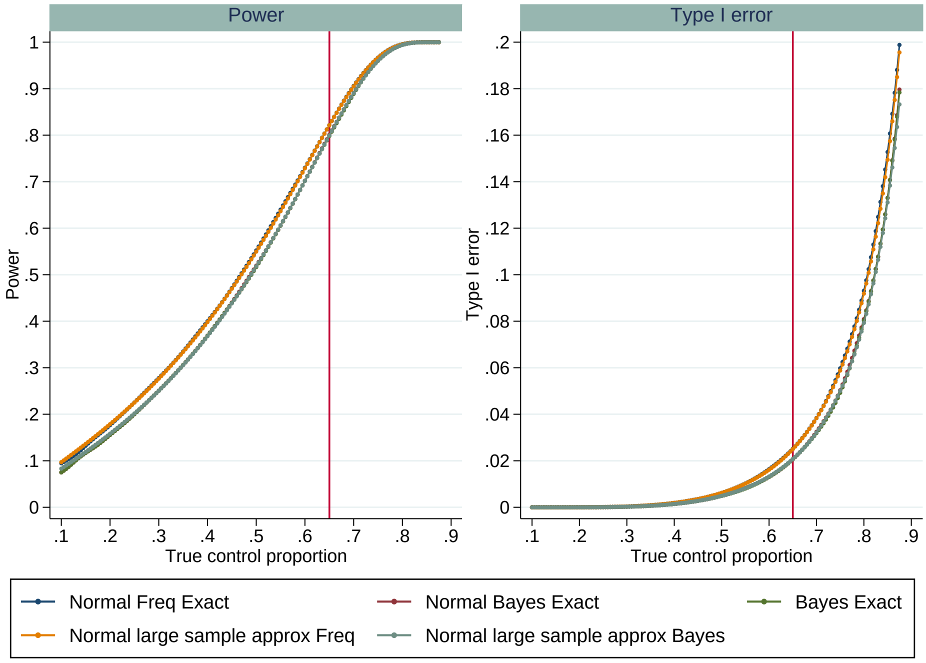
$$1-\beta = \Phi\left(\frac{(p_c - p_0) - \Phi^{-1}(0.975) \times \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_c+n_h w} + \frac{1}{n_t}\right)}}{\sqrt{\left(\frac{p_t(1-p_t)}{n_t} + \frac{n_c p_c(1-p_c)}{(n_c+n_h w)^2}\right)}}\right), \tag{2.31}$$

where $p_t = p_c + 0.12$ and $\bar{p} = \frac{n_c p_c + n_h w p_h + n_t p_t}{n_c + n_h w + n_t}$.

Operating characteristics comparison for the Viele example incorporating historical data with a fixed weight of 0.4

The example used here assumes the historical data are incorporated as additional information into the control arm of the current study and given a weight of 0.4. There are 200 patients in each of the current trial control and treatment arm and the treatment effect is assumed to be 12%. Figure 2.26 compares the exact Bayesian design operating characteristics (Bayes Exact calculated using the posterior distributions in Equation 2.27), the normal approximation using the Bayesian variance (Normal Bayes exact calculated using the test statistic in Equation 2.29), the frequentist approach using the exact operating characteristics (Normal Freq calculated using the test statistic in Equation 2.26), the frequentist large sample approximations and finally the large sample formula approximation of the Bayesian design operating characteristics (the Bayesian large sample approximations are calculated from Equations 2.30 and 2.31). The approximations give very similar results but the Bayesian and freqentist approaches differ due to the way the variance is calculated.

Figure 2.26: Operating characteristics comparison for the fixed weight historical data design. Viele example, $n_c = n_t = 200$, $\Delta = 12\%$, historical data, 65/100 responses, $w = 0.4$. The vertical red lines represent complete agreement between the historical and current control proportions.



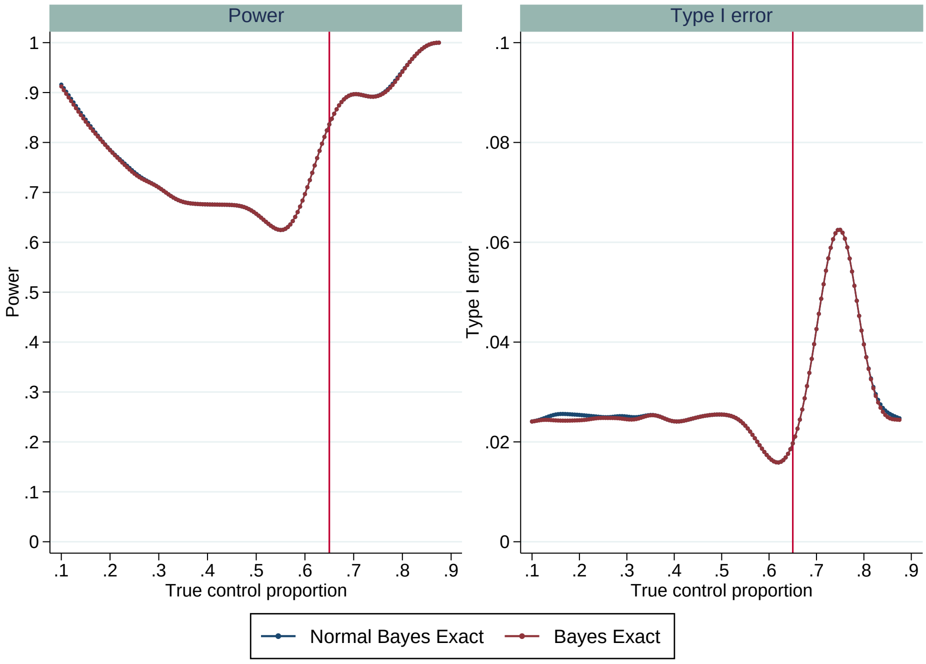
2.6.3 Incorporating historical data with a variable weight

Returning to the additional informative historical data design described in Section 2.4. For this design, the weight given to the historical data varies depending on the agreement between the historical and current controls. The weight used in this section for illustration is the one-sample equivalence weight calculated using Equation 2.11 and the posterior distributions are given in Equation 2.13. Since the weight varies, the operating characteristics can only be determined using the exact calculation based on the beta posterior distributions directly using the iterative procedure proposed by Cook [55] or from constructing a test statistic based on approximating the beta posterior distributions for the control and treatment response probabilities by a normal distribution, as in Equation 2.29. The weight is now dependent on the observed control responses and the historical data instead of being a fixed value. The normal test statistic approach is slightly quicker for calculating the operating characteristics than the exact approach using the method by Cook and is required when the method proposed by Cook can not be used to calculate the $\Pr(p_t > p_c)$, for example if none of the beta distribution parameter values are integers. A similar approach can be used for the adaptive design.

Operating characteristics comparison for the Viele example incorporating historical data with a variable weight - additional information design

Figure 2.27 shows the operating characteristics from the exact Bayesian approach and the normal approximation based on the construction of a z-statistic from the normal approximation of the posterior distribution of the control and treatment response probabilities. The one-sample equivalence weight approach with 8% equivalence bounds is used to calculate the weight to give to the historical data. The exact Bayesian approach and the normal approximation give similar operating characteristics.

Figure 2.27: Operating characteristics comparison for the exact Bayesian approach and the normal approximation of the posterior distributions approach using the one-sample equivalence weight with 8% equivalence bounds. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



2.7 Discussion

In this chapter, two intuitive and computationally tractable approaches for assessing agreement between the historical and current control data are proposed. A design which incorporates historical data as additional information or an adaptive design that replaces current controls with historical controls can then be utilised which allow the possibility of borrowing historical data when it is in agreement with the current control data. We propose using the method described by Cook [55] to calculate $\Pr(p_t > p_c)$ which allows the operating characteristics of the proposed designs to be calculated exactly. The equivalence weight approach is flexible and allows control over how quickly the historical data are discounted when there is disagreement between the historical and current controls. The maximum inflation in the type I error rate across all possible current control response probabilities can be calculated and the maximum possible type I error rate can be controlled by the choice of equivalence bounds. The equivalence weight is an intuitive way to think about discounting historical data and should provide an approach that is easy to discuss with clinicians about utilising historical data and the effect that using historical data has on the design characteristics of the current study.

We do not advocate the use of these methods in trials that can easily recruit the required number of patients under a standard design. But where recruitment is slow or the patient population is small, these methods could potentially increase the power or reduce the duration of trials that are not feasible to run under a standard design.

Comparing the probability and equivalence weight approaches to historical data methods proposed in the literature. All historical data methods work in a similar way, discounting the historical data when there is disagreement between the current and historical controls, but the methods differ in the amount of information borrowed and the rate of discounting as conflict increases. In this example, we found the power prior with a prior on the power did not give an intuitive weight to the historical data and the commensurate prior was difficult to use in this adaptive setting. The commensurate prior approach was computationally intensive. The priors on the parameters that govern the borrowing in both the commensurate prior and power prior approaches has a large effect on the down-weighting of the historical data and these prior require careful thought. The robust mixture prior approach works well. However, calculating the ESS is computationally intensive and negative values can be obtained which is not intuitive and requires setting the ESS to zero when used in an adaptive design setting. Otherwise the design would require more controls to be randomised than a standard trial design not incorporating any historical data. For one historical study, the effect of the weight given to the informative component of the robust mixture prior needs exploration but can be used to control the maximum type I error rate across a range of true control response probabilities.

The probability and equivalence approaches were chosen to give a large weight when there is complete agreement between the historical and current controls. If there is a substantial amount of historical data, it may not be desirable to give the historical data a large weight because the historical data would have a greater influence on the control parameter estimate than the current control data. In this case a maximum weight may be chosen that caps the amount of historical data incorporated into the final analysis. A large difference in current and historical control sample sizes is unlikely when there is only one historical study.

For the analysis of the power prior approach, the fully Bayesian approach was only considered briefly in Appendix D, throughout the thesis the modified power approach is used taking a summary measure of the marginal posterior distribution of the power as a fixed weight. These two approaches to using the modified power will give slightly different operating characteristics, the summary measure approach was used in this thesis as the fully Bayesian version is computationally intensive.

For all methods, only one interim analysis was considered, multiple comparisons of the historical controls with the current control data as it accumulates would give more power to detect a difference at the expense of computation time.

For all historical data methods, careful thought is required for what are appropriate historical data. Control data may be available from multiple studies spanning many years. It is better to include a few studies that are carefully selected following Pocock's criteria for incorporating historical data [11], than to include all studies where there is a lot of heterogeneity. Selecting a few historical studies with low heterogeneity will result in a larger prior effective sample size than using all historical studies where there is lots of heterogeneity. In this chapter, only one historical study has been considered but methods for incorporating multiple historical studies have been proposed for the robust mixture prior, commensurate prior and power prior methods [14, 19, 23].

Chapter 3

Historical data methods for the design and analysis of a trial with a normally distributed outcome

3.1 Introduction

This chapter is structured in a similar way to Chapter 2. The aim of this chapter is to address the five questions about the use of historical data posed by Pocock [11] when the outcome data are normally distributed, these questions are described in Section 2.1. Throughout this chapter it is assumed that the historical data chosen are relevant to the current study taking into account Pocock's six acceptability criteria for using historical data in the design and analysis of a new study [11], the acceptability criteria are listed in Section 1.2. In this chapter, it is assumed that only one relevant historical study is available, with data available on the control arm only. This chapter is structured as follows: how to assess agreement between the historical and current controls is discussed in Sections 3.2 and 3.3; how to incorporate historical data into the design of a current study is discussed in Section 3.4; and finally how to incorporate the historical data in the analysis of a current study is addressed in Sections 3.5.1 and 3.5.4.

The two approaches proposed in Chapter 2 for assessing agreement between historical and current controls, the probability weight and the equivalence probability weight, described in Section 2.3, are extended to handle normally distributed outcome data. The analysis approach of the power prior is used to incorporate the historical data into the final analysis of the current study. The probability weight and equivalence probability weight are compared to two historical data methods proposed in the literature: the modified power prior and the robust mixture prior. The weight given to the historical data is compared between these approaches. The operating characteristics for the power prior approach with a summary measure of the marginal distribution of the power used as a fixed weight and the probability and equivalence weights used as a fixed weight are compared. The Commensurate prior approach is also discussed.

It is assumed throughout this chapter that there is one relevant historical study available and that this historical study provides information on the control arm only. The maximum number of additional patients that the historical study may provide is the sample size of the historical study. Depending on the agreement between the historical and current control data, the historical data may be down-weighted and therefore the additional number of patients that the historical data provides will be reduced. The aim here is to assess the conflict between the historical and current control data to determine how much weight to give the historical data in the final analysis.

When outcome data are binary, the distribution of the response probability in the historical data is compared to the distribution of the response probability in the current control data. There is only one parameter, the response probability, that summarises the data in each sample and the variance is completely dependent on the mean for the true underlying response probability. For normally distributed outcome data, two parameters are required to describe the distribution of the outcome data, the mean and the variance. The final analysis for a trial with normal outcome data, which compares a treatment group to control, usually compares only the means of the two samples. The sample size is calculated to achieve a desired power and type I error rate for the comparison of the means in the control and treatment group. For the final analysis and the sample size calculation, comparing the treatment mean to the control mean, it is assumed that the variances are either: known and equal; known and unequal; or unknown in both the treatment and control group.

When incorporating historical data into the control arm of the current trial, it is important to consider the whole distribution of the data. A difference in either the mean or the variance could indicate that the historical and current control data represent two different populations for which we would want to discount the historical data in the current trial analysis.

3.1.1 Notation

Let x_{hi}, x_{ci} and x_{ti} denote the observed outcome value for patient i in the historical, current control and treatment group respectively. Let $\bar{x}_h$ denote the sample mean of the historical control data and $\hat{\sigma}_h^2$ the sample variance of the historical data. Let $\bar{x}_c$, $\hat{\sigma}_c^2$ and $\bar{x}_t$, $\hat{\sigma}_t^2$ be the corresponding sample estimates for the current controls and treatment group respectively. Let n_h, n_c and n_t denote the sample sizes of the historical, current control and treatment group respectively. Let μ_c be the true underlying mean in the current control arm and μ_t the true underlying mean in the treatment arm. Let σ_c^2 and σ_t^2 be the true underlying variance in the control and treatment arms respectively. Let τ_c denote the precision for the current control arm $\tau_c = 1/\sigma_c^2$ and τ_t denote the precision for the treatment arm. Where it is assumed that the true underlying mean and variance in the current and historical controls may differ, the true underlying mean, variance and precision in the historical controls are denoted by μ_h, σ_h^2 and τ_h, respectively.

3.1.2 Illustrative example

Throughout this chapter, unless otherwise stated, the main example used is an adaptation of the example from Viele et al. [42] for normally distributed outcome data. We consider this design representative of a confirmatory trial. The primary analysis of interest is a hypothesis test of $H_0 : \mu_c = \mu_t$ against $H_1 : \mu_c < \mu_t$. A standard two-arm randomised controlled trial incorporating no historical data would require 200 patients per treatment arm to detect a mean difference of 12 with a one-sided type I error rate of 2.5% and approximately 76% power, assuming the standard deviation in the control and treatment group are known and equal to 45. In addition, for the historical data designs, there are 100 historical control patients available with a sample mean of 65 and sample standard deviation 45.

Throughout the chapter, for the primary analysis comparing the treatment mean to control mean in the current trial, it is assumed that the variance in the treatment group is the same as the variance in the control group. Differences in both the means and variances in the current and historical control data are explored.

3.2 Published methods for assessing agreement between historical and current control data

For the historical data methods discussed in this section, each method assesses the agreement between the historical and current control data and calculates either a weight w or a prior effective sample size (ESS). For methods that calculate a weight, this weight is used to down-weight the historical data and we define the effective historical sample size ($EHSS$) to be the weight times the historical sample size wn_h. In Chapter 2 a distinction was made between the $EHSS$ which is based only on the historical data and the prior ESS which also incorporates the information contained in the prior before the historical data are observed. In this chapter the initial priors assumed for the parameters before the historical data are observed are reference priors and therefore $ESS \approx EHSS$.

3.2.1 Modified power prior

The modified power prior assumes that the current and historical data are estimating the same underlying parameters of interest ($\mu_h = \mu_c$ and $\sigma_h^2 = \sigma_c^2$). A reference prior [58] is chosen as the initial joint prior for the mean and variance in the control arm, μ_c and σ_c^2. The reference prior is, $\pi(\mu_c, \sigma_c^2) \propto 1/\sigma_c^2$ [43], this prior is derived in Section 3.3.2. The prior for the power, denoted $\pi(\alpha_0)$ is assumed to be a Beta(a,b) distribution. The same priors proposed in Section 2.2.1 for binary data are also considered here. These priors are: Beta(1,1); Beta(0.5,0.5); and Beta(0.3,0.3). For normally distributed outcome data and only one historical study, the joint modified power prior distribution of $(\mu_c, \sigma_c^2, \alpha_0)$ is given by [15, 43],

$$\begin{aligned}\pi(\mu_c, \sigma_c^2, \alpha_0 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h) &\propto \frac{(\sigma_c^2)^{\frac{\alpha_0 n_h}{2}-1}\exp\left\{-\frac{\alpha_0 n_h}{2\sigma_c^2}[\hat{\sigma}_h^2 + (\mu_c - \bar{x}_h)^2]\right\}\alpha_0^{a-1}(1-\alpha_0)^{b-1}}{\int\limits_0^{\infty}\int\limits_{-\infty}^{\infty}(\sigma_c^2)^{\frac{\alpha_0 n_h}{2}-1}\exp\left\{-\frac{\alpha_0 n_h}{2\sigma_c^2}[\hat{\sigma}_h^2 + (\mu_c - \bar{x}_h)^2]\right\}d\mu_c d\sigma_c^2} \\ &\propto \frac{\alpha_0^{\frac{\alpha_0 n_h}{2}+a-1}(1-\alpha_0)^{b-1}}{\left(\frac{2\sigma_c^2}{n_h\hat{\sigma}_h^2}\right)^{\frac{\alpha_0 n_h}{2}+1}\Gamma(\frac{\alpha_0 n_h - 3}{2}+1)}\exp\left\{-\frac{\alpha_0 n_h}{2\sigma_c^2}[\hat{\sigma}_h^2 + (\mu_c - \bar{x}_h)^2]\right\},\end{aligned} \tag{3.1}$$

where $\alpha_0 \in (1/n_h, 1]$, $\mu_c \in (-\infty, \infty)$, $\sigma_c^2 \in [0, \infty)$ and,

$$\hat{\sigma}_h^2 = \frac{1}{n_h}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2,\ \hat{\sigma}_c^2 = \frac{1}{n_c}\sum_{i=1}^{n_c}(x_{ci} - \bar{x}_c)^2,\ \bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi} \text{ and } \bar{x}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}x_{ci}.$$

Note that here the sample variance estimators are the maximum likelihood estimates as specified in [15]. The joint prior defined in Equation 3.1 is zero outside the range $\alpha_0 \in (1/n_h, 1]$. The joint prior is only defined in the region of $\alpha_0 \in (1/n_h, 1]$ since this is the region where the double integral in the denominator of Equation 3.1 is finite. The lower bound of α_0 implies that some historical data are automatically taken into account, depending on the availability of historical data. When there are no historical data available $n_h = 0$ and the joint prior for the current study control arm would be the reference prior $\pi(\mu_c, \sigma_c^2) \propto 1/\sigma_c^2$.

Combining the joint power prior given in Equation 3.1 with the normal likelihood for the current study control data and integrating out μ_c and σ_c^2, the marginal posterior distribution for the power, α_0 is given by [43],

$$\pi(\alpha_0 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c) \propto \frac{\alpha_0^{\frac{\alpha_0 n_h}{2}+a-1}(1-\alpha_0)^{b-1}\Gamma\left(\frac{\alpha_0 n_h+n_c-3}{2}+1\right)}{\left[\frac{\alpha_0 n_c}{\alpha_0 n_h+n_c}\frac{(\bar{x}_h-\bar{x}_c)^2}{\hat{\sigma}_h^2}+\alpha_0+\frac{n_c}{n_h}\frac{\hat{\sigma}_c^2}{\hat{\sigma}_h^2}\right]^{\frac{\alpha_0 n_h+n_c-3}{2}+1}\Gamma(\frac{\alpha_0 n_h-3}{2}+1)},$$

in the range $\alpha_0 \in (1/n_h, 1]$.

When α_0 is assumed to be unknown and given a distribution, the marginal distributions of μ_c and σ_c^2 cannot be written in closed form. Here, we calculate the mean, median or mode of the power distribution, using formulae similar to those given in Section 2.2.1, and use this as a fixed power prior weight. The conditional distribution of μ_c given a fixed value of α_0 can then be used for inference. This is described in Section 3.5.1.

Figure 3.1 shows three different priors for α_0, a Beta(1,1), Beta(0.5,0.5) and a Beta(0.3,0.3), and the posterior distributions of α_0 for each prior, given different levels of agreement between the historical and current controls. It is assumed that there are 100 current control patients and 100 historical patients. The historical control data are fixed with sample mean, $\bar{x}_h = 65$ and sample standard deviation, $\hat{\sigma}_h = 45$. The initial joint prior $\pi(\mu_c, \sigma_c^2) \propto (1/\sigma_c^2)$ is assumed before the historical data are observed.

Figure 3.1: Marginal distributions of α_0 for different observed current control sample means and standard deviations and different priors on the power. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls.

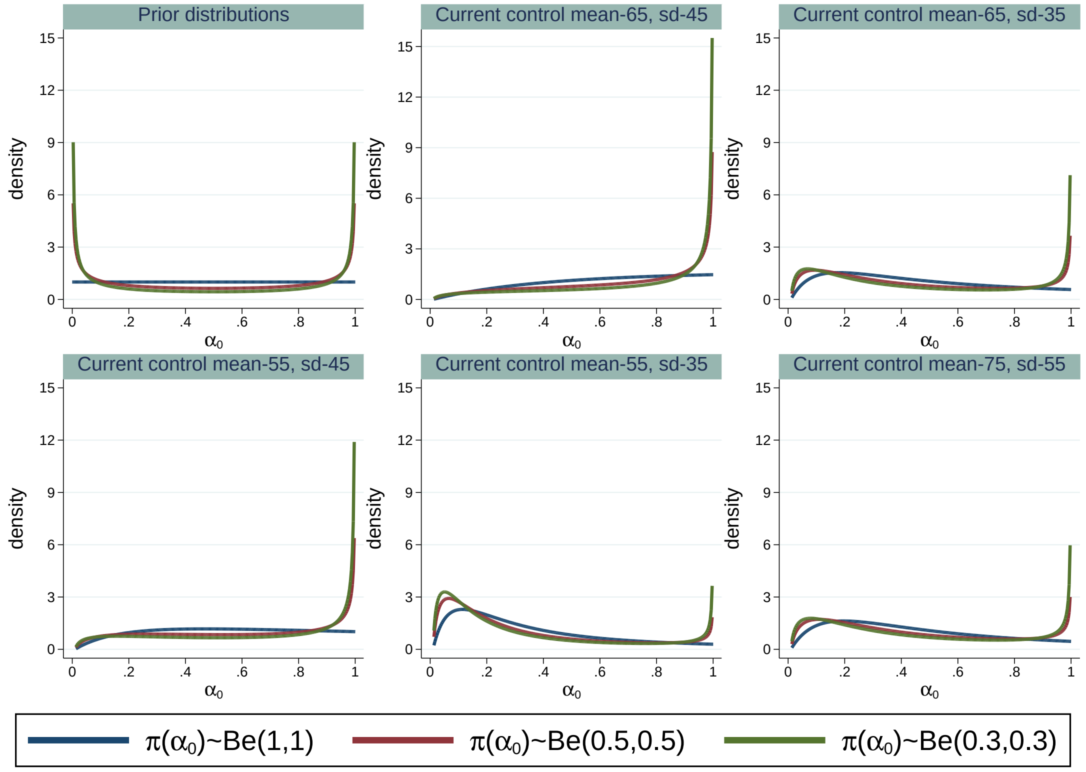


Figure 3.1 illustrates that under a Beta(1,1) prior for α_0, the marginal posterior distribution of α_0 is flat at complete agreement between the historical and current data ($\bar{x}_c = \bar{x}_h = 65$ and $\hat{\sigma}_c = \hat{\sigma}_h = 45$). The quasi-dichotomous priors result in a posterior distribution for α_0 with more mass near zero and more mass near one depending on the level of agreement between the current and historical data. Choosing smaller parameter values for the quasi-dichotomous priors governs the amount of mass at the tails of the posterior distribution. Small parameter values can cause problems when calculating summary measures of the power parameter using numerical integration because of the amount of mass at the tails of the distribution.

Figure 3.2 shows the mean of the posterior distribution of α_0 for priors Beta(1,1), Beta(0.5,0.5) and Beta(0.3,0.3) for a range of observed current control means and standard deviations in the current trial. Figure 3.2 also shows the mode of the posterior distribution of α_0 for a Beta(1,1) prior for a range of observed control means and standard deviations in the current trial. These summary measures are used as a fixed weight to down-weight the historical data.

Table 3.1: Mean and mode of the marginal distribution of α_0 for different observed current control means and standard deviations and different priors on the power. Example, historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$, $n_h = 100$ and $n_c = 100$.

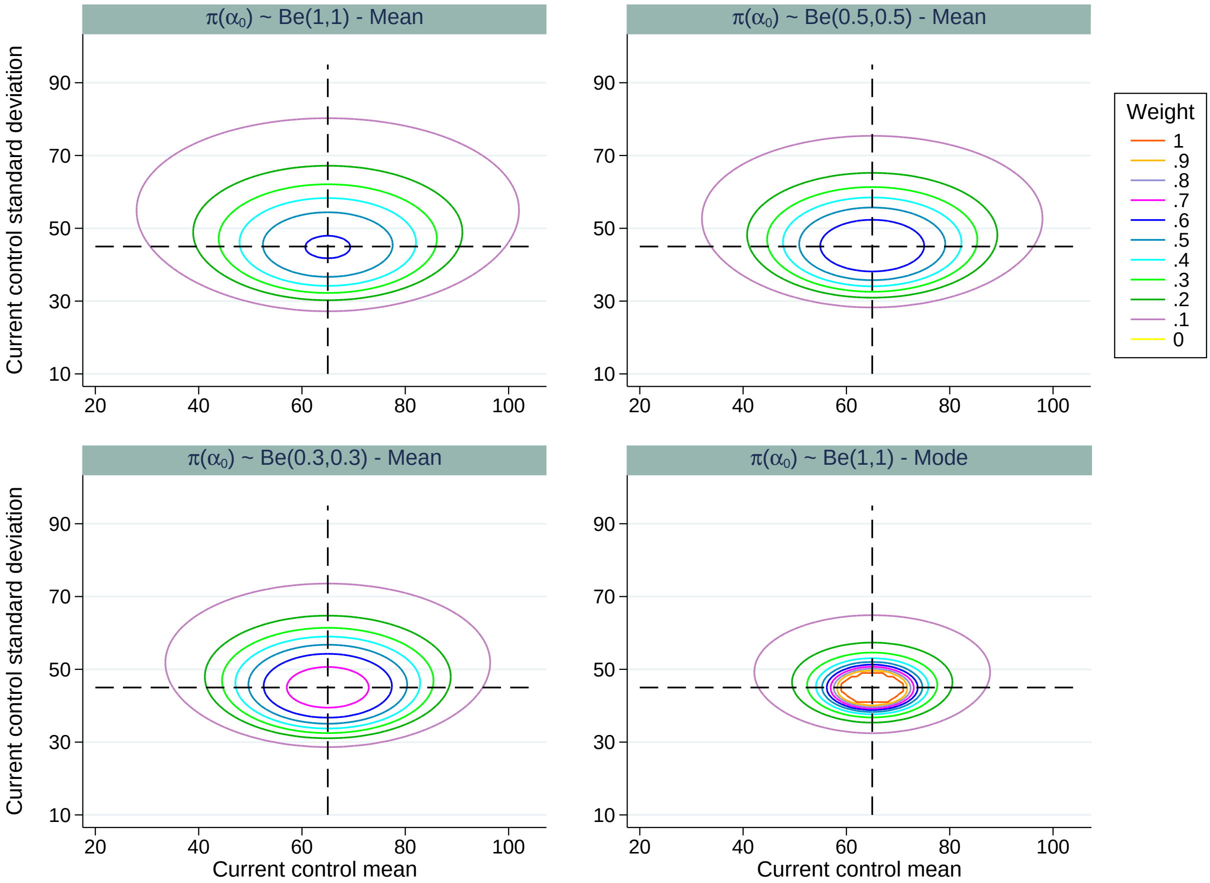
μ_c	σ_c		Mean (α_0)		Mode (α_0)
		Beta(1,1)	Beta(0.5,0.5)	Beta(0.3,0.3)	Beta(1,1)
65	45	0.613	0.689	0.757	1
65	35	0.438	0.459	0.496	0.179
55	45	0.543	0.601	0.662	.471
55	35	0.343	0.332	0.345	0.114
75	55	0.423	0.437	0.469	0.196

Table 3.2: Median and credible interval of the marginal distribution of α_0 for different observed current control means and standard deviations and different priors on the power. Example, historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$, $n_h = 100$ and $n_c = 100$.

			Median(95% credible interval)	
μ_c	σ_c	Beta(1,1)	Beta(0.5,0.5)	Beta(0.3,0.3)
65	45	0.637 (0.124,0.983)	0.760 (0.101,0.999)	0.875 (0.089,1)
65	35	0.395 (0.054,0.957)	0.383 (0.022,0.996)	0.393 (0,1)
55	45	0.543 (0.088,0.975)	0.628 (0.058,0.999)	0.742 (0.034,1)
55	35	0.276 (0.037,0.920)	0.217 (0.001,0.985)	0.176 (0,0.999)
75	55	0.379 (0.056,0.947)	0.359 (0.024,0.994)	0.358 (0,1)

Using the modified power prior approach, the mean weight given to the historical data for all priors on α_0 is maximised at complete agreement between the historical and current

Figure 3.2: Contour plots of the modified power prior weight for different observed control means and standard deviations in the current trial and different priors on α_0. Example, historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$, $n_h = 100$ and $n_c = 100$. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



controls, when the sample means and variances are the same. The weight decreases as the difference in either the means increases, which is measured in absolute terms or the differences in the variances increases, which is measured in relative terms. Taking the mode of the posterior distribution, assuming a Beta(1,1) prior gives a weight of one at complete agreement and also for a range of current control means and standard deviations around the historical sample estimates of the mean and standard deviation, the mode then decreases to zero as the mean and standard deviation differences between the current and historical data increase, this is why the contour line for the mode giving a weight of one is not smooth. Similar to the binary data case, as seen in Section 2.2.1, using the mean of the marginal distribution of α_0 gives a lower weight at complete agreement than using the mode or median of the marginal posterior distribution, as illustrated in Tables 3.1 and 3.2. The weight at complete agreement using the mean ranges from 0.61 to 0.76 and for the median ranges from 0.64 to 0.88, for the different priors on α_0. As illustrated in Figure 3.2, using the mean of the posterior distribution of the power discounts slowly as the difference between the historical and current controls increases, compared to using the mode as a weight. The weights for the power prior are symmetric around the historical data mean but not around the standard deviation. At agreement in the historical and current controls the posterior distribution of α_0 is relatively flat. Therefore, although the mode is one at complete agreement, which is desirable, this is taking the maximum of an almost flat distribution.

3.2.2 Robust mixture prior

Assuming known and equal variance in the current and historical control groups

Initially, the robust mixture prior proposed by Schmidli et al. [23] is considered where it is assumed that the control data follow a distribution from the regular one-parameter exponential family and the variance is assumed to be known and the same in the historical and current control data ($\sigma_h^2 = \sigma_c^2$). For one historical study, the robust mixture prior for μ_c is a two-component mixture distribution, with a mixture component that is conjugate and based on the historical data and a weakly-informative conjugate prior component. For normally distributed outcome data, the weakly-informative prior component of the mixture distribution is a unit information prior [23, 59]. The unit information prior is a data dependent prior with the mean as the maximum likelihood estimate (MLE) of the current control data and precision equal to the information provided by one observation [59]. The unit information prior is a weakly-informative prior since it has a larger variance than the variance of the historical data component of the mixture prior but it is still centred around the location of the current control data [59]. The prior for μ_c is given by,

$$\pi(\mu_c \mid \bar{x}_h, \sigma_h^2, n_h, \bar{x}_c, \sigma_c^2, n_c) = w\mathrm{N}(\bar{x}_h, \sigma_h^2/n_h) + (1-w)\mathrm{N}(\bar{x}_c, \sigma_c^2), \tag{3.2}$$

where $\sigma_h^2 = \sigma_c^2$, because we are assuming known and equal variance in the current and historical control groups. The weight w is pre-specified, chosen based on the prior belief of how similar the historical data are to the current control data. When the historical data mean is not in agreement with the mean in the current control group, a large weight is given to the weakly-informative component of the robust mixture prior. When the historical and current control means are in agreement, a large weight is given to the mixture component based on the historical data. In Equation 3.2, $\bar{x}_c$ is the MLE of the current controls and σ_c^2 is the variance of μ_c representing one observation of the data. Mutsvari et al. [60] discuss in more detail how to choose the variance of the weakly-informative component of the mixture prior distribution and show that choosing a variance that is too large can lead to a large weight being given to the component of the mixture distribution based on the historical data, even when there is conflict between the current control and historical data.

The Bayesian update of the mixture prior to posterior distribution is presented for the general case where the historical control data variance (σ_h^2) may differ from the current control variance (σ_c^2) and both variances are assumed to be known, however, in this section they are assumed to be the same value. The posterior distribution for μ_c is then a mixture of normal distributions with updated weights and parameter values,

$$\begin{aligned}\pi(\mu_c \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c) =&\tilde{w}\mathrm{N}\left(\frac{n_c\tau_c\bar{x}_c + n_h\tau_h\bar{x}_h}{n_c\tau_c + n_h\tau_h}, \frac{1}{n_c\tau_c + n_h\tau_h}\right)\\ &+ (1-\tilde{w})\mathrm{N}\left(\frac{n_c\tau_c\bar{x}_c + \tau_c\bar{x}_c}{n_c\tau_c + \tau_c}, \frac{1}{n_c\tau_c + \tau_c}\right),\end{aligned} \tag{3.3}$$

where, $\tilde{w} = \dfrac{wf_1(x_c \mid \bar{x}_h, \tau_h, n_h, \tau_c, n_c)}{wf_1(x_c \mid \bar{x}_h, \tau_h, n_h, \tau_c, n_c) + (1-w)f_2(x_c \mid \bar{x}_c, \tau_c, n_c)}$,

$$f_1(x_c \mid \bar{x}_h, \tau_h, n_h, \tau_c, n_c) = \int_{-\infty}^{\infty} \mathrm{N}(x_c \mid \mu_c, 1/(n_c\tau_c))\mathrm{N}(\mu_c \mid \bar{x}_h, 1/(n_h\tau_h))d\mu_c,$$

$$f_2(x_c \mid \bar{x}_c, \tau_c, n_c) = \int_{-\infty}^{\infty} \mathrm{N}(x_c \mid \mu_c, 1/(n_c\tau_c))\mathrm{N}(\mu_c \mid \bar{x}_c, 1/\tau_c)d\mu_c,$$

and $\mathrm{N}(x_c \mid \mu_c, 1/(n_c\tau_c))$ is the likelihood of the current control data. $f_1(x_c \mid \bar{x}_h, \tau_h, n_h, \tau_c, n_c)$ and $f_2(x_c \mid \bar{x}_c, \tau_c, n_c)$ are the marginal likelihood of the data, the probability of observing the current trial data given the prior information of each mixture component. $f_1(x_c \mid \bar{x}_h, \tau_h, n_h, \tau_c, n_c)$ and $f_2(x_c \mid \bar{x}_c, \tau_c, n_c)$ can be determined through numerical integration, however for this example the marginal likelihoods are available in closed form and are given by [25],

$$\begin{aligned}f_1 &= \left(\frac{n_h\tau_h}{n_c\tau_c + n_h\tau_h}\right)^{\frac{1}{2}} \exp\left\{\frac{1}{2}\left(n_h\tau_h\bar{x}_h^2 + n_c\tau_c\bar{x}_c^2 - (n_c\tau_c + n_h\tau_h)\left(\frac{n_h\tau_h\bar{x}_h + n_c\tau_c\bar{x}_c}{n_c\tau_c + n_h\tau_h}\right)^2\right)\right\},\\ f_2 &= \left(\frac{\tau_c}{n_c\tau_c + \tau_c}\right)^{\frac{1}{2}} \exp\left\{\frac{1}{2}\left(\tau_c\bar{x}_c^2 + n_c\tau_c\bar{x}_c^2 - (n_c\tau_c + \tau_c)\left(\frac{\tau_c\bar{x}_c + n_c\tau_c\bar{x}_c}{n_c\tau_c + \tau_c}\right)^2\right)\right\}.\end{aligned}$$

Determining the effective sample size of the robust mixture prior

Under the assumption that the population variance is known, the effective sample size can be approximated. Here we illustrate the approximation that is given in Hobbs et al. [20], which is based on the Morita algorithm [26].

This approximation compares the precision of the posterior distribution of the mean in the control group under the model where historical data are incorporated to the model where historical data are not incorporated. The effective historical sample size is given by,

$$EHSS \approx n_c \left\{ \frac{prec(\pi(\mu_c \mid \bar{x}_h, \sigma_h^2, n_h, \bar{x}_c, \sigma_c^2, n_c))}{prec(\pi(\mu_c \mid \bar{x}_c, \sigma_c^2, n_c))} - 1 \right\},$$

where $prec()$ denotes the precision of the distribution.

This effective historical sample size approximation is derived from comparing the information of the posterior mixture distribution for μ_c with the information of a created distribution with known sample size [26]. The derivation for the $EHSS$ is similar to the derivation of the $EHSS$ given in Chapter 2, Equation 2.7. The information of a normal likelihood with sample size n_c and known variance σ_c^2 is given by $\frac{n_c}{\sigma_c^2}$, this is the precision of μ_c under the model where no historical data are incorporated.

Assuming that the population variance in the historical and current controls are known, the variance of the posterior distribution of μ_c, and therefore the precision of μ_c, from the model incorporating historical data, which is a mixture distribution, given in Equation 3.3, can be calculated analytically using,

$$\text{Var}(\mu_c \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c, w) = \mathbf{E}(\mu_c^2 \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c, w) - (\mathbf{E}(\mu_c \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c, w))^2,$$

where,

$$\mathbf{E}(\mu_c^2 \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c, w) = \tilde{w}\left(\frac{1}{n_c\tau_c + n_h\tau_h} + \left(\frac{n_c\tau_c\bar{x}_c + n_h\tau_h\bar{x}_h}{n_c\tau_c + n_h\tau_h}\right)^2\right) + (1-\tilde{w})\left(\frac{1}{n_c\tau_c + \tau_c} + \left(\frac{n_c\tau_c\bar{x}_c + \tau_c\bar{x}_c}{n_c\tau_c + \tau_c}\right)^2\right)$$

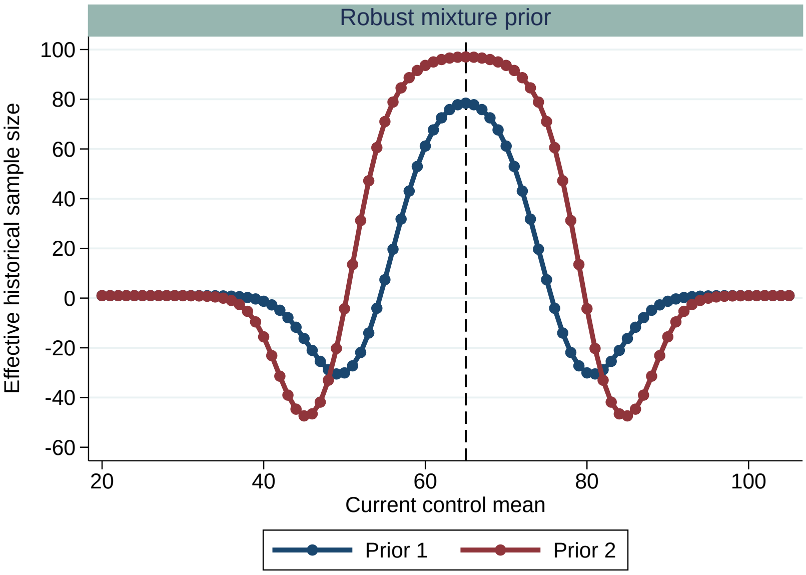
and

$$\mathbf{E}(\mu_c \mid \bar{x}_h, \tau_h, n_h, \bar{x}_c, \tau_c, n_c, w) = \tilde{w}\left(\frac{n_c\tau_c\bar{x}_c + n_h\tau_h\bar{x}_h}{n_c\tau_c + n_h\tau_h}\right) + (1-\tilde{w})\left(\frac{n_c\tau_c\bar{x}_c + \tau_c\bar{x}_c}{n_c\tau_c + \tau_c}\right).$$

For the example considered in this chapter there are 100 (n_h) historical control patients available with a mean of 65 and standard deviation 45. Assuming there are also 100 current controls available and the standard deviation in the current controls is assumed to be known and the same as in the historical data, $\sigma_c = 45$. Figure 3.3 illustrates the $EHSS$ for varying means in the current control data under two different initial priors for μ_c. These priors are,

$$\begin{aligned}\textbf{Prior 1}: \pi(\mu_c) &= 0.5 \times \mathrm{N}(65, 45^2/100) + 0.5 \times \mathrm{N}(\bar{x}_c, 45^2),\\ \textbf{Prior 2}: \pi(\mu_c) &= 0.9 \times \mathrm{N}(65, 45^2/100) + 0.1 \times \mathrm{N}(\bar{x}_c, 45^2).\end{aligned} \tag{3.4}$$

Figure 3.3: Robust mixture posterior $EHSS$ for a range of current control means, assuming a common historical and current control standard deviation $\sigma_h = \sigma_c = 45$, $n_c = 100$, historical data $\bar{x}_h = 65, n_h = 100$. Two different priors for μ_c are explored, given in Equation 3.4. The vertical dashed line represents complete agreement between the historical and current control means.



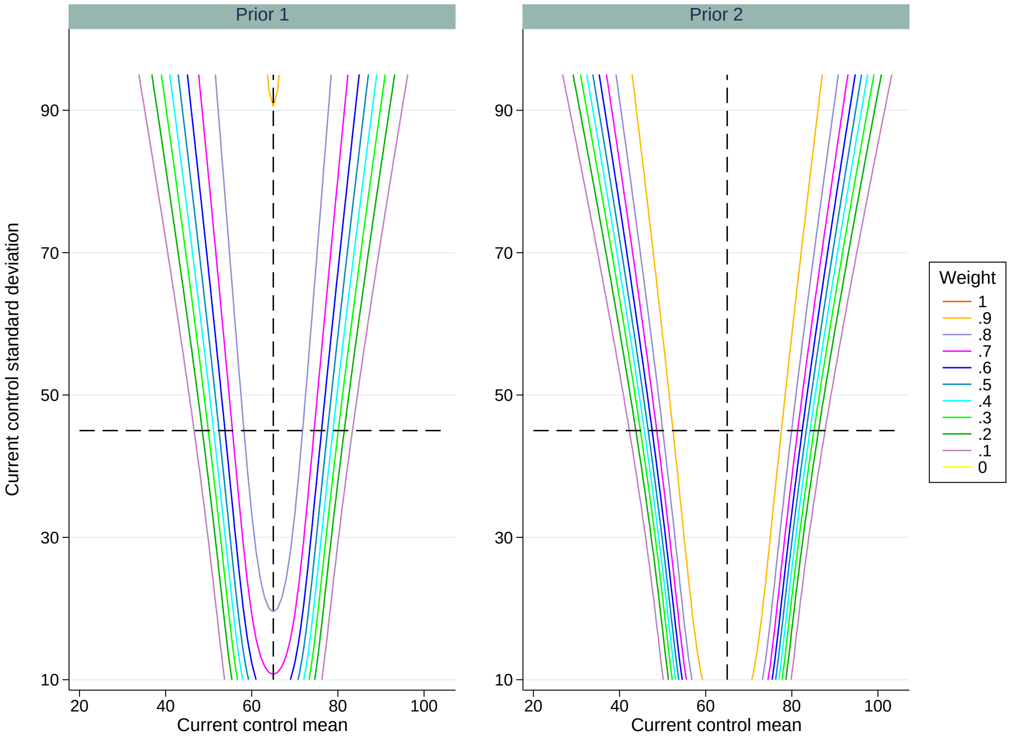
The assumption of a common variance in the historical and current control group is unlikely to be a realistic assumption. The population variance is unlikely to be known and if the sample variance estimates are used in place of the known variances, the sample variance estimates are unlikely to be equal in the current and historical control data.

Figure 3.4 shows the posterior weight ($\tilde{w}$) given to the mixture component based on the historical data when the sample mean and variance in the current controls vary but the known variance formulae are used for the analysis. The sample estimates of the variances are used in place of the known variances. The robust mixture prior distributions are given by,

$$\begin{aligned}\textbf{Prior 1}: \pi(\mu_c) &= 0.5 \times \mathrm{N}(65, 45^2/100) + 0.5 \times \mathrm{N}(\bar{x}_c, \hat{\sigma}_c^2),\\ \textbf{Prior 2}: \pi(\mu_c) &= 0.9 \times \mathrm{N}(65, 45^2/100) + 0.1 \times \mathrm{N}(\bar{x}_c, \hat{\sigma}_c^2).\end{aligned} \tag{3.5}$$

Figure 3.4 shows that a large weight is given to the historical data when the means in the historical and current controls agree, even though the estimated variances differ. This is expected since we do not assume a distribution for the variance parameter. The next section explores whether it is possible to have a robust mixture prior on the joint distribution of the control mean and variance.

Figure 3.4: Contour plots of the robust mixture prior weight on the informative component of the posterior distribution of μ_c for a range of current control sample means and variances when using the current control variance sample estimates in the known variance formulae for the analysis, $n_c = 100$, historical data $\bar{x}_h = 65, \sigma_h = 45, n_h = 100$. The two priors given in Equation 3.5 are explored. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



Assuming unknown and not necessarily equal variances in the current and historical control data

Ideally, the weight given to the historical data component of the robust mixture prior would decrease if either of the estimated means or the variances in the current and historical data differ. To detect differences in both the means and the variances of the current and historical control data, a joint robust mixture prior is required for the mean and variance parameters. We explore whether it is possible to have a similar robust mixture prior on the mean and variance as was used for just the mean parameter in the previous section. Let n_0 denote the belief in the prior estimate of the mean for the weakly-informative component of the robust mixture prior and v_0 denote the belief in the prior estimate of the variance for the weakly-informative component of the robust mixture prior, n_0 and v_0 are pre-specified to give a prior of the form,

$$\begin{aligned}\pi(\mu_c, \sigma_c^2 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2) = {} & w(\mathrm{N}(\bar{x}_h, \sigma_c^2/n_h) SI\chi^2(n_h - 1, \hat{\sigma}_h^2)) \\ & + (1-w)(\mathrm{N}(\bar{x}_c, \sigma_c^2/n_0) SI\chi^2(v_0, \hat{\sigma}_c^2)),\end{aligned} \tag{3.6}$$

where $SI\chi^2(v, \tau^2)$ denotes a scaled inverse chi-squared distribution with v degrees of freedom and scale parameter τ^2. The scaled inverse chi-squared distribution is chosen here as it is a conjugate distribution for the variance. The parameters of the second component of the mixture distribution are chosen to be weakly-informative. The prior estimate of the mean for the weakly-informative component is taken to be the sample mean of the current control data. This is to place the prior in the correct location, but the belief in this value (n_0) will be small to represent the vague component of the mixture prior. The prior estimate of the variance in the weakly-informative component of the robust mixture distribution is chosen to be the sample variance estimate of the current control data and the belief in this value (v_0) will be small to represent the vague component of the mixture prior, this is a data dependent prior. The aim of this prior is to discount both the mean and variance of the historical data when there is a difference in either the mean or the variance between the historical and current control data. When the historical and current control data agree, the historical data component of the mixture distribution should be given a large weight and information should be borrowed for both the mean and variance parameters.

To obtain the joint posterior distribution for the control mean and variance, each component of the mixture distribution is updated as a standard conjugate Bayesian update. Each component of the posterior mixture distribution is then a normal scaled inverse chi-squared distribution with updated parameter values. The weights are updated using the marginal distribution of the data. The posterior distribution for the mean and variance in the control arm is given by,

$$\begin{aligned}\pi(\mu_c, \sigma_c^2 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c) = \tilde{w}(\mathrm{N}(\bar{x}_{hc}, \sigma_c^2/n_{hc}) SI\chi^2(n_{hc}, \hat{\sigma}_{hc}^2)) \\ + (1-\tilde{w})(\mathrm{N}(\bar{x}_{0c}, \sigma_c^2/n_{0c}) SI\chi^2(v_{0c}, \hat{\sigma}_{0c}^2)),\end{aligned} \tag{3.7}$$

where $\bar{x}_{hc} = \dfrac{n_c\bar{x}_c + n_h\bar{x}_h}{n_c + n_h}$, $\bar{x}_{0c} = \dfrac{n_c\bar{x}_c + n_0\bar{x}_c}{n_c + n_0}$, $n_{hc} = n_c + n_h - 1$, $n_{0c} = n_c + n_0$, $v_{0c} = n_c + v_0$, $\hat{\sigma}_{hc}^2 = \frac{1}{n_c+(n_h-1)}\left((n_h-1)\hat{\sigma}_h^2 + n_c\hat{\sigma}_c^2 + \frac{n_c n_h}{n_c+n_h}(\bar{x}_h - \bar{x}_c)^2\right)$, $\hat{\sigma}_{0c}^2 = \frac{1}{n_c+v_0}\left(v_0\hat{\sigma}_c^2 + n_c\hat{\sigma}_c^2 + \frac{n_c n_0}{n_c+n_0}(\bar{x}_0 - \bar{x}_c)^2\right)$,

and, $\tilde{w} = \dfrac{w f_1(x_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_c, n_h)}{w f_1(x_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_c, n_h) + (1-w) f_2(x_c \mid \bar{x}_c, \hat{\sigma}_c^2, n_c, n_0, v_0)}$,

where,

$$f_1(x_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_c, n_h) = \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x_c \mid \mu_c, \sigma_c^2)\mathrm{N}(\mu_c \mid \bar{x}_h, \sigma_c^2/n_h) SI\chi^2(\sigma_c^2 \mid n_h - 1, \hat{\sigma}_h^2) d\mu_c d\sigma_c^2, \tag{3.8}$$

$$f_2(x_c \mid \bar{x}_c, \hat{\sigma}_c^2, n_c, n_0, v_0) = \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x_c \mid \mu_c, \sigma_c^2)\mathrm{N}(\mu_c \mid \bar{x}_c, \sigma_c^2/n_0) SI\chi^2(\sigma_c^2 \mid v_0, \hat{\sigma}_c^2) d\mu_c d\sigma_c^2, \tag{3.9}$$

where $\mathrm{N}(x_c \mid \mu_c, \sigma_c^2)$ denotes the likelihood of the current control data. For normally distributed outcome data and the joint mixture prior for the mean and variance given in Equation 3.6, the marginal likelihoods given in Equations 3.8 and 3.9 have a closed form solution [61],

$$f_1 = \frac{\Gamma(n_{hc}/2)}{\Gamma((n_h-1)/2)}\sqrt{\frac{n_h}{n_{hc}}}\frac{((n_h-1)\hat{\sigma}_h^2)^{(n_h-1)/2}}{(n_{hc}\hat{\sigma}_{hc}^2)^{n_{hc}/2}}\frac{1}{\pi^{n_c/2}} \text{ and } f_2 = \frac{\Gamma(v_{0c}/2)}{\Gamma(v_0/2)}\sqrt{\frac{n_0}{n_{0c}}}\frac{(v_0\hat{\sigma}_c^2)^{v_0/2}}{(v_{0c}\hat{\sigma}_{0c}^2)^{v_{0c}/2}}\frac{1}{\pi^{n_c/2}}.$$

The difficultly in using the robust mixture prior approach where each mixture component is a joint distribution of the mean and variance is what parameters should be chosen for the weakly-informative component of the mixture. The robust mixture prior is used to create a prior that has heavy tails. Heavy-tailed priors have been shown to be discarded when there is increasing conflict between the prior and the current data [24]. The weights of the prior distribution are updated based on how likely the observed data are under each of the mixture components of the prior.

Figure 3.5 illustrates the posterior weight on the informative component of the mixture distribution for four different priors. The first component of the mixture prior is based on the historical data and is therefore fixed. The four priors considered are:

$$\begin{aligned}
&\textbf{Prior 1}: \pi(\mu_c,\sigma_c^2) = 0.5 \times (\mathrm{N}(65, 45^2/100) SI\chi^2(100-1, 45^2)) + 0.5 \times (\mathrm{N}(\bar{x}_c, \sigma_c^2/1) SI\chi^2(1, \hat{\sigma}_c^2));\\
&\textbf{Prior 2}: \pi(\mu_c,\sigma_c^2) = 0.9 \times (\mathrm{N}(65, 45^2/100) SI\chi^2(100-1, 45^2)) + 0.1 \times (\mathrm{N}(\bar{x}_c, \sigma_c^2/1) SI\chi^2(1, \hat{\sigma}_c^2));\\
&\textbf{Prior 3}: \pi(\mu_c,\sigma_c^2) = 0.5 \times (\mathrm{N}(65, 45^2/100) SI\chi^2(100-1, 45^2)) + 0.5 \times (\mathrm{N}(\bar{x}_c, \sigma_c^2/10) SI\chi^2(10, \hat{\sigma}_c^2));\\
&\textbf{Prior 4}: \pi(\mu_c,\sigma_c^2) = 0.9 \times (\mathrm{N}(65, 45^2/100) SI\chi^2(100-1, 45^2)) + 0.1 \times (\mathrm{N}(\bar{x}_c, \sigma_c^2/10) SI\chi^2(10, \hat{\sigma}_c^2)).
\end{aligned} \tag{3.10}$$

Figure 3.5: Contour plots of the joint robust mixture prior weight on the informative component of the joint posterior distribution of μ_c and σ_c^2 for a range of current control sample means and standard deviations, $n_c = 100$, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$. The four priors explored are given in Equation 3.10. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.

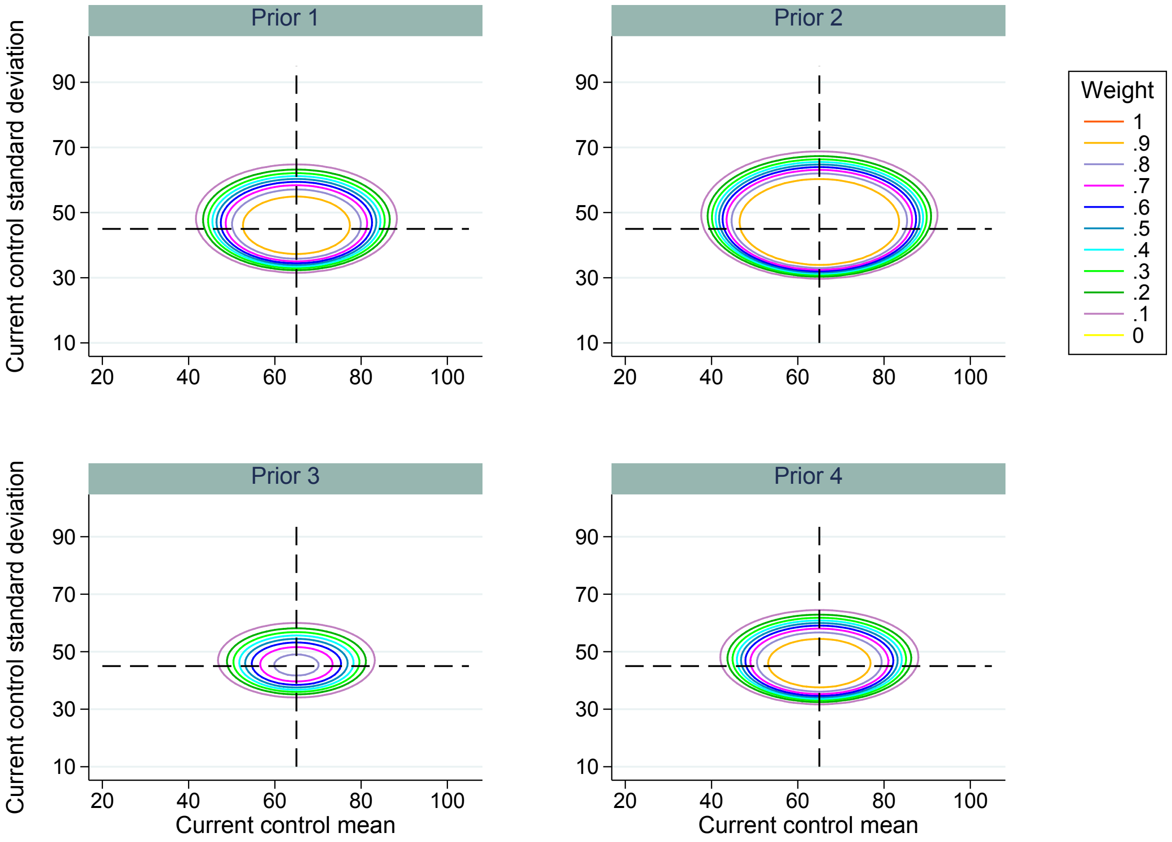


Figure 3.5 shows the weight ($\tilde{w}$) given to the informative component of the posterior mixture distribution for a range of observed current control means and standard deviations

under four different joint priors for μ_c and σ_c^2. Priors 1 and 2 have $n_0 = v_0 = 1$, which gives a flat distribution for the weakly-informative component of the mixture prior and results in a posterior distribution which gives a high weight to the historical data component of the mixture for a wide range of agreement between the current and historical controls, the weight then discounts to zero quickly as the difference between the current and historical controls gets too large. Choosing a larger value for n_0 and v_0, $n_0 = v_0 = 10$ as in priors 3 and 4 gives a lower weight to the historical data component of the mixture posterior distribution at complete agreement in the historical and current controls compared to priors 1 and 2. The prior weight parameter w can be thought of as the belief in the historical data being similar to the current control data. The prior weight $(1-w)$ works in a similar way to n_0 and v_0, all of these parameters effect the posterior mixture distribution weights. A large initial prior weight gives more posterior weight to the historical data component of the mixture prior at agreement in the estimated historical control means and variances and also gives a high weight for a range of current control means and variance around complete agreement with the historical data. As the difference between the historical and current controls gets too large, the posterior weight decreases quickly to zero. An initial lower prior weight for the historical data component of the mixture prior gives a lower weight to the historical data at agreement and discounts more slowly to zero as the difference between the current and historical controls increases compared to a high initial prior weight. The shape of the weight contours are similar to other historical data approaches considered in this chapter, depending on the values chosen for n_0, v_0 and w.

3.2.3 Limitations of published historical data methods

The commensurate prior approach was not considered in this chapter due to the limitations of the approach discussed in Chapter 2. These limitations were that the commensurate prior approach requires MCMC, which is computationally intensive in a trial design setting, especially for adaptive designs, a strong prior is required on the commensurability parameter to induce sufficient borrowing from the historical data and the choice of prior is not intuitive, and the cut function is required to stop feedback from the current data model to the historical data parameter estimates. When the outcome data are normally distributed, the commensurate prior model can be fitted to assess the agreement in the means and the variances of the historical and current control data by having a commensurate prior on both the current mean and variance parameter. Each parameter would be centred at the historical data parameter estimate, with τ_μ, the commensurability parameter measuring the discrepancy between the historical and current data mean and τ_{σ^2} measuring the discrepancy between the historical and current data variance [18].

In Section 3.2.2 we proposed a joint robust mixture prior on the control mean and variance parameters that will discount the historical data when there are differences in either the means or the variances between the current and historical controls. The joint robust mixture prior works as desired but is not considered further in this chapter due to the difficulty in calculating a prior ESS from this joint prior and because the choice of the parameters for the weakly-informative component of the joint mixture prior requires further exploration. As illustrated in Chapter 2 with the robust mixture prior for binary data, it is difficult to determine how much historical data are incorporated into the final analysis using the robust mixture prior approach. Calculating the effective sample size of a joint prior or posterior distribution has not been considered here and this makes the joint robust mixture prior difficult to use in an adaptive design setting. In the specification of the robust mixture prior given in Equation 3.6, the prior parameters for the weakly-informative component of the mixture and the initial weights chosen for the mixture prior all interact to determine the amount of discounting of the historical data. A non-conjugate distribution could be considered for the weakly-informative component of the robust mixture prior, however this would make the analysis computationally intensive and would require MCMC or other similar techniques. Using a non-conjugate distribution for the weakly-informative component of the mixture prior would make an adaptive design implausible due to the computation time.

For the modified power prior approach, the limitations are the same as for binary data (see Section 2.2.4): The choice of prior on the power and the power prior weights obtained for different levels of agreement in the historical and current controls are not intuitive; a fully Bayesian approach, using the whole distribution of the power prior rather than taking a summary measure of the posterior distribution is computationally intensive in the design setting; and finally a strong prior is required on the power parameter to induce sufficient borrowing from the historical data.

3.3 Assessing agreement between historical and current control data – overlap, probability and equivalence probability weight

We now explore three new approaches to assess agreement between historical and current control data which is normally distributed. The three methods explored here allow weights to be between zero and one, where zero represents no historical data borrowing and one represents pooling of the historical and current data. The aim of these approaches is to obtain a high weight when there is agreement in both the means and the variances in the historical and current data and to give a low weight when there is disagreement in the historical and current control means or variances.

3.3.1 Weights based on the amount of overlap in distributions

Overlap in the probability density functions of the historical and current control data

This approach considers using the proportion of overlap in the historical data distribution and the current control data distribution as a weight to down-weight the historical data in the current trial analysis. The overlapping coefficient is a measure of agreement between two probability distributions [62]. Let $\pi_h(x \mid \bar{x}_h, \hat{\sigma}_h^2)$ represent the probability density function (PDF) of the historical data and $\pi_c(x \mid \bar{x}_c, \hat{\sigma}_c^2)$ the PDF of the current control data.

The overlapping coefficient is given by,

$$w = \int_{-\infty}^{\infty} min[\pi_h(x \mid \bar{x}_h, \hat{\sigma}_h^2), \pi_c(x \mid \bar{x}_c, \hat{\sigma}_c^2)]dx,$$

which for normally distributed data is equal to,

$$\int_{-\infty}^{\infty} min\left[\frac{1}{\hat{\sigma}_h\sqrt{2\pi}}\exp\left\{\frac{-(x-\bar{x}_h)^2}{2\hat{\sigma}_h^2}\right\}, \frac{1}{\hat{\sigma}_c\sqrt{2\pi}}\exp\left\{\frac{-(x-\bar{x}_c)^2}{2\hat{\sigma}_c^2}\right\}\right] dx.$$

When the two normal densities have unequal variances, which is likely to be the case in the historical and current sample data. The two densities, $\pi_h(x \mid \bar{x}_h, \hat{\sigma}_h^2)$ and $\pi_c(x \mid \bar{x}_c, \hat{\sigma}_c^2)$ will cross at two points. The value of x at these points can be determined from [62],

$$\frac{\bar{x}_h\hat{\sigma}_c^2 - \bar{x}_c\hat{\sigma}_h^2 \pm \hat{\sigma}_h\hat{\sigma}_c\left[(\bar{x}_h-\bar{x}_c)^2 + (\hat{\sigma}_c^2-\hat{\sigma}_h^2)\log\left(\frac{\hat{\sigma}_c^2}{\hat{\sigma}_h^2}\right)\right]^{\frac{1}{2}}}{\hat{\sigma}_c^2-\hat{\sigma}_h^2}.$$

Letting x_1 denote the smaller of these two points and x_2 the larger, then the overlapping coefficient is given by,

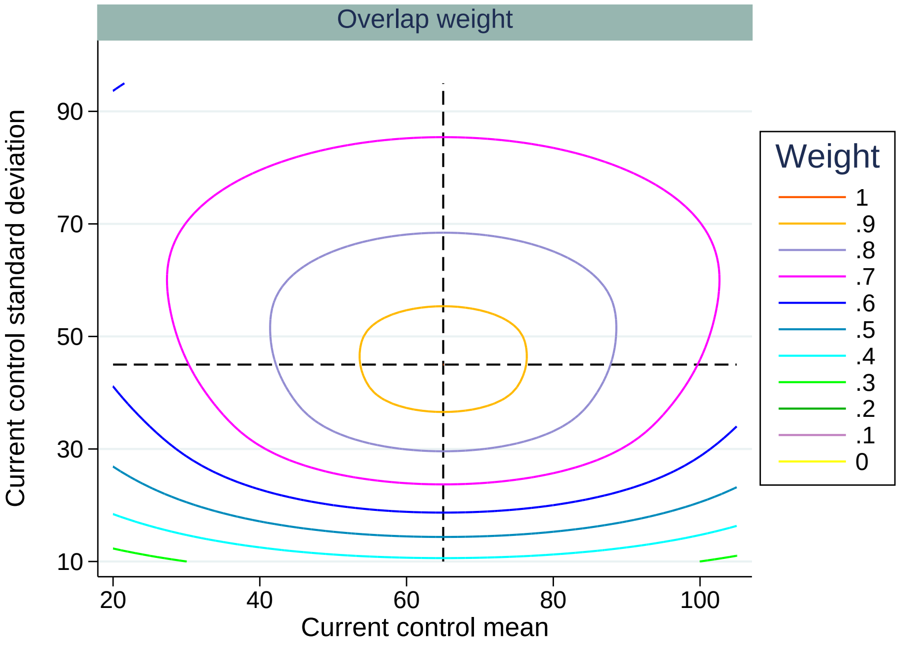
$$w = \Phi\left(\frac{x_1-\bar{x}_h}{\hat{\sigma}_h}\right) + \Phi\left(\frac{x_2-\bar{x}_c}{\hat{\sigma}_c}\right) - \Phi\left(\frac{x_1-\bar{x}_c}{\hat{\sigma}_c}\right) - \Phi\left(\frac{x_2-\bar{x}_h}{\hat{\sigma}_h}\right) + 1,$$

where the observed sample variances in the current and historical data are equal, the normal density functions intersect at a single value, $x = (\bar{x}_h + \bar{x}_c)/2$. The weight is then given by [62],

$$w = \Phi\left(\frac{- \mid \bar{x}_h - \bar{x}_c \mid}{2\hat{\sigma}}\right),$$

where $\hat{\sigma} = \hat{\sigma}_h = \hat{\sigma}_c$.

Figure 3.6: Contour plot of the overlap weights for different observed current control means and standard deviations using the data distributions. Example, $n_c = 100$ and historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$ and $n_h = 100$. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



The overlap weight is one when the observed sample means and variances in the historical and current controls completely agree, as illustrated in Figure 3.6. The weights are symmetric about the sample mean of the historical data but the overlap weight discounts slowly as the difference in the means between the historical and current data increases. At a mean difference of 25, when the standard deviations are the same in the historical and current controls, a weight of 0.78 is given to the historical data. When the current control standard deviation is much larger than the historical standard deviation, there is still overlap in the distributions and quite a large weight is given to the historical data. The weight decreases to around 0.3 for a large difference in the means between the historical and current controls of 35 when the current control standard deviation is small, at around 10.

Overlap in the marginal distributions of the historical and current control parameters

The overlap weight could also be calculated using the posterior distributions of the mean and variance parameters in the current and historical controls. Assuming reference priors, $\pi(\mu_h, \sigma_h^2) \propto 1/\sigma_h^2$, $\pi(\mu_c, \sigma_c^2) \propto 1/\sigma_c^2$. The posterior marginal distributions of the means and variances in the historical and current controls are given by [61],

$$\mu_h \sim t_{n_h-1}\left(\bar{x}_h, \frac{\hat{\sigma}_h^2}{n_h}\right), \quad \sigma_h^2 \sim SI\chi^2(n_h - 1, \hat{\sigma}_h^2),$$
$$\mu_c \sim t_{n_c-1}\left(\bar{x}_c, \frac{\hat{\sigma}_c^2}{n_c}\right), \quad \sigma_c^2 \sim SI\chi^2(n_c - 1, \hat{\sigma}_c^2),$$

where the sample variances are the unbiased estimates of the variance given by, $\hat{\sigma}_h^2 = \frac{1}{n_h-1}\sum\limits_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2$, $\hat{\sigma}_c^2 = \frac{1}{n_c-1}\sum\limits_{i=1}^{n_c}(x_{ci} - \bar{x}_c)^2$, to be consistent with the estimators used by Murphy and Grieve [61, 63].

The weights obtained from the overlap in the posterior distributions of the means and the overlap in the posterior distributions of the variances are multiplied to obtain an overall weight. The overall weight is used to down-weight the historical data in the current trial analysis.

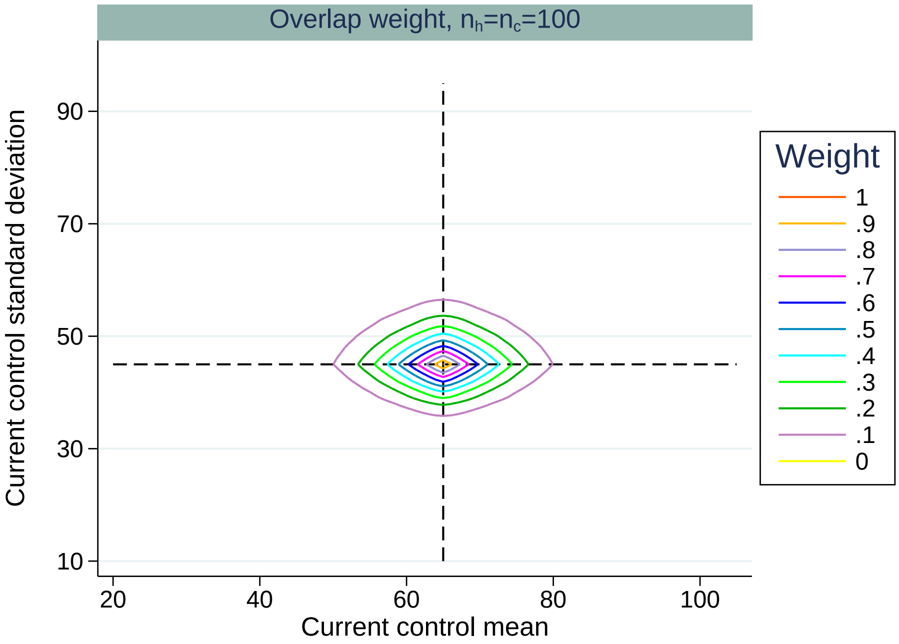
The overlapping coefficient is then given by,

$$w = \int_{-\infty}^{\infty} min\left[t_{n_h-1}\left(x \,\middle|\, \bar{x}_h, \frac{\hat{\sigma}_h^2}{n_h}\right), t_{n_c-1}\left(x \,\middle|\, \bar{x}_c, \frac{\hat{\sigma}_c^2}{n_c}\right)\right] dx$$
$$\times \int_{0}^{\infty} min[SI\chi^2(s \mid n_h - 1, \hat{\sigma}_h^2), SI\chi^2(s \mid n_c - 1, \hat{\sigma}_c^2)]ds.$$

When using the sampling distributions of the mean and variance to calculate the overlap weight, if the sample sizes in the historical and current control data differ, a weight of one is not obtained, even if the sample means and variances in the historical and current control data are the same, as shown in Figure 3.8. When $\mu_c = 65$, $\sigma_c = 45$ and $n_c = 500$ a weight of 0.396 is obtained. Figure 3.7 illustrates the overlap weight calculated using the posterior marginal distributions of the means and variances when the sample sizes in the historical and current controls agree and Figure 3.8 gives the overlap weight when the sample sizes in the current and historical data differ.

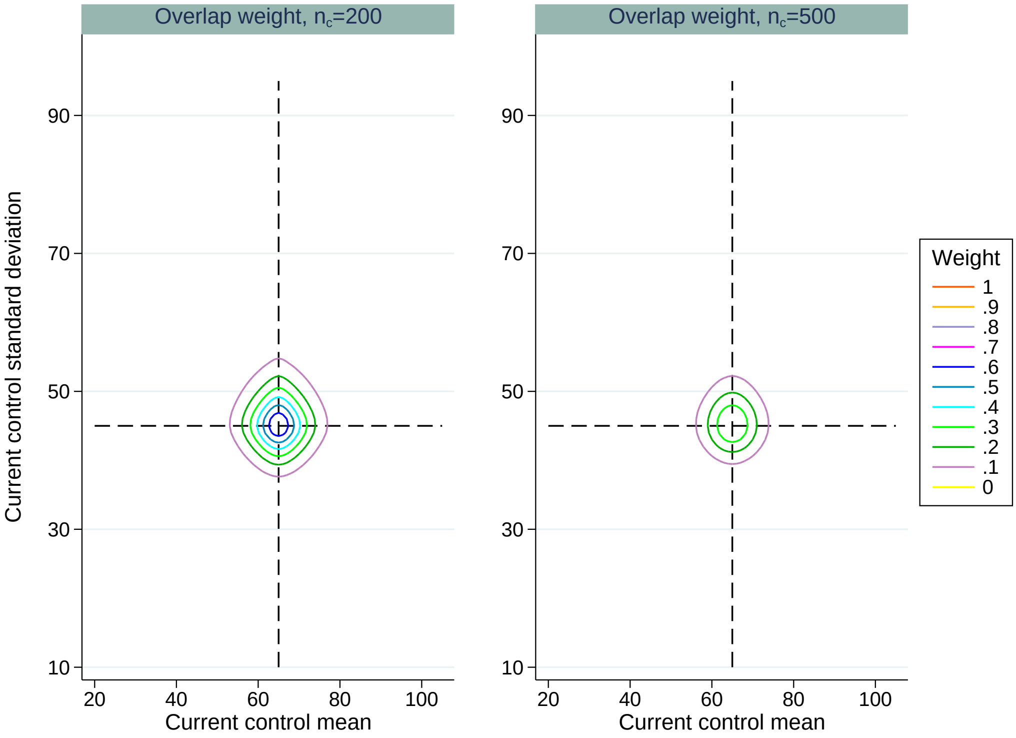
The overlap weight was not considered further in this chapter due to the slow discount-

Figure 3.7: Contour plots of the overlap weights for different observed current control means and standard deviations using the posterior marginal distributions of the historical and current control means and variances. Example, $n_c = 100$ and historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$ and $n_h = 100$. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



ing of the weight when there are differences between the current and historical control data means and variances using the PDF of the data to calculate the overlap weight, and because of the dependency of the overlap weight on the historical and current control sample sizes when using the marginal distributions of the mean and variance to calculate the overlap weight.

Figure 3.8: Contour plots of the overlap weights for different observed current control means and standard deviations using the posterior marginal distributions of the parameters for different current control sample sizes. Example, historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$ and $n_h = 100$. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



3.3.2 Probability weight

Probability weight using the probability density functions of the historical and current data

Assuming normal distributions for both the historical and current control data. The probability weight uses the probability that the PDF of the current control data is greater than the PDF of the historical data to calculate the weight. Let,

$X_h \sim \mathrm{N}(\bar{x}_h, \hat{\sigma}_h^2)$ and $X_c \sim \mathrm{N}(\bar{x}_c, \hat{\sigma}_c^2)$, then,

$$w = 2 \times min\{\Pr(X_c > X_h), 1 - \Pr(X_c > X_h)\},$$

where,

$$X_c - X_h = Z \sim \mathrm{N}(\bar{x}_c - \bar{x}_h, \hat{\sigma}_c^2 + \hat{\sigma}_h^2).$$

Then,

$$\begin{aligned}\Pr(X_c > X_h) &= \int_0^\infty \frac{1}{\sqrt{(\hat{\sigma}_c^2 + \hat{\sigma}_h^2)}\sqrt{2\pi}} \exp\left\{\frac{-\left(z - (\bar{x}_c - \bar{x}_h)\right)^2}{2(\hat{\sigma}_c^2 + \hat{\sigma}_h^2)}\right\} dz \\ &= \Phi\left(\frac{\bar{x}_c - \bar{x}_h}{\sqrt{\hat{\sigma}_c^2 + \hat{\sigma}_h^2}}\right),\end{aligned}$$

where $\hat{\sigma}_g^2 = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(x_{gi} - \bar{x}_g)^2$, $\bar{x}_g = \dfrac{1}{n_g}\sum_{i=1}^{n_g} x_{gi}$ for g = h,c.

Figure 3.9: Contour plot of the probability weights when applied to the data distributions for different observed current control means and standard deviations. Example, $n_c = 100$ and historical data $\bar{x}_h = 65$, $\hat{\sigma}_h = 45$ and $n_h = 100$. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.

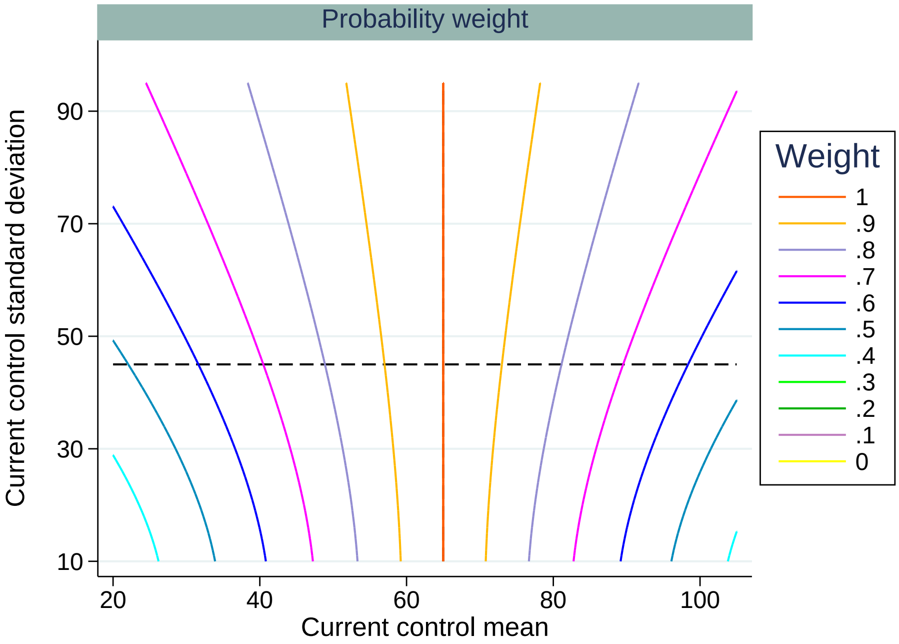


Figure 3.9 illustrates that the probability weight based on the data PDF will give a weight of one when the means in the historical and current control data are the same, even if the variances differ. The contour line for a weight of one lies on top of the dashed line of complete agreement between the current and historical control means in Figure 3.9. Since this probability weight is based on the data distribution rather than the sampling distributions of the mean and variance parameters, the weight discounts slowly as the difference between the historical and current controls increases. The probability weight based on the data PDF is not considered further in this chapter.

Probability weight using the posterior distribution of the difference in means and the posterior distribution of the ratio of the variances between the current and historical controls

It is assumed that the historical data, x_{hi} $(i = 1, \ldots, n_h)$ and current control data x_{ci} $(i = 1, \ldots, n_c)$ are independent samples from $\mathrm{N}(\mu_h, \sigma_h^2)$ and $\mathrm{N}(\mu_c, \sigma_c^2)$ respectively. A scaled inverse chi-squared distribution is chosen as the prior for σ_h^2 and σ_c^2, this is a conjugate prior for the variance and is a more general parameterisation of the inverse gamma prior distribution which is often used as a conjugate prior for the variance. Normal conjugate priors are assumed for μ_h and μ_c. The joint prior is then given by,

$$\begin{aligned}\pi(\mu_h, \mu_c, \sigma_h^2, \sigma_c^2) =& \mathrm{N}(\mu_h \mid \mu_{h0}, \sigma_h^2/n_{h0})\mathrm{N}(\mu_c \mid \mu_{c0}, \sigma_c^2/n_{c0}) \\ & \times SI\chi^2(\sigma_h^2 \mid v_{h0}, \sigma_{h0}^2) SI\chi^2(\sigma_c^2 \mid v_{c0}, \sigma_{c0}^2),\end{aligned}$$

where μ_{g0}, σ_{g0}^2, n_{g0} and v_{g0} denote the prior values for the mean, variance, belief in the prior mean and belief in the prior variance, respectively, for group g, where $g = h, c$. Here, the priors are chosen to be minimally-informative, also known as the reference prior [17]. The parameter values are chosen to be $n_{h0} = n_{c0} = 0$, $v_{h0} = v_{c0} = -1$ and $\sigma_{h0} = \sigma_{c0} = 0$. To give,

$$\begin{aligned}\pi(\mu_h, \mu_c, \sigma_h^2, \sigma_c^2) \propto & \, \sigma_h^{-1}(\sigma_h^2)^{-(v_{h0}/2+1)} \exp\left(-\frac{1}{2\sigma_h^2}[v_{h0}\sigma_{h0}^2 + n_{h0}(\mu_{h0} - \mu_h)^2]\right) \\ & \times \sigma_c^{-1}(\sigma_c^2)^{-(v_{c0}/2+1)} \exp\left(-\frac{1}{2\sigma_c^2}[v_{c0}\sigma_{c0}^2 + n_{c0}(\mu_{c0} - \mu_c)^2]\right) \\ & \propto \sigma_h^{-2}\sigma_c^{-2}.\end{aligned} \tag{3.11}$$

Under the reference prior, given in Equation 3.11. The joint posterior distribution, given the historical and current control data is,

$$\begin{aligned}\pi(\mu_h, \mu_c, \sigma_h^2, \sigma_c^2 \mid \bar{x}_h, \bar{x}_c, \hat{\sigma}_h^2, \hat{\sigma}_c^2, n_h, n_c) = & \, N(\bar{x}_h, \sigma_h^2/n_h) N(\bar{x}_c, \sigma_c^2/n_c) \\ & \times SI\chi^2(n_h - 1, \hat{\sigma}_h^2) SI\chi^2(n_c - 1, \hat{\sigma}_c^2),\end{aligned} \tag{3.12}$$

where $\hat{\sigma}_g^2 = \frac{1}{n_g - 1}\sum\limits_{i=1}^{n_g}(x_{gi} - \bar{x}_g)^2$ is the unbiased sample variance for $g = h, c$.

From Grieve [63], the joint posterior distribution in Equation 3.12 can be re-written in terms of the parameters of interest, the difference in means and the ratio of the variances. Let $\Delta = \mu_c - \mu_h$ and $\phi = \sigma_c^2/\sigma_h^2$, then [63],

$$\begin{aligned}\pi(\Delta, \phi \mid \bar{x}_h, \bar{x}_c, \hat{\sigma}_h^2, \hat{\sigma}_c^2, n_h, n_c) = D &\left\{ \left(\frac{1}{n_h} + \frac{\phi}{n_c} \right) \left(v_h \hat{\sigma}_h^2 + \frac{v_c \hat{\sigma}_c^2}{\phi} \right) \right\}^{-1/2} \\ &\times \left\{ 1 + \frac{(\Delta - \bar{x}_c + \bar{x}_h)^2}{(n_h^{-1} + \phi n_c^{-1})(v_h \hat{\sigma}_h^2 + v_c \hat{\sigma}_c^2 \phi^{-1})} \right\}^{-(n_h+n_c-1)/2} \\ &\times E \phi^{-(n_c+1)/2} \left\{ 1 + \frac{v_c \hat{\sigma}_c^2}{v_h \hat{\sigma}_h^2 \phi} \right\}^{-(n_h+n_c-2)/2}, \end{aligned} \tag{3.13}$$

where, $v_h = n_h - 1$, $v_c = n_c - 1$, $D = \mathrm{B}^{-1}\left(\frac{1}{2}, \frac{v_h+v_c}{2}\right)$, $E = \mathrm{B}^{-1}\left(\frac{v_h}{2}, \frac{v_c}{2}\right)\left(\frac{v_c \hat{\sigma}_c^2}{v_h \hat{\sigma}_h^2}\right)^{v_c/2}$ and $\mathrm{B}^{-1}(a, b)$ denotes the reciprocal of the beta function with parameters a and b.

Equation 3.13 is given in [63] and derived by rewriting Equation 3.12 in terms of Δ, ϕ, $\theta = \mu_h$, and $\psi = \sigma_h^2$ with Jacobian ψ, and integrating out θ and ψ respectively.

The properties of this joint distribution that we will use for making inference are [63],

$$\delta \sim \text{Behrens-Fisher} \quad \text{and} \quad \phi \frac{\hat{\sigma}_h^2}{\hat{\sigma}_c^2} \sim F_{v_h, v_c},$$

where F_{v_h,v_c} denotes the F-distribution with v_h and v_c degrees of freedom. We assess the agreement between the current and historical controls means and variances separately using the marginal distributions of the difference in means and the ratio of the variances. A probability weight is obtained for the difference in means between the historical and current controls and a probability weight is obtained for the difference in variances between the historical and current controls. These weights are then combined by simply multiplying them to obtain an overall weight to discount the historical data. Multiplying the weights, gives a weight of one when both of the individual weights are one and a weight of zero when both of the individual weights are zero.

Probability weight for the difference in means

The probability weight for the difference in means between the historical and current controls is calculated as,

$$w_\Delta = 2 \times min\{\Pr(\mu_c > \mu_h), 1 - \Pr(\mu_c > \mu_h)\}.$$

The marginal distribution of the difference in mean parameters between the historical and current controls $\Delta = \mu_c - \mu_h$ follows a Behrens-Fisher distribution which has been approximated by Welsch [64].

It is assumed $\Delta = \mu_c - \mu_h$ has the approximate distribution [64],

$$\pi(\Delta \mid Data) \approx t\left(\bar{x}_c - \bar{x}_h, \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_h^2}{n_h}, \frac{\left(\frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_h^2}{n_h}\right)^2}{\frac{(\hat{\sigma}_c^2/n_c)^2}{n_c - 1} + \frac{(\hat{\sigma}_h^2/n_h)^2}{n_h - 1}}\right). \tag{3.14}$$

We calculate the probability that this distribution is greater than zero. To simplify the calculation, the t-distribution given in Equation 3.14 can be standardised to,

$$t = \frac{\Delta - (\bar{x}_c - \bar{x}_h)}{\left(\frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_h^2}{n_h}\right)^{\frac{1}{2}}} \sim t_v,$$

where t_v denotes the standard t-distribution with v degrees of freedom, given in Equation 3.14. Then the probability weight for the mean difference is given by,

$$w_\Delta = 2 \times min(\Pr(t_v \leq t), 1 - \Pr(t_v \leq t)).$$

The probability weight for the mean is doubled because of symmetry, as we are interested in a difference between the historical and current control means in either direction.

Probability weight for the difference in variances

To calculate the probability weight based on the ratio of the variances in the historical and current controls, the marginal posterior distribution of the ratio of variances is used. The posterior probability that $1/\phi$ is less than one is given by,

$$\Pr(1/\phi \leq 1) = \Pr\left(f_{(n_c-1, n_h-1)} \leq \frac{\hat{\sigma}_c^2}{\hat{\sigma}_h^2}\right),$$

and the probability weight for the agreement in variances is given by,

$$w_\phi = \begin{cases} \Pr\left(F_{(n_c-1,n_h-1)} \leq \frac{\hat{\sigma}_c^2}{\hat{\sigma}_h^2}\right) + \Pr\left(F_{(n_c-1,n_h-1)} \geq \frac{\hat{\sigma}_h^2}{\hat{\sigma}_c^2}\right) & \text{if} \quad \hat{\sigma}_h^2 > \hat{\sigma}_c^2 \\ \Pr\left(F_{(n_c-1,n_h-1)} \leq \frac{\hat{\sigma}_h^2}{\hat{\sigma}_c^2}\right) + \Pr\left(F_{(n_c-1,n_h-1)} \geq \frac{\hat{\sigma}_c^2}{\hat{\sigma}_h^2}\right) & \text{if} \quad \hat{\sigma}_c^2 > \hat{\sigma}_h^2. \end{cases}$$

Again, a difference in the variances in the historical and current control in either direction is of interest, we therefore calculate the upper and lower tail area probabilities to determine the probability weight for the difference in the variances.

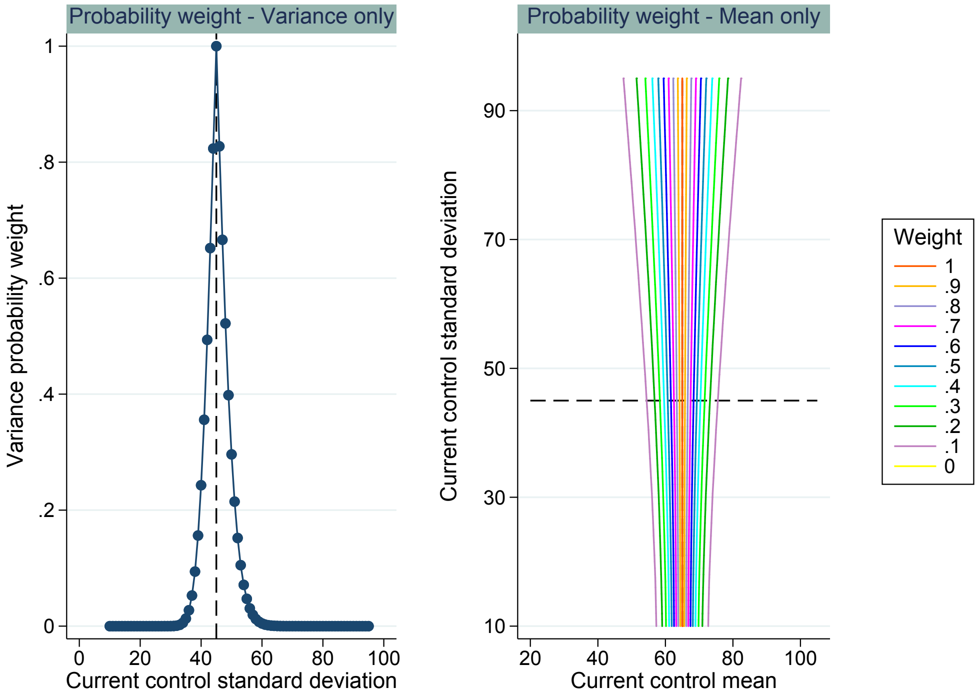
The probability weights for the mean and the variance are then combined. If the weights are averaged, when either of the historical and current control means or variances completely agree and the other is in conflict a weight of a half will be obtained. When the difference in either the means or the variances gets too large, we want to completely discount the historical data. Taking the minimum of the probability weight based on the difference in means and the probability weight from the ratio of the variances quickly discounts the historical data as any differences between the current and historical data are observed. A weighted combination of the two weights could also be used where there is an a priori opinion about the relative importance of the variance and mean conflicts. This parameter could be elicited from experts. Alternatively, the variance and mean probabilities weights can be multiplied. For some trials, agreement in variances may not be as important as agreement in means and a weighted combination of the mean and variance probability weight may be used.

Illustrative example for the probability weight based on the posterior distribution of the historical and current parameters

Figure 3.10 shows the probability weights comparing the historical and current control means and variances separately. The variance weight is independent of the observed sample means. The maximum weight of one is obtained when the historical and current control sample variances are in complete agreement ($\hat{\sigma}_h^2 = \hat{\sigma}_c^2 = 45^2$) and discounts quickly as the difference in variances increases. The probability weight for the difference in means is calculated using the Welsch approximation to the Behrens-Fisher distribution, given in Equation 3.14. This approximation is dependent on the observed historical and current control sample variances. However, this dependence has little effect when calculating the probability weight. As with the probability weight for the variance, a weight of one is obtained when the sample means in the historical and current controls are in complete agreement ($\bar{x}_h = \bar{x}_c = 65$) and this weight discounts quickly as the difference in means increases.

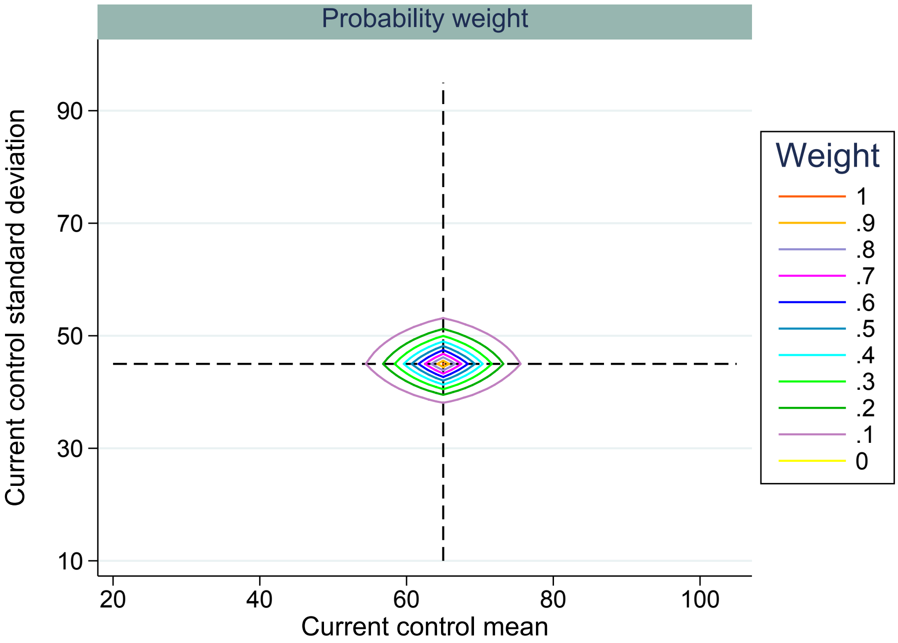
Figure 3.11 shows the overall historical data weight obtained, taking the product of the weights for the mean (w_Δ) and the variance (w_ϕ). The probability weight approach, multiplying the variance and mean probability weights, always gives a weight of one to the historical data at complete agreement between the historical and current controls. This is because both of the individual probabilities give a weight of one at complete agreement. However, the overall probability weight discounts the historical data quickly as either differences in the means or the variances between the historical and current

Figure 3.10: Probability weight from the posterior distribution of the difference in means and the ratio of the variances in the current and historical controls (mean and variance separately) for different observed current control means and standard deviations. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed line represents complete agreement between the current and historical control standard deviations and the vertical dashed line represents complete agreement between the historical and current control means.



controls are observed. When the current control sample size is 200, the historical data are discounted at a quicker rate as the difference between the current and historical control means or variances increases, compared to when the current control sample size is 100.

Figure 3.11: Probability weight (multiplying the mean and variance probability weights) for different observed current control means and standard deviations. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed line represents complete agreement between the current and historical control standard deviations and the vertical dashed line represents complete agreement between the historical and current control means.



3.3.3 Equivalence probability weight

To compare the historical and current control data, both the equivalence of the means and the variances are considered. Testing both equivalence of location and dispersion simultaneously was considered by Bauer and Bauer [65]. Grieve considers a Bayesian approach based on the posterior probability that the parameters lie within a pre-specified region of equivalence [63]. Here, we use the marginal distributions of Δ and ϕ given by [63],

$$\delta \sim \text{Behrens-Fisher} \quad \text{and} \quad \phi\frac{\hat{\sigma}_h^2}{\hat{\sigma}_c^2} \sim F_{v_h,v_c},$$

to calculate an equivalence probability weight for the difference in means and the ratio of the variances separately. We calculate, $\Pr(\delta_{\phi l} < \phi < \delta_{\phi u})$ using the CDF of the F-distribution, where $\delta_{\phi l}$ and $\delta_{\phi u}$ denote the lower and upper equivalence bounds for the ratio of the variances, and we calculate $\Pr(\delta_{\Delta l} < \Delta < \delta_{\Delta u})$ using an approximation for the cumulative probabilities of the Behrens-Fisher distribution, the Welsch approximation is used here, where $\delta_{\Delta l}$ and $\delta_{\Delta u}$ denote the lower and upper equivalence bounds for the difference in the means in the current and historical control data. As in previous sections, we then obtain a joint equivalence probability weight by multiplying the equivalence probability weight for the mean difference by the equivalence probability weight for the difference in variances. This is similar to the approach used to calculate the probability weight in Section 3.3.2.

Here, we initially explore the equivalence probabilities for the difference in means and ratio of the variances separately. It is assumed that the mean and variance equivalence bounds are chosen independently based on reasonable deviations in the difference in means or the ratio of variances between the historical and current controls. Figure 3.12 illustrates the equivalence weight obtained for the ratio of variances component for different observed current control standard deviations and different equivalence bounds for the ratio of the variances.

Figure 3.12: Variance equivalence weights for different current control standard deviations. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. Equivalence bounds on the ratio of variances of (0.5,1/0.5), (0.6,1/0.6), (0.7,1/0.7) and (0.8,1/0.8). The vertical dashed line represents complete agreement between the current and historical control standard deviations.

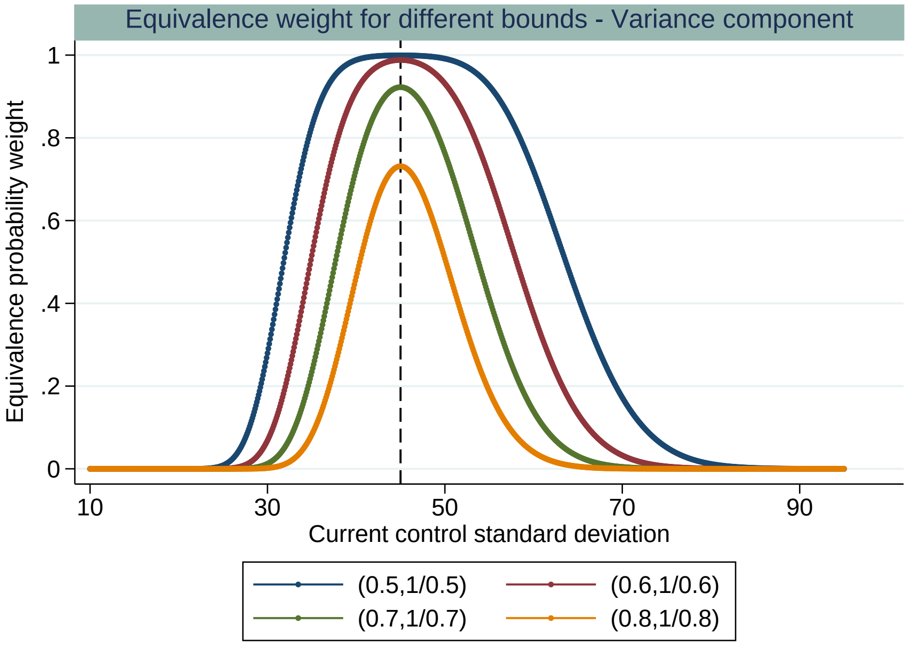


Figure 3.12 shows that for this example the maximum weight is obtained when there is complete agreement in the observed current and historical standard deviations. Wider equivalence bounds give a higher weight for a wider range of current control standard deviations around agreement between the historical and current controls. The F-distribution used to calculate the weight is independent of the sample means. The maximum weight is attained at complete agreement in the historical and current sample variances since in this example the sample sizes in the historical data and the current control group are the

same. When the sample sizes differ, the ratio of the variances follows an F-distribution with unequal degrees of freedom. If bounds of the form $(\delta_{\phi l}, 1/\delta_{\phi l})$ are then used for the equivalence bounds, the maximum weight will not occur at complete agreement in the sample variances. Figure 3.13 shows the equivalence weight for the variance component when there are 100 historical control patients and 500 patients in the current trial control group. The maximum weight occurs when the current control sample standard deviation is slightly above the historical standard deviation of 45. Where the maximum occurs varies depending on the equivalence bounds chosen. However, the example used in Figure 3.13 illustrated the equivalence weights for the variance component for a large difference in sample size between the current and historical controls and there is little difference in the maximum equivalence weight across all current control standard deviations and the equivalence weight obtained at complete agreement in the historical and current control sample standard deviations.

Figure 3.13: Variance equivalence weights for different current control standard deviations. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 500 current controls. Equivalence bounds on the ratio of variances of (0.5,1/0.5), (0.6,1/0.6), (0.7,1/0.7) and (0.8,1/0.8). The vertical dashed line represents complete agreement between the current and historical control standard deviations.

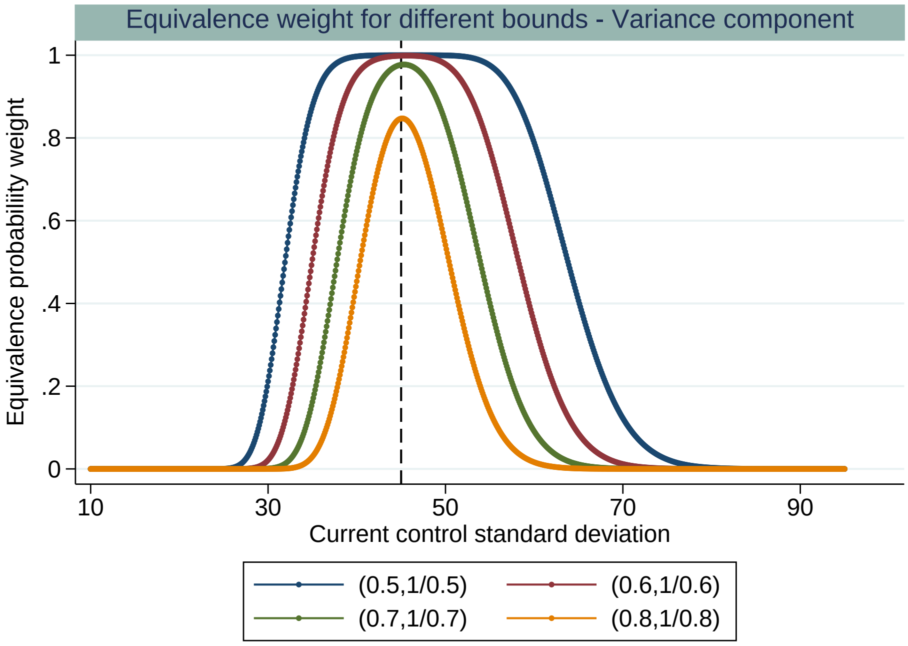
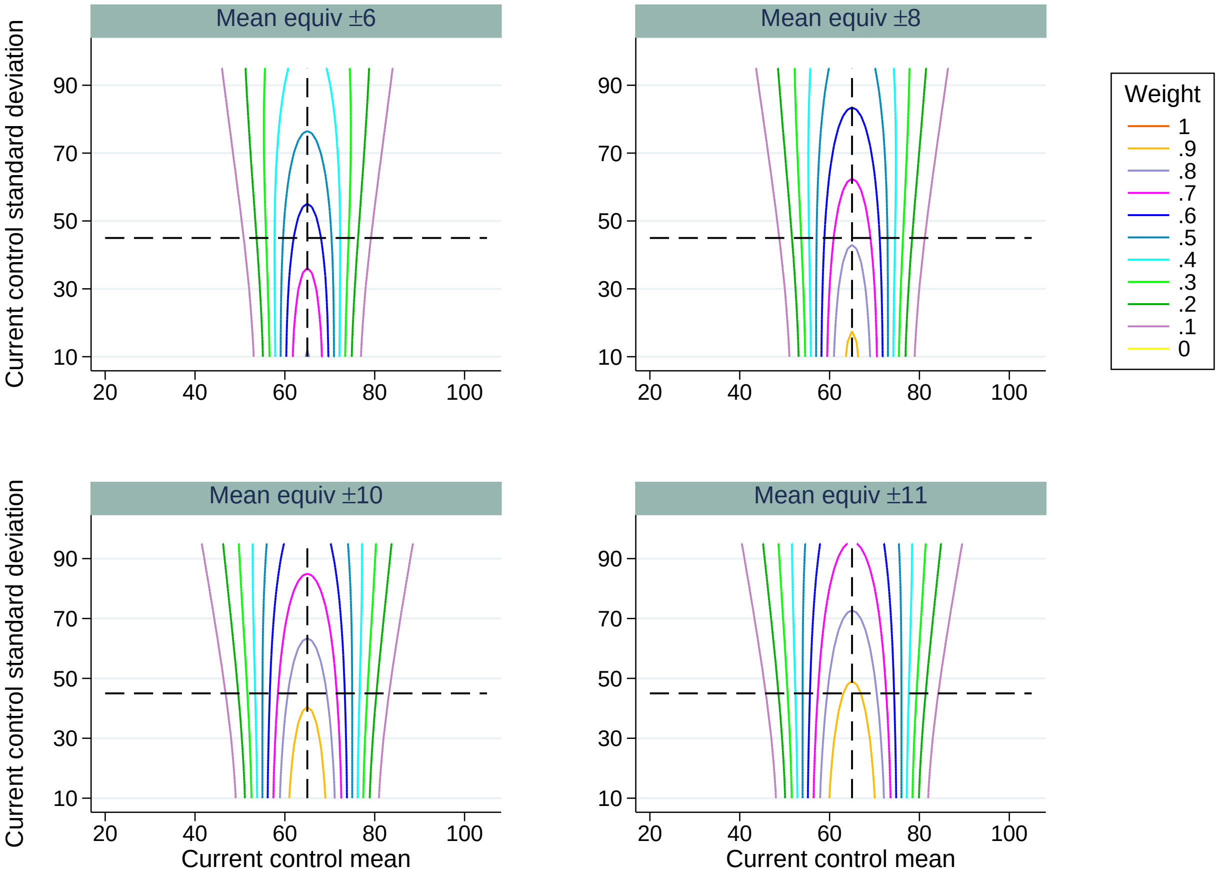


Figure 3.14 shows the contour plots for the equivalence weight from the difference in means component. The Welsch approximation to the Behrens-Fisher distribution is used. The Welsch approximation contains the sample variances of the historical and current control data, as shown in Equation 3.14. The variance of the difference in means from the Welsch approximation is a weighted average of the historical and current control sample

variances. From Figure 3.14 we can see that when the estimated current control standard deviation is smaller than the estimated historical standard deviation and the estimated current control and historical sample means are in complete agreement, a higher weight is given to the historical data compared to when the estimated current control standard deviation is larger than the estimated historical standard deviation and the estimated means are still in complete agreement. This is because the weighted average of the sample variances gives a smaller variance around the difference in means and a larger proportion of the distribution lies within the equivalence bounds.

Figure 3.14: Mean equivalence weights for different observed current control means and standard deviations and equivalence bounds on the mean difference of $\pm 6, \pm 8, \pm 10$ and ± 11. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.

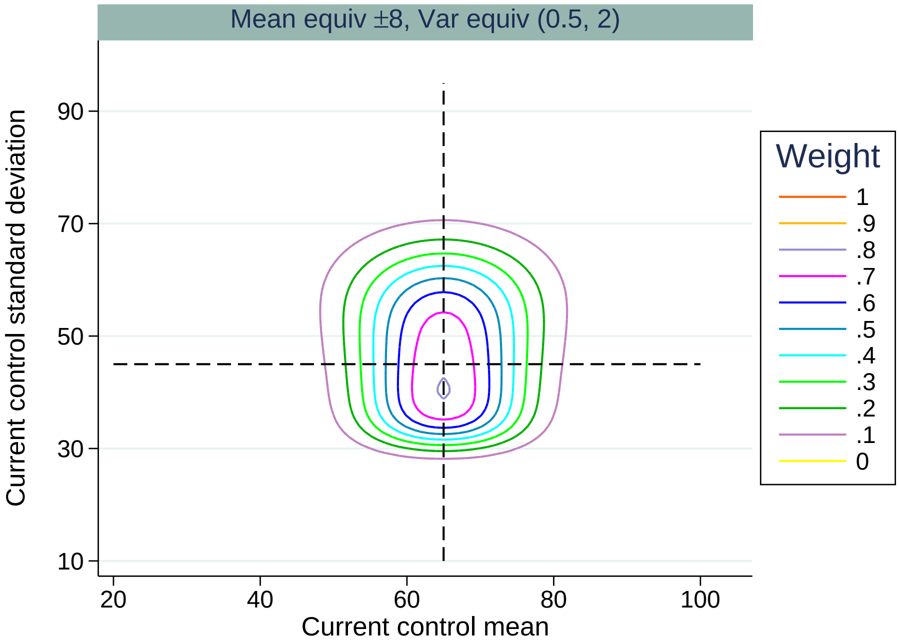


The overall equivalence weight is obtained from multiplying the equivalence weight based on the ratio of the variances and the equivalence weight based on the difference in means. However, the overall equivalence weight does not give maximum weight to the historical data when the sample means and variances in the historical and current control data completely agree, even when the sample sizes in the historical data and the current control group are equal. The overall equivalence weight is not maximised at complete agreement because the equivalence weight component for the difference in means is dependent on the estimated sample variances in the current control and historical data. The maximum weight for the equivalence in means does not occur when the current and historical variances are in complete agreement, even though for the individual mean component weight, the maximum weight is at complete agreement in the means and for the individual variance component weight the maximum weight is given at complete agreement in the variances. The variance component of the joint equivalence weight does not decrease quickly enough to offset the higher weight given for the mean component at complete agreement when the estimated current control standard deviation is smaller than the estimated historical standard deviation. This is illustrated in Figure 3.15 where the equivalence bounds on the mean difference are assumed to be ±8 and the equivalence bounds on the ratio of the variances are assumed to be (0.5,2). The contour of maximum weight lies on complete agreement in the estimated means but where the estimated current control standard deviation is below the observed historical standard deviation.

The value of the observed current control standard deviation that gives the maximum equivalence weight for the mean component of the joint equivalence weight varies. When the current and historical control means are estimated to be similar, the maximum equivalence weight for the difference in means component occurs at the lowest current control standard deviation explored. For large differences in means between the current and historical data, the maximum equivalence weight for the difference in means component occurs at the largest standard deviation explored. This is intuitive since the variance of the posterior distribution for the difference in means between the historical and current data is given by, $\frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_h^2}{n_h}$. When the difference in the historical and current control sample means is large and the current control variance is small, less of the posterior distribution for the difference in the historical and current control means will lie around zero and a lower weight is obtained. For a difference distribution where the current control variance is large, some of the distribution is likely to lie within the equivalence bounds and a larger weight is obtained. This is illustrated in Figure 3.16 for mean equivalence bounds of ±8. Therefore, since the population variances are unknown, the difference in the sample variances between the historical and current control data needs to be incorporated into the mean component of the joint equivalence weight.

Mathematically, obtaining an equivalence weight that is larger at complete agreement in sample means but not in variances compared to when there is complete agreement in

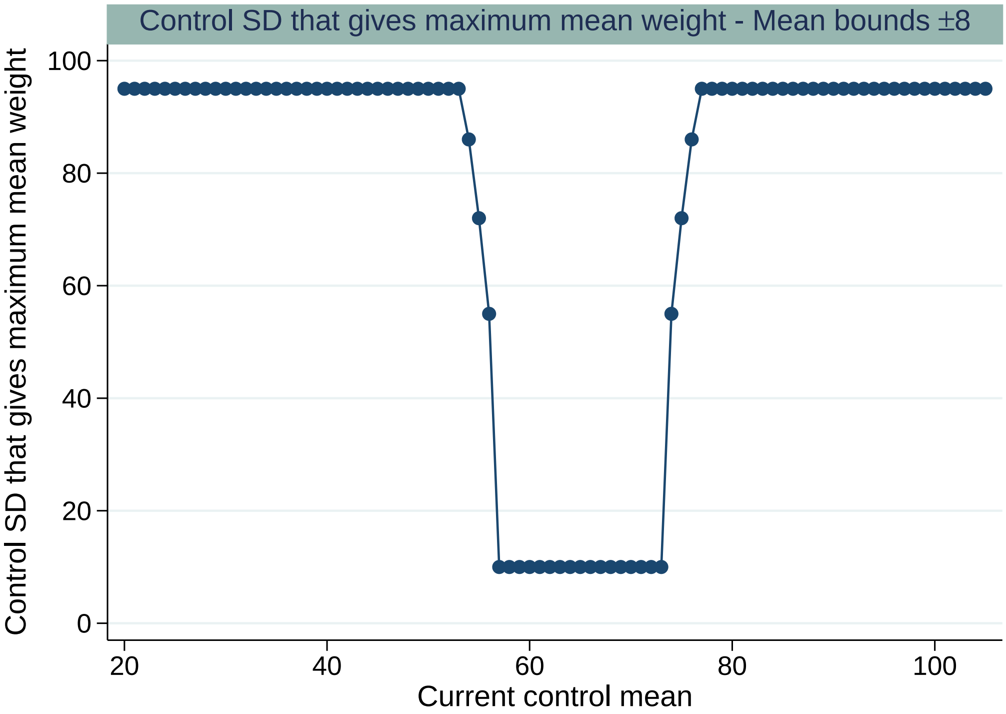
Figure 3.15: Contour plot of the joint equivalence weight with mean equivalence bounds of ±8 and variance equivalence bounds of (0.5,2) for different observed current control means and standard deviations. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed line represents complete agreement between the current and historical control standard deviations and the vertical dashed line represents complete agreement between the historical and current control means.



both means and variances between the historical and current controls is correct. However, ideally we want a weight that is maximised when both the sample mean and variance estimates in the historical and current control completely agree. This problem occurs since the sample standard deviation estimates are used in calculating the mean weight component of the joint equivalence weight. We therefore obtain a different weight for each observed current control mean and standard deviation combination. However, we do not know the true variance ratio between the historical and current controls.

Bauer and Bauer observed a similar problem in the frequentist framework when testing the joint equivalence of mean and variances, in their paper [65] they consider using a corrected t-statistic which replaces the unknown true variance in the t-statistic with the maximum or minimum variance equivalence bound depending on the observed ratio of variances, as a way to get a conservative estimate. Bauer and Bauer link the choice of mean and variance equivalence bounds based on the power of the tests of both the means and variances. Further, Bauer and Bauer discuss an alternative to the corrected t-statistic, originally proposed by Barnard [66] which calculated the p-value of the test statistic by averaging the p-values of the test statistics over the confidence distribution of

Figure 3.16: Current control standard deviation that gives the maximum equivalence probability weight for the mean component for different observed current control means over the range of current control standard deviations explored - Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls.



the true ratio of the variances, given the observed ratio of the sample variances.

A similar approach is used here. The true ratio of variances is unknown. However, in choosing the equivalence bounds for the ratio of the variances, we have specified an acceptable range of difference between the historical and current control variances. The mean equivalence weight component of the joint equivalence weight can then be calculated by averaging the weights obtained for each combination of sample mean and variance over the range of current control standard deviations that lie within the chosen equivalence bounds of the ratio of the variances. This gives a single weight for the mean component of the joint equivalence weight. The variance component of the equivalence weight is calculated using the CDF of the F-distribution as before. The corrected equivalence probability weight for the mean component is given by,

$$\Pr(\delta_{\Delta l} < \Delta < \delta_{\Delta u} \mid \bar{x}_h, \bar{x}_c, \hat{\sigma}^2_h, \hat{\sigma}^2_c, n_h, n_c) = \\ \frac{\displaystyle\int\limits_{\hat{\sigma}^2_h/\delta_{\phi u}}^{\hat{\sigma}^2_h/\delta_{\phi l}} \Pr\left(t_v \leq \frac{\delta_{\Delta u} - \mid \bar{x}_c - \bar{x}_h \mid}{\left(\dfrac{\hat{\sigma}^2_c}{n_c} + \dfrac{\hat{\sigma}^2_h}{n_h}\right)^{\frac{1}{2}}} \right) - \Pr\left(t_v \leq \frac{\delta_{\Delta l} - \mid \bar{x}_c - \bar{x}_h \mid}{\left(\dfrac{\hat{\sigma}^2_c}{n_c} + \dfrac{\hat{\sigma}^2_h}{n_h}\right)^{\frac{1}{2}}} \right) d\hat{\sigma}^2_c}{\hat{\sigma}^2_h/\delta_{\phi l} - \hat{\sigma}^2_h/\delta_{\phi u}}, \tag{3.15}$$

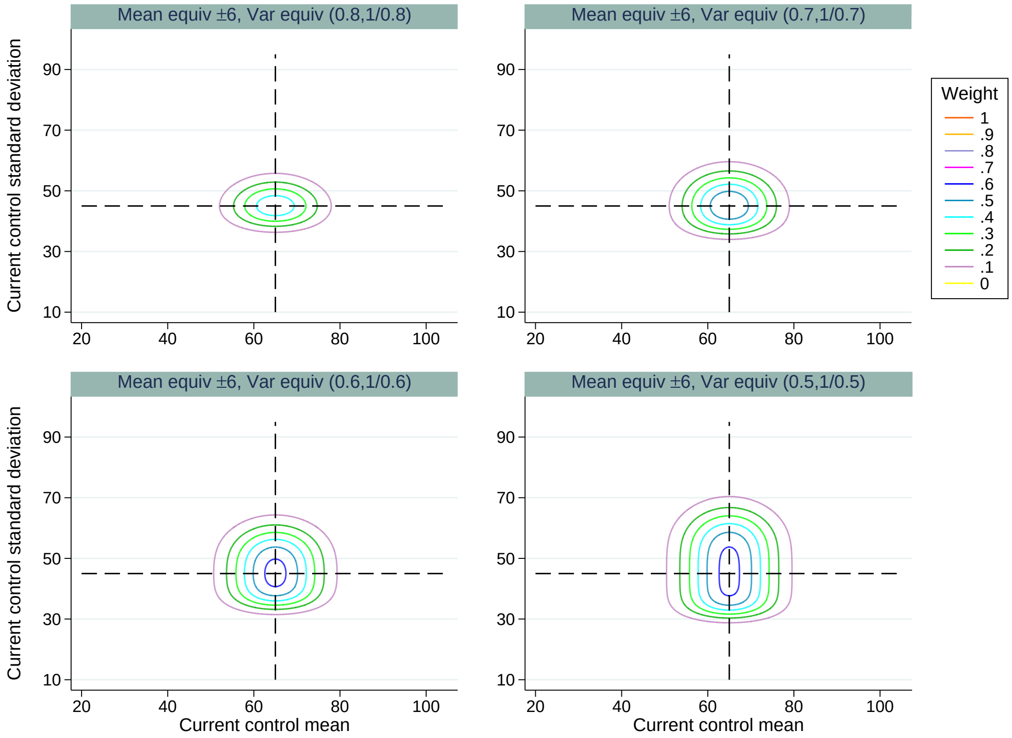
where,

$$v = \frac{\left(\dfrac{\hat{\sigma}^2_c}{n_c} + \dfrac{\hat{\sigma}^2_h}{n_h}\right)^2}{\dfrac{(\hat{\sigma}^2_c/n_c)^2}{n_c - 1} + \dfrac{(\hat{\sigma}^2_h/n_h)^2}{n_h - 1}}.$$

If the population variances in the historical and current controls are known, there would not be a problem of obtaining the maximum weight when the historical and current control estimated sample means and variances do not completely agree. When the population variances are known, if the ratio of the variances lay outside the equivalence range, the variance component of the joint equivalence weight would be zero and within the equivalence range, the joint equivalence weight would be one. Assuming the variances are known, there would only be a single weight for the difference in means component of the joint weight, which would always be maximised at complete agreement in the historical and current control sample means. The joint equivalence weight obtained from multiplying the variance equivalence weight and the mean equivalence weight averaged over the range of observed current control standard deviations that lie within the specified variance equivalence bounds is denoted the corrected joint equivalence probability weight. Figures 3.17, 3.18 and 3.19 illustrate the corrected joint equivalence probability weights obtained for different choices of equivalence bounds on the difference in means and ratio of variances. The mean equivalence bounds shown are ±6, ±8 and ±10 and the variance equivalence bounds are (0.8, 1/0.8), (0.7, 1/0.7), (0.6, 1/0.6) and (0.5,1/0.5).

For the example used here, the sample sizes are the same in the current control and the historical data, therefore the maximum weight is obtained at complete agreement in the sample means and variances. When the equivalence bounds on the difference in means and the ratio of the variances are small (mean equiv ±6, Var equiv (0.8,1/0.8), Figure 3.17) a low equivalence weight of 0.48 is obtained at complete agreement in the sample

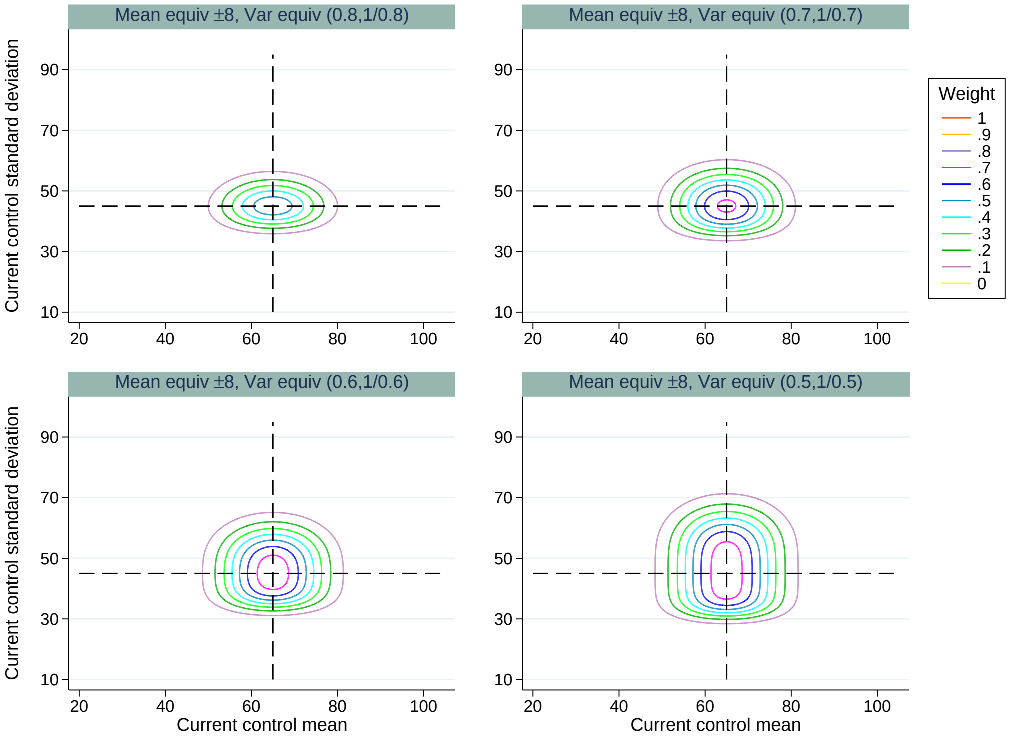
Figure 3.17: Contour plots of the corrected joint equivalence probability weights for different observed current control means and standard deviations. Equivalence bounds on the difference in means of ±6 and varying equivalence bounds on the ratio of the variances. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



means and variances ($\bar{x}_h = \bar{x}_c = 65, \hat{\sigma}_h^2 = \hat{\sigma}_c^2 = 45^2$). Assuming larger equivalence bounds (mean equiv ±10, Var equiv (0.5,2), Figure 3.19), the equivalence weight at complete agreement is 0.86. Larger equivalence bounds on the mean and variance take longer to completely discount the historical data as the difference in the sample means and variances increases. The rate of discounting is also dependent on the historical and current control data sample sizes. For larger sample sizes the weight will decrease more quickly as the differences between the historical and current control data increases. This is because the posterior distribution for the difference in means and the ratio of the variances will be more peaked for larger sample sizes.

The factors that affect the equivalence weight are: the difference in the sample means and variances in the current control and historical data; the equivalence bounds chosen; the historical data sample size; and the current control sample size. The choice of equivalence bounds is discussed further in Section 3.3.4.

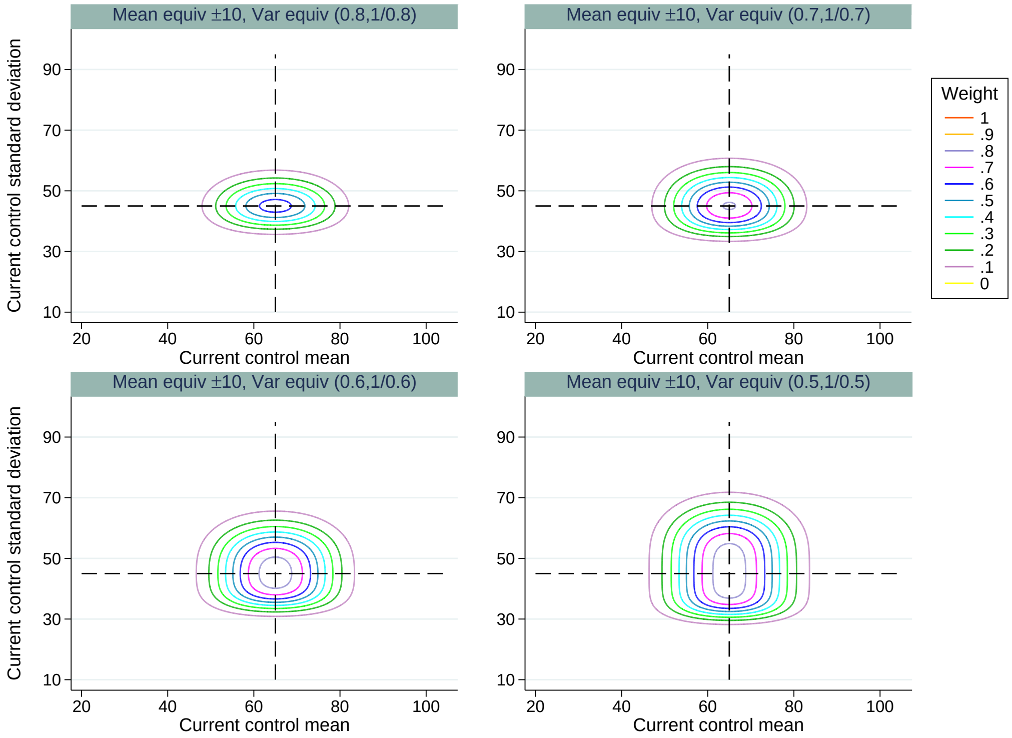
Figure 3.18: Contour plots of the corrected joint equivalence probability weights for different observed current control means and standard deviations. Equivalence bounds on the difference in means of ±8 and varying equivalence bounds on the ratio of the variances. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



As previously noted, when the sample sizes in the historical data and the current control arm differ. Choosing equivalence bounds of the form $(\delta_{\phi l},\, 1/\delta_{\phi l})$ results in a joint equivalence weight that is not maximised at complete agreement in the current control and historical sample estimates of the mean and variance. However, for the example used above, assuming the current control sample size is 500, the maximum possible equivalence weight is very similar to the equivalence weight calculated at complete agreement.

Only the corrected equivalence probability weight is considered for the remainder of this chapter and is denoted by the equivalence probability weight. The probability weight and equivalence probability weight are used as a fixed weight to discount the historical data, following the same justification as given in Section 2.3.3. For one historical study, the $EHSS$ for the probability and equivalence weight approaches is then wn_h, the same as for the power prior approach.

Figure 3.19: Contour plots of the corrected joint equivalence probability weights for different observed current control means and standard deviations. Equivalence bounds on the difference in means of ±10 and varying equivalence bounds on the ratio of the variances. Example, historical data mean 65, historical data standard deviation 45, 100 historical controls and 100 current controls. The horizontal dashed lines represent complete agreement between the current and historical control standard deviations and the vertical dashed lines represent complete agreement between the historical and current control means.



3.3.4 Choosing the equivalence bounds

The equivalence bounds chosen need to represent a clinically relevant equivalence distance, where it is assumed that within these bounds the historical and current controls are compatible. An expert is likely to have knowledge or an opinion about acceptable equivalence bounds for the difference in means between the historical and current controls, and in the absence of knowledge about acceptable equivalence bounds for the differences in means, the equivalence bounds can be chosen based on statistical properties of the study design. However, choosing an equivalence bound for the ratio of the variances of the historical and current trial control data is less intuitive, there may be less information to inform these bounds or the expert may have less knowledge on this parameter. Allowing equivalence bounds on both the differences in means and the ratio of the variances allows flexibility in the design. The study designer may be more willing to accept a larger difference in the variances than in the means of the historical and current control data and the equivalence approach can incorporate this into the design. A difference in variances may occur due to slightly different inclusion criteria for the current and historical studies for example, however the average response to treatment may be expected to be similar in the control groups. The final analysis of the current trial will focus only on the difference in means between the combined weighted historical data and current control data and the current trial treatment group.

Further considerations for the choice of equivalence bounds are:

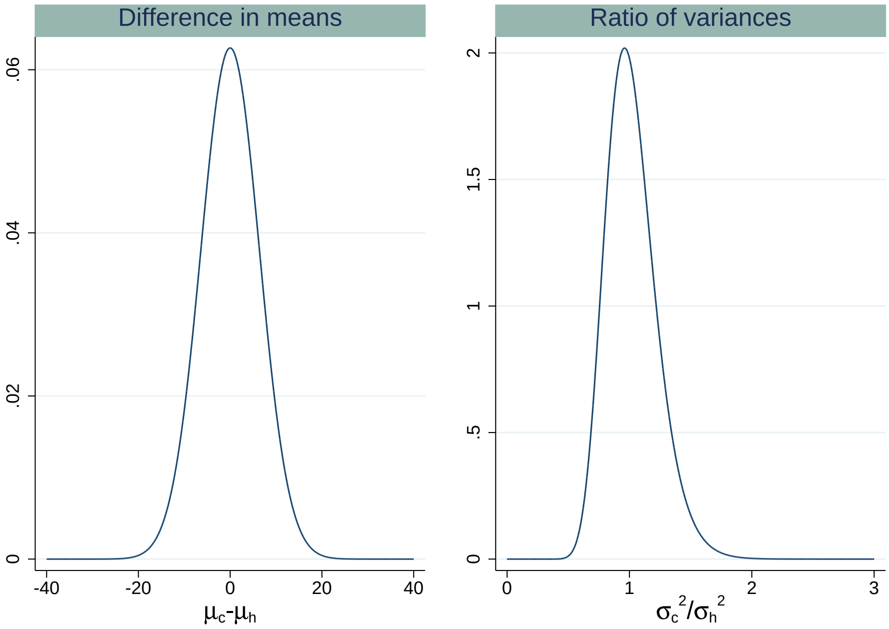
- The equivalence bound for the mean difference between the historical and current controls δ_{Δ_u} should be less than the treatment effect to be detected in the current trial.
- Narrow equivalence bounds will require large amounts of data to achieve a high weight even when the historical and current controls are in agreement.
- Choosing symmetric equivalence bounds on the difference in means for the current and historical data and averaging the equivalence weight obtained over the chosen equivalence range for the ratio of the variances will give maximum weight to the historical data when the sample means in the current and historical data are in complete agreement.
- Choosing bounds of the form $(\delta_{\phi l},\ 1/\delta_{\phi l})$ for the ratio of the variances when the historical and current control sample sizes differ will give a maximum equivalence weight when there is a slight difference in the historical and current control sample variances. However, in all examples considered, the value of the maximum equivalence weight across all ratios of the sample variances and the equivalence weight at complete agreement in variances have been very similar and in practice this should make little difference.
- The equivalence bounds can be chosen to govern how quickly the discounting of the historical data occurs as the difference between the historical and current controls increases.
- It is possible to fix the equivalence bounds on the ratio of the variances and choose the equivalence bounds on the mean difference that control the maximum possible type I error rate for the final analysis of treatment to control.
- Where the historical data sample size is much larger than the current trial control group, the equivalence bounds may be chosen so that the effective sample size of the historical data is not greater than the current control sample size at complete agreement.

Finally, to aid the choice of sensible equivalence bounds for both the difference in means and the ratio of the variances, the marginal posterior distributions or an approximation of the marginal posterior distributions of these parameters are plotted. For the example here, we plot the posterior distributions under the assumption of agreement in the historical and current control sample means and agreement in the historical and current control sample variances. As an approximation we have,

$$\begin{aligned} \mu_c - \mu_h &\sim \mathrm{N}\left(\bar{x}_c - \bar{x}_h, \frac{\hat{\sigma}_c^2}{n_c} + \frac{\hat{\sigma}_h^2}{n_h}\right), \\ \frac{\sigma_c^2}{\sigma_h^2}\frac{\hat{\sigma}_h^2}{\hat{\sigma}_c^2} &\sim \mathrm{F}(n_h - 1, n_c - 1), \end{aligned} \tag{3.16}$$

where $\bar{x}_c - \bar{x}_h = 0$, $\hat{\sigma}_c = \hat{\sigma}_h = 45$, $n_c = 100$ and $n_h = 100$.

Figure 3.20: Guidance posterior distributions, $\mu_c - \mu_h$ and σ_c^2/σ_h^2 follow the distributions given in Equation 3.16, where $\bar{x}_c - \bar{x}_h = 0$, $\hat{\sigma}_c = \hat{\sigma}_h = 45$, $n_c = 100$ and $n_h = 100$.



From Figure 3.20 the equivalence bounds explored that cover a range of different equivalences between the current and historical control data are: ±6; ±8; and ±10 for the mean difference and equivalence bounds of: $(0.8, 1/0.8)$; $(0.7, 1/0.7)$; $(0.6, 1/0.6)$; and $(0.5, 1/0.5)$ for the ratio of the variances.

3.4 Design

3.4.1 Additional information design

The primary analysis of interest is a hypothesis test of $H_0 : \mu_c = \mu_t$ against $H_1 : \mu_c < \mu_t$ where μ_c is the true mean in the current control arm and μ_t is the true mean for the treatment arm. We assume that the control and treatment data are distributed normally with unknown population variances, which are not necessarily equal.

The sample size of the current trial is fixed to detect a given treatment difference at a specified power and type I error rate. At the end of the current study, the historical and current control data are compared, the weight and effective sample size for the historical data are calculated. The control sample size is then increased from n_c to $n_c + wn_h$ ($n_c + EHSS$) and the combination of the current control data and the weighted historical data are used in the final analysis, comparing treatment to control.

The aim of this design is to increase the power of the current study by increasing the sample size of the control arm when there is agreement between the historical and current control data.

3.4.2 Adaptive design with a single interim analysis

The primary analysis of interest is a hypothesis test of $H_0 : \mu_c = \mu_t$ against $H_1 : \mu_c < \mu_t$. A two-stage adaptive design, originally proposed by Schmidli et al. [23] and Hobbs et al. [20] is considered here. The allocation ratio is adapted after the first stage:

Stage one: Randomise n_{t1} to treatment and n_{c1} to control;

Interim analysis: Calculate the weight at the interim analysis, w_1, using the first stage controls and the historical data

Stage two: Randomise $(n_t - n_{t1})$ to treatment and $\max(n_c - n_{c1} - EHSS\ ; nmin)$ to control.

Final analysis: Re-calculate the weight, w_2 using all current control data and the historical data to determine the weight to be given to the historical data in the final analysis.

Where $EHSS = w_1 \times n_h$. n_t and n_c are the desired sample sizes at the end of the trial for the treatment and control group, respectively. n_t and n_c are chosen as the sample sizes required to detect a given treatment difference at a specified power and type I error rate in a standard design not incorporating historical data. $nmin$ is a pre-specified minimum number of patients to be randomised in stage two.

The adaptive design replaces current controls yet to be randomised with historical controls when the historical and current controls are in agreement. The aim of the adaptive design is to reduce the duration of the current study and the number of control patients to be randomised in the current study when the historical and current controls agree.

3.5 Analysis

3.5.1 Analysis approach and operating characteristics for the additional information design using the power prior, probability and equivalence probability weight

The primary analysis of interest is a hypothesis test of $H_0 : \mu_c = \mu_t$ against $H_1 : \mu_c < \mu_t$. A reference prior is assumed for the joint prior on the control mean and variance before the historical data are observed $\pi(\mu_c, \sigma_c^2) \propto 1/\sigma_c^2$. A reference prior is also assumed for the joint prior on the mean and variance in the current trial treatment group, $\pi(\mu_t, \sigma_t^2) \propto 1/\sigma_t^2$.

A weight is calculated that assesses the agreement between the current and historical control data using: the modified power prior and taking a summary measure of the marginal posterior distribution of α_0; the probability weight approach; or the equivalence probability weight approach. The weight assesses the difference in both the means and variances of the current and historical control data. The power prior approach with a fixed power is then used for the analysis [14, 15]. This approach combines the weighted historical data with the current trial control data to compare to the current treatment group in the final analysis.

The joint prior for μ_c and σ_c^2 is formed from the initial reference prior, updated with the likelihood of the historical data. The posterior distribution from the reference prior updated with the historical data forms the prior for the current control data and is given by,

$$\pi(\mu_c, \sigma_c^2 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, w) \propto \sigma_c^{-2} \left((\sigma_c^2)^{-n_h/2} \exp\left\{ -\frac{n_h}{2\sigma_c^2} [\hat{\sigma}_h^2 + (\mu_c - \bar{x}_h)^2] \right\} \right)^w ,$$

where w is either a summary measure of the marginal posterior distribution of the power prior, the probability weight or the equivalence probability weight, and is therefore a fixed value and $\hat{\sigma}_h^2 = \frac{1}{n_h} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ and $\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$.

Since there is no prior information on the treatment parameters in the current trial, the prior for the current trial treatment group is the reference prior, $\pi(\mu_t, \sigma_t^2) \propto (1/\sigma_t^2)$.

The likelihood distributions for the current trial data are,

$$p(\mathbf{x_c} \mid \mu_c, \sigma_c^2, n_c) = \frac{1}{(2\pi)^{n_c/2}}(\sigma_c^2)^{-n_c/2}\text{exp}\left\{-\frac{n_c}{2\sigma_c^2}[\hat{\sigma}_c^2 + (\mu_c - \bar{x}_c)^2]\right\},$$
$$p(\mathbf{x_t} \mid \mu_t, \sigma_t^2, n_t) = \frac{1}{(2\pi)^{n_t/2}}(\sigma_t^2)^{-n_t/2}\text{exp}\left\{-\frac{n_t}{2\sigma_t^2}[\hat{\sigma}_t^2 + (\mu_t - \bar{x}_t)^2]\right\},$$

for the control and treatment group respectively, where $\mathbf{x_c}$ and $\mathbf{x_t}$ denote the vectors of outcome data in the current control and treatment group, respectively, and,

$\bar{x}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} x_{ci}$, $\bar{x}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{ti}$, $\hat{\sigma}_t^2 = \frac{1}{n_t}\sum_{i=1}^{n_t}(x_{ti} - \bar{x}_t)^2$ and $\hat{\sigma}_c^2 = \frac{1}{n_c}\sum_{i=1}^{n_c}(x_{ci} - \bar{x}_c)^2$. Here, the sample variance estimates are the maximum likelihood estimates, to be consistent with the sample estimates used in the modified power prior derivation [15].

The marginal posterior distribution of the treatment mean in the current trial, given the reference prior is then given by [61],

$$\pi(\mu_t \mid \bar{x}_t, \hat{\sigma}_t^2, n_t) = t_{n_t-1}(\mu_t \mid \bar{x}_t, \hat{\sigma}_t^2/(n_t - 1)). \tag{3.17}$$

The marginal posterior distribution for the control mean is then derived as follows. The joint posterior distribution of the control mean and variance (the initial reference prior, updated with the historical likelihood raised to a fixed power, updated again with the likelihood of the current control data) is given by,

$$\pi(\mu_c, \sigma_c^2 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c, w) \propto$$
$$(\sigma_c^2)^{\left(-\dfrac{wn_h}{2}-\dfrac{n_c}{2}-1\right)}\text{exp}\left\{-\frac{wn_h}{2\sigma_c^2}[\hat{\sigma}_h^2 + (\mu - \bar{x}_h)^2] - \frac{n_c}{2\sigma_c^2}[\hat{\sigma}_c^2 + (\mu - \bar{x}_c)^2]\right\}.$$

Integrating out σ_c^2 to obtain the marginal posterior distribution of μ_c, let $\phi = \sigma_c^2$, $\beta = \frac{wn_h}{2} + \frac{n_c}{2}$,

and $A = wn_h[\hat{\sigma}_h^2 + (\mu_c - \bar{x}_h)^2] + n_c[\hat{\sigma}_c^2 + (\mu_c - \bar{x}_c)^2]$, $x = \frac{A}{2\phi}$, $\frac{d\phi}{dx} = -\frac{A}{2}x^{-2}$.

Then,

$$\pi(\mu_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c, w) = \int \phi^{-(\beta+1)} e^{-A/2\phi} d\phi$$

$$= -\left(\frac{A}{2}\right) \int \left(\frac{A}{2x}\right)^{-(\beta+1)} e^{-x} x^{-2} dx$$

$$\propto A^{-\beta-1+1} \int x^{\beta+1-2} e^{-x} dx \propto A^{-\beta}.$$

$$A^{-\beta} = (wn_h[\hat{\sigma}_h^2 + (\mu - \bar{x}_h)^2] + n_c[\hat{\sigma}_c^2 + (\mu - \bar{x}_c)^2])^{-\left(\frac{wn_h}{2} + \frac{n_c}{2}\right)},$$

which can be re-arranged to give,

$$\pi(\mu_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c, w) \propto$$
$$\left(1 + \frac{1}{wn_h + n_c - 1}\left(\frac{\left(\mu_c - \frac{wn_h\bar{x}_h + n_c\bar{x}_c}{wn_h + n_c}\right)^2}{\frac{wn_hn_c(\bar{x}_h - \bar{x}_c)^2 + (wn_h\hat{\sigma}_h^2 + n_c\hat{\sigma}_c^2)(wn_h + n_c)}{(wn_h + n_c)^2(wn_h + n_c - 1)}}\right)\right)^{-\left(\frac{wn_h}{2} + \frac{n_c}{2}\right)},$$

which is a t-distribution of the form,

$$\pi(\mu_c \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c, w) \propto \left(1 + \frac{1}{v}\left(\frac{\mu_c - \bar{x}_0}{\hat{\sigma}_0}\right)^2\right)^{-\left(\frac{v+1}{2}\right)}, \tag{3.18}$$

where,

$v = wn_h + n_c - 1$, $\bar{x}_0 = \dfrac{wn_h\bar{x}_h + n_c\bar{x}_c}{wn_h + n_c}$, and,

$$\hat{\sigma}_0^2 = \frac{wn_hn_c(\bar{x}_h - \bar{x}_c)^2 + (wn_h\hat{\sigma}_h^2 + n_c\hat{\sigma}_c^2)(wn_h + n_c)}{(wn_h + n_c)^2(wn_h + n_c - 1)}.$$

This is the same distribution as is stated in [15] for the conditional distribution of μ_c given α_0 using the modified power prior approach. The above derivation of the marginal distribution of μ_c is using the standard power prior approach, assuming the power w is a fixed value.

The final analysis compares the difference in means of the current treatment group to the combined control group. The posterior distribution of the difference between μ_t and

μ_c can be calculated using: (1) simulation, (2) generating data from each of the marginal distributions defined in Equations 3.17 and 3.18 and calculating the difference, (3) numerical integration or, (4) using an approximation to the distribution of the difference.

Assuming the variances are unknown in the control and treatment groups and that the variances are not necessarily equal, the distribution of the difference in means can be approximated using Welsch's approximation, given by,

$$\mu_t - \mu_c \approx t\left(\bar{x}_t - \bar{x}_0, \left(\frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_0^2}{wn_h + n_c}\right)^{\frac{1}{2}}, \frac{\left(\frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_0^2}{wn_h + n_c}\right)^2}{\frac{(\hat{\sigma}_t^2/n_t)^2}{n_t - 1} + \frac{(\hat{\sigma}_0^2/(wn_h + n_c))^2}{wn_h + n_c - 1}}\right). \tag{3.19}$$

Trial success is declared when $\Pr(\mu_t - \mu_c > 0) > 0.975$.

To determine the operating characteristics of the proposed historical data designs, data are generated from normal distributions for the control and treatment arm. n_c and n_t observations are generated for the current control arm and treatment arm, respectively, for each simulation. For the simulations considered here, the variance is assumed to be the same in the control and treatment arm, $\sigma_c^2 = \sigma_t^2 = \sigma^2$.

Under the alternative hypothesis, the data are generated from,

$$X_{ci} \sim \mathrm{N}(\mu_c, \sigma^2), X_{ti} \sim \mathrm{N}(\mu_t, \sigma^2), \tag{3.20}$$

and under the null hypothesis, the data are generated from,

$$X_{ci} \sim \mathrm{N}(\mu_c, \sigma^2), X_{ti} \sim \mathrm{N}(\mu_c, \sigma^2). \tag{3.21}$$

The power and type I error rate are then calculated using,

$$\left(\sum_{i=1}^{n_{sims}} \mathbb{1}(\Pr(\mu_t - \mu_c > 0) > 0.975 \mid \bar{x}_h, \hat{\sigma}_h^2, n_h, \bar{x}_c, \hat{\sigma}_c^2, n_c, \bar{x}_t, \hat{\sigma}_t^2, n_t, w)\right) \Big/ n_{sims}, \tag{3.22}$$

under the distributions given in Equation 3.20 for the power and 3.21 for the type I error rate. Where n_{sims} is the number of simulations performed.

The expected $EHSS$ for a given true underlying control mean and variance can be calculated using,

$$n_h\mathbf{E}(w \mid \mu_c, \sigma_c^2, n_c, \bar{x}_h, \hat{\sigma}_h^2, n_h) = n_h \sum_{i=1}^{n_{sims}} (w \mid \mu_c, \sigma_c^2, n_c, \bar{x}_h, \hat{\sigma}_h^2, n_h) \Big/ n_{sims}. \qquad (3.23)$$

The expected sample size in the control group for a given true control mean and variance is then $n_c + n_h\mathbf{E}(w \mid \mu_c, \sigma_c^2, n_c, \bar{x}_h, \hat{\sigma}_h^2, n_h)$,

The mean squared error for μ_c is given by,

$$\mathbf{E}(\bar{x}_0 - \mu_c)^2 = \sum_{i=1}^{n_{sims}} \left(\left(\frac{wn_h\bar{x}_h + n_c\bar{x}_c}{wn_h + n_c} \right) - \mu_c \right)^2 \Big/ n_{sims}. \qquad (3.24)$$

In the formulae used for the analysis we have assumed that the control and treatment variances may differ. Assuming that the variances are equal and using a pooled variance for the t-test would assume that the combined historical and current control variance is equal to the treatment variance, which may not be a realistic assumption.

3.5.2 Operating characteristics for a standard trial design with no historical data

As a comparison for the methods proposed which incorporate historical data, the operating characteristics are calculated for a trial design where the data are assumed to be normally distributed and the means are compared between a treatment and control group. The standard deviations are assumed to be unknown but equal in each of the treatment groups and the sample sizes are assumed to be equal in each treatment group. No historical data are incorporated into the analysis.

For this design, the test statistic assumed is,

$$t = \frac{(\bar{x}_t - \bar{x}_c)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_t} + \frac{1}{n_c} \right)}}, \qquad (3.25)$$

where $\hat{\sigma}$ is the pooled estimate of the assumed common standard deviation in the treatment and control group. The power for the one-sided test is then given by,

$$1-\beta = T_{n_t+n_c-2}\left(\frac{\mu_t-\mu_c}{\sqrt{\frac{2\sigma^2}{n}}} - t^{-1}_{(n_t+n_c-2),0.975}\right), \tag{3.26}$$

where $t^{-1}_{(n_t+n_c-2),0.975}$ is the 97.5th quantile of the t-distribution with $n_t + n_c - 2$ degrees of freedom. The type I error rate, by design, is 2.5%.

Under a standard trial design, not incorporating any historical data, the power and type I error rate are distributed binomially with probabilities α and $1-\beta$. Note that the power varies depending on the assumed true standard deviation, since here a range of true standard deviations in the current trial are considered. The confidence interval around the power can be used as a guide as to how many simulations are required.

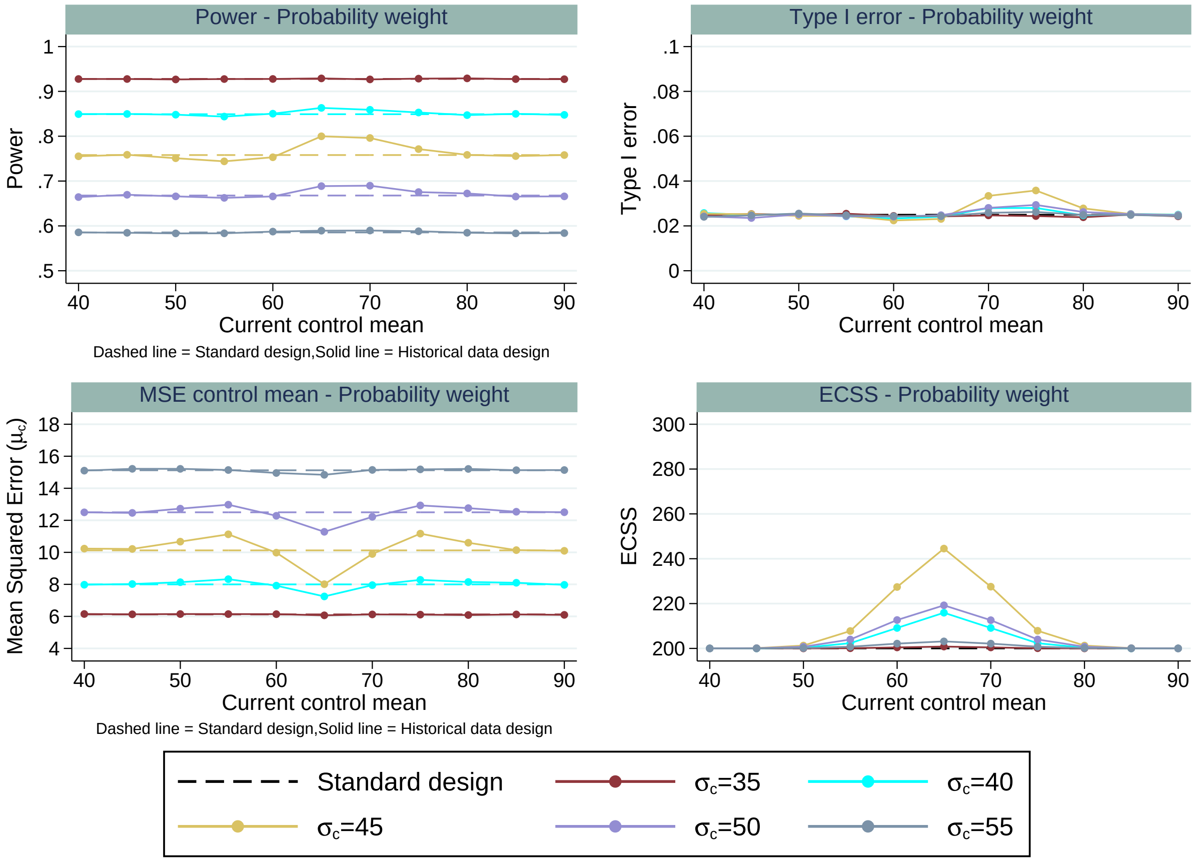
3.5.3 Additional information design – frequentist operating characteristics example using the probability and equivalence probability weight

For the example considered here, $n_c = n_t = 200$ and depending on the weight given to the historical data, there are up to a further 100 control patients from the historical data. The treatment effect, which is the difference in the mean in the treatment group compared to the control group, is assumed to be 12. The historical control mean is assumed to be 65 and the standard deviation, 45. The true underlying mean and variance in the current control data is varied. For the simulations of the power, the treatment mean is always assumed to be 12 higher than the current control mean and the standard deviation in the treatment group is assumed to be the same as the current control group. For simulations of the type I error rate, the treatment mean and standard deviation are assumed to be the same as the current control group. In the simulations, the range of the mean in the current control is varied from 10 to 120 in steps of 5 and the range of the standard deviation in the current control group is varied from 20 to 70 in steps of 5, only the relevant ranges where the historical data affects the design characteristics are shown in the figures. We simulated 100000 trials per scenario. The standard design, assuming a control mean of 65 and standard deviation of 45, not incorporating any historical data, has 76% power and one-sided type I error rate of 2.5% when there are 200 patients in each treatment group. Approximately one-half the width of a 95% confidence interval for the estimated power is then 0.0026 and for the estimated type I error rate is 0.00098. This will vary depending on the assumed standard deviation in the current trial and the amount of historical data incorporated into the final analysis for each of the approaches. However, 100000 simulations should give a reasonable approximation to all designs explored.

Probability weight

Figure 3.21 illustrates the operating characteristics for the probability weight approach for the design where the historical data are incorporated as additional information.

Figure 3.21: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the probability weight approach and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.

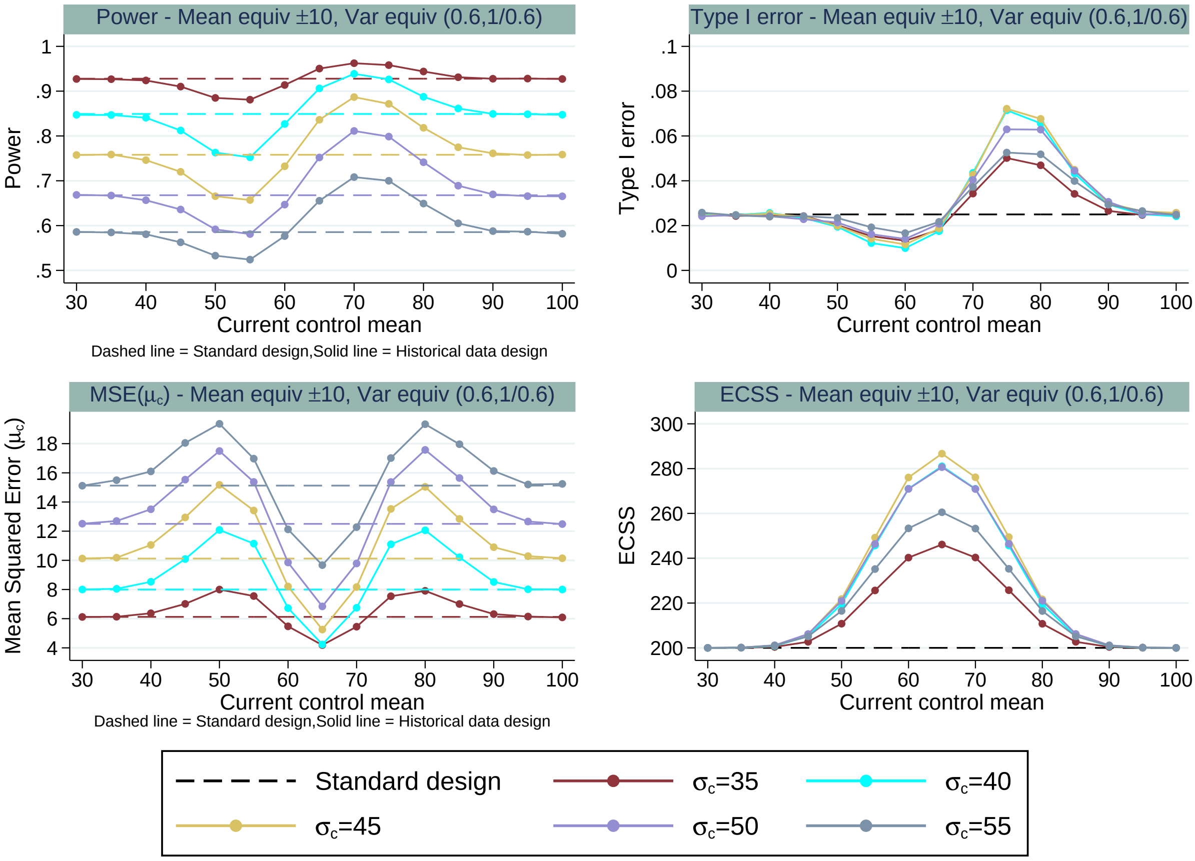


The probability weight approach quickly discounts the historical data when there is disagreement in either the means or the variances in the current controls and historical data. When the true underlying mean and variance in the current control data is in complete agreement with the historical data, on average the control arm sample size is increased by approximately 44.5 patients. This is the expected effective historical sample size at complete agreement. There is only a small gain in power, from 0.758 for the standard design to 0.800 for the probability weight design incorporating historical data at complete agreement. The final analysis is only a comparison of the mean in the treatment to the mean in the control group, the variances are not of interest. Therefore, for the comparison of means the maximum inflation in the type I error rate is most likely to occur at complete agreement in the variances between the current and historical control data, where a higher weight is given to the historical data. For the range of true underlying current control means explored here, the maximum type I error rate is 0.0358 and occurs when the true underlying current control mean is 75. There is only a small inflation in the type I error rate when there is disagreement between the historical and current control data since the probability weight discounts quickly. A disadvantage of the probability weight approach is that there is no flexibility to control the rate at which the historical data are discounted as the difference in either of the means or variances between the historical and current control data increases. There is also no flexibility in how much historical data are incorporated into the final analysis at complete agreement between the historical and current control data using the probability weight approach.

Equivalence probability weight

Figure 3.22 shows the design characteristics of the equivalence weight approach with equivalence bounds of (-10,10) on the difference in means and (0.6,1/0.6) on the ratio of the variances. The mean treatment effect to detect between the treatment arm and control is 12, therefore the equivalence bounds of ±10 on the mean difference between the current and historical controls is quite large. Here, the equivalence bounds on the variance were also chosen to be quite large. This is reflected in the amount of historical data borrowed when there is complete agreement between the historical and current controls. The expected effective historical sample size at complete agreement in the historical and current controls is 86.69 and the power is 0.836. The maximum type I error rate across the true current control means explored is 0.0722. The increase of power at complete agreement therefore increases the risk of a higher type I error rate if there is disagreement in the current and historical controls. The equivalence approach with these equivalence bounds also takes longer to completely discount the historical data and revert back to the standard design operating characteristics as the difference between the historical and current controls increases. When the variances in the historical and current data are substantially different ($\sigma_c = 35$), only a small amount of historical data are borrowed, even if the means are in complete agreement. Narrower equivalence bounds borrow less historical data both when the historical and current controls agree and disagree. For equivalence bounds of (-8,8) on the difference in means and (0.7,1/0.7) on the ratio of the variances, the maximum type I error rate across the current control means explored was 0.0595. However, the power and $EHSS$ at complete agreement are reduced compared to the design with larger equivalence bounds. The power at complete agreement was 0.826 and the $EHSS$ 72.6. The choice of bounds allows control over the maximum inflation in the type I error rate. This is explored further in Section 3.5.6.

Figure 3.22: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the corrected equivalence probability weight approach with mean equivalence bounds ±10 and variance equivalence bounds (0.6,1/0.6) and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.



Expert opinion on an acceptable range of equivalence for the agreement in variances between the current and historical control data is unlikely. It is therefore an option to combine the probability weight and equivalence weight approaches. The overall weight used to discount the historical data is the product of a weight comparing the means and a weight comparing the variances. The probability weight could be used for the variance component and an equivalence weight for the means component. However, the probability weight discounts quickly. If a difference in variances between the historical and current controls is thought to not be a problem, then choosing wide equivalence bounds for the ratio of the variances and focusing on the mean difference may be the best approach.

The design characteristics of the additional information design using the equivalence weight approach with equivalence bounds of (-8,8) on the difference in means and (0.7,1/0.7) on the ratio of the variances are given in Appendix F.

3.5.4 Analysis approach and operating characteristics for the adaptive design using the power prior, probability and equivalence probability weight

For the adaptive design considered here. The analysis approach described in Section 3.5.1 is used to calculate $\Pr(\mu_t - \mu_c > 0)$. The number of patients randomised to the control group now varies for each simulated trial depending on the weight calculated at the interim analysis, w_1. For each simulated trial, a fixed number of patients are randomised to control in the first stage of the trial, n_{c1}. The weight, w_1 is then calculated, where w_1 assesses the agreement between the historical and current controls using the n_{c1} control patients randomised in stage one and the historical controls. A further $R_c = \max(n_c - n_{c1} - w_1 n_h; nmin)$ are then randomised to control in stage two. Where n_c is the total number of control patients required. The weight, w_2 is then re-calculated at the end of the study using the historical data and all of the control patients from stages one and two, denoted by $n_{ctotal} = n_{c1} + R_c$.

To determine the operating characteristics of the adaptive design. Data are generated from normal distributions for the first stage controls, second stage controls and the treatment group using the distributions given in Equations 3.20 and 3.21. n_t observations are generated for the treatment group, n_{c1} observations for the first stage control group and the second stage number of observations generated depends on the weight w_1 calculated at the interim analysis (R_c denotes the second stage number of observations generated for a given simulation). The summary statistics in the control group at the final analysis are calculated by pooling the first and second stage controls. The power, type I error rate and mean squared error of the adaptive design are calculated using Equations 3.22 and 3.24 with,
$\hat{\sigma}_c^2 = \frac{1}{n_{ctotal}} \sum_{i=1}^{n_{ctotal}} (x_{ci} - \bar{x}_c)^2$, $\bar{x}_c = \frac{1}{n_{ctotal}} \sum_{i=1}^{n_{ctotal}} x_{ci}$, $n_c = n_{ctotal}$ and $w = w_2$ in each simulation.

The expected current control sample size for a true underling μ_c and σ_c^2 is given by,

$$ECCSS = n_{c1} + \sum_{i=1}^{n_{sims}} (R_c \mid \mu_c, \sigma_c^2, n_c, \bar{x}_h, \hat{\sigma}_h^2, n_h, w_1) \Big/ n_{sims}.$$

Since only the control data are used at the interim analysis, the treatment effect can remain blinded until the end of the trial.

3.5.5 Adaptive design – frequentist operating characteristics example using the probability and equivalence probability weight

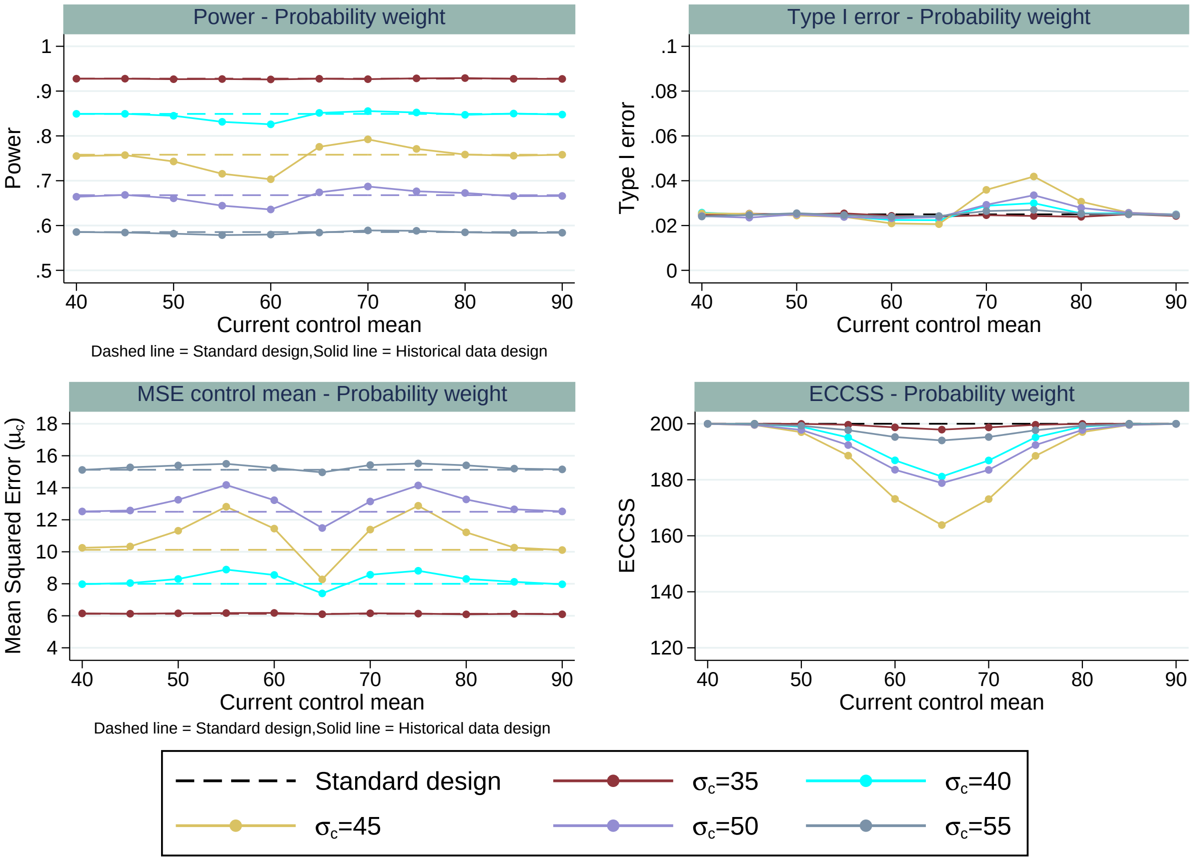
We explore the operating characteristics of the adaptive design proposed in Section 3.4.2 for a range of true control means and variances in the current study. The treatment effect is assumed to be a difference of 12 in the treatment mean compared to control. There are 100 historical control patients available with a mean of 65 and a standard deviation of 45. The interim analysis is conducted after 100 patients have been randomised to the current control arm ($n_{c1} = 100$) and the minimum number of control patients to be randomised in stage two of the trial is fixed to be 20 ($nmin = 20$). We simulated 100000 trials per scenario. The design characteristics of the adaptive design incorporating historical data are compared to the standard design not incorporating any historical data described in Section 3.5.2.

Probability weight

Figure 3.23 shows the operating characteristics of the adaptive design using the probability weight approach. Compared to the additional information design, the power at complete agreement in the historical data and the current control means and variances is slightly lower for the adaptive design. The maximum type I error rate across the current control means explored is slightly higher at 0.0419. This is due to the historical controls replacing current controls yet to be randomised. At complete agreement between the historical and current controls there is a small gain in power, a reduction in type I error rate and a saving of on average 36 patients in the control group compared to a standard trial design not incorporating any historical data.

Figure 3.24 shows the expected probability weight calculated at the interim analysis, where 100 current control patients are available, and the expected probability weight calculated at the final analysis, using all of the current control data. The probability weight is estimated with a higher accuracy at the final analysis, since there are more current control patients available to compare to the historical data. At complete agreement ($\mu_c = \bar{x}_h$ and $\sigma_c^2 = \hat{\sigma}_h^2$) the probability weight is higher at the final analysis. When there is disagreement in the means or the variances between the historical and current controls, on average the probability weight typically decreases from the interim to the final analysis. As seen with the additional information design, the probability weight discounts the historical data quickly for either a difference in the means or variances between the historical and current controls.

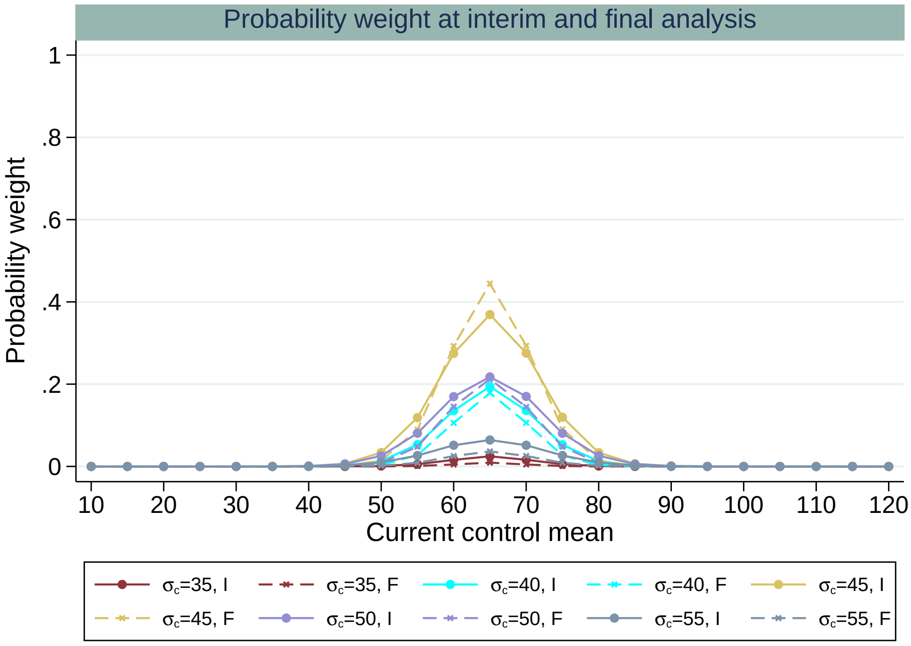
Figure 3.23: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the probability weight approach and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.



Equivalence probability weight

Figure 3.25 shows the operating characteristics of the adaptive design using the equivalence probability weight approach. As with the additional information design, since the equivalence bounds chosen for the mean and the variance are quite large, at complete agreement in the means and variances of the control and historical data, there is a slight increase in power, a reduction in the type I error rate and a substantial saving in the expected number of control patients required. The expected sample size in the current control arm at complete agreement is 126. However, this comes at the cost of the maximum type I error rate across the current controls means explored being 0.0944. Choosing narrower equivalence bounds would increase the expected current control sample size at complete agreement but also reduce the maximum possible type I error rate. The operating characteristics of the adaptive design using the equivalence probability weight

Figure 3.24: Expected historical data probability weight at the interim analysis and at the end of the study for the adaptive design. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$. I denotes the interim analysis and F denotes the final analysis.



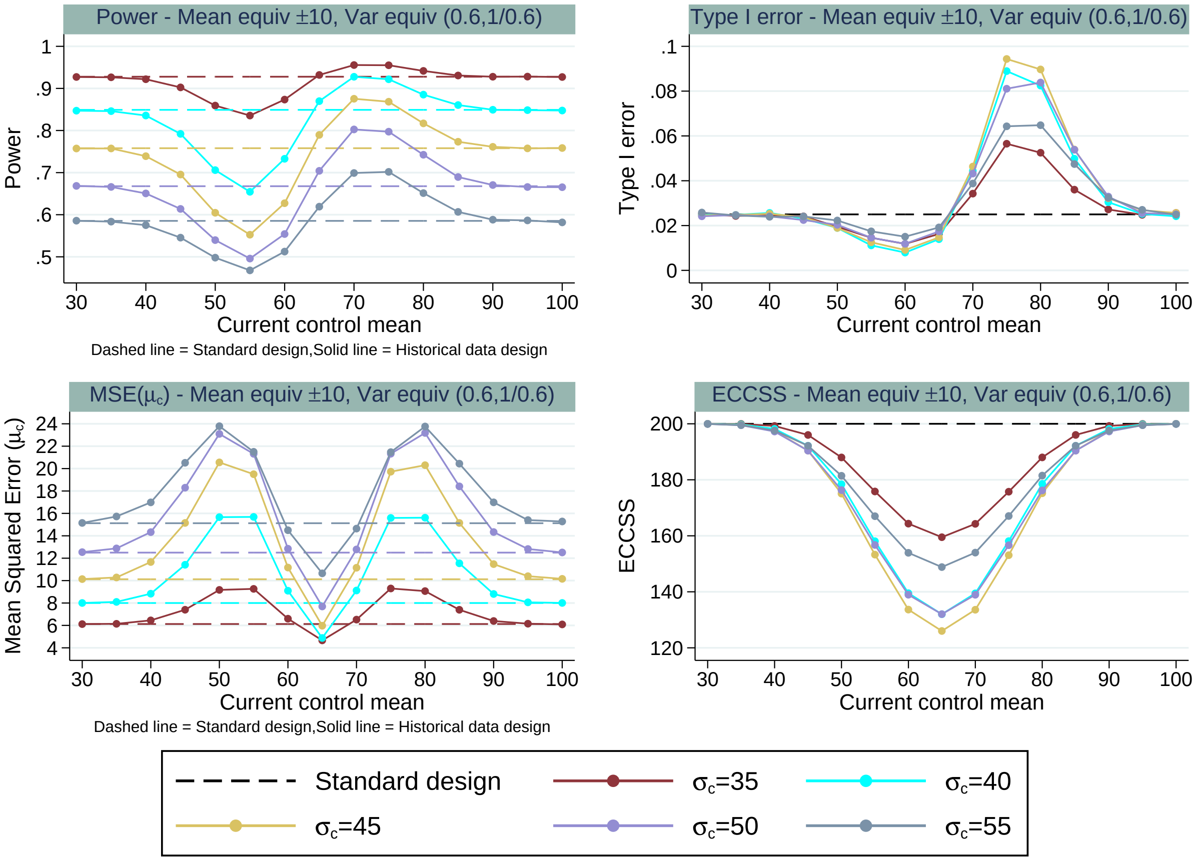
approach with equivalence bounds of (-8,8) on the difference in means and (0.7,1/0.7) on the ratio of the variances are given in Appendix F.

Figure 3.26 illustrates the expected equivalence probability weights at the interim analysis and final analysis for two different sets of equivalence bounds, (-10,10) on the difference in means and (0.6,1/0.6) on the ratio of the variances and (-8,8) on the difference in means and (0.7,1/0.7) on the ratio of the variances. Narrower equivalence bounds borrow less at agreement and discount the historical data more quickly as differences occur in either the means or the variances in the current and historical data. Depending on the equivalence bounds chosen and the level of disagreement, the expected weight given to the historical data can either increase or decrease from the interim to the final analysis.

3.5.6 Mean difference equivalence bounds that control the maximum type I error rate

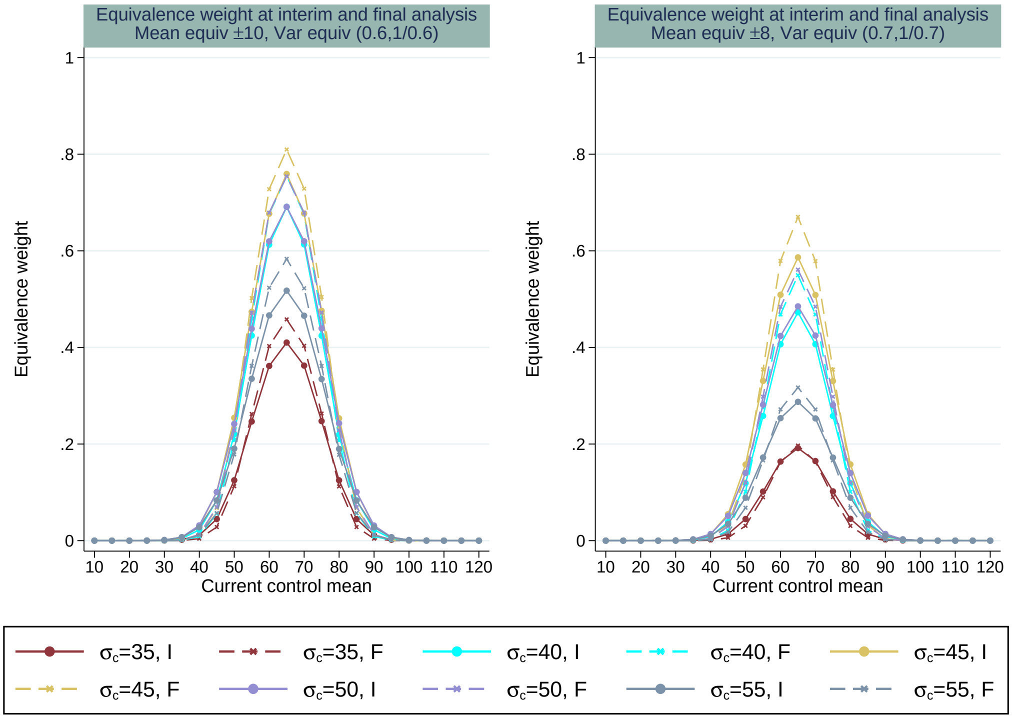
The equivalence bounds chosen have a large effect on how much of the historical data are borrowed and how quickly the historical data are discounted when there is disagreement between the current and historical controls. The equivalence bounds are chosen to represent a clinically acceptable deviation between the historical and current control data in both the means and the variances. However, the effect of the chosen equivalence

Figure 3.25: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the corrected equivalence probability weight approach with mean equivalence bounds ±10 and variance equivalence bounds (0.6,1/0.6) and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.



bounds on the operating characteristics of the current study are important. The final analysis considered here compares the mean in the treatment group to the mean in the control group. The type I error is defined as the probability of declaring the mean in the treatment group to be greater than the control group when in fact there is no difference in the means. The maximum type I error rate, by this definition is likely to occur where the variances in the historical and current control group are in agreement or close to agreement, since a larger weight is typically given to the historical data when the variances agree than when they do not. Therefore, to find the maximum type I error rate, we only search over the range of the true control means in the current trial needs to be searched over for the case where the variances in the historical and current control data agree.

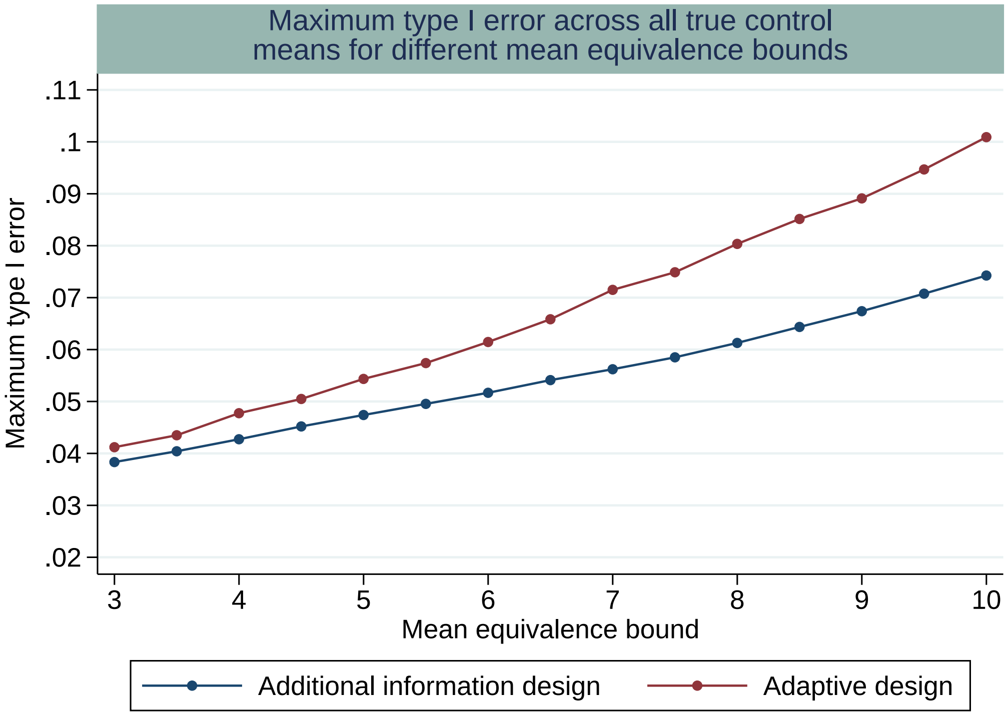
Figure 3.26: Expected historical data corrected equivalence probability weight at the interim analysis and at the end of the study for the adaptive design and two sets of equivalence bounds. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$. I denotes the interim analysis and F denotes the final analysis.



Here, it is assumed that the comparison of the means in both the historical and current controls and the control group to the treatment group are generally more important than the comparison of the variances. The equivalence bounds for the ratio of the variances are fixed and the equivalence bounds for the difference in means that control the maximum type I error rate at the desired level are found. Since the maximum inflation in the type I error rate always occurs when the current control data mean is larger than the historical data mean, only a range of current control means need to be explored for each mean equivalence bound to reduce the computation time.

Figure 3.27 illustrates the maximum type I error rate obtained across a range of possible true control means in the current study for both the additional information design and the adaptive design, for a range of mean equivalence bounds. The variance equivalence bounds are fixed at (0.6,1/0.6). The adaptive design always has a higher type I error rate than the additional information design, because in the adaptive design historical controls replace current controls yet to be randomised when there is agreement between the current and historical control data. Whereas, in the additional information design, the sample size is increased when the historical and current controls agree. The

Figure 3.27: Distribution of the maximum type I error rate across a range of equivalence bounds for the mean for the additional information and adaptive design using the corrected equivalence probability weight approach with fixed equivalence bounds of (0.6,1/0.6) on the ratio of the variances



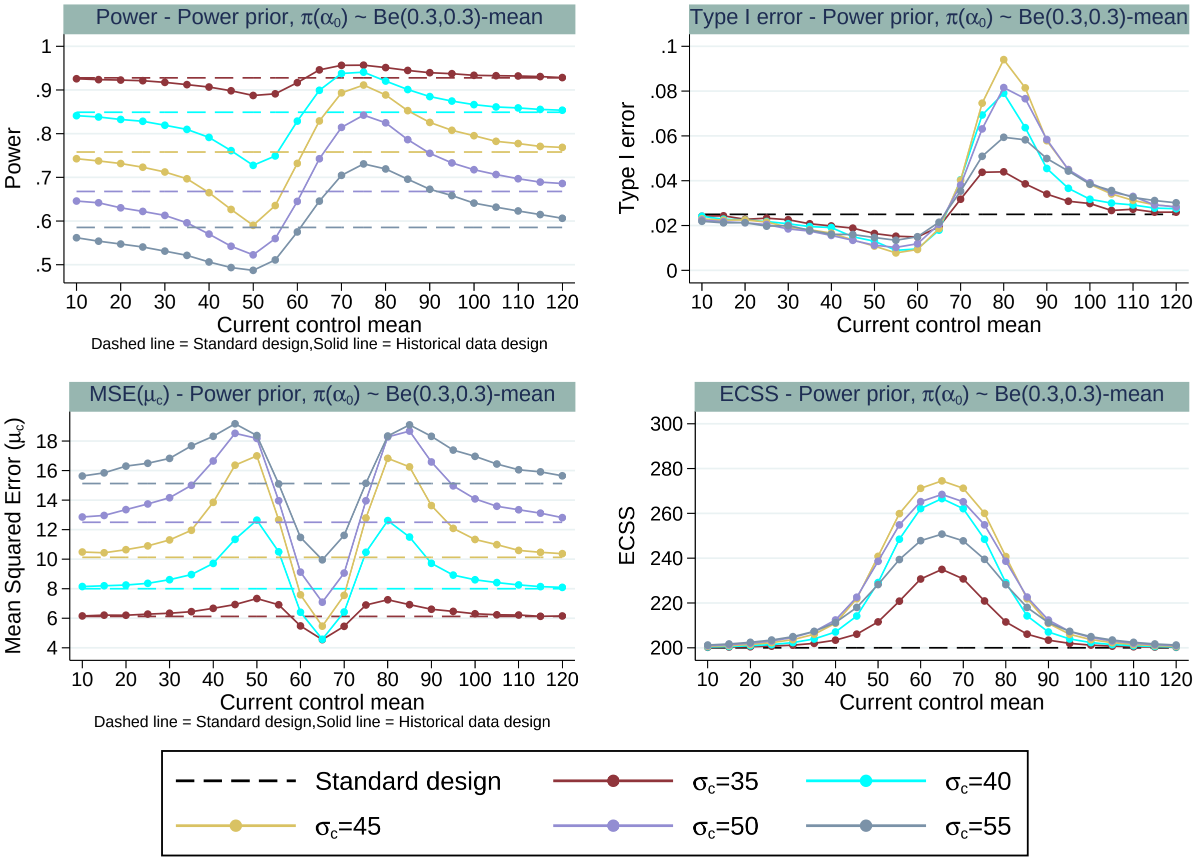
mean equivalence bounds can be chosen to control the maximum possible type I error rate at a desired level. The power at complete agreement between the historical and current controls for different mean equivalence bounds and fixed equivalence bounds of (0.6,1/0.6) on the ratio of the variances is illustrated in Appendix G for both the additional information and adaptive design.

3.5.7 Additional information and adaptive design – frequentist operating characteristics example using the power prior

Figures 3.28 and 3.29 show the operating characteristics for the additional information and the adaptive design using the power prior approach with a Beta$(0.3, 0.3)$ prior on the power parameter. The mean of the marginal posterior distribution of the power is used as a fixed weight. The operating characteristics for the other priors considered for the power parameter: Beta(1,1) with the mean or the mode as a fixed weight and Beta(0.5,0.5) with the mean as a fixed weight are given in the Appendices H and I.

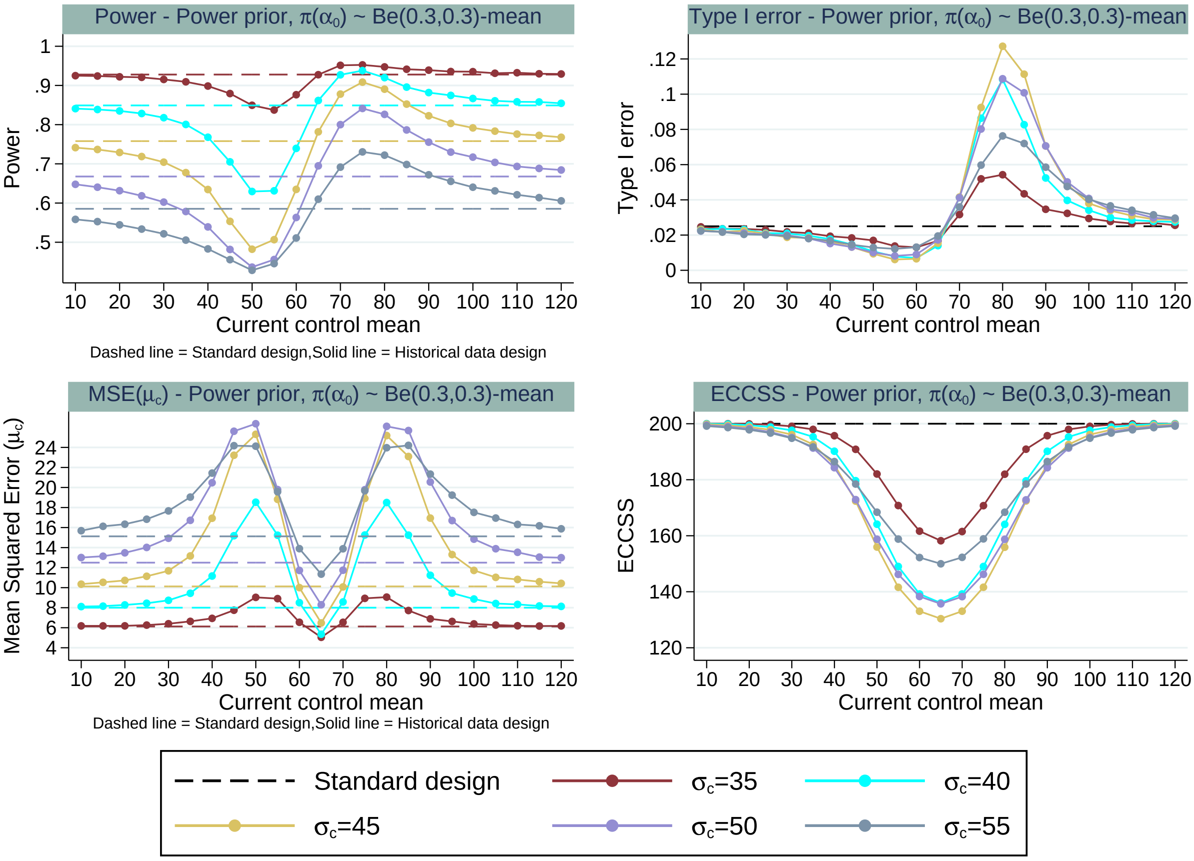
The weights using the power prior approach are slow to discount the historical data when there are differences between the historical and current controls, which results in a

Figure 3.28: Comparison of the power, type I error rate, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the modified power prior with a Beta(0.3, 0.3) prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.



large inflation in the type I error rate for both the additional information and the adaptive design, this is with a strong Beta(0.3,0.3) prior on the power parameter.

Figure 3.29: Comparison of the power, type I error rate, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the using the modified power prior with a Beta$(0.3, 0.3)$ prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.



3.6 Discussion

In this chapter we have extended the probability weight and the equivalence probability weight approaches proposed for binary data in Chapter 2 to when the outcome data from the trial are normally distributed. The main difference with normally distributed data is that both differences in means and variances between the historical and current controls are considered. The additional information design or the adaptive design can then be utilised which allow the possibility of borrowing the historical data when it is in agreement with the current trial control data. For normally distributed data, the probability weight discounts the historical data quickly for small differences in either of the means or the variances between the current and historical data. A design using the probability weight therefore has a low risk of inflating the type I error rate, however, there is also only a small gain in power for the additional information design and only a small reduction in the number of controls required in the adaptive design. The equivalence probability weight approach is flexible and allows control over how quickly the historical data are discounted when there is disagreement between the historical and current controls. The maximum inflation in the type I error rate can be controlled by the choice of equivalence bounds on the mean difference when the equivalence bounds on the ratio of the variances are fixed. As with the equivalence approach for binary outcome data, the aim of this approach is to be intuitive and easy to discuss with clinicians.

For normally distributed outcome data, it was only possible to compare the probability and equivalence weight approaches with the modified power prior weight approach. The commensurate prior was not considered because of the limitations discussed in Chapter 2. The robust mixture prior is easy to implement under the assumption that the variances are known and equal in the historical and current control data. However this is not a realistic assumption and when the variances are allowed to differ, formulating the robust mixture prior is difficult and it is unclear how the effective sample size of the historical data would be determined. The modified power prior approach, where a summary measure of the posterior distribution of the power is used as a fixed weight was easy to implement. However, the disadvantages of the power prior approach that arise for binary data, also apply for normally distributed data. The main disadvantage of the power prior approach is that the choice of prior on the power parameter is not intuitive and careful thought is required to choose a prior that gives the desired discounting of the historical data.

The probability and equivalence approaches were chosen to give a high weight when there is complete agreement between the historical and current controls. If there is a substantial amount of historical data, it may be desirable to choose narrow equivalence bounds to limit the amount of historical data borrowed.

Chapter 4

Adding a treatment arm to an ongoing trial

4.1 Introduction

In this chapter we explore how to design a trial where it is planned to add a new treatment arm part way through the study. Current practice in trials when a new treatment arm has been added is to compare the new treatment only to controls randomised concurrently [32] and this is the setting we consider here. The aim of our proposed design is to control the family-wise error rate (FWER), the probability of declaring at least one treatment is better than control when in fact there is no difference between any of the experimental treatments and control, in the strong sense, whilst maintaining the marginal power of each comparison at the level of the original study. Furthermore, for standard multi-arm trials, optimal allocation randomises a larger number of patients to the control arm than to each experimental treatment arm. In this chapter optimal allocation is explored for designs where a treatment arm is added with the aim of increasing the overall power of the study. We define the overall power as the probability of detecting all treatments that are better than control.

Alongside the administrative advantages of adding a treatment arm to an already up and running and established trial discussed in Chapter 1, statistically, having one control arm for multiple experimental treatments results in a design that requires fewer patients than running a separate trial for each experimental treatment versus control when the type I error rate is controlled at the same level for each treatment to control pairwise comparison.

Our main considerations for the design of a trial where a new treatment arm is introduced are:

- Control of the FWER for multiple comparisons of treatment to control. The number of multiple comparisons will change during the trial.

- Maintaining or increasing the power.
- The optimal allocation to each treatment arm before and after a new treatment arm is added.

Furthermore, we will discuss the use of historical controls from external studies and non-concurrent controls from within the current study.

Section 4.2.1 discusses an approach to designing a standard multi-arm trial using the test statistics and critical values proposed by Dunnett [39]. The change in correlation between these test statistics varies for different ratios of treatment to control sample size and this is explored. In Section 4.2.3 the correlation between test statistics is derived when a treatment arm is added during the trial. It is assumed that the original trial was a standard two-arm RCT and that only concurrent controls are used for each treatment comparison. An adaptive design is proposed in Section 4.2.4 where the sample size of all treatment groups is increased to control the FWER and maintain the marginal power of the original design. The adaptive design initially assumes 1:1 allocation to all treatment groups. Sections 4.3.1 and 4.3.2 then cover optimal allocation for a standard multi-arm trial and a trial where a treatment arm is added, respectively. For a trial adding a treatment arm, optimal allocation is considered both when all treatments finish recruiting simultaneously and when the treatments finish recruiting at different time points.

4.1.1 Motivation and example

STAMPEDE

(Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy.)

The STAMPEDE trial [67] provides the motivation for looking at the design of trials where treatment arms are added during the ongoing trial. STAMPEDE is a multi-arm, multi-stage (MAMS) randomised trial for the treatment of patients with prostate cancer. From the outset STAMPEDE had 5 experimental treatment arms and one control arm with 2:1:1:1:1:1 randomisation, control to each treatment and is a seamless phase II/III design.

Three additional arms were added over the duration of the study, with adjustments made for multiple testing in terms of the multiple stages of the trial, but no adjustment was made for multiple testing of each treatment with control. The study designers considered adding a new treatment to be a new trial within the STAMPEDE protocol. In terms of the FWER, all comparisons of an experimental treatment to control were not considered a family of hypotheses and each comparison was considered separately. Treatments added during the trial were only compared to controls randomised concurrently and treatment arms finished recruiting at different time points. A simplified design for adding a treatment arm is considered throughout this chapter.

Illustrative example

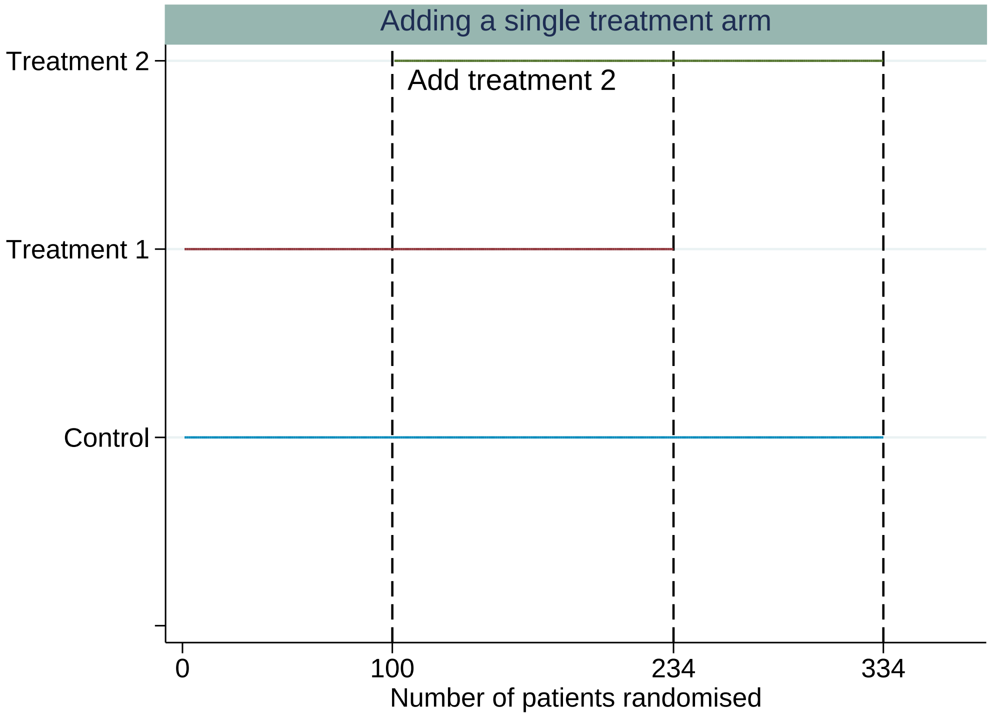
The following example is used throughout to provide a practical example of the methodology. Assuming the primary outcome variable is normally distributed. An initial two-arm confirmatory RCT trial is designed to detect a treatment difference of three in the experimental arm compared to control. The standard deviation is assumed to be 10 in both treatment arms and allocation is 1:1, treatment to control. This design requires 234 patients per treatment group to achieve 90% power with a one-sided type I error of 2.5%, representative of a confirmatory trial. It is assumed that the treatment arm added during the ongoing trial is also looking to detect a treatment difference of three compared to control and also has a standard deviation of 10. This example is illustrated in Figure 4.1. A trial starts with a single experimental treatment (treatment 1) and a control group, with 1:1 allocation. In this example, after 100 patients have been randomised to both treatment 1 and control, a new treatment (treatment 2) is added to the trial. The first dashed vertical line represents when the new treatment arm is added (100 patients randomised per group), the second dashed line represents when the original trial would have ended (234 patients randomised to control and 234 patients randomised to treatment 1) and the third dashed vertical line indicates when the trial ends after adding a new treatment arm during the trial (treatment 2). The outcome data are assumed to be normally distributed and the standard deviation is assumed to be known and the same for all treatment arms.

4.1.2 Notation

Let $j = 0, \ldots, J$ represent the treatment group, where $j = 0$ represents the control group and there are J experimental treatment groups. Let $k = 1, \ldots, K$ denote the stages of the trial. For the initial example considered, illustrated in Figure 4.1, stage one represents before the new treatment is added, stage two when all treatments are being randomised to and stage three when only treatment 2 and control are being randomised to. Let i represent the patients within each stage and treatment group, $i = 1, \ldots, n_{jk}$. n_{jk} denotes the number of patients in treatment group j in stage k.

For the standard multi-arm trial design described in Section 4.2.1. Let $\bar{X}_j$ denote the sample mean for treatment j, μ_j the true population mean for treatment j and σ^2 the true variance. The variance is assumed to be common across treatment groups. n_j

Figure 4.1: Example of adding a single experimental treatment arm to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when the original treatment (treatment 1) finishes recruitment and the third dashed vertical line represents when the control and treatment 2 finish recruiting patients.



and n_0 are the sample sizes of the experimental and control treatment groups respectively.

For the design where a treatment arm is added during the trial, described in Section 4.2.2. Let $\bar{X}_{jk}$ denote the sample mean of treatment j in stage k. Let $k = \cdot$ denote all stages. Therefore, $\bar{X}_{j\cdot}$ denotes the sample mean of treatment group j across all stages of the trial. Let n_{jk} denote the sample size of treatment j in stage k. Furthermore, when considering the control arm, let $k_{(j)}$ denote the set of stages for which controls are randomised concurrently to treatment j. Let X_{ijk} denote the observed outcome for patient i on treatment j in stage k. Let Δ_j denote the true treatment effect comparing experimental treatment j to control. Note that if all treatments finish recruiting at the same time there will be no stage 3 in the trial. As with the standard multi-arm design, μ_j denotes the true population mean for treatment j and σ is the common standard deviation for all treatment groups.

The sample means are assumed to be independently and normally distributed. The methods presented in this chapter make probability statements about the true underlying treatment differences.

For the standard multi-arm trial, let $\lambda : 1$ denote the allocation ratio of control to each experimental treatment group. In Section 4.3.2, where designs are proposed for adding a treatment arm and the optimal allocation of patients to each treatment arm is determined once the new treatment arm has been added, let λ_{jk} denote the allocation ratio of treatment j in stage k to treatment J in stage k.

In addition to the above notation, for the Bayesian design considered in Section 4.5. Let $v = 1/\sigma^2$ denote the common precision across treatment groups. The superscripts (0) and (1) indicate prior and posterior parameters, respectively. For example, $n_j^{(0)}$ denotes the prior effective sample size for treatment j and $n_j^{(1)}$ denotes the posterior effective sample size for treatment j.

4.2 Dunnett test

4.2.1 Design

For a multi-arm trial, with J experimental treatments and a control treatment, the test statistics comparing each experimental treatment to control are given by [39],

$$Z_j = \frac{\bar{X}_j - \bar{X}_0}{\sqrt{\dfrac{1}{n_j} + \dfrac{1}{n_0}}}, \tag{4.1}$$

with $j = 1, \ldots, J$ hypotheses being tested. Under the null hypothesis of no treatment difference, $Z_j \sim \mathrm{N}(0, \sigma^2)$.

The joint distribution of the J test statistics is multivariate normal with Z_j having mean 0, variance σ^2 and correlation between any two test statistics, Z_1 and Z_2 for example, given by [39],

$$\rho_{Z_1 Z_2} = 1 \bigg/ \sqrt{\left(\frac{n_0}{n_1} + 1\right)\left(\frac{n_0}{n_2} + 1\right)}\,. \tag{4.2}$$

Let σ_{Z_1} and σ_{Z_2} denote the standard deviation of test statistics Z_1 and Z_2 respectively. The correlation between any two test statistics, Z_1 and Z_2 for example, is derived as follows,

$$\rho_{Z_1 Z_2} = \frac{Cov(Z_1, Z_2)}{\sigma_{Z_1}\sigma_{Z_2}},$$

where,

$$
\begin{aligned}
Cov(Z_1, Z_2) &= Cov\left(\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_0}}}, \frac{\bar{X}_2 - \bar{X}_0}{\sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_0}}}\right) \\
&= \frac{1}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_0}}\sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_0}}} \times \underbrace{Cov(\bar{X}_0, \bar{X}_0)}_{Var(\bar{X}_0)=\sigma^2/n_0}
\end{aligned}
$$

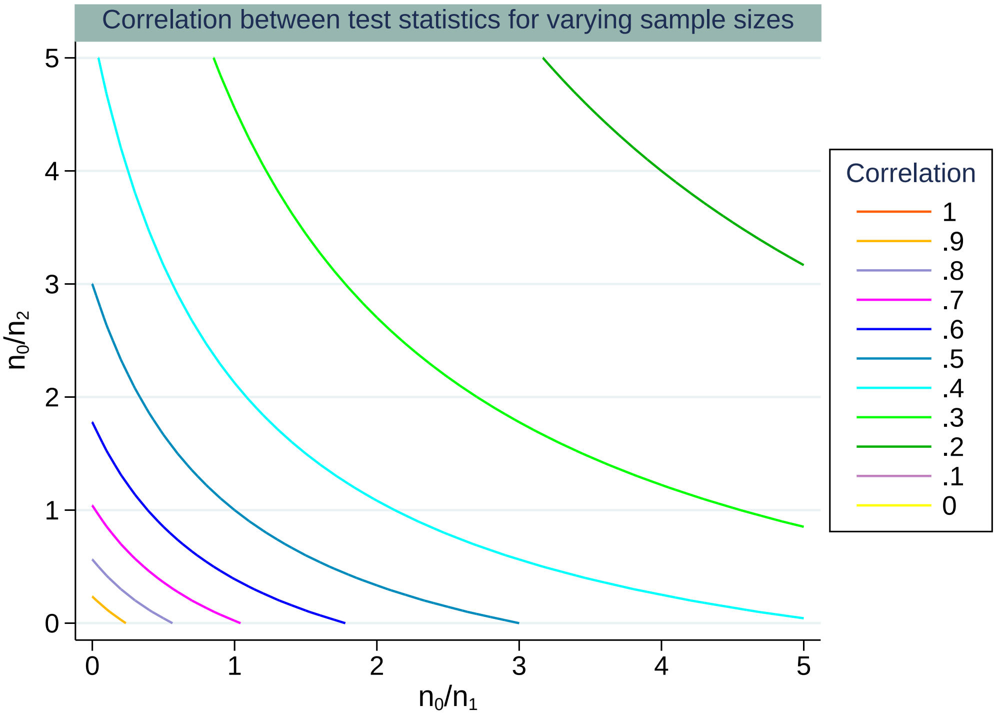
$$
\implies \rho_{Z_1Z_2} = \frac{\dfrac{\sigma^2}{\sqrt{\dfrac{(n_0+n_1)(n_0+n_2)n_0^2}{n_1 n_0^2 n_2}}}}{\sigma^2} = \frac{1}{\sqrt{\left(\dfrac{n_0}{n_1}+1\right)\left(\dfrac{n_0}{n_2}+1\right)}},
$$

where Cov denotes the covariance. Figure 4.2 illustrates how the correlation between the test statistics varies as the ratio of the control to treatment sample sizes (n_0/n_1 and n_0/n_2) vary. The correlation is 0.5 for equal sample sizes per treatment group. Where the number of patients in each experimental treatment group is larger than the number of controls, the correlation is greater than 0.5 and tends to one as the number of experimental treatment group patients increases relative to the number of control patients. When there are fewer patients in each experimental treatment group than the control group, the correlation is less than 0.5 and tends to zero as the number of experimental treatment group patients decreases relative to the number of control patients. This is not an intuitive result, since it is intuitive to think that the correlation would increase as the number of common controls increases.

Considering the standardised test statistics, Z_j/σ, under the null hypothesis, the joint distribution of the J test statistics is multivariate normal with Z_j having mean 0, variance 1 and correlation between any two test statistics, Z_1 and Z_2 for example, denoted by $\rho_{Z_1Z_2}$ and defined in Equation 4.2. If a single critical value c is used to declare significance, then the probability of not rejecting any null hypothesis is given by,

$$
\int_{-\infty}^{c}\int_{-\infty}^{c} \cdots \int_{-\infty}^{c} \pi_Z((z_1, z_2, \ldots, z_J)', \mathbf{0}, \mathbf{\Sigma}) dz_J dz_{J-1} \ldots dz_1, \tag{4.3}
$$

Figure 4.2: Correlation between Dunnett test statistics for different sample size ratios of control to treatment in each test statistic



where $\pi_Z(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of a multivariate normal (MVN) distribution with mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$.

To control the FWER at level α, c is chosen such that the integral in Equation 4.3 is equal to $1 - \alpha$. Controlling the FWER under the global null in this case ensures strong FWER control.

4.2.2 Extension to the Dunnett test when a treatment arm is added during the trial

In this section we consider a trial design where a treatment arm is added to an ongoing trial. The proposed design adapts the trial at the time point when the new treatment arm is added, with the aim of controlling the FWER. Throughout this chapter it is assumed that the decision to add a new treatment arm is driven by external reasons, independent of what is happening in the current trial, for example, a new treatment finishes phase II development and becomes available for the confirmatory trial. Adding a treatment arm based on external reasons does not require looking at the current trial data and therefore the FWER is not inflated due to interim analyses [32]. The aim of the trial is to identify whether any or all treatments are better than control and finally, it is assumed that only controls randomised concurrently are used in the analysis for each treatment. A further assumption throughout this section is that the allocation ratio is equal for all treatments throughout the study.

The main difference here to the standard Dunnett design is that the controls used for each treatment to control comparison no longer completely overlap and this changes the correlation between the test statistics.

The test statistics of interest at the end of the study, comparing each experimental treatment to control are given by,

$$Z_j = \frac{\bar{X}_{j\cdot} - \bar{X}_{0k_{(j)}}}{\sqrt{\dfrac{1}{n_{j\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(j)}} n_{0k}}}} \sim N(0, \sigma^2),$$

where $\bar{X}_{0k_{(j)}} = \dfrac{\sum\limits_{k \in k_{(j)}} \sum\limits_{i=1}^{n_{0k}} X_{i0k}}{\sum\limits_{k \in k_{(j)}} n_{0k}}$ and $k_{(j)}$ denotes the set of stages for which controls are randomised concurrently to treatment j.

4.2.3 Correlation between test statistics when adding a treatment arm

When a treatment arm is added during the trial, only some controls are used in both treatment comparisons and therefore the correlation between test statistics is reduced.

The joint distribution of the J test statistics is multivariate normal with Z_j having mean 0, variance σ^2 and correlation between any two test statistics, Z_1 and Z_2 for example, given by,

$$\rho_{Z_1 Z_2} = \frac{1}{\sqrt{\dfrac{1}{n_{1\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(1)}} n_{0k}}}\sqrt{\dfrac{1}{n_{2\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(2)}} n_{0k}}}} \frac{\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}}{\left(\sum\limits_{k \in k_{(1)}} n_{0k}\right)\left(\sum\limits_{k \in k_{(2)}} n_{0k}\right)}, \tag{4.4}$$

where $\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}$ represents the number of overlapping controls and for equal sample sizes per treatment group, the correlation simplifies to,

$$\frac{\sum\limits_{k\in k_{(1)}\cap k_{(2)}} n_{0k}}{2\times\sum\limits_{k\in k_{(1)}} n_{0k}}. \tag{4.5}$$

The correlation is 0.5 when there is complete overlap in the controls for each treatment group and zero when there is no overlap in controls, which is in agreement with the correlation from the Dunnett test statistics for a standard multi-arm trial.

Derivation of the correlation between test statistics when adding a treatment arm

$$\rho_{Z_1Z_2} = \frac{Cov\left(\frac{\bar{X}_{1\cdot}-\bar{X}_{0k_{(1)}}}{\sqrt{\frac{1}{n_{1\cdot}}+\frac{1}{\sum\limits_{k\in k_{(1)}} n_{0k}}}}, \frac{\bar{X}_{2\cdot}-\bar{X}_{0k_{(2)}}}{\sqrt{\frac{1}{n_{2\cdot}}+\frac{1}{\sum\limits_{k\in k_{(2)}} n_{0k}}}}\right)}{\sigma_{z_1}\sigma_{z_2}}.$$

Let $\kappa = \frac{1}{\sqrt{\frac{1}{n_{1\cdot}}+\frac{1}{\sum\limits_{k\in k_{(1)}} n_{0k}}}\ \sqrt{\frac{1}{n_{2\cdot}}+\frac{1}{\sum\limits_{k\in k_{(2)}} n_{0k}}}}$. Then,

$$Cov\left(\frac{\bar{X}_{1\cdot}-\bar{X}_{0k_{(1)}}}{\sqrt{\frac{1}{n_{1\cdot}}+\frac{1}{\sum\limits_{k\in k_{(1)}} n_{0k}}}}, \frac{\bar{X}_{2\cdot}-\bar{X}_{0k_{(2)}}}{\sqrt{\frac{1}{n_{2\cdot}}+\frac{1}{\sum\limits_{k\in k_{(2)}} n_{0k}}}}\right) = \kappa\ Cov(\bar{X}_{0k_{(1)}}, \bar{X}_{0k_{(2)}})$$

$$= \kappa\ Cov\left(\frac{\sum\limits_{k\in k_{(1)}}\sum\limits_{l=1}^{n_{0k}} X_{l0k}}{\sum\limits_{k\in k_{(1)}} n_{0k}}, \frac{\sum\limits_{k\in k_{(2)}}\sum\limits_{m=1}^{n_{0k}} X_{m0k}}{\sum\limits_{k\in k_{(2)}} n_{0k}}\right)$$

$$= \kappa\ \frac{1}{\sum\limits_{k\in k_{(1)}} n_{0k}\sum\limits_{k\in k_{(2)}} n_{0k}}\left(\sum_{k\in k_{(1)}}\sum_{l=1}^{n_{0k}}\right)\left(\sum_{k\in k_{(2)}}\sum_{m=1}^{n_{0k}}\right)\ \underbrace{Cov(X_{l0k}, X_{m0k})}_{\mathbf{E}(X_{l0k}X_{m0k})-\mathbf{E}(X_{l0k})\mathbf{E}(X_{m0k})}$$

$(k \notin k_{(1)} \cap k_{(2)} \implies \mathbf{E}(X_{l0k}X_{m0k}) = \mathbf{E}(X_{l0k})\mathbf{E}(X_{m0k}))$

$$= \kappa \frac{1}{\sum\limits_{k \in k_{(1)}} n_{0k} \sum\limits_{k \in k_{(2)}} n_{0k}} \sum_{m \in k_{(1)} \cap k_{(2)}} \sum_{m=1}^{n_{0k}} \underbrace{\mathbf{E}(X_{m0k}X_{m0k}) - \mathbf{E}(X_{m0k})\mathbf{E}(X_{m0k})}_{= \mathrm{Var}(X_{m0k}) = \sigma^2}$$

$(\text{Since } k \in k_{(1)} \cap k_{(2)} \implies X_{l0k} = X_{m0k})$

$$= \kappa\sigma^2 \frac{\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}}{\sum\limits_{k \in k_{(1)}} n_{0k} \sum\limits_{k \in k_{(2)}} n_{0k}}.$$

$$\rho_{Z_1 Z_2} = \frac{\dfrac{1}{\sqrt{\dfrac{1}{n_{1\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(1)}} n_{0k}}}} \dfrac{1}{\sqrt{\dfrac{1}{n_{2\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(2)}} n_{0k}}}} \dfrac{\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}}{\sum\limits_{k \in k_{(1)}} n_{0k} \sum\limits_{k \in k_{(2)}} n_{0k}} \times \sigma^2}{\sigma^2}$$

$$= \frac{1}{\sqrt{\dfrac{1}{n_{1\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(1)}} n_{0k}}}} \frac{1}{\sqrt{\dfrac{1}{n_{2\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(2)}} n_{0k}}}} \frac{\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}}{\sum\limits_{k \in k_{(1)}} n_{0k} \sum\limits_{k \in k_{(2)}} n_{0k}}.$$

This simplifies if the number of patients in each treatment group are equal,

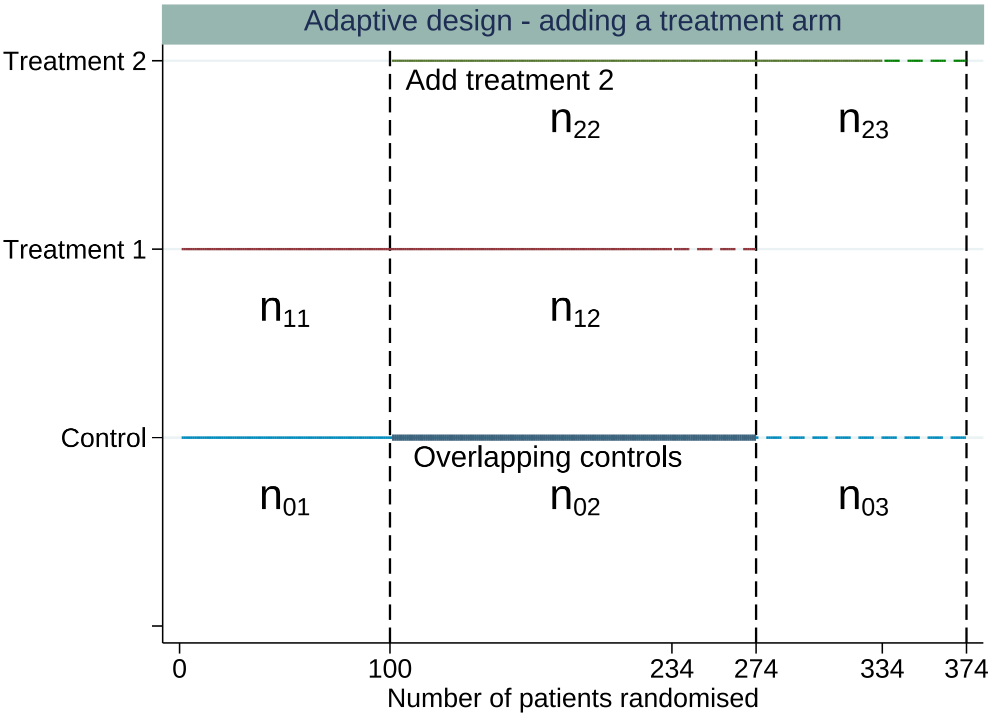
$$\sum_{k \in k_{(1)}} n_{0k} = \sum_{k \in k_{(2)}} n_{0k} = n_{1\cdot} = n_{2\cdot} = n \implies \rho_{Z_1 Z_2} = \frac{\sum\limits_{k \in k_{(1)} \cap k_{(2)}} n_{0k}}{2n}.$$

4.2.4 An adaptive design for adding a treatment arm with the Dunnett correction

Initially, adding a single treatment to an ongoing two-arm trial is considered. Figure 4.3 illustrates the adaptive design which is described in detail in this section.

The trial design considered here starts with two treatment arms, an experimental treatment and a control. The null hypothesis is, $H_0 : \mu_1 = \mu_0$ and the alternative hypothesis, $H_1 : \mu_1 > \mu_0$.

Figure 4.3: Example of adding a single experimental treatment arm to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when the original treatment (treatment 1) finishes recruitment and the third dashed vertical line represents when the control and treatment 2 finish recruiting patients. The horizontal dashed lines represent the additional patients required per treatment group above the original sample size estimate to control the FWER while maintaining randomisation of 1:1:1 to all treatment arms.



The test statistic is given by,

$$Z_1 = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}.$$

The sample size required per treatment group for this design is calculated using the standard formula,

$$n = \frac{2\sigma^2(Z_{1-\alpha} - Z_\beta)^2}{(\mu_1 - \mu_0)^2},$$

where α is the significance level, β the probability of a type II error and $Z_{1-\alpha}$ and Z_β are the $(1-\alpha)$th and the βth quantiles of the standard normal distribution.

The power is given by,

$$1-\beta=\Phi\left(\frac{(\mu_1-\mu_0)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_0}\right)}}-Z_{1-\alpha}\right).$$

Assuming that a new treatment arm is added after n_{01} patients have been randomised to control. The test statistics of interest are now,

$$Z_1=\frac{\bar{X}_{1\cdot}-\bar{X}_{0k_{(1)}}}{\sqrt{\dfrac{1}{n_{11}+n_{12}}+\dfrac{1}{n_{01}+n_{02}}}}, Z_2=\frac{\bar{X}_{2\cdot}-\bar{X}_{0k_{(2)}}}{\sqrt{\dfrac{1}{n_{22}+n_{23}}+\dfrac{1}{n_{02}+n_{03}}}},$$

where $k_{(1)}=\{1,2\}$ and $k_{(2)}=\{2,3\}$. Z_1 and Z_2 are distributed $N(0,\sigma^2)$ under the null hypothesis of no treatment difference.

We assume equal randomisation to all treatment arms throughout. All treatment groups are of the same size ($n=n_{01}+n_{02}=n_{02}+n_{03}=n_{11}+n_{12}=n_{22}+n_{23}$). Then the correlation between Z_1 and Z_2 is $n_{02}/(2n)$, which can be written in terms of the ratio of the first stage control sample size to the second stage control sample size, $\rho_{Z_1Z_2}=1/2(\frac{n_{01}}{n_{02}}+1)$. Here it is assumed that the true treatment effect in both of the experimental treatment arms compared to control are the same ($\mu_1=\mu_2=\mu$). This could be the case when two experimental treatments are explored that are similar but contain different compounds or have differences in cost or risk profiles, but the efficacy effects are thought to be similar. However, it could be assumed that the treatment effect for each experimental treatment differs and therefore different sample sizes would be required. The null hypothesis for the second treatment is, $H_0:\mu_2=\mu_0$ and the alternative hypothesis is, $H_1:\mu_2>\mu_0$. To control the FWER, the number of patients randomised to every treatment arm is increased when the new treatment arm is added to the trial, using the following steps:

1. Estimate the correlation using the original sample size calculation;

   The correlation is given by, $\rho_{Z_1Z_2}=\dfrac{1}{2\left(\frac{n_{01}}{n_{02}}+1\right)}$

2. Increase the sample size of all treatment arms to control the FWER at level α and maintain the marginal power for each pairwise comparison;

   Determine the critical value that controls the FWER at level α using,

$$= \int_{-\infty}^{c} \int_{-\infty}^{c} \pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma}) dz_2 dz_1 = 1 - \alpha$$

where $\pi(\boldsymbol{z}, \mathbf{0}, \mathbf{\Sigma})$ is the probability density function of a multivariate normal distribution with means 0, and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{Z_1 Z_2} \\ \rho_{Z_2 Z_1} & 1 \end{pmatrix}$.

The critical value here is determined using multivariate normal integration and a root finding algorithm in Stata which implements Brent's method [41].

The sample size required per group is then given by,

$$n = \frac{2\sigma^2 (Z_{\Phi(c)} - Z_\beta)^2}{(\mu - \mu_0)^2}$$

3. Re-estimate the correlation using the sample sizes obtained in step 2 and repeat step two using this correlation until the sample size does not change.

The estimate of the correlation between test statistics at the end of the study using the original sample size calculation will be an underestimate, since, when the sample size of all treatment groups is increased, the number of overlapping controls will increase and therefore the correlation will increase. This is why an iterative approach is required.

4.3 Optimal allocation with a fixed trial sample size

In multi-arm trials there are various ways to define the power of a study. The marginal power was considered in Section 4.2, which is the probability of rejecting a particular false null hypothesis. However, in a single trial, we are making multiple comparisons and the main aim of the trial may be to determine whether any or all treatments are better than control, or at least one treatment is greater than control. When there are multiple treatment comparisons, interest may be in the overall power of the study. Here, the overall power of a multi-arm trial is defined to be the probability of rejecting all false null hypotheses, correctly determining all treatments that are better than control. The overall power is considered in this section, with the aim of determining the optimal sample size which gives the highest probability of finding all treatments that are better than control. It is assumed that all treatments are of equal importance. Further possible definitions of power in a multi-arm trial are: weighted power – a weighted sum of the marginal powers; minimal power – the probability of rejecting at least one false null hypothesis and average power – the average proportion of false null hypotheses that are rejected. A comprehensive review of power definitions in multi-arm trials is given in [36]. Firstly, the power and optimal allocation for the original Dunnett test are considered [39].

4.3.1 Optimal allocation for the original Dunnett test

Dunnett states that the optimal allocation for the standard Dunnett test is $n_0 = n\sqrt{J}$. Where n_0 is the number of controls, n is the number in each treatment group and J is the number of experimental treatments [39].

When using optimal allocation the number of patients per group are not equal. This alters the correlation between the test statistics and therefore the critical value used for each test.

Two methods are described in the next section for deriving the optimal allocation ratio for the standard Dunnett design. Minimising the total variance to determine the optimal allocation was proposed in [68]. The criterion we propose for determining the optimal allocation is to maximise the overall power, this is the criterion we use to determine the optimal allocation for our extension to the Dunnett design where a new treatment arm is added during the study.

Optimal allocation derivation

Minimising the total variance

The optimal allocation ratio can be approximated by minimising the total variance (TV), assuming a common variance among treatments and equal sample sizes for all experimental treatment groups, $n_j = n$ for $j > 1$. The total variance is the sum of the variances of the differences in means ($\bar{X}_j - \bar{X}_0$) and is given by [68],

$$TV = J\sigma^2 \left(\frac{1}{n} + \frac{1}{n_0} \right).$$

The values of n and n_0 are determined that minimise the total sample variance subject to the constraint that the total sample size is $N = Jn + n_0$. This makes the assumption that all treatments are equally important.

To minimise $J\sigma^2 \left(\frac{1}{n} + \frac{1}{n_0} \right)$ where $N = n_0 + Jn$. This is a Lagrange multiplier problem [68], $\min\{TV + \lambda(N - Jn - n_0)\}(*)$:

Solve,

$$\frac{\partial(*)}{\partial n} = \frac{-J\sigma^2}{n^2} - \lambda J = 0, \tag{4.6}$$

and

$$\frac{\partial(*)}{\partial n_0} = \frac{-J\sigma^2}{n_0^2} - \lambda = 0. \tag{4.7}$$

From Equation 4.7, $\lambda = \dfrac{-\,J\sigma^2}{n_0^2}$, substituting into 4.6 gives [68],

$$\frac{-J\sigma^2}{n^2} = \lambda J = \frac{-J\sigma^2}{n_0^2} \implies n^2 = \frac{n_0^2}{J} \implies n = \frac{n_0}{\sqrt{J}} \implies n_0 = n\sqrt{J}.$$

This gives the simple rule $n_0 = n\sqrt{J}$ for optimal allocation and substituting n_0 into $N = Jn + n_0$ gives $n = \dfrac{N}{J+\sqrt{J}}$ [68].

Maximising the overall power

An alternative approach to determining the optimal allocation ratio is to maximise the overall power. The overall power is the probability of rejecting all false null hypotheses. For a three arm design, comparing each experimental treatment to control, the overall power is defined by the bivariate normal distribution,

$$1 - \beta = \int_{c^*}^{\infty} \int_{c^*}^{\infty} \pi_Z((z_1, z_2)', \boldsymbol{\mu}, \boldsymbol{\Sigma}) dz_2 dz_1$$

where $\pi_Z((z_1, z_2)', \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a bivariate normal distribution with mean $\boldsymbol{\mu} = (0,0)'$ and variance covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

where $c^* = \dfrac{\left(c\sqrt{\frac{\sigma^2}{n_0^*} + \frac{\sigma^2}{n^*}}\right) - \Delta}{\sqrt{\frac{\sigma^2}{n_0} + \frac{\sigma^2}{n}}}$, c is the Dunnett critical value and n^* and n_0^* in the numerator of c^* are fixed as the sample sizes calculated from 1:1 allocation of patients to treatment and control. The values of n and n_0 in the denominator that maximise the overall power are then found by searching over a range of allocation ratios to determine the allocation ratio that maximises the overall power. This allocation ratio will randomise more patients to control than each of the treatment groups. This changes the correlation between the test statistics which affects the FWER. Based on this optimal allocation ratio, $\boldsymbol{\Sigma}$ is updated and the Dunnett critical value that controls the FWER is re-calculated. The optimal allocation that maximises the overall power is then re-calculated, with the numerator n^* and n_0^* fixed at the values from the previous iteration, c is the updated Dunnett critical value, and n and n_0 in the denominator are found that maximise the overall power, this procedure is repeated until the sample sizes obtained for the control and the treatment groups both maximise the overall power and also control the FWER.

The two approaches to determining the optimal allocation ratio, minimising the total variance and maximising the overall power, both give an optimal allocation ratio of approximately $n_0 = n\sqrt{J}$. The overall power and formula approximation of the optimal allocation are calculated assuming both treatments have the same treatment effect over control.

Both approaches to determining the optimal allocation ratio can be applied when there are multiple treatment arms.

4.3.2 Optimal allocation when adding a treatment arm

When adding a treatment arm to an ongoing trial, the optimal allocation depends on whether adding a new arm was pre-planned, to be added at a specific time point, or planned to be included when the treatment becomes available at an unknown time point. If adding the arm was pre-planned it may be desirable to use unequal randomisation from the beginning of the trial. However, the situation considered here is that at the design stage of the initial trial, it is planned that a new treatment arm will be added, but the time point is only known once the two-arm trial has begun recruiting. In this case, optimal allocation for the initial trial is 1:1, treatment:control.

A further consideration is that optimal allocation will differ depending on whether, once a new treatment is added, randomisation continues to all treatment arms until the end of the study or whether the original treatment finishes recruitment earlier. Elm et al. [30] suggest that once a new treatment arm is added, randomisation should continue to all treatment arms until the end of the study. In this section we consider treatments finishing recruitment simultaneously and at different times.

Finally, it is assumed here that the patient population is homogeneous across stages. A stage is defined when there is a change in the design of the trial. For example, a treatment arm is added, a treatment finishes recruitment or the allocation ratio is changed. The assumption of homogeneity across stages is also the primary assumption for group sequential test procedures and for platform trials [69, 70]. The limitations of this assumption are considered in the discussion.

The overall power calculations throughout this section are the power to detect a treatment effect in both groups and the treatment effect is assumed to be the same in both groups. Only controls randomised concurrently to the treatment being tested are used in the analysis. As above, it is assumed that the addition of a treatment arm is based on external reasons and not on the results of the current trial.

Optimal allocation when all treatments finish recruiting simultaneously

In this section optimal allocation is considered when a treatment arm is added to a two-arm study that was initially randomising 1:1, treatment to control. It is assumed that adding a treatment arm was planned and that recruitment will continue to all arms until the end of the trial. The null hypotheses being tested are: $H_0 : \mu_j = \mu_0$ and the alternative hypotheses are, $H_1 : \mu_j > \mu_0$. Where μ_j are assumed to be common across treatment groups.

Minimising the total variance

It is not possible to use the approach of minimising the total variance described in [68] to determine the optimal allocation in the design when a treatment is added during the trial. When adding a treatment arm the total variance is given by,

$$TV = \left(\frac{\sigma^2}{n_{01} + n_{02}} + \frac{\sigma^2}{n_{11} + n_{12}} + \frac{\sigma^2}{n_{22}} + \frac{\sigma^2}{n_{02}} \right),$$

and we wish to minimise the total variance under the constraint, $N = n_{01} + n_{02} + n_{11} + n_{12} + n_{22}$.

This approach does not account for the fact that more controls $(n_{01} + n_{02})$ are used in the treatment 1 to control comparison than the treatment 2 to control comparison (n_{02}). It therefore allocates an equal number of patients to both treatment groups. Whereas, for optimal allocation, a larger number of patients should be randomised to the new treatment 2 arm. There is not a simple formula for the optimal allocation here as there is for the standard Dunnett test.

Maximising the overall power

To determine the optimal allocation ratio when adding a treatment arm during the trial, it is possible to numerically maximise the overall power for a fixed total sample size using the following steps.

The test statistics of interest at the end of the study, comparing treatments 1 and 2 (where treatment 2 is added during the trial) to control, are given by,

$$Z_1 = \frac{\bar{X}_{1\cdot} - \bar{X}_{0k_{(1)}}}{\sqrt{\dfrac{1}{n_{1\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(1)}} n_{0k}}}}, \tag{4.8}$$

$$Z_2 = \frac{\bar{X}_{2\cdot} - \bar{X}_{0k_{(2)}}}{\sqrt{\dfrac{1}{n_{2\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(2)}} n_{0k}}}}, \tag{4.9}$$

where $\bar{X}_{0k_{(1)}} = \frac{\sum\limits_{k \in k_{(1)}} \sum\limits_{i=1}^{n_{0k}} X_{i0k}}{\sum\limits_{k \in k_{(1)}} n_{0k}}$, $\bar{X}_{0k_{(2)}} = \frac{\sum\limits_{k \in k_{(2)}} \sum\limits_{i=1}^{n_{0k}} X_{i0k}}{\sum\limits_{k \in k_{(2)}} n_{0k}}$, $k_{(1)} = \{1, 2\}$ and $k_{(2)} = \{2\}$. Z_1 and Z_2 are distributed $\mathrm{N}(0, \sigma^2)$ under the null hypothesis of no treatment difference.

1. Follow the steps in Section 4.2.4 to determine the required total sample size, N, when adding a treatment arm during the trial using 1:1:1 allocation to all treatment arms.

   The total sample size is fixed at N. It is then determined how best to allocate the remaining patients once the new treatment arm has been added to optimise the overall power. The new treatment arm is added at a fixed time point after $n_{01} = n_{11}$ patients have been randomised to the original treatment and control.

2. Determine the optimal allocation ratios in stage two of the trial that maximise the overall power. The estimate of the correlation and critical value from step one are used.

   The correlation based on the total sample size calculation in step one is given by, $\rho_{Z_1Z_2} = 1/2(\frac{n_{01}}{n_{02}} + 1)$ where n is the number of patients used per group for each treatment comparison and n_{02} is the number of overlapping controls. The remaining patients to be randomised can then be written as, $R = N - (n_{01} + n_{11})$. The total number of patients is given by, $N = n_{01} + n_{02} + n_{11} + n_{12} + n_{22}$ (there is no stage 3 since it is assumed randomisation continues to all treatment arms until the end of the study).

   To determine the optimal allocation for the remaining patients once the new treatment arm is added, the randomisation allocation ratio to the new treatment is fixed at one, the remaining patients to be randomised at the time point that the new treatment arm is added is written as,

$$R = \underbrace{\lambda_{02} n_{22}}_{n_{02}} + n_{22} + \underbrace{\lambda_{12} n_{22}}_{n_{12}},$$

and the values of λ_{02} and λ_{12} are determined that maximise the overall power, which is defined by,

$$1-\beta = \int_{c1^*}^{\infty}\int_{c2^*}^{\infty} \pi_Z((z_1,z_2)', \mathbf{0}, \mathbf{\Sigma})dz_2dz_1, \tag{4.10}$$

where $\pi_Z((z_1,z_2)', \mathbf{0}, \mathbf{\Sigma})$ is the bivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{z_1z_2} \\ \rho_{z_2z_1} & 1 \end{pmatrix},$$

where $\rho_{z_1z_2}$ is the correlation calculated in step one, defined in Equation 4.5.

$c1^*$ and $c2^*$ in Equation 4.10 are defined by,

$$c1^* = \frac{\left(c\sqrt{\dfrac{2\sigma^2}{n}}\right) - (\mu_1-\mu_0)}{\sqrt{\dfrac{\sigma^2}{n_{01}+n_{02}} + \dfrac{\sigma^2}{n_{11}+n_{12}}}}, \tag{4.11}$$

$$c2^* = \frac{\left(c\sqrt{\dfrac{2\sigma^2}{n}}\right) - (\mu_2-\mu_0)}{\sqrt{\dfrac{\sigma^2}{n_{02}} + \dfrac{\sigma^2}{n_{22}}}}, \tag{4.12}$$

where c is the Dunnett critical value and n the sample size per comparison group estimated in step one. $(\mu_j - \mu_0)$ is the treatment effect under the alternative hypothesis and,

$$n_{22} = \frac{R}{\lambda_{02}+1+\lambda_{12}},\ n_{12} = \frac{R}{\lambda_{02}+1+\lambda_{12}} \times \lambda_{12},\ n_{02} = \frac{R}{\lambda_{02}+1+\lambda_{12}} \times \lambda_{02}.$$

where λ_{02} and λ_{12} are the allocation ratios that maximise the overall power. This change in allocation ratio will alter the correlation. As in Section 4.2.4, iteration is required to control the FWER at the desired level.

3. Re-estimate the correlation and critical value that control the FWER based on the sample sizes per comparison group calculated in step two. The sample sizes per

comparison group are now unequal and Equation 4.4 is required to estimate the correlation.

4. Repeat step 2 (replacing the Dunnett critical value and variance in the numerator of Equations 4.11 and 4.12 with the critical value and variance calculated using the sample sizes that maximise the overall power. Replace the correlation in Equation 4.10 with the correlation calculated using the optimal sample sizes) and repeat step 3 until the optimal allocation is determined that maximises the overall power and where the FWER is also controlled at the desired level.

Optimal allocation when treatments finish recruiting at different time points

The previous section looks at determining the optimal allocation for the remaining patients when a new treatment arm is added part way through the trial, assuming patients are recruited to all treatments until the end of the study. This results in a reduced allocation ratio to the original experimental treatment arm compared to the new experimental treatment and control arms. It may be undesirable to reduce the allocation ratio to a treatment arm during the trial.

In this section the aim is to determine a better allocation ratio than 1:1:1, whilst allowing the original treatment arm to recruit at the same rate as the new treatment arm. This design will randomise more controls in stage two and fewer in stage three, which will also increase the overlapping controls and therefore the correlation. The optimal allocation is determined by maximising the overall power under the constraint that $\lambda_{12} = \lambda_{22} = 1$. An alternative approach would be to fix the minimum allocation to the original treatment arm compared to the new treatment arm for the remainder of the trial and again determine the optimal allocation by maximising the overall power.

The test statistics here are the same as those given in Equations 4.8 and 4.9. However, there are now three stages to the design, $k_{(1)} = \{1, 2\}$ and $k_{(2)} = \{2, 3\}$. When all treatments finish recruiting simultaneously, there are only two stages.

1. The total sample size is fixed at N, the number of patients required to control the FWER at level α using equal randomisation when adding a treatment arm during the trial, described in Section 4.2.4.

   It is then determined how best to allocate the remaining patients, $R = N-(n_{01}-n_{11})$, once the new treatment arm has been added to optimise the overall power. The new treatment arm is added at a fixed time point after $n_{01} = n_{11}$ patients have been randomised to the original treatment and control.

2. Determine the optimal allocation ratios that maximise the overall power. The estimate of the correlation and critical value from step one are used.

Fixing $\lambda_{12} = \lambda_{22} = \lambda_{23} = 1$, the remaining patients to be randomised, R, can be written as,

$$R = \underbrace{\lambda_{02} n_{22}}_{n_{02}} + \underbrace{n_{22}}_{n_{12}} + n_{22} + n_{23} + \underbrace{\lambda_{03} n_{23}}_{n_{03}}.$$

Optimisation is required to determine the values of $\lambda_{02}, \lambda_{03}$ and n_{22} or n_{23} that maximise the overall power. Only one of n_{22} or n_{23} are required since N, n_{01} and n_{11} are all fixed values. However, numerical optimisation of these parameters will always allocate zero patients to control in the third stage and randomise all to control in stage two. This is because all controls will then be used for the treatment 1 to control comparison. However, some randomised controls are required in stage three. Therefore, λ_{03} is also fixed, adding a constraint on the minimum allocation to control in stage three of the trial. Numerical optimisation is then used to find the values of λ_{02} and n_{23} that maximise the overall power. The number of controls available for the first treatment comparison will vary depending on when the original treatment arm finishes recruiting.

The overall power is defined by,

$$1 - \beta = \int_{c1^*}^{\infty} \int_{c2^*}^{\infty} \pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma}) dz_2 dz_1, \tag{4.13}$$

where $\pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma})$ is the bivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix,

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_{Z_1 Z_2} \\ \rho_{Z_2 Z_1} & 1 \end{pmatrix},$$

where $\rho_{Z_1 Z_2}$ is the correlation calculated in step one, defined in Equation 4.5.

$c1^*$ and $c2^*$ in Equation 4.13 are defined by,

$$c1^* = \frac{\left(c\sqrt{\dfrac{2\sigma^2}{n}}\right) - (\mu_1 - \mu_0)}{\sqrt{\dfrac{1}{n_{1\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(1)}} n_{0k}}}}, \tag{4.14}$$

and,

$$c2^{*} = \frac{\left(c\sqrt{\dfrac{2\sigma^{2}}{n}}\right) - (\mu_2 - \mu_0)}{\sqrt{\dfrac{1}{n_{2\cdot}} + \dfrac{1}{\sum\limits_{k \in k_{(2)}} n_{0k}}}}, \tag{4.15}$$

where $k = 1, \ldots, 3$, c is the Dunnett critical value and n the sample size per comparison group estimated in step 1. $(\mu_j - \mu_0)$ is the treatment effect under the alternative hypothesis and,

$$n_{22} = \frac{R - (\lambda_{03}+1)n_{23}}{\lambda_{02}+2}, \; n_{1\cdot} = n_{11} + n_{12}, \; \sum_{k \in k_{(1)}} n_{0k} = n_{01} + n_{22}\lambda_{02}, \; n_{2\cdot} = n_{22} + n_{23},$$
$$\sum_{k \in k_{(2)}} n_{0k} = n_{22}\lambda_{02} + n_{23}\lambda_{03}.$$

where λ_{03} is fixed and n_{23} and λ_{02} are found via optimisation. This change in allocation ratio will alter the correlation. As in Section 4.2.4, iteration is required to control the FWER at the desired level.

3. Re-estimate the correlation and critical value that control the FWER based on the sample sizes per comparison group, calculated in step two. The sample sizes per comparison group are now unequal and Equation 4.4 is required to estimate the correlation.

4. Repeat step 2 (replacing the Dunnett critical value and variance in the numerator of Equations 4.14 and 4.15 with the critical value and variance calculated using the sample sizes that maximise the overall power. Replace the correlation in Equation 4.13 with the correlation calculated using the optimal sample sizes) and repeat step 3 until the optimal allocation is determined that maximises the overall power and where the FWER is also controlled at the desired level.

4.4 Example – comparing methodology

This section returns to the main example, described in Section 4.1.1. Outcome data are assumed to be normally distributed. The treatment effect to detect between any experimental treatment and control is assumed to be three. The standard deviation is assumed to be known and equal to 10 in all treatment groups. It is assumed that the decision to add a treatment arm to the trial is driven by external reasons, independent of what is happening in the current trial, the aim of the trial is to identify whether any or all treatments are better than control and finally, it is assumed that only controls randomised concurrently are used in the analysis for each treatment.

4.4.1 Two experimental treatments – independent trials

In this section, trials comparing multiple experimental treatments to control in a single trial are compared to running separate trials, each with an independent control arm. For separate, independent trials, the standard approach is to control the marginal type I error rate of each study. In a single study with multiple treatment arms, the aim is to control the FWER.

For an individual trial, comparing one experimental treatment to control, 234 patients are required per treatment group for a one-sided test at level $\alpha = 0.025$ and 90% power to detect a treatment difference of three, with a known and equal standard deviation of 10 in both treatment groups. Therefore, for two independent trials, the total sample size required is $N = 2Jn = 936$ patients, where $n = 234$, the number of patients in each treatment arm. The critical value for each test is approximately 1.96. For two independent trials, the FWER is $1 - (1 - 0.025)^2 = 0.0494$. For independent trials, the number of type I errors is a binomial random variable with J trials and probability of success α [33].

4.4.2 Three arm trial design – no multiplicity correction

To assess two experimental treatments in a single trial, there are two experimental arms and a single control arm. Making no adjustment for multiplicity, the total number of patients required is $N = (J + 1)n = 702$.

The probability of rejecting at least one null hypothesis here is given by,

$$= 1 - \left(\int_{-\infty}^{\Phi^{-1}(1-0.025)} \int_{-\infty}^{\Phi^{-1}(1-0.025)} \pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma}) dz_1 dz_2 \right) = 0.0454.$$

Assuming Z_1 and Z_2 are both $\sim N(0, 1)$, equal sample sizes per group and $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

The FWER for this design is 0.0454, slightly less than the FWER when running two separate trials. However, the total sample size for a three arm trial with no multiplicity correction (702) is significantly smaller than for the separate trials design (936).

4.4.3 Three arm trial design – Dunnett multiplicity correction

This section explores a standard multi-arm trial comparing two experimental treatments to a control treatment. All treatment arms recruit from the beginning of the study and

Dunnett's method is used to adjust for multiple testing.

To control the FWER at 2.5% and the marginal power of each comparison at 90%, 272 patients are required per treatment group for a one-sided test to detect a treatment difference of three between each experimental treatment and control when the standard deviation is known to be 10 in all treatment groups (3 groups = 816 total).

The critical value for each individual test is now increased to 2.21 to control the FWER at 2.5%. Since,

$$1 - \left(\int_{-\infty}^{2.21} \int_{-\infty}^{2.21} \pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma}) dz_1 dz_2 \right) = 0.025.$$

Whereas, in the previous two designs in Sections 4.4.1 and 4.4.2, a standard normal 2.5% critical value of $\Phi^{-1}(1 - 0.025) \approx 1.96$ was used.

Here, the FWER is controlled at the desired level of 2.5% and the sample size is reduced from 936 in separate trials, to 816 patients in a single trial.

Note that for low powered studies ($<$ 50% power for individual studies) it is often better to run separate trials than a single trial with a Dunnett correction, in terms of the number of patients required. However, for more common situations this is not usually the case [33].

4.4.4 Adding an arm during the trial – no multiplicity correction

A trial starts with two treatment arms initially and part way through the trial a new treatment arm is added. It is assumed that no correction is made to the critical value for multiple comparisons (the critical value chosen at the design stage is used for both treatment comparisons at the end of the study). There are now two comparisons to be made at the end of the study. We explore the effects of adding a treatment arm on the FWER.

Continuing with the main example described in Section 4.1.1. A new treatment arm is added after 100 patients have been randomised to each treatment group, 234 patients are allocated to the new treatment and a further 100 patients are randomised to control. The number of concurrent controls randomised for each treatment arm are then equal. The number of controls used in both treatment comparisons is 134. The total number of patients per treatment group for each of the treatment comparisons is 234 (134 of the controls are used in both treatment comparisons, overlapping controls). The correlation is given by,

$$\rho_{Z_1 Z_2} = \frac{134}{2 \times 234} = 0.286.$$

The FWER is then calculated as,

$$1 - \left(\int_{-\infty}^{\Phi^{-1}(1-0.025)} \int_{-\infty}^{\Phi^{-1}(1-0.025)} \pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma}) dz_1 dz_2 \right) = 0.0477,$$

where $\pi_Z((z_1, z_2)', \mathbf{0}, \mathbf{\Sigma})$ is the probability density function of a MVN distribution with mean $\mathbf{0}$, and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.286 \\ 0.286 & 1 \end{pmatrix}$.

An extra 234 patients are randomised to treatment 2 and an extra 100 concurrent controls for the treatment 2 comparison when adding a treatment arm. Here, the sample size is not increased to account for the multiplicity and the FWER is 0.0477. The total number of patients required for this design is 802 ($3 \times 234 + 100$ extra controls).

The FWER is less for this design (0.0477) than the FWER obtained from two separate trials (0.0494) and requires 134 fewer controls than two separate trials. However, some may argue that for the single trial, the FWER should be controlled at the 2.5% α level of a single independent trial. This design is considered in the next section.

4.4.5 Adding an arm during the trial – Dunnett multiplicity correction

Following the steps described in Section 4.2.4.

1. Estimating the correlation using the original sample size per group of 234 gives a correlation of 0.286.

2. Using this correlation, the critical value that controls the FWER at 0.025 is 2.2295, giving a sample size per group of 273.941.

3. Re-estimating the correlation and sample size until they are constant results in a final correlation of 0.317, critical value for each test of 2.2277 and a final sample size per group of 273.659.

Since we require integer sample sizes, the change in correlation has no impact on the required sample size for this design. The increase in the number patients required comes from controlling the FWER because the treatment comparisons are within the same trial rather than separate trials. The total sample size required per group to control the FWER at 2.5% based on this correlation estimate would be 274 (an additional 40 patients per group). Three groups of 274 patients plus 100 extra non-overlapping controls, gives a total sample size of N=922, just below the 936 patients required to run two separate trials. However, here we are controlling the FWER at 2.5%, whereas for two separate studies, only the marginal type I error rate for each comparison is controlled at 2.5%. This design is illustrated in Figure 4.4.

Figure 4.4: Example of adding a single experimental treatment arm to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when the original treatment (treatment 1) finishes recruitment and the third dashed vertical line represents when the control and treatment 2 finish recruiting patients. The horizontal dashed lines represent the additional patients required per treatment group above the original sample size estimate to control the FWER while maintaining randomisation of 1:1:1 to all treatment arms.

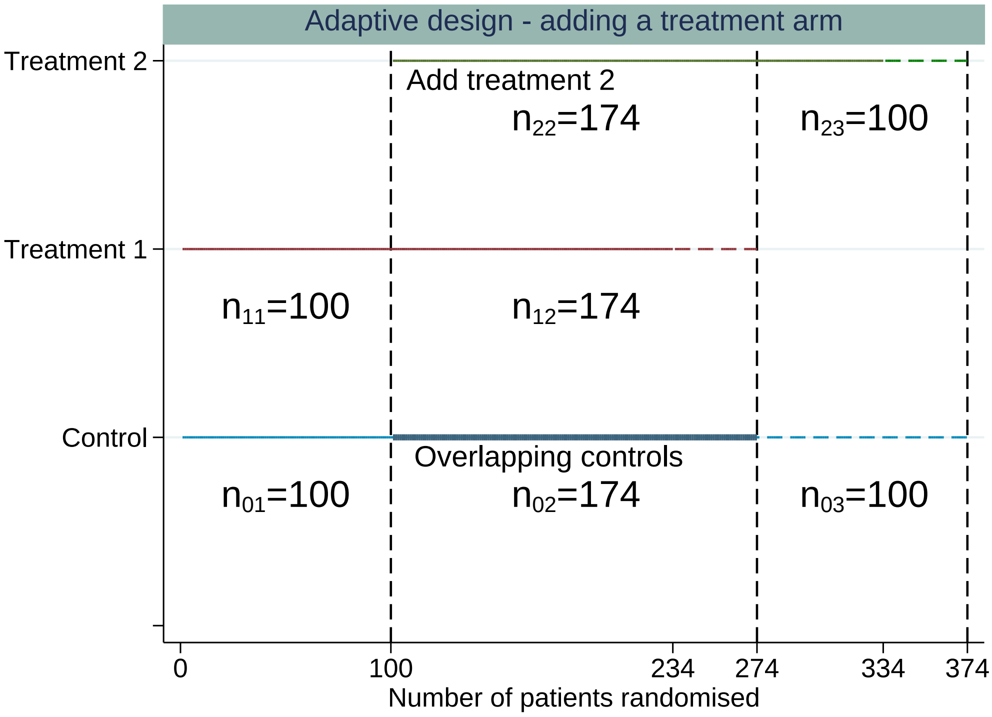


Figure 4.4 illustrates the trial design where a treatment arm is added and the sample size of all treatments arms are increased to control the FWER using the Dunnett correction. Treatment 2 is added after 100 patients have been randomised to control and treatment 1 (the first dashed vertical line). The sample size for each group is then increased by 40 patients from 234 to 274, shown in Figure 4.4 by the horizontal dashed lines. The second dashed vertical line shows when the original experimental treatment arm finishes recruitment. The number of overlapping controls is now 174 and the correlation between test statistics is 0.32.

As illustrated in this example, small changes in the correlation between test statistics results in small changes to the critical value and therefore only small changes to the total number of patients required. Figure 4.5 shows the critical value obtained for different values of the correlation and FWER. For larger FWER the change in correlation has a larger impact on the change in critical values. Therefore, for phase II trials, the change in correlation from adding a treatment arm may increase the number of patients required.

Figure 4.5: Effect of correlation between test statistics on the critical value required to control the FWER at different levels for a standard 3 arm trial.

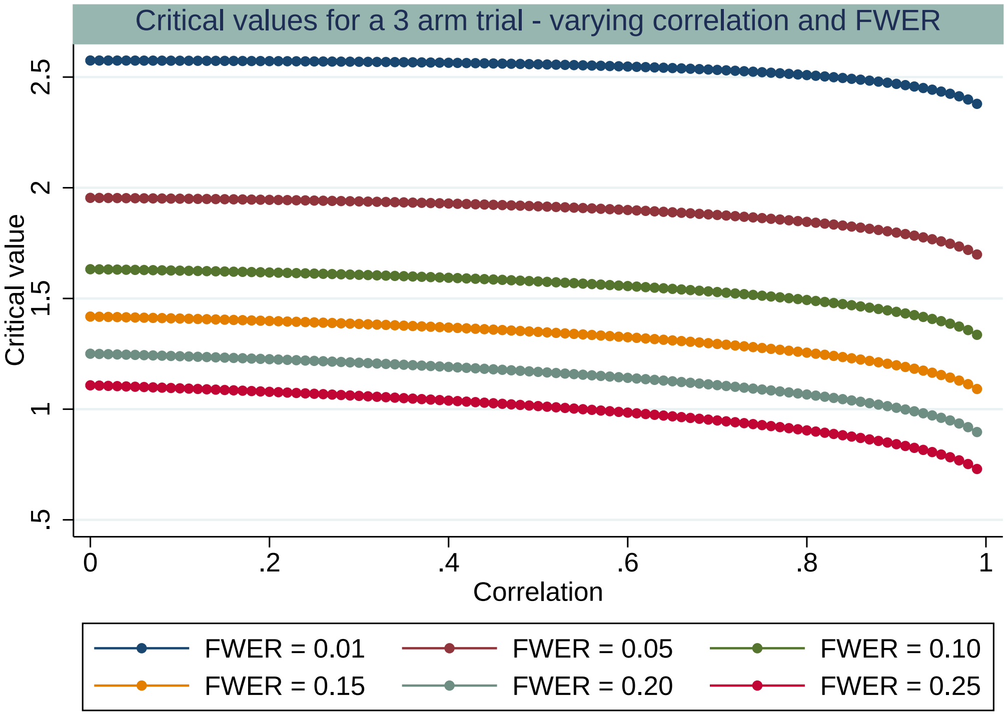


Table 4.1 compares the FWER and the number of patients required to have 90% marginal power for each treatment to control comparison for the designs discussed in Sections 4.4.1 to 4.4.5. From Table 4.1, for this example, in terms of the number of patients required, a single trial design adding a treatment arm requires fewer patients than a separate trial design, even when controlling the FWER at the level of the type I error rate for a two-arm trial.

Table 4.1: FWER and sample size comparisons for 90% marginal power

Design	FWER	Total sample size
Two separate trials	0.0494	936
Single trial – no multiplicity adjustment	0.0454	702
Single trial – Dunnett adjustment for multiplicity	0.0250	816
Adding an arm – no multiplicity adjustment	0.0477	802
Adding an arm – Dunnett adjustment for multiplicity	0.0250	922

4.4.6 Comparing a single study to separate studies in terms of the number of patients required

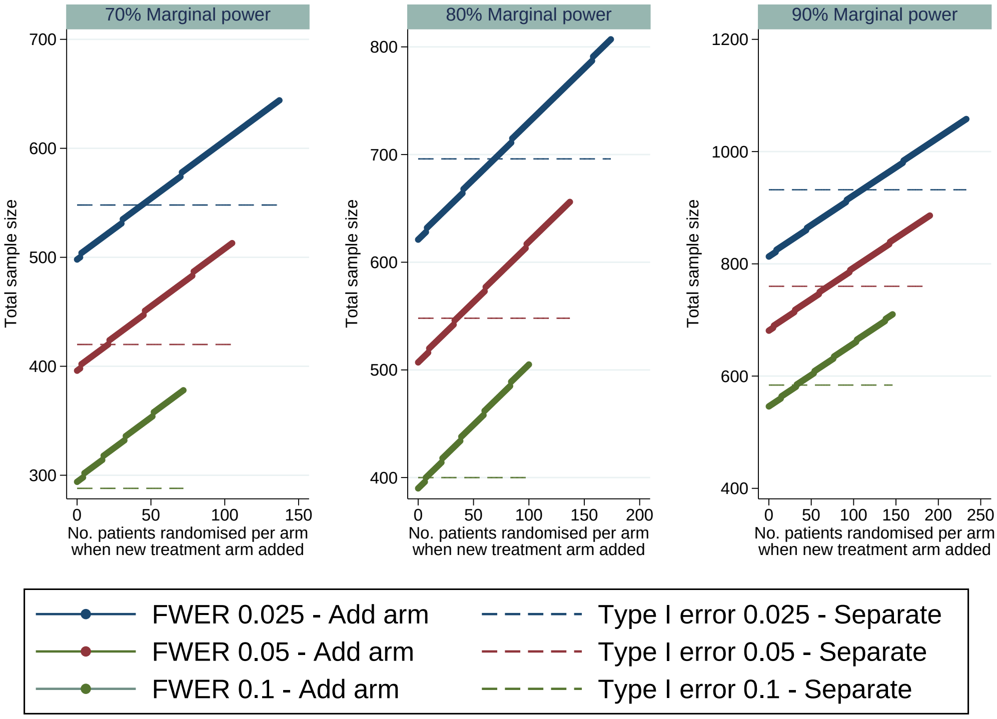
As mentioned above, the main contributing factors to the total sample size required in a single trial when adding a treatment arm are, the control of the FWER and when the new treatment arm is added and the extra concurrently randomised controls required (for the treatment 2 comparison when using 1:1:1 randomisation). Therefore, if control of the FWER is required in a single study and not in separate trials, it is of interest to determine the time point at which it is better to start a separate trial rather than add an arm to an ongoing trial with respect to the total sample size required. This is comparing a single trial controlling the FWER at 2.5% to separate trials, each with a 2.5% type I error rate.

The following example compares a single trial to separate trials when a treatment arm is added at different time points during the trial, for varying power (70%, 80% and 90%) and FWER (2.5%, 5% and 10%). Figure 4.6 shows that for 90% power and 2.5% error rate, once 111 patients have been randomised to control, it would be better to run a new trial than add a new treatment to the current trial, in terms of the number of patients required, if in a single trial control of the FWER is required at the 2.5% level. The statistical advantage of adding a treatment arm and running a single trial decreases as the power decreases and FWER increases. Figure 4.6 shows that for 10% error rate and 70% power it is better to run separate studies than a multi arm trial, even if all treatment arms are recruiting from the beginning of the study. It is assumed here that randomisation is 1:1:1 and the treatment effect to be detected is the same for all treatments.

4.4.7 Adding multiple treatment arms

So far, adding only one new treatment arm has been considered. We now consider adding two treatment arms during the trial, correcting for multiple comparisons using the Dunnett procedure. There are now three comparisons of interest, comparing three experimental treatments to control. The correlation between test statistics is given in Equation 4.4. The three test statistics follow a J-variate normal distribution.

Figure 4.6: Comparing the total sample size required for a single trial versus separate trials when adding a new treatment arm at different time-points for varying error rates and marginal power.



It is assumed that the second experimental treatment arm is added before the original experimental treatment arm finishes recruitment and therefore there is correlation between all test statistics. However, the correlations between all test statistics can never be equal.

Continuing the main example described in Section 4.1.1. Two new treatment arms are added, the first after 200 patients have been randomised to either the original experimental treatment or control with 1:1 allocation.

The variance-covariance matrix using the original sample size calculation of 234 patients per group is given by,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.286 & 0.073 \\ 0.286 & 1 & 0.286 \\ 0.073 & 0.286 & 1 \end{pmatrix}$$

Based on the above correlation matrix, the sample size required per group to control the FWER at 2.5% and the correlation are re-estimated until they converge. The final sample size required per group is then 298 patients and the correlation matrix is given by,

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.332 & 0.164 \\ 0.332 & 1 & 0.332 \\ 0.164 & 0.332 & 1 \end{pmatrix}$$

Figure 4.7 illustrates a design where two treatment arms are added. This design has balanced overlap in the number of controls used between treatment 1 and treatment 2 and between treatment 2 and treatment 3. The sample size of all groups is increased to control the FWER using the Dunnett correction. Treatment 2 is added after 200 patients have been randomised to either control or treatment 1. Treatment 3 is then added after a further 100 patients have been randomised to each of the groups. The sample size for each group is increased by 64 patients from 234 to 298, shown in the Figure by the horizontal dashed lines. The dashed vertical lines show where new treatment arms are added and treatments finish recruiting. The number of overlapping controls now differs between test statistics. Between Z_1 and Z_2, and Z_3 and Z_2 the number of overlapping controls is 198 (as shown by the navy and pink lines) and between Z_3, Z_1 is 98 controls (as shown by the yellow line).

The total sample size required here is 1392 compared to the total sample size of 1404 when running three separate studies. However, this is under control of the FWER at the 2.5% level, the same as the marginal type I error rate for a single study comparing one experimental treatment to control. If this strict control of the FWER is required in a single study compared to running separate trials. The benefits of adding treatment arms to a trial will decrease as the number of treatment arms added increases, if multiple test correction is required for the multi-arm trial but not the separate trials.

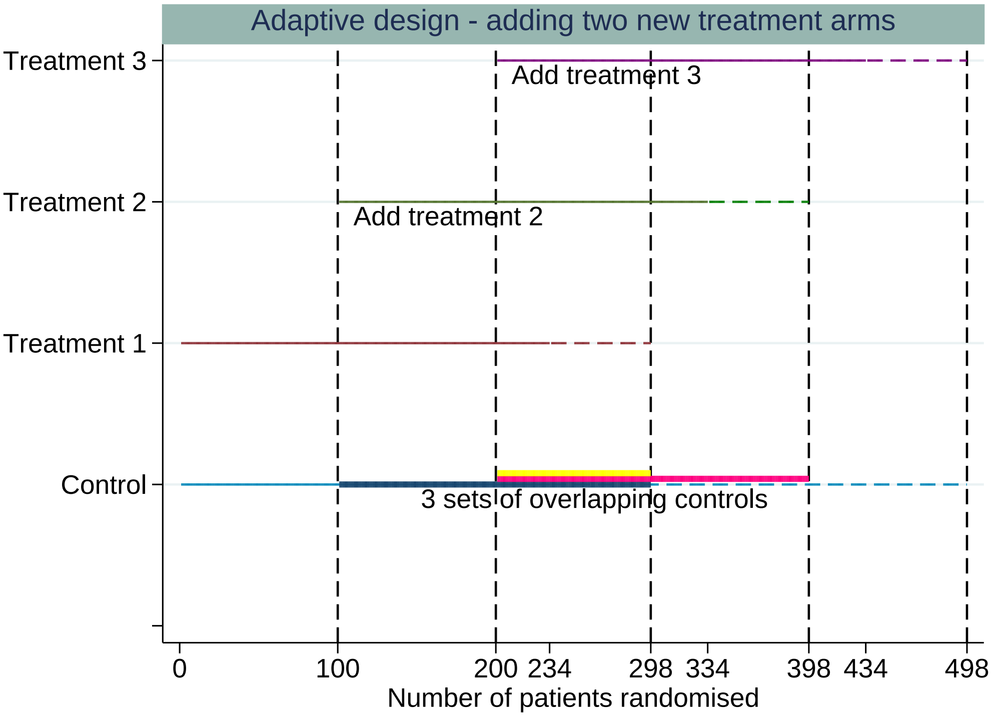
So far, only equal allocation to all treatment arms has been considered and power in terms of controlling the marginal power for each treatment to control comparison. We now look at optimal allocation to maximise the overall power of the study.

4.4.8 Optimal allocation – standard Dunnett design

Returning to the main example, assuming all treatment arms recruit from the beginning of the study and Dunnett's method is used to adjust for multiple testing.

Table 4.2 compares the optimal allocation ratio obtained from maximising the overall power (max ratio) and the approximation from minimising the total variance ($n\sqrt{J}$), given in Section 4.3.1 for a standard multi-arm trial with different numbers of experimental treatment arms. The optimal allocation is calculated based on the fixed total sample

Figure 4.7: Example of adding two experimental treatment arms (treatment 2 and treatment 3) to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the first new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when the second new treatment arm (treatment 3) is added to the trial. The remaining vertical lines are when treatments finish recruiting. The horizontal dashed lines represent the additional patients required per treatment group above the original sample size estimate to control the FWER while maintaining randomisation of 1:1:1 to all treatment arms.



size calculated assuming 1:1 allocation to all treatment arms (N). Table 4.2 then gives the overall power based on 1:1 allocation ($1-\beta$ (1:1)), the overall and marginal power using the allocation ratio $n\sqrt{J}$ ($1-\beta$ ($n\sqrt{J}$)) and the overall and marginal power using the allocation ratio that maximises the overall power ($1-\beta$ (max)).

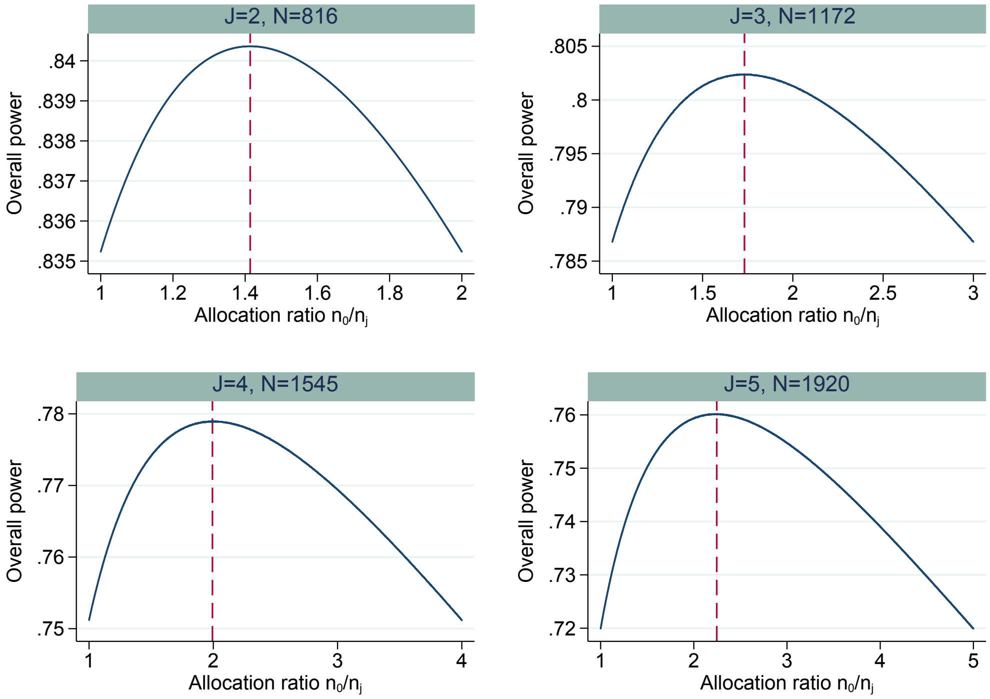
In this example, adapting the allocation ratio increases both the marginal and overall power of the study. The approximation of the optimal allocation, randomising $n\sqrt{J}$ patients to control gives similar results to maximising the overall power of the study as shown in Table 4.2. Figure 4.8 shows the overall power for different allocation ratios and different numbers of experimental arms for a standard multi-arm trial design.

Table 4.2: Optimal allocation ratio for the standard Dunnett design when using the optimal allocation proposed by Dunnett $n\sqrt{J}$ or maximising the overall power. The overall power for the allocation ratios are also given.

J	max ratio	$n\sqrt{J}$	N (1:1)	$1-\beta$ (1:1)	$1-\beta$ ($n\sqrt{J}$)	$1-\beta$ (max)*
	($n_0 : n$)	($n_0 : n$)			(marginal)	(marginal)
2	1.414 (338:240)	1.414 (340:240)	816	0.834	0.840 (0.910)	0.840 (0.909)
3	1.732 (429:248)	1.732 (430:248)	1172	0.782	0.802 (0.918)	0.802 (0.918)
4	1.995 (515:258)	2 (516:258)	1545	0.743	0.779 (0.928)	0.780 (0.928)
5	2.252 (597:265)	2.236 (595:266)	1920	0.709	0.760 (0.936)	0.760 (0.935)

* Calculated based on controlling the FWER at the 2.5% level. N is fixed based on 1:1 randomisation for 90% marginal power (treatment arm sample sizes may be increased due to rounding based on the optimal allocation.)

Figure 4.8: Overall power by allocation ratio for a standard multi-arm trial for trials with a different number of treatment arms and each with fixed sample size. The dashed vertical red lines represent the optimal allocation ratio that maximises the overall power.



4.4.9 Optimal allocation – adding a treatment arm – treatments finish recruiting simultaneously

Assuming the trial starts with two treatment arms, an experimental and a control arm, randomising 1:1. An additional treatment arm is added after 200 patients have been randomised to the control or the original treatment group. Based on this design, following the steps in Section 4.2.4, 922 patients are required.

Up to the time point that the new treatment arm is added, the optimal allocation is 1:1, treatment to control. Since randomisation continues to all arms until the end of the study, controls randomised before the new treatment arm is added will not be used in the new treatment to control comparison, but all controls will be used for the original treatment to control comparison.

The total sample size is fixed at 922 patients (274 patients per group +100 extra controls). This is the sample size required to obtain 90% marginal power for both treatment comparisons and control the FWER at 2.5% when using 1:1:1 allocation. This design has an overall power of 0.822.

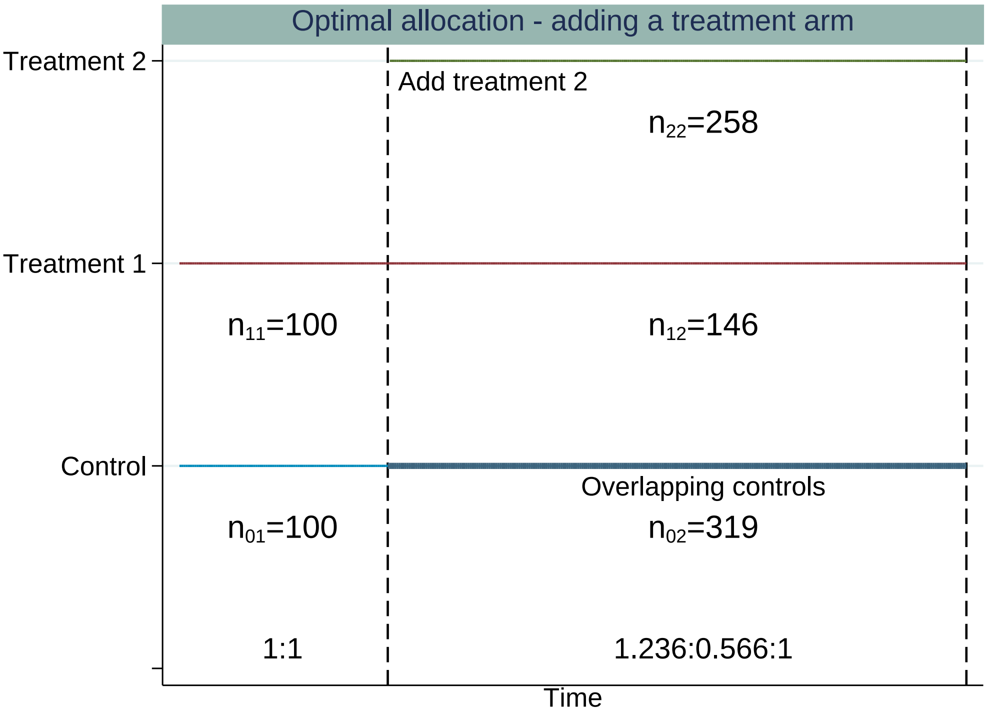
For the optimisation we initially use a correlation of 0.317 which gives a critical value using the Dunnett correction of 2.229. Taking into account the 200 patients that have already been randomised, the optimal allocation ratio is calculated from when the new treatment arm is added. The remaining number of patients to be randomised when the new treatment arm is added are 722. Numerically optimising the overall power and iterating gives a final allocation ratio for the second stage of 1.236:0.566:1, control:original treatment:new treatment. This gives the following sample sizes in each of the treatment groups: 258 for the new treatment; 246 for the original treatment and 419 controls (319 for the treatment 2 to control comparison). The overall power here is 86.24% and the marginal power is 91.23% for the new treatment to control comparison and 93.43% for the original treatment to control comparison. This design controls the FWER at the 2.5%. The Dunnett critical value used in the final analysis is 2.225. Figure 4.9 illustrates this design.

Figure 4.10 illustrates the overall power obtained for different allocation ratios to control and the original treatment arm with respect to the new treatment arm. This is the allocation used once the new treatment arm is added, prior to this, allocation is 1:1 original treatment:control.

4.4.10 Optimal allocation – adding a treatment arm – treatments finish recruiting at different times

Reducing allocation to the original experimental treatment arm when the new treatment arm is added delays learning about the effect of the original treatment arm and may affect patient accrual. Patients may be more willing to participate in the trial if they have equal chance of receiving either of the experimental treatments. This section assumes the trial starts with two treatment arms randomising 1:1. An additional arm is added after 200

Figure 4.9: Example of adding a single experimental treatment arm to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when all treatments finish recruiting. The allocation ratio is adapted when treatment 2 is added and all treatments finish recruiting simultaneously. The allocation ratios and sample sizes in each stage for each treatment are displayed.



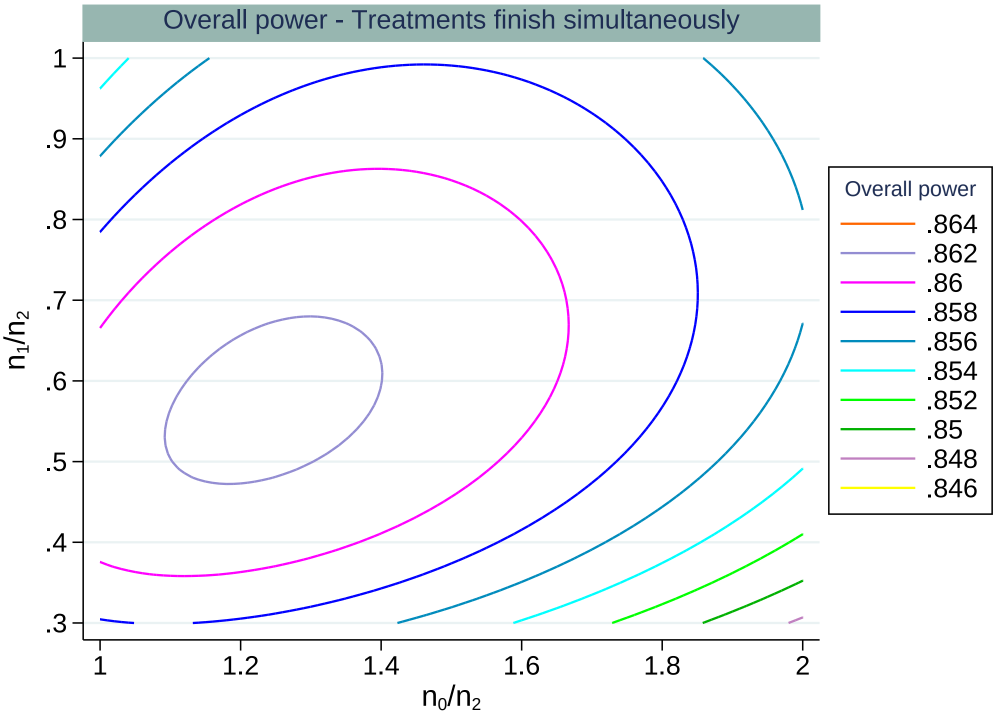
patients have been randomised to either the control or the original treatment arm. Here, once the new treatment arm is added, randomisation continues at the same rate to both experimental treatment arms.

The total sample size is fixed at 922 patients (274 patients per group +100 extra controls). This is the sample size required to obtain 90% marginal power for both treatment comparisons and control the FWER at 2.5% when using 1:1:1 allocation.

For the optimisation, the initial correlation estimate is 0.317, which gives a critical value using the Dunnett correction of 2.229. Taking into account the 200 patients that have already been randomised, the optimal allocation ratio is calculated from when the new arm is added. The remaining number of patients to be randomised when the new treatment arm is added is 722. The optimal allocation ratio to control in stage two and the number of patients to be randomised to treatment two in stage three are then found that maximise the overall power.

After iteration, the overall power is maximised when the allocation ratio in the sec-

Figure 4.10: Overall power for different sample size ratios of control to treatment 2 and treatment 1 to treatment 2 in stage 2 of the trial when all treatments finish recruiting simultaneously and the total sample size is fixed at 922 patients. Allocation is 1:1 treatment 1 to control prior to the addition of treatment 2 to the trial.



ond stage is 1.758:1:1, control:treatment 1:treatment 2 and 71 patients are randomised to treatment 2 with 1:0.5 allocation, treatment 2 to control in stage three. This gives the following sample sizes in each stage of the trial:

Stage 1 – 100:100, control:treatment 1

Stage 2 – 289:164:164, control:treatment 1:treatment 2

Stage 3 – 36:71, control:treatment 2

With 389 controls for the treatment 1 comparison and 325 controls for the treatment 2 comparison. The overall power is 85.20%. The marginal power is 89.83% for the new treatment to control comparison and 93.74% for the original treatment to control comparison. This design is illustrated in Figure 4.11. For a larger number of treatment arms where changes in the correlation have a larger impact on the critical values obtained, a larger number of iterations may be required, only a few iterations were required for the examples considered here.

All of the optimal allocation methods considered achieve higher overall power compared to an equal allocation (overall power for equal allocation is 82.2%) design for the

Figure 4.11: Example of adding a single experimental treatment arm to a two-arm trial comparing treatment 1 to control. The first dashed vertical line represents when the new treatment arm (treatment 2) is added to the trial. The second dashed vertical line represents when treatment 1 finishes recruiting and the third dashed vertical line represents when treatment 2 and control finish recruiting. The allocation ratios are adapted when treatment 2 is added to the trial and again when treatment 1 finishes recruiting. The allocation ratios and sample sizes in each stage for each treatment are displayed.

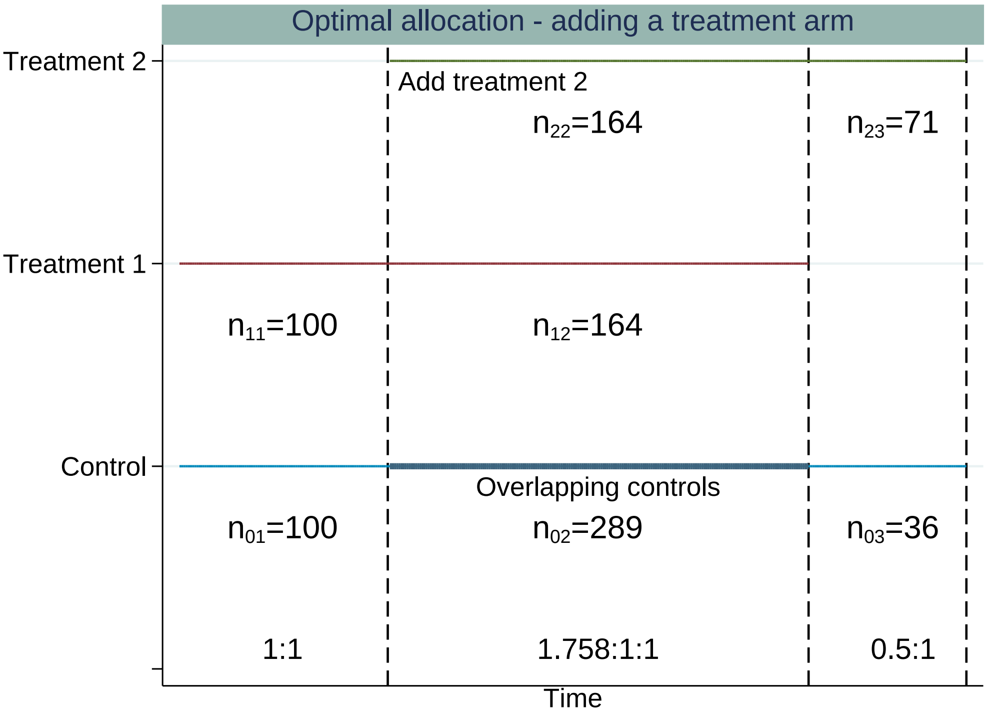


Table 4.3: Allocation ratios, overall power and sample size comparisons for the adding a treatment arm design using optimal allocation when treatments finish recruitment at different times. The total sample size of the trial is based on an adding a treatment arm design with 1:1:1 allocation design with marginal power of 90% and FWER of 2.5%.

Stage 3 ratio	Stage 2 ratio	n_{23}	Correlation	Overall power
1:0.2	2.015:1:1	100	0.170	0.852
1:0.3	1.932:1:1	93	0.176	0.850
1:0.4	1.856:1:1	85	0.181	0.848
1:0.5	1.791:1:1	78	0.186	0.847
1:0.6	1.730:1:1	71	0.191	0.845
1:0.7	1.678:1:1	64	0.197	0.844
1:0.8	1.632:1:1	58	0.202	0.843

same sample size. The power calculations are based on the unrounded optimal sample size estimates. The sample sizes are then rounded for each stage which may affect the operating characteristics slightly. An alternative approach, not considered here, would be to try integer sample sizes around the non-integer sample size found to determine which gives the highest overall power for the design. Table 4.3 compares the overall power for different fixed stage 3 allocation ratios and Figure 4.11 illustrates the design for the stage 3 fixed allocation ratio of 1:0.5, treatment to control. The marginal power for the treatment 2 to control comparison can be less than the marginal power obtained when using 1:1:1 allocation under this design where the allocation ratios are adapted and the original experimental treatment can finish recruitment before the new experimental treatment arm finishes recruitment, since the optimal allocation is determined that maximises the overall power. The gain in power from the treatment 1 to control comparison can outweigh that of the treatment 2 to control comparison. Furthermore, placing a higher constraint on the number of controls that are to be randomised in stage 3 of the trial, results in less patients being randomised to both treatment 2 and control in stage three and a larger number of patients being randomised in stage 2.

4.5 Bayesian design – adding a treatment arm during the trial with the Dunnett correction

Adaptive trial designs and the idea of utilising all relevant information considered in earlier chapters fit more comfortably into a Bayesian framework. It is therefore of interest to consider the multiple testing approaches and methods for adding a treatment arm into an ongoing trial in a Bayesian framework.

A Bayesian approach comparable to the frequentist approach proposed by Dunnett [39] for comparing multiple experimental treatments with a common control has recently been proposed in the literature [40]. This approach is described in more detail in Chapter 1. It is shown in [40] that the sample sizes required under a Bayesian design are analogous to a frequentist multi-arm trial design where the total sample size per treatment group is formed from the prior effective sample size plus the number of patients in the current trial for that treatment group.

Assuming that the sample sizes (prior + current sample size) in each experimental treatment arm are equal, in a Bayesian framework inference is made on the posterior distribution of the treatment difference. For a standard three arm trial design with all treatments recruiting from the start of the trial and all controls overlapping, the covariance of the treatment difference is given by,

$$Cov(\mu_1^{(1)} - \mu_0^{(1)}, \mu_2^{(1)} - \mu_0^{(1)}) = Cov(\mu_0^{(1)}, \mu_0^{(1)}) = \text{Var}(\mu_0^{(1)}) = ((n_0^{(0)} + n_0)v)^{-1},$$

where $\mu_j^1 = (\mu_j^0 n_j^0 + n_j \bar{x}_j)/(n_j^0 + n_j)$ is the posterior mean for treatment j. $n_j^{(0)}$ denotes the prior effective sample size for treatment j and n_j denotes the sample size in the current trial for treatment j. $v = 1/\sigma^2$ denotes the precision. This is analogous to the covariance of the treatment differences from the frequentist design.

Assuming the same design for adding a treatment arm as in Section 4.2.4. The main difference in the Dunnett adjustment for a standard multi-arm trial and when adding a treatment arm to the trial is the covariance between the mean differences (or the correlation as considered in previous sections). If the sample sizes in all treatment groups are assumed to be equal and allocation is 1:1 or 1:1:1 to all treatments throughout the study, it can be shown that in the Bayesian framework, when adding a treatment arm during the trial, the covariance between treatment mean differences is given by (using the notation in Figure 4.3 for the current trial sample sizes),

$$Cov(\mu_1^{(1)} - \mu_0^{(1)}, \mu_2^{(1)} - \mu_0^{(1)}) = \frac{v^{-1}(n_0^{(0)} + n_{02})}{(n_0^{(0)} + n_{01} + n_{02})(n_0^{(0)} + n_{02} + n_{03})}$$

where $n_0^{(0)} + n_{01} + n_{02}$ is the number of controls used for the first treatment comparison (prior+current data) and $n_0^{(0)} + n_{02} + n_{03}$ is the number of controls used for the second treatment comparison. $n_0^{(0)} + n_{02}$ represents the number of controls used in both treatment comparisons. This is consistent with the frequentist approach, given in Equation 4.5, with the control sample sizes replaced by the total information (prior + trial sample sizes) rather than just the current trial sample size. For unequal sample sizes in each treatment group, the covariance is given by,

$$Cov(\mu_1^{(1)} - \mu_0^{(1)}, \mu_2^{(1)} - \mu_0^{(1)}) = \frac{v^{-1}\left(n_0^{(0)} + \sum_{k \in k_{(1)} \cap k_{(2)}} n_{0k}\right)}{\left(n_0^{(0)} + \sum_{k \in k_{(1)}} n_{0k}\right)\left(n_0^{(0)} + \sum_{k \in k_{(2)}} n_{0k}\right)}$$

All results for adding a treatment arm considered earlier in this chapter can be formulated in a Bayesian framework. The only difference is that the current trial sample sizes are replaced by the total information (prior + trial sample sizes). For the sample size calculation, only the prior sample sizes are used, however this approach allows the prior estimates to be used in the final analysis. In this design the standard deviation is assumed known, but the final analysis will use the sample variance estimates. Therefore, the properties of the design may not hold if the assumptions about the standard deviation were incorrect.

We did not explore this design in much detail since it is a bit strange to only use concurrent controls in a Bayesian framework because of the likelihood principle [17]. However, it is interesting to note that the results are the same in a Bayesian framework as in the frequentist framework.

4.6 Discussion

In this chapter we have explored how to design a trial where a treatment arm is added part way through. It was assumed that the trial initially started as a standard two-arm randomised controlled trial comparing an experimental treatment to a standard of care with 1:1 allocation. The setting considered here is a confirmatory trial where it is recommended that for a single trial with two or more experimental treatments, control of the FWER is required [35]. In this chapter we have assumed that the new treatment arm is added based on external reasons and not based on outcome data from the current trial. Furthermore, it was assumed interest lies in the comparison of each experimental treatment to control, the experimental treatments are not directly compared. Whether control of the FWER is required under these assumptions has been widely discussed [29, 33, 34]. The main argument for not adjusting for multiplicity is that if the same two comparisons were made in separate trials, controlling the FWER would not be required. The use of the same control group in both treatment comparisons in a multi-arm trial is an argument for controlling the FWER. It has been considered that due to random chance the control group in a multi-arm trial could over or underestimate the true treatment effect. This control group is then used for all treatment to control comparisons [33]. An increase in sample size to achieve control of the FWER will reduce some of the sample variation in the controls. Since the regulatory advice is to control the FWER in this setting, this was the approach considered in this chapter.

To determine the optimal allocation ratio, it was assumed that the total sample size of the study was fixed. Based on this total sample size, the optimal allocation for the remaining patients to be randomised after the treatment arm was added was determined. The allocation ratio was determined that maximised the probability of detecting a treatment effect in both arms. This was chosen because we assumed that the expected treatment effect in all experimental arms was the same. In this case, the marginal power will either increase or be similar to the marginal power in the design assuming equal allocation to all treatment arms. The marginal power may be lower than desired if the overall power is used for optimisation of the allocation ratio and the treatment effects are assumed to differ in each of the experimental arms. Where the treatment effects are assumed to differ, optimisation can be based on maximising the probability of there being no type I errors as in the original Dunnett paper [39] or the probability that there is a treatment effect in at least one of the treatment arms. The methods described in Section 4.3.2 can be used in both of these situations.

In this chapter, two definitions of power in multi-arm trials have been considered, detecting a particular effective treatment compared to control, the marginal power and the overall power, the probability of rejecting all false null hypotheses, where both treatments have been effective. The definition of power will depend on the study objectives. Whether that is to determine all treatments better than control, any treatment better than control or the best treatment.

4.6.1 Concurrent controls

Current practice in clinical trials that have added a treatment arm is to use only concurrently randomised controls for each treatment to control comparison [32]. One reason for this is to preserve randomisation. Incorporating control data from the first stage in the second treatment control comparison is utilising non-randomised information into that comparison. A second reason is that the patient population may differ before and after the treatment arm has been added. Incorporating the first stage control data into the second treatment comparison could then bias the treatment effect estimate for treatment two. Depending on the possible types of change that occur when a new treatment arm is added, this may affect the original treatment to control comparison. However, the use of only concurrent controls guards against some possible biases or loss of power that could occur from adding a treatment arm to an ongoing study. This is discussed in Section 4.6.2.

Possible reasons for a change in the patient population when the new treatment arm is added are: a change in patient characteristics, patients may be more willing to be randomised with the possibility of receiving two experimental treatments; a change in baseline characteristics over time; and clinicians are more willing to randomise patients higher or lower risk patients into the trial because of their expectations of the new treatment.

4.6.2 Analysis

For the designs considered in this chapter it was assumed that the patient population is homogeneous across stages. The final analysis assumed a z-test for each treatment comparison, pooling data across stages but only using concurrent controls and assuming that the population variance is known. In reality, the population variance would not be known, however for large samples as are likely in a confirmatory trial, the z-test is adequate. However, if the approximation of the standard deviation was inaccurate this will affect the operating characteristics of the trial.

If it is thought that adding a treatment during the trial may alter the trial in some way, causing a stage effect, adjustment for stage or treatment by stage interactions may be required using a linear regression approach. This method is described by Elm et al. [30] and may reduce the power of the study. The effect of different stage effects was not explored here but was considered in [30]. This paper shows that a linear model adjusting for stage was a more powerful analysis approach when a stage effect was present compared to a pooled analysis. This paper considered a random stage effect and the allocation ratio was adjusted when the new treatment arm was added so all treatments finished recruiting at the same time.

A few considerations on stage effects are, if 1:1 allocation is maintained to all treatment arms throughout the study, for a stage effect that changes all treatment groups when the new treatment arm is added, a pooled analysis using concurrent controls will not bias the treatment effect estimates. If 1:1 allocation is used and there is a stage effect in the control group only, only the original treatment effect estimate will be affected. A stage effect in the control arm only could be caused by patient drift, as is seen in antibacterial trials. In this case a treatment by stage interaction would be required. If there are any stage effects and the allocation ratio is adapted, adjusting for stage and a stage by treatment interaction will be required to obtain an unbiased treatment effect estimate for the original treatment to control comparison.

4.6.3 Bayesian design

As described in Section 4.5, the theory of designing a trial where a treatment arm is added to an ongoing study and the design is adapted to control the operating characteristics in a frequentist framework hold in the Bayesian methodology. However, the Bayesian paradigm is to combine all sources of evidence and relevant information. In the frequentist design we only consider the use of concurrently randomised controls in each of the treatment comparisons, whereas a fully Bayesian design in this setting would consider all trial information and prior information before the trial started. In a fully Bayesian design the controls randomised in stage one would be used for the second stage treatment to control comparison. This is equivalent to pooling the data across stages of the trial. Throughout the thesis we have used the Bayesian framework to borrow information whilst considering the consequences on the operating characteristics. Since control of the frequentist operating characteristics is required for regulatory approval. The Bayesian design was considered in Section 4.5 with the use of only concurrent controls since there is an advantage of the Bayesian approach in terms of the ease of interpretation of the parameters values in a Bayesian setting.

4.6.4 Historical data and non-concurrent controls

When adding a treatment arm to an ongoing clinical trial there are two types of relevant non-randomised data. External historical data from a previously run trial (usually for the control treatment only) and the first stage control data before the new treatment arm is added. Both of which could be considered historical data. Following current practice, the first stage controls would not be used in the second treatment comparison.

Ideally, controls randomised before the new treatment arm is added would be used in the new treatment control comparison to gain power. If the original treatment arm finishes recruitment before the new treatment arm, there will also be non-concurrent controls for the original treatment.

Historical data methods compare the current and historical data. When there is disagreement, the historical data are down-weighted in favour of the results from the current trial. This approach is not applicable with non-randomised comparison data from within the same trial.

However, based on the assumption that it is pre-planned to add a treatment arm during the study. An interesting design that could be explored would be a two-stage design where external historical control data replace some current controls yet to be randomised up to the time point the new treatment arm or arms are added. The agreement between the first stage controls and historical data will determine how many current controls are randomised in stage 1. Using this approach more controls will be randomised in stage 2 of the trial and therefore more concurrent controls will be available for all treatments. The second stage control data can then be compared with the combined first stage current control data and the historical data. Where the data from the first and second stage are similar they can be pooled and when they differ, the first stage data can be discounted. The historical data used in the first stage could then be used in the final analysis if it is in agreement with the second stage data or discounted using historical data methods if not. If population drift is anticipated in the control arm, as is the case in antibiotic trials where patients become resistant to standard treatments, then this approach may be beneficial.

Chapter 5

Summary and future research

5.1 Summary of thesis

This thesis focuses on designs to improve efficiency of clinical trials over the standard confirmatory two-arm randomised controlled trial. The main aims of the proposed designs are to either increase the power of the current trial or reduce the number patients required in the current trial and therefore the duration of the trial. Reducing the number of patients required in a confirmatory trial allows treatments to proceed through the development process more quickly and therefore patients can benefit from treatments earlier than if a standard trial design were used. In disease areas where recruitment is challenging, these designs may allow trials to be run that were previously infeasible. We consider two approaches for improving the efficiency of clinical trial designs, utilising historical data and adding a treatment arm to an ongoing trial. The uptake of historical data methods has been low, possible reasons for this are: the risk of a loss of power or inflation in type I error in the current trial when the historical and current control data do not agree; a lack of understanding of historical data methods by clinicians and statisticians; and a lack of software available to implement historical data methods. Another example of gaining efficiency in a clinical trial is to add an extra treatment arm to an ongoing trial. Recent trials have implemented adding a treatment arm, however, the design and analysis implications of adding a treatment arm have not been fully explored [32].

Chapters 2 and 3 focus on historical data methods, specifically when there is one historical study available. The aim of the adaptive design proposed is to replace current control patients yet to be randomised with historical control data when the current and historical control data agree, reducing the number of controls to be randomised in the current study. As the disagreement between the historical and current controls increases, the adaptive design discounts the historical data and reverts back to a standard trial design to minimise the loss of power and the inflation in type I error.

Chapter 2 initially explores three methods proposed in the literature for incorporating historical data into the design and analysis of the current trial when the outcome data are binary. The methods considered are: power priors [15, 71] using a fully Bayesian approach or a summary measure of the marginal posterior distribution of the power as a fixed weight; the commensurate prior [19] and the robust mixture prior [23]. The limitations of these approaches were discussed extensively in Chapter 2. The main limitations were: the choice of prior on the "borrowing" parameter; calculating the effective historical sample size when using the adaptive design; and the computation time in calculating the operating characteristics of the design. The choice of prior for each approach was not intuitive in the amount of historical data that was incorporated into the final analysis and strong priors were required on the "borrowing" parameters to ensure a sufficient amount of historical data was incorporated into the final analysis when the historical and current controls were in agreement. For the adaptive design, at the interim analysis, it is required to calculate the effective historical sample size, but this is difficult for the fully Bayesian version of the power prior, the commensurate prior and the robust mixture prior. Finally, determining the operating characteristics for the commensurate prior and the fully Bayesian power prior and the effective historical sample size for the robust mixture prior were computationally intensive. Of the methods considered, the power prior approach with a fixed power weight was the simplest and most intuitive way to incorporate historical data into the current trial design and analysis. Choosing the weight for the historical data requires careful consideration.

From exploring historical data methods previously proposed in the literature, the most important factors for the design of a trial incorporating historical data were: a method to assess the agreement between historical and current control data that is intuitive to describe to clinicians and easy to calculate; the operating characteristics of the design need to be quick to calculate and the approach needs to allow control over the maximum possible type I error rate across all true control response probabilities. Control of the maximum type I error is important for the design to obtain regulatory approval, although in rare diseases the maximum type I error can be relaxed. The proposed equivalence weight approach for assessing agreement along with the analysis approach of the power prior with a fixed weight meets these criteria. The equivalence bounds allow discussion with clinicians about the acceptable range of agreement between historical and current controls and allows control over how much historical data are borrowed at complete agreement and how quickly the historical data are discounted as the difference between the current and historical controls increases. The equivalence weight is used as a fixed power in the analysis approach of the power prior. The initial priors for the response probabilities are chosen to be beta distributions with integer parameter values. The probability that the response probability in the treatment group is greater than the response probability in the control group can then be calculated quickly and exactly using the iterative procedure described by Cook [55]. The operating characteristics for the adaptive design incorporating historical data can then be calculated exactly and quickly. Finally, equivalence bounds can be determined that control the maximum type I error rate at the desired level.

In Chapter 2 the historical data methods explored are: power priors; commensurate priors and robust mixture priors. Chapter 3 explores these historical data methods when the outcome data are normally distributed. For normally distributed outcome data, both differences in the means and the variances between the historical and current control data are considered important when assessing agreement. Chapter three explores the limitations of published historical data methods when the outcome data are normally distributed. Many of the limitations of historical data methods found for binary data hold for normally distributed outcome data. The commensurate prior and the robust mixture prior seem to be "black box" approaches. Both models can be fitted for normally distributed data, however the choice of prior on the borrowing parameter for the commensurate prior and the prior parameters for the robust mixture prior are difficult to choose. The approaches for determining the effective sample size of the historical data have not been explored for the commensurate prior and robust mixture prior approaches when assuming unknown mean and variance in the current and historical data. Again, it was found that the most intuitive approach to incorporate historical data into the design and analysis of the current trial was to use the power prior approach with a fixed power.

The equivalence weight approach for assessing agreement between the historical and current controls is slightly more complicated for normally distributed data than the binary data case. For normally distributed outcome data, the equivalence weight is based on the marginal posterior distributions of the difference in the means and the ratio of the variances in the current and historical control data. The equivalence weight allows control over the amount of historical data borrowed and the rate of discounting. By fixing the equivalence bounds on the ratio of the variances, the mean equivalence bound can be chosen to cap the maximum possible inflation in the type I error rate across all true control means.

There are a few disadvantages to the equivalence weight approach for normally distributed outcome data: the operating characteristics cannot be calculated exactly, therefore simulation is required; it can be difficult to choose and interpret equivalence bounds for a variance parameter; and when the sample sizes in the historical and current controls differ, choosing equivalence bounds of the form $(\delta_{\phi_l}, 1/\delta_{\phi_l})$ does not give maximum weight to the historical data when there is complete agreement in the sample variances. However, in practice, the maximum weight obtained for the variance component of the equivalence weight and the weight obtained at complete agreement in the sample variances have been very similar in the examples considered. With one historical study, the difference in sample sizes between the current and historical controls is likely to be small.

Finally, Chapter 4 looked at adding a treatment arm to an ongoing trial. An iterative procedure is proposed that re-calculates the sample size required in all treatment groups of the trial when a treatment arm is added during the study with the aim of controlling the family-wise error rate. The Dunnett procedure [39] is used which compares multiple experimental treatments to a single control and the correlation is adjusted to account for only concurrent controls being used in each treatment to control comparison. Adding a treatment arm to an ongoing study is compared to running a separate trial for each experimental treatment in terms of the total number of patients required if the error rate is required to be the same for each separate trial and the multi-arm trial. Finally, given a fixed number of patients, the optimal allocation ratios to each treatment are explored that maximise the overall power of the study when a treatment arm is added during the trial.

5.2 Future research

The methodology presented in Chapters 2 and 3 describes the equivalence weight approach for assessing agreement between historical and current controls for binary and normally distributed outcome data. Further work is needed to apply this approach to other types of outcome data. Neither ordinal nor survival data have been considered. For survival data, there is the added complication of censoring and equivalence could be assessed using the Kaplan Meier curves, the logrank test or the hazard ratios. The concept of the equivalence approach is the same for all types of outcome data, however each requires a different implementation and careful thought is required on how to define equivalence for each type of outcome.

In this thesis we have focused on the specific problem of having only one relevant historical study. Both having one historical study and having multiple historical studies have their own complications. For only one historical study, it is not possible to obtain an estimate of the between study heterogeneity. Where there are multiple relevant historical studies, there will not only be heterogeneity between the historical studies and the current control group but also between the historical studies themselves. The historical studies need to be combined and the heterogeneity between the studies determined. When there is one historical study, the effective historical sample size is simply the sample size of the historical study. When there are multiple historical studies, the effective historical sample size incorporates the heterogeneity between the historical studies, increased heterogeneity reduces the effective historical sample size [12]. How to combine multiple historical studies and given the heterogeneity between them determine how much information is contained in the historical data is a complicated process. The combined historical data then has to be compared to the current control data and incorporated into the current trial analysis.

Schmidli et al. [23] propose first performing a meta-analysis of the historical studies and using the meta-analytic predictive distribution as a prior for the current study. One historical distribution is then compared to the current controls. This approach takes into account the heterogeneity between the historical studies by a reduced prior effective sample size and as with the robust mixture prior, the meta analysis predictive distribution is discounted when there is prior data conflict. An alternative approach to incorporating multiple historical control studies into the design and analysis of the current trial is to consider each historical study separately. The agreement between each historical study and the current control arm is assessed. The historical studies can then be incorporated into the final analysis each with their own weighting parameter, this approach is proposed as a way to incorporate multiple historical studies using the power prior approach [71]. Further work will explore the advantages and disadvantages of these two approaches and how they compare in terms of how much historical data are incorporated into the final analysis using each approach. The best approach may differ depending on the number of historical studies. For a small number of historical studies, where estimating the between study heterogeneity is difficult, incorporating the studies individually may be the best approach. For many historical studies a meta-analysis approach may be best.

The advantage of incorporating historical data into both the design and analysis of a current study is that it allows fewer control patients to be randomised in the current study if the historical and current controls agree. Sequential designs that allow early stopping for efficacy and/or futility also allow the possibility for fewer patients to be randomised in the current trial. An area of interesting future work would be to compare the expected total sample size for the equivalence approach utilising historical data to both frequentist and Bayesian sequential designs. Historical data methods and sequential designs could potentially be used in combination.

The examples considered here were consistent with a confirmatory trial design in terms of the operating characteristics chosen and therefore the required sample size. We did not explore how the methods worked when the current study was an early phase trial design, which typically have a smaller sample size and allow a higher type I error rate. Further, we did not consider multiple historical studies and a setting where there is a large amount of historical data compared to the sample size of the current control arm. Therefore further work would look at the effect of different sample sizes in the historical and current trial and trials with different design characteristics. The adaptive design considered in Chapters 2 and 3 only incorporated one interim analysis. As data on the current control arm accrues, the accuracy of the estimate of the agreement between the historical and current controls will also improve. Possible future work could look at the optimal time to perform the interim analyses and the optimal number of interim analyses that are required to improve the efficiency of the design but still allow the design to be logistically plausible.

In the recent literature, alternative approaches have been proposed for determining the power when using the power prior method. An empirical Bayes approach has been proposed by Gravestock and Held [72]. This approach is similar to deriving the weight as the mode of the marginal posterior distribution of the power under a flat Beta(1,1) prior. Ibrahim et al. [71] propose various methods to facilitate the choice of a fixed power for a normal linear model, these are: a penalized likelihood type criterion; marginal likelihood criterion; deviance information criterion; and pseudo-marginal likelihood criterion. Finally, Pan et al. [73] propose the calibrated power prior where the power is defined as a function of a congruence measure between the historical and current data. Future work would compare these approaches to the equivalence weight approach proposed in this thesis.

In Chapter 4, it was assumed that the treatment effect to detect, comparing each of the experimental treatments to control, was the same for all treatments. This may not be a plausible assumption for some designs and it is of interest to determine the optimal allocation ratios to each treatment group in terms of power when the assumed treatment effects differ. A further assumption made in Chapter 4 was that all the treatments were of equal importance and the aim was to detect whether any or all treatments were better than control. If the aim was to detect the best treatment then the power would depend on the mean effect of the best treatment and the mean effect of all other experimental treatments. Future work could look at adding a treatment arm and optimal allocation to detect the best treatment.

To determine the optimal allocation ratio after a new treatment arm has been added to the trial, we have assumed that the patient population is homogeneous across stages. Elm et al. [30] looked at different analysis methods, specifically: pooling data across stages; regression adjusting for stage; and combination p-value methods. Elm et al. [30] explored the performance of these methods when the assumption that the patient population is homogeneous across stages was not true. They consider the stage effect to be random and affecting all treatment groups. It is of interest to see how stage effects may affect the operating characteristics when a treatment arm is added during the trial. Finally, the design and analysis for adding a treatment arm to a trial considered in Chapter 4 only used concurrently randomised controls for each treatment comparison. “Historical” control data from within the same trial meets all of Pocock’s [11] criteria for acceptable historical data and future work will look at ways to incorporate this data into the new treatment control comparisons when a treatment arm has been added during an ongoing trial.

5.3 Conclusion

This thesis explores two approaches to improve the efficiency of clinical trial designs over the standard two-arm randomised controlled trial. The first approach is to utilise historical data. An equivalence weight approach is proposed that assesses agreement between historical and current control data and allows the study designer to specify acceptable ranges of agreement. An adaptive design is then used which allows historical controls to replace current controls when there is agreement between the current and historical data. This reduces the number of controls to be randomised in the current study. When there is disagreement, the historical data are discounted and the trial reverts back to a standard design to safeguard against a large reduction in the power or a large inflation in the type I error rate of the current study. The equivalence approach is explored for both binary and normally distributed outcome data. The second approach explored to improve efficiency is to add a treatment arm to an ongoing study. Multiple experimental treatments are compared to concurrently randomised controls from a single control group. The sample size of the study is increased to control the family-wise error rate using the Dunnett procedure [39], where the correlation is derived to account for the use of concurrent controls only in each treatment comparison. The use of a single control group typically results in a reduction in the number of controls required compared to running a separate trial for each experimental treatment, depending on when the new treatment arm is added and the desired operating characteristics of the design. Further, there are substantial savings in both time and money from not having to initiate a new trial. The work of this thesis has made the use of historical data in trial design more approachable to clinicians and addressed concerns over family-wise error rate inflation when adding a treatment arm to an ongoing study.

Bibliography

[1] US Food and Drug Administration. Guidance for Industry: E9 Statistical Principles for Clinical Trials, 1998. URL `https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf`.

[2] Aylin Sertkaya, Hui-Hsing Wong, Amber Jessup, and Trinidad Beleche. Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials*, 13(2): 117–126, 2016.

[3] Alberto Grignolo and Sy Pretourius. Phase III Trial Failures: Costly, But Preventable. *Applied Clinical Trials*, 25(8), 2016.

[4] US Food and Drug Administration. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products, 2004. URL `https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm113411.pdf`.

[5] US Food and Drug Administration. Guidance for Industry and FDA Staff: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, 2010. URL `https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf`.

[6] US Food and Drug Administration. Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics, 2010. URL `https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf`.

[7] Nancy J Nelson. Adaptive Clinical Trial Design: Has Its Time Come? *Journal of the National Cancer Institute*, 102(16):1217–1218, 2010.

[8] Thomas Bayes, Richard Price, and John Canton. An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions*, 53:370–418, 1763.

[9] Donald A Berry. [Investigating Therapies of Potentially Great Benefit: ECMO]: Comment: Ethics and ECMO. *Statistical Science*, 4(4):306–310, 1989.

[10] Ian Wadsworth, Lisa V Hampson, and Thomas Jaki. Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods. *Statistical Methods in Medical Research*, 2016.

[11] Stuart J Pocock. The Combination of Randomized and Historical Controls in Clinical Trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976.

[12] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010.

[13] N.W Galwey. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion? *Statistics in Medicine*, 36(6):899–916, 2017.

[14] Joseph G Ibrahim and Ming-Hui Chen. Power Prior Distributions for Regression Models. *Statistical Science*, 15(1):46–60, 2000.

[15] Yuyan Duan. *A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation.* PhD thesis, Virginia Polytechnic Institute and State University, 2005.

[16] Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, 2009.

[17] David Lunn, Chris Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis.* CRC press, 2012.

[18] Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. *Biometrics*, 67(3):1047–1056, 2011.

[19] Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis*, 7(3):639–674, 2012.

[20] Brian P Hobbs, Bradley P Carlin, and Daniel J Sargent. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10(3):430–440, 2013.

[21] Andrew Gelman. Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[22] Eleanor M Pullenayegum. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine*, 30(26):3082–3094, 2011.

[23] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

[24] Anthony O'Hagan and Luis Pericchi. Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, 26(4):372–401, 2012.

[25] M.Farrow. Mas3301 bayesian statistics, 2009. URL `http://www.mas.ncl.ac.uk/~nmf16/teaching/mas3301/week11.pdf`.

[26] Satoshi Morita, Peter F Thall, and Peter Müller. Determining the Effective Sample Size of a Parametric Prior. *Biometrics*, 64(2):595–602, 2008.

[27] Donald J Schuirmann. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Pharmacodynamics*, 15(6):657–680, 1987.

[28] Tie-Hua Ng. Choice of delta in equivalence testing. *Drug Information Journal*, 35 (4):1517–1527, 2001.

[29] Matthew R Sydes, Mahesh KB Parmar, Malcolm D Mason, Noel W Clarke, Claire Amos, John Anderson, Johann de Bono, David P Dearnaley, John Dwyer, Charlene Green, et al. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. *Trials*, 13(1):168, 2012.

[30] Jordan J Elm, Yuko Y Palesch, Gary G Koch, Vanessa Hinson, Bernard Ravina, and Wenle Zhao. Flexible Analytical Methods for Adding a Treatment Arm Mid-Study to an Ongoing Clinical Trial. *Journal of Biopharmaceutical Statistics*, 22(4):758–772, 2012.

[31] James Wason, Dominic Magirr, Martin Law, and Thomas Jaki. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 2012.

[32] Dena R Cohen, Susan Todd, Walter M Gregory, and Julia M Brown. Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials*, 16(1):179, 2015.

[33] Michael A Proschan and Dean A Follmann. Multiple Comparisons with Control in a Single Experiment versus Separate Experiments: Why Do We Feel Differently? *The American Statistician*, 49(2):144–149, 1995.

[34] James MS Wason, Lynne Stecher, and Adrian P Mander. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*, 15(1):364, 2014.

[35] Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials, 2002. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf`.

[36] Alex Dmitrienko, Ajit C Tamhane, and Frank Bretz. *Multiple Testing Problems in Pharmaceutical Statistics*. CRC Press, 2009.

[37] US Food and Drug Administration. Guidance for industry: Multiple Endpoints in Clinical Trials, 2017. URL `https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf`.

[38] Zbyněk Šidák. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[39] Charles W Dunnett. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[40] John Whitehead, Faye Cleary, and Amanda Turner. Bayesian sample sizes for exploratory clinical trials comparing multiple experimental treatments with a control. *Statistics in Medicine*, 34(12):2048–2061, 2015.

[41] StataCorp. Stata statistical software: Release 13, 2013. URL `https://www.stata.com/stata13/`.

[42] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2013.

[43] Yuyan Duan, Keying Ye, and Eric P Smith. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106, 2006.

[44] Mu Zhu and Arthur Y Lu. The Counter-intuitive Non-informative Prior for the Bernoulli Family. *Journal of Statistics Education*, 12(2):1–10, 2004.

[45] David Dejardin, Joost van Rosmalen, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials using the modified power prior: Practical considerations for survival models. Bayes Pharma Conference, 2014. URL `http://www.bayes-pharma.org/bayes2014docs/Day1/Dejardin.pdf`.

[46] StataCorp. Mata reference manual release 13, 2013. URL https://www.stata.com/manuals13/m.pdf.

[47] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

[48] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[49] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis.* CRC press, 1986.

[50] StataCorp. Stata 13 base reference manual, 2013. URL https://www.stata.com/manuals13/u.pdf.

[51] William R Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933.

[52] STA 250: Statistics-Notes 11. Laplace Approximation to the Posterior. URL http://www2.stat.duke.edu/~st118/sta250/laplace.pdf.

[53] Vose Software. Normal approximation to the beta posterior distribution. URL http://www.vosesoftware.com/riskwiki/NormalapproximationtotheBetaposteriordistribution.php.

[54] John Cook. Error in the normal approximation to the beta distribution. URL https://www.johndcook.com/blog/normal_approx_to_beta/.

[55] John D Cook. Exact Calculation of Beta Inequalities. Technical report, UT MD Anderson Cancer Center Department of Biostatistics, 2005. URL https://www.johndcook.com/exact_beta_inequalities.pdf.

[56] Robert L Cuffe. The inclusion of historical control data may reduce the power of a confirmatory study. *Statistics in Medicine*, 30(12):1329–1338, 2011.

[57] NCSS. Tests for two proportions, 2007. URL https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Tests_for_Two_Proportions.pdf.

[58] JO Berger and JM Bernardo. On the development of the reference prior method. *Bayesian Statistics*, 4, 1992.

[59] Robert E Kass and Larry Wasserman. A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

[60] Timothy Mutsvari, Dominique Tytgat, and Rosalind Walley. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 66(5):979–996, 2016.

[61] Kevin P. Murphy. Conjugate bayesian analysis of the gaussian distribution, 2007. URL `https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf`.

[62] Henry F Inman and Edwin L Bradley Jr. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-Theory and Methods*, 18(10): 3851–3874, 1989.

[63] Andrew P Grieve. Joint equivalence of means and variances of two populations. *Journal of Biopharmaceutical Statistics*, 8(3):377–390, 1998.

[64] Bernard L Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362, 1938.

[65] Peter Bauer and Michael M Bauer. Testing Equivalence Simultaneously for Location and Dispersion of two Normally Distributed Populations. *Biometrical Journal*, 36 (6):643–660, 1994.

[66] GA Barnard. Comparing the Means of Two Independent Samples. *Applied Statistics*, 33(3):266–271, 1984.

[67] London MRC Clinical Trials Unit. Stampede trial, January 2010. URL `http://www.stampedetrial.org`.

[68] The Pennsylvania State University. The optimum allocation for the dunnett test, June 2015. URL `https://onlinecourses.science.psu.edu/stat503/node/16`.

[69] Huaibao Feng, Jun Shao, and Shein-Chung Chow. Adaptive Group Sequential Test for Clinical Trials with Changing Patient Population. *Journal of Biopharmaceutical Statistics*, 17(6):1227–1238, 2007.

[70] Scott M Berry, Jason T Connor, and Roger J Lewis. The Platform Trial: An Efficient Strategy for Evaluating Multiple Treatments. *JAMA*, 313(16):1619–1620, 2015.

[71] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015.

[72] Isaac Gravestock and Leonhard Held. Adaptive power priors with empirical bayes for clinical trials. *Pharmaceutical Statistics*, 2017.

[73] Haitao Pan, Ying Yuan, and Jielai Xia. A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 10(4):325–337, 2016.

Appendix A

Expected control sample size for the adaptive design with a binary outcome

Expected total control sample size across different true current control proportions for the adaptive design using the probability and equivalence probability weight

Section 2.5.5 explores the design characteristics of the Viele example for the adaptive design proposed in Section 2.4.2. Figures A.1 and A.2 show the $ECSS$ across a range of true current control proportions for the Viele example using the adaptive design with the probability weight and equivalence probability weight with 8% equivalence bounds, respectively. For a range of current control response probabilities around complete agreement between the current and historical data, the $ECSS$ for the adaptive design is slightly above the 200 patients required under a standard trial design. This increase in sample size is due to the minimum requirement of 20 controls to be randomised in stage two of the trial and also due to the weight given to the historical data changing from the interim analysis to the final analysis. The $ECSS$ is also slightly below 200 for a range of current control proportions due to the change in the weight calculated at the interim analysis and the final analysis. The change in the weight from the interim to the final analysis is illustrated in Appendix B.

Figure A.1: Expected control sample size across different true current control proportions for the adaptive design using the probability weight approach and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red line represents complete agreement between the historical and current control proportions.

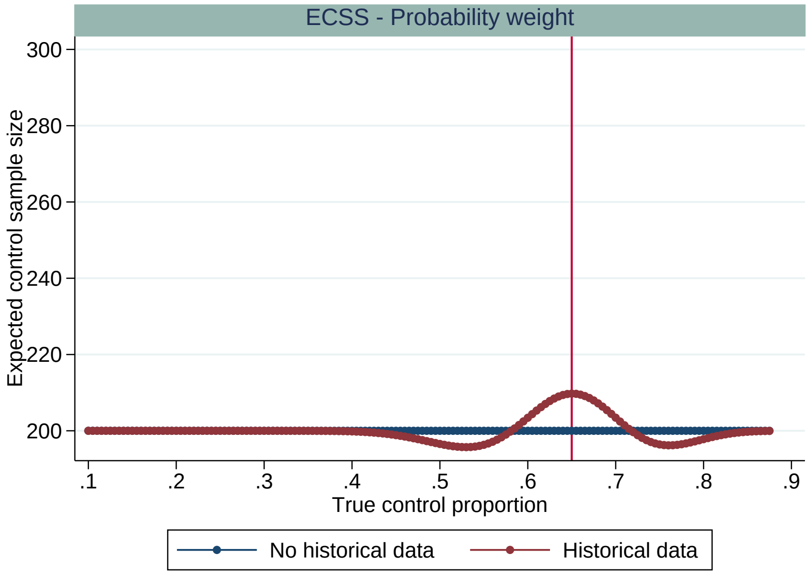
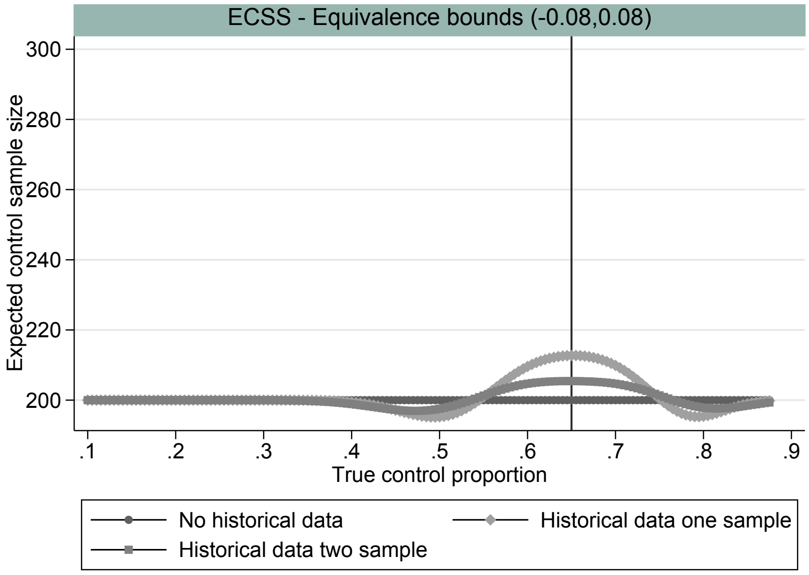


Figure A.2: Expected control sample size across different true current control proportions for the adaptive design using the one-sample and two-sample equivalence probability weight approaches with 8% equivalence bounds and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red line represents complete agreement between the historical and current control proportions.



Appendix B

Expected weights at the interim and final analysis with a binary outcome

Comparison of the expected probability and equivalence probability weights across different true current control proportions at the interim and final analysis for the adaptive design

For the adaptive design proposed in Section 2.4.2, the weight is calculated at the interim analysis and this weight is used to choose how many current controls to randomise in stage two of the trial. The weight given to the historical data is re-calculated at the end of the trial and this weight is used to discount the historical data in the final analysis. Figures B.1 and B.2 illustrate the difference in the expected weights at the interim analysis and the final analysis for the Viele example using the probability weight and equivalence probability weight.

At the end of the study, when the sample size of the controls is larger than at the interim analysis, on average the weight is slightly higher at complete agreement and discounts to zero at a quicker rate than at the interim analysis. The difference between the equivalence weight at the interim and final analysis differs depending on the equivalence bounds chosen.

Figure B.1: Expected probability weight at the interim analysis and at the end of the study for the adaptive design. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red line represents complete agreement between the historical and current control proportions.

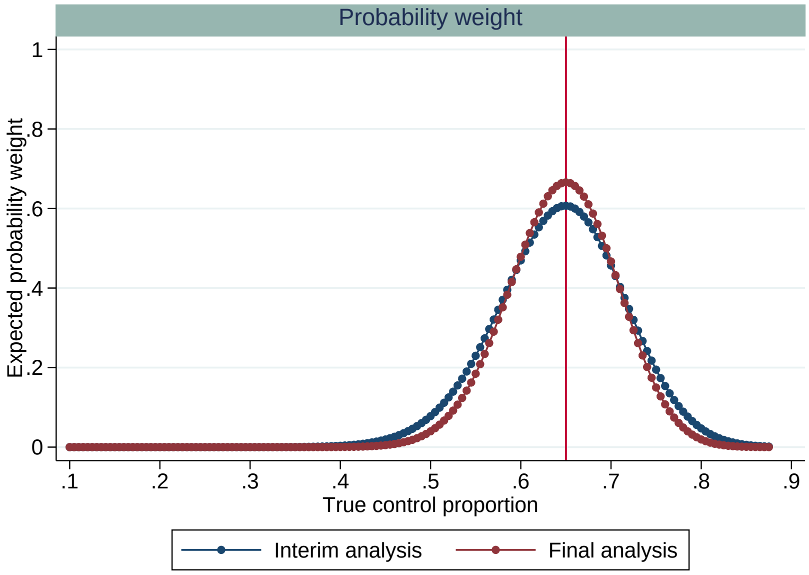
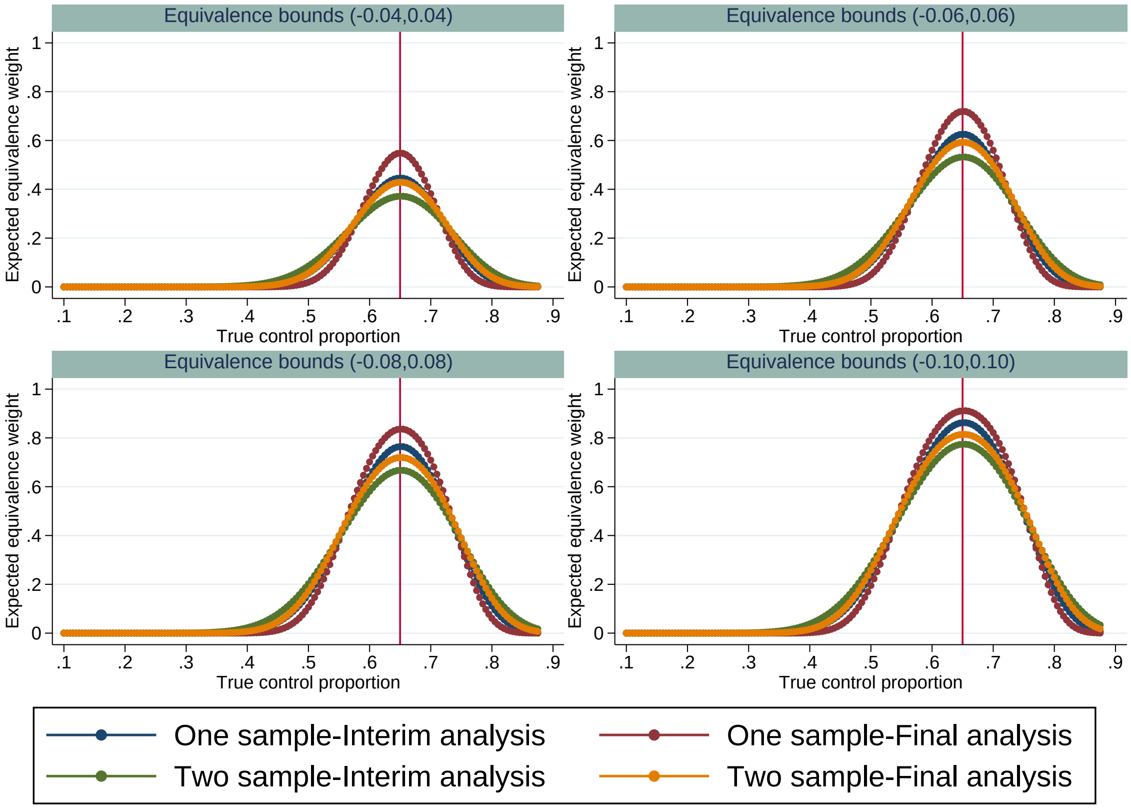


Figure B.2: Expected equivalence probability weight at the interim analysis and at the end of the study for different equivalence bounds for the adaptive design. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



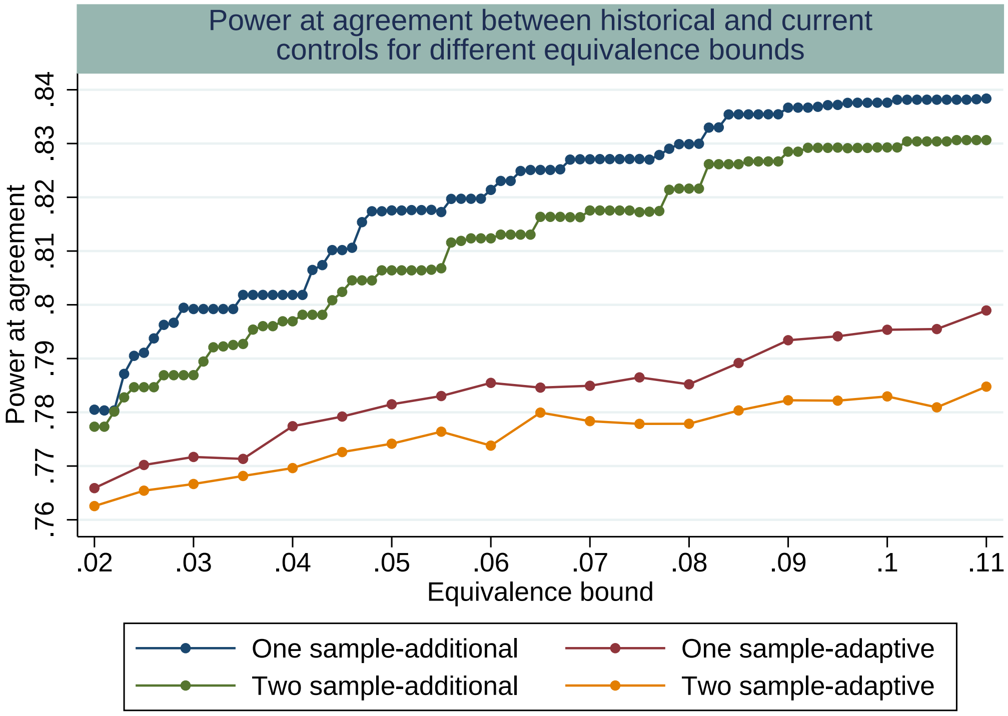
Appendix C

Power at agreement for different equivalence bounds with a binary outcome

Power at agreement in the historical and current controls for different equivalence bounds with a binary outcome

Section 2.5.6 explores how to choose the equivalence bounds to control the maximum type I error across all true control response probabilities in the current study. Figure C.1 shows the power at complete agreement between the historical and current controls for different equivalence bounds. Smaller equivalence bounds result in less gain in power at agreement between the current and historical controls but also result in a lower maximum type I error when there is disagreement between the current and historical controls. The power at complete agreement in the current and historical controls versus the maximum type I error across all true current control response probabilities is the trade off in choosing the equivalence bounds. The curves in Figure C.1 are not smooth due to the rounding up of the effective historical sample size in the additional information design and the rounding up in calculating both the remaining controls to be randomised in stage two of the adaptive design and when re-calculating the effective historical sample size at the end of the study for the adaptive design. Typically, larger equivalence bounds should give a larger power at complete agreement between the current and historical controls. This is at the expense of a higher maximum type I error across all true control proportions.

Figure C.1: Power at complete agreement in the historical and current controls for different equivalence bounds for the additional information and adaptive design, Viele example.



Appendix D

Fully Bayesian power prior with a binary outcome

Comparison of the power and type I error of the fully Bayesian power prior and the power prior taking the mean of the marginal distribution of the power as a fixed weight

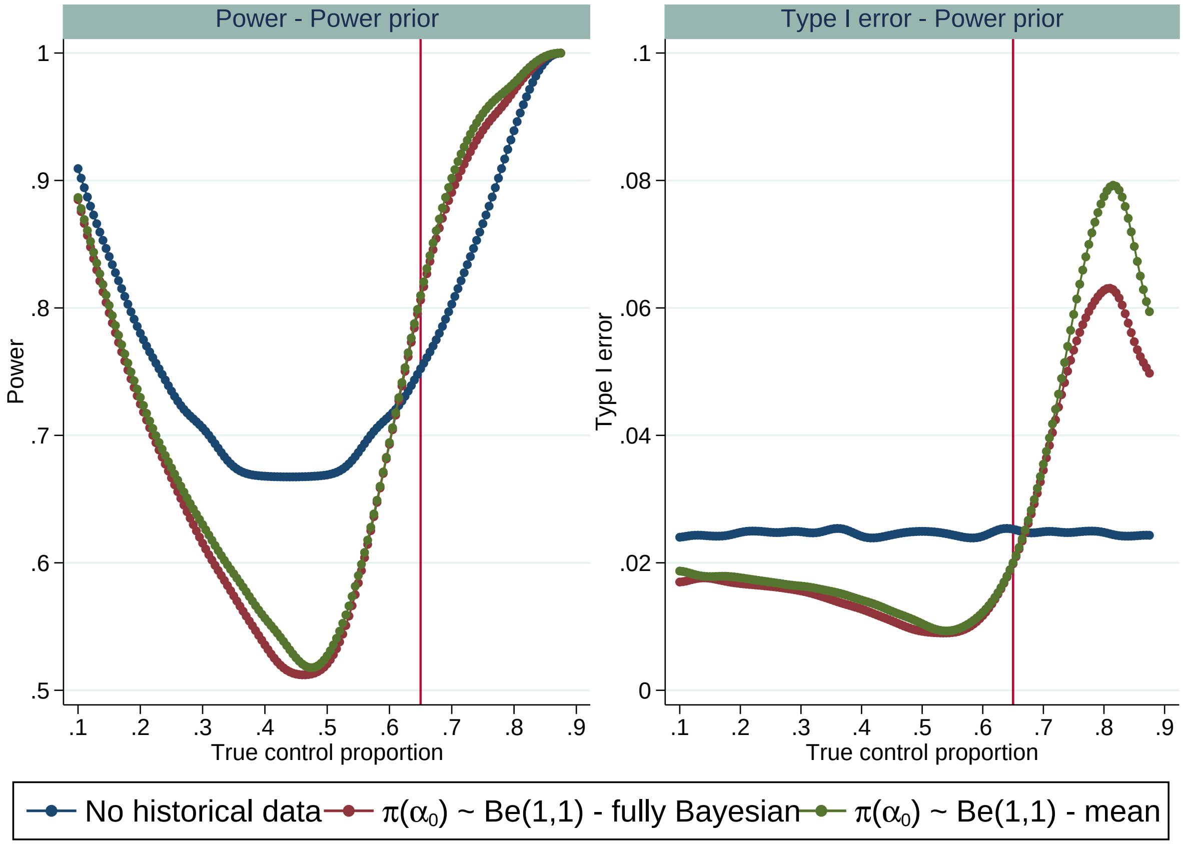
Figure D.1 compares the power and type I error of the additional information design for the Viele example when using the fully Bayesian version of the modified power prior and when taking the mean of the marginal distribution of α_0 as a fixed power, both modified power prior approaches are described in Section 2.2.1. The power is assumed to have a Beta(1,1) prior. Beta(1,1) priors are also assumed for the control and treatment response probabilities. The joint posterior distribution for p_c and α_0 is given in Equation 2.2. The $\Pr(p_t > p_c)$ is then calculated using numerical integration,

$$\Pr(p_t > p_c) = \int\limits_0^1 \int\limits_{p_c}^1 \int\limits_0^1 \pi(p_c, \alpha_0 \mid x_h, y_h, x_c, y_c)\pi(p_t \mid x_t, y_t) d\alpha_0 dp_t dp_c$$

The power and type I error are then calculated using Equation 2.14. Using this process to determine the operating characteristics of a design is computationally intensive.

From Figure D.1, the operating characteristics have a similar pattern for the fully Bayesian modified power prior and the modified power prior using the mean of the marginal distribution of the power as a fixed weight, when a Beta(1,1) prior is assumed for the power. The fully Bayesian modified power prior has a lower maximum type I error across all true current control proportions compared to the modified power prior using the mean of the marginal distribution of the power as a fixed weight, this is likely to be due to the large uncertainty in estimating α_0 as shown by the 95% credible intervals for α_0 in Table 2.2. Using the fully Bayesian version of the power prior does not provide a

Figure D.1: Comparison of the power and type I error across different true current control proportions for the additional information design using the fully Bayesian modified power prior approach, the modified power prior taking the mean of the marginal distribution of α_0 as a fixed power and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 198$, $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



measure of how much historical data are incorporated into the final analysis. The adaptive design was not considered using the fully Bayesian version of the power prior due to the computation time and because the adaptive design requires an estimate of the $EHSS$ at the interim analysis.

Appendix E

Adaptive design example using the robust mixture prior and the power prior with a binary outcome

Adaptive design – frequentist operating characteristics for the Viele example using the robust mixture prior and the power prior

Figure E.1 illustrates the operating characteristics for the adaptive design using the robust mixture prior approach. Calculating the expected control sample size (incorporating both the current controls and historical data) at the end of the study is computationally intensive using the robust mixture prior and therefore only the expected current control group sample size (which only requires calculating the ESS at the interim analysis) is calculated. Figure E.2 shows the operating characteristics for the adaptive design using the power prior approach with different priors on the power. Finally, Figure E.3 shows the expected weight given to the historical data at the interim analysis and the final analysis for the Viele example using the modified power prior with different priors on the power.

Figure E.1: Comparison of the power, type I error, mean squared error and expected current control sample size across different true current control proportions for the adaptive design using the robust mixture prior approach with 0.9 and 0.5 initial weight on the informative component of the mixture prior and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.

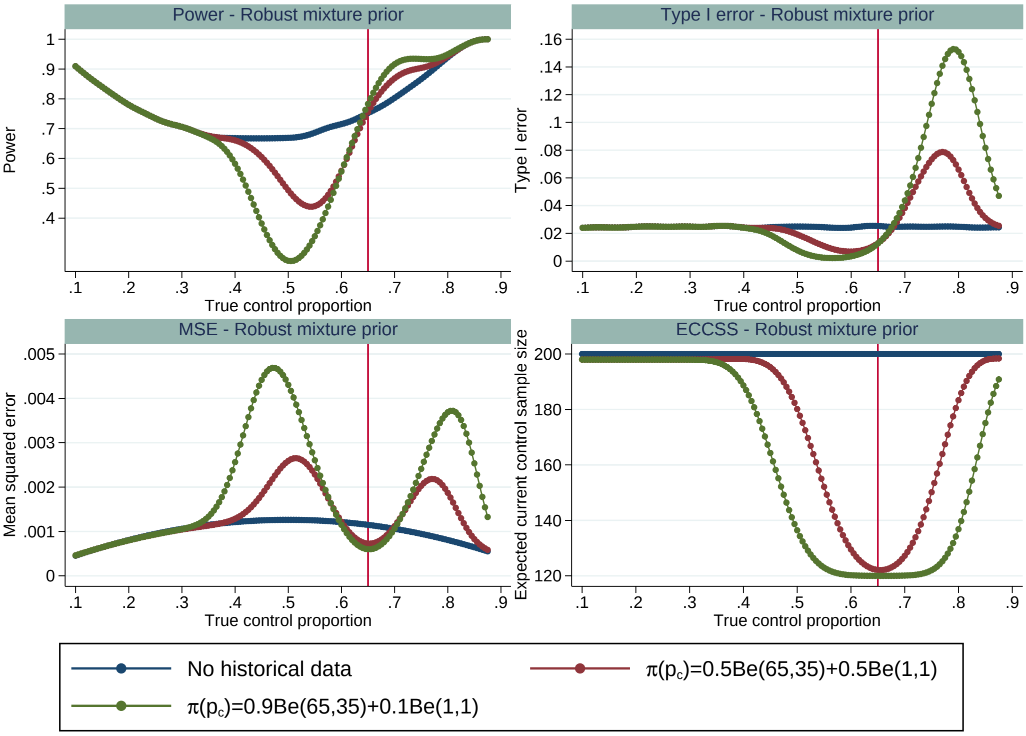


Figure E.2: Comparison of the power, type I error, mean squared error and expected current control sample size across different true current control proportions for the adaptive design using the power prior, assuming different priors on the power and a standard design incorporating no historical data. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.

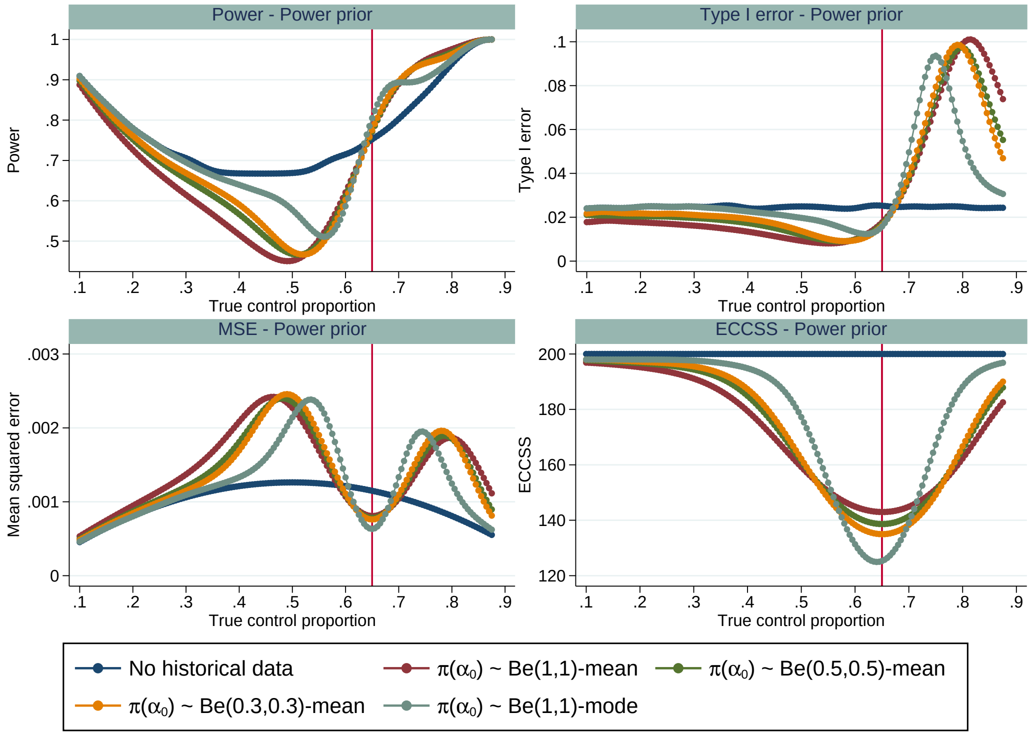
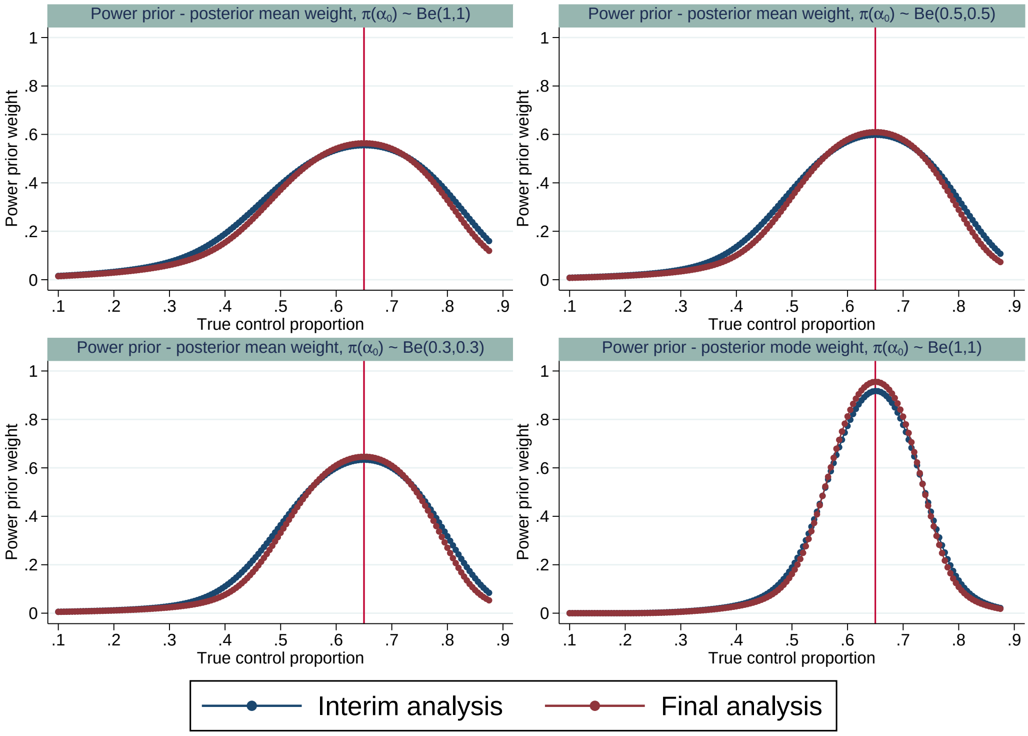


Figure E.3: Expected power prior weight at the interim analysis and at the end of the study for the adaptive design with different priors on the power. Viele example, historical data 65/100 responses, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and $\Delta = 12\%$. The vertical red lines represent complete agreement between the historical and current control proportions.



Appendix F

Equivalence probability weight example with a normally distributed outcome

Additional information and adaptive design - frequentist operating characteristics example using the equivalence probability weight with a normally distributed outcome

The following figures illustrate the design characteristics for the additional information and adaptive design using the equivalence probability weight approach with mean equivalence bounds of ±8 and variance equivalence bounds of (0.7,1/0.7), analogous to Figures 3.22 and 3.25, respectively.

Figure F.1: Comparison of the power, type I error, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the corrected equivalence probability weight approach with mean equivalence bounds ±8 and variance equivalence bounds (0.7,1/0.7) and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.

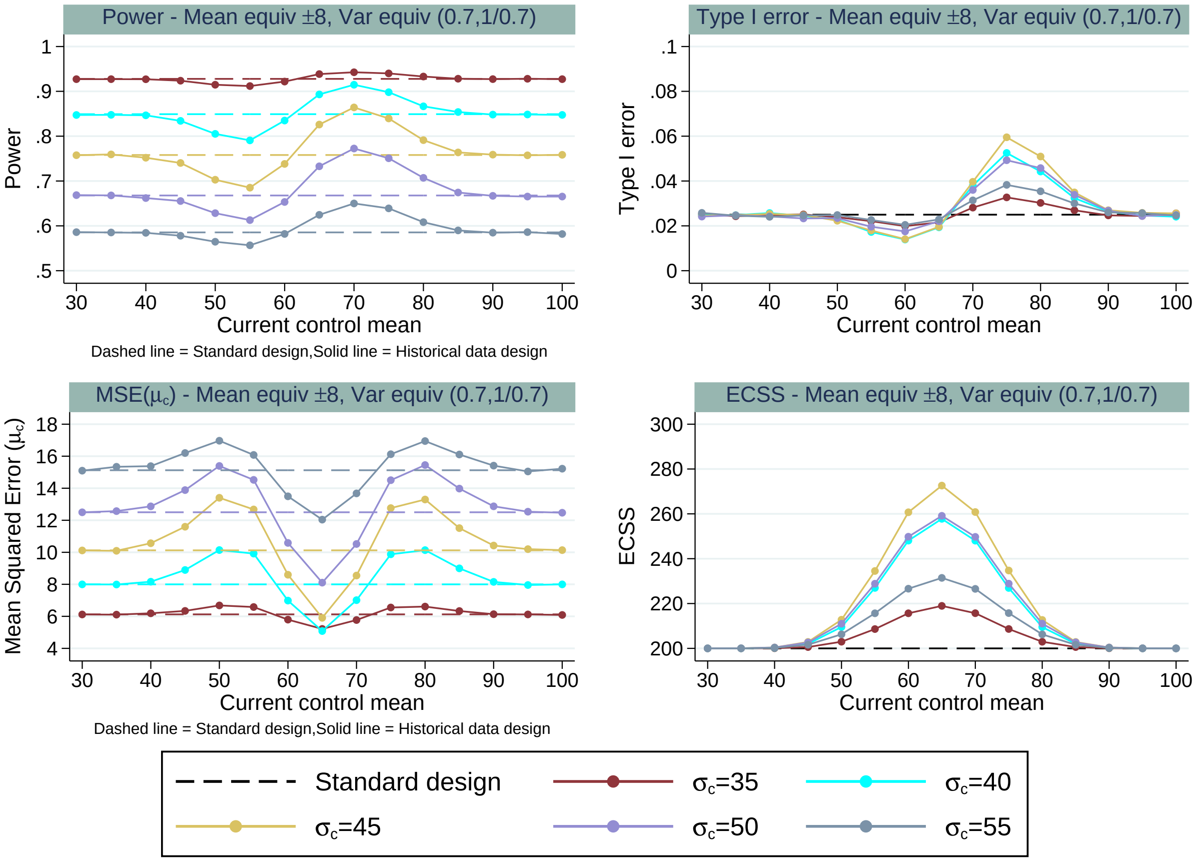
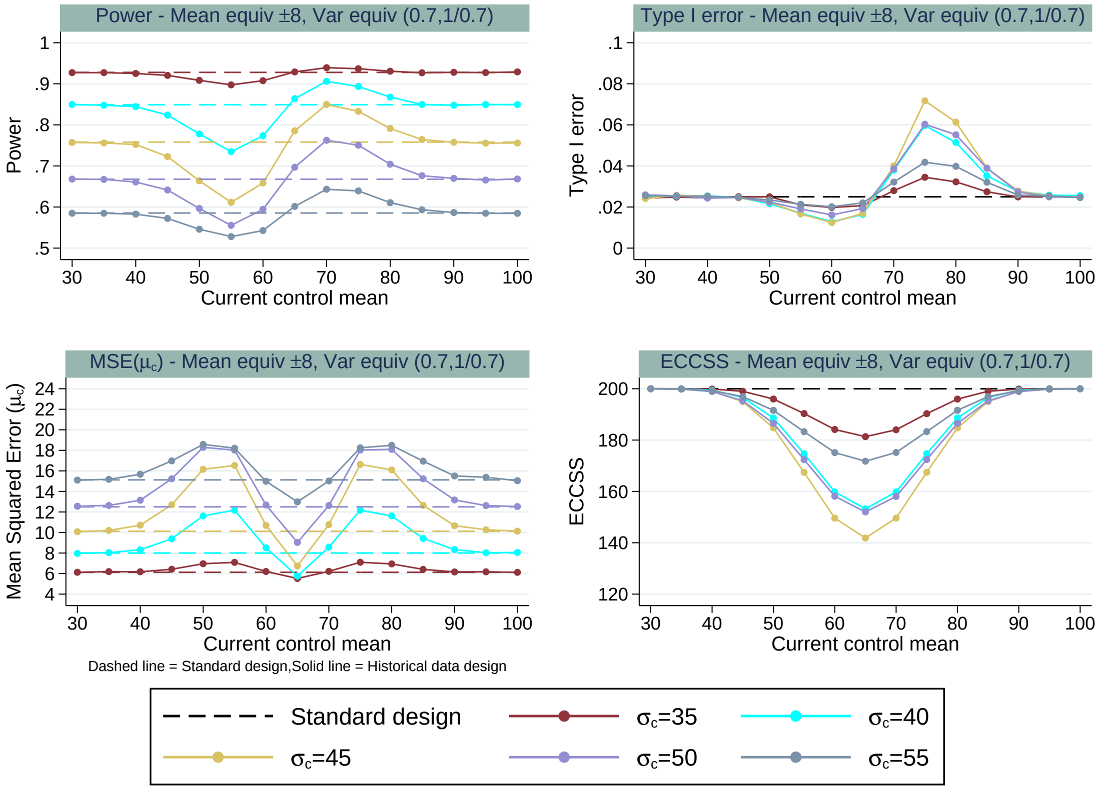


Figure F.2: Comparison of the power, type I error, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the corrected equivalence probability weight approach with mean equivalence bounds ±8 and variance equivalence bounds (0.7,1/0.7) and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.



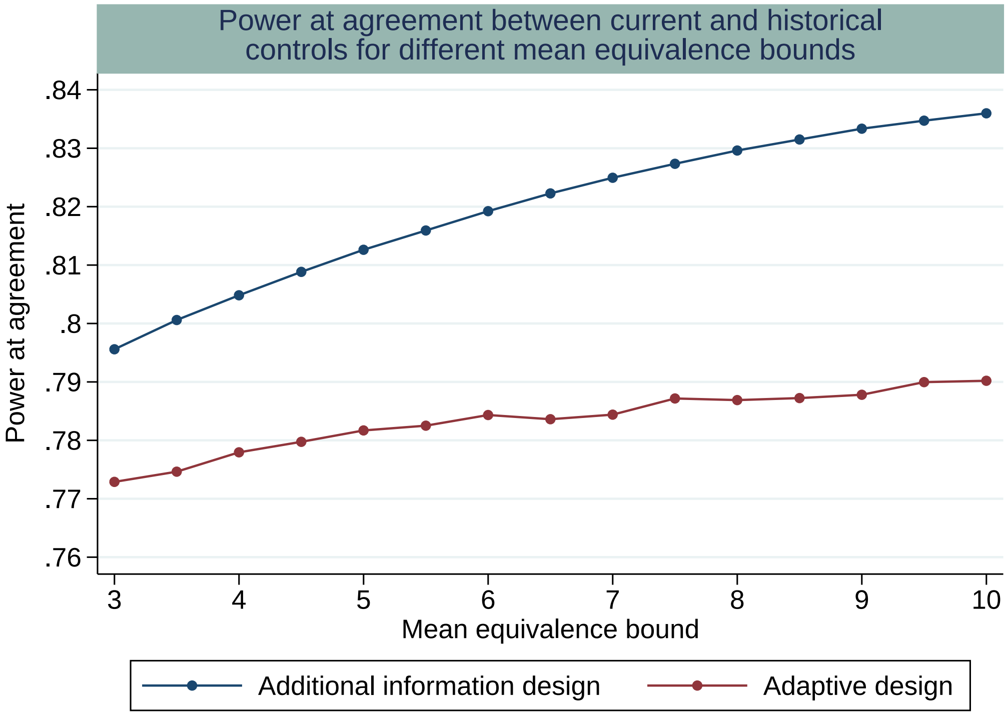
Appendix G

Power at agreement for different equivalence bounds with a normally distributed outcome

Power at complete agreement between the historical and current controls for different equivalence bounds on the mean difference with a normally distributed outcome

Sections 3.3.4 and 3.5.6 discuss how to choose the mean equivalence bounds to control the maximum type I error across all true current control means. Figure 3.27 shows the maximum type I error across a range of equivalence bounds for the mean when the equivalence bounds on the ratio of the variances are fixed at (0.6,1/0.6). Figure G.1 shows the power at complete agreement between the historical and current control means and standard deviations for different mean equivalence bounds when the equivalence bounds on the ratio of the variances are fixed at (0.6,1/0.6). Smaller equivalence bounds result in less gain of power at agreement between the current and historical controls but also have a lower risk with a smaller maximum type I error when there is disagreement between the current and historical controls. The power at complete agreement in the current and historical controls versus the maximum type I error across all true means in the current controls is the trade off in choosing the equivalence bounds.

Figure G.1: Power at complete agreement in the historical and current controls for different equivalence bounds on the mean difference when the equivalence bounds on the ratio of the variances are fixed at (0.6,1/0.6) for the additional information and adaptive design.



Appendix H

Additional information design example using the power prior with a normally distributed outcome

Additional information design - frequentist operating characteristics example using the modified power prior

The following figures illustrate the design characteristics for the additional information design using the modified power prior with a Beta(1,1) prior on the power using the mean of the marginal distribution of the power as a fixed weight, a Beta(0,5,0,5) prior on the power using the mean of the marginal distribution of the power as a fixed weight and a Beta(1,1) prior on the power using the mode of the marginal distribution of the power as a fixed weight. The following figures are analogous to Figure 3.28.

Figure H.1: Comparison of the power, type I error, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the modified power prior with a Beta(1, 1) prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.

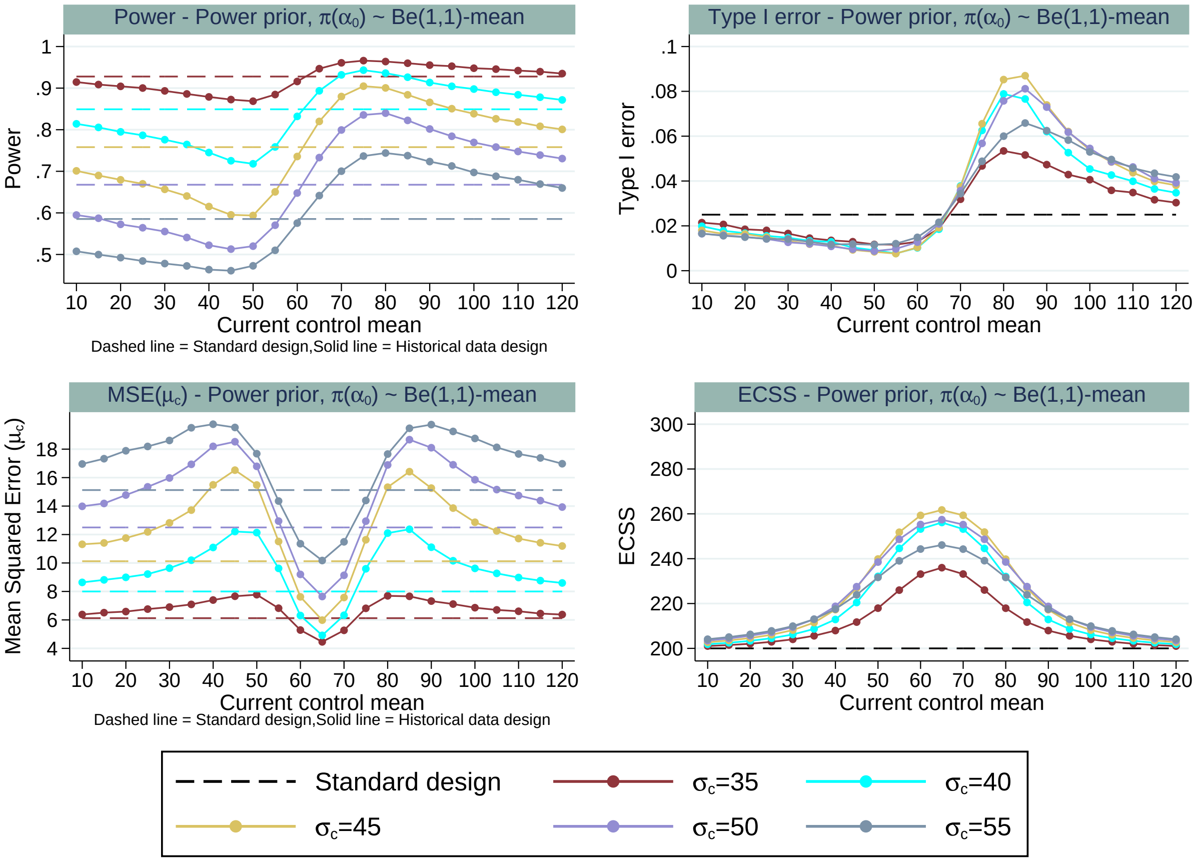


Figure H.2: Comparison of the power, type I error, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the modified power prior with a Beta$(0.5, 0.5)$ prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.

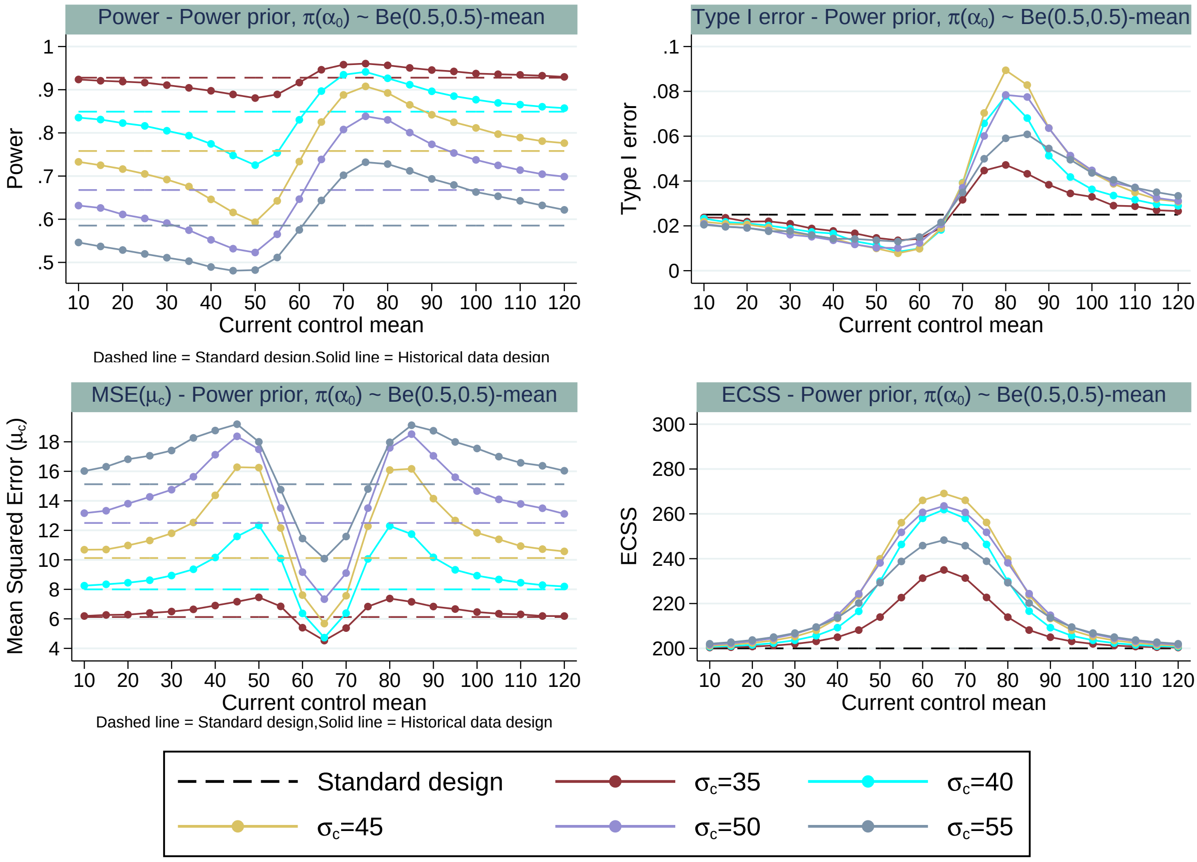
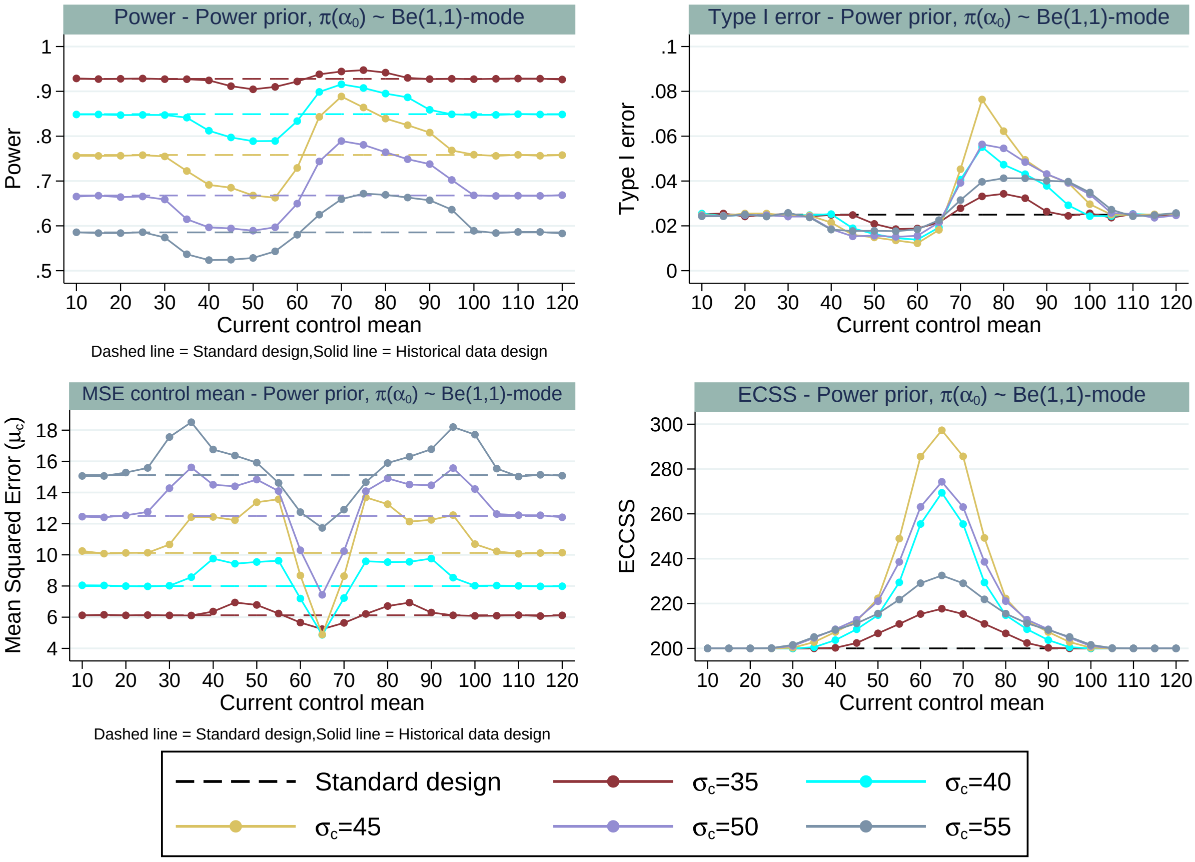


Figure H.3: Comparison of the power, type I error, mean squared error and expected control sample size across different true means and standard deviations in the current trial control arm for the additional information design using the modified power prior with a Beta(1, 1) prior on α_0 taking the mode of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$ and treatment effect $\mu_t = \mu_c + 12$.



Appendix I

Adaptive design example using the power prior with a normally distributed outcome

Adaptive design - frequentist operating characteristics example using the modified power prior

The following figures illustrate the design characteristics for the adaptive design using the modified power prior with a Beta(1,1) prior on the power using the mean of the marginal distribution of the power as a fixed weight, a Beta(0,5,0,5) prior on the power using the mean of the marginal distribution of the power as a fixed weight and a Beta(1,1) prior on the power using the mode of the marginal distribution of the power as a fixed weight. The following figures are analogous to Figure 3.29.

Figure I.1: Comparison of the power, type I error, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the using the modified power prior with a Beta(1, 1) prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.

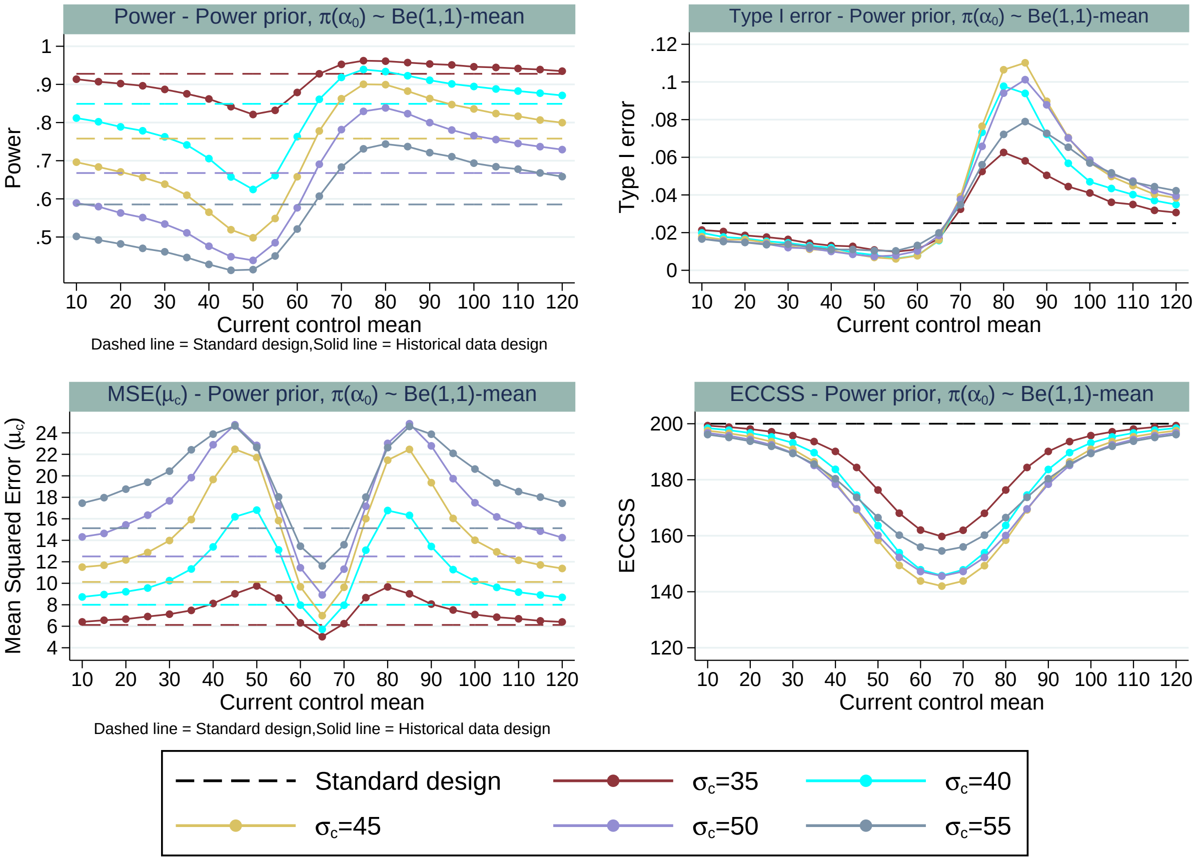
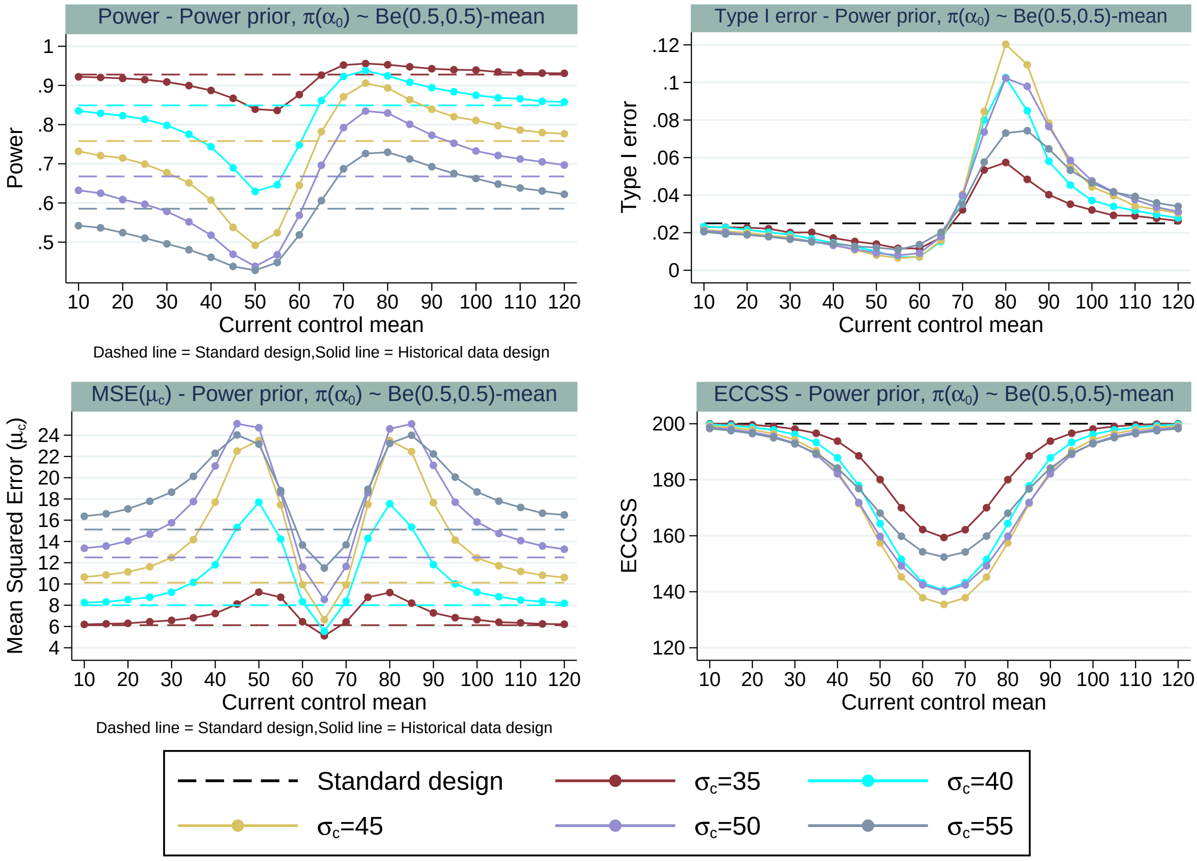


Figure I.2: Comparison of the power, type I error, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the using the modified power prior with a Beta(0.5, 0.5) prior on α_0 taking the mean of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.



Power - Power prior, $\pi(\alpha_0)$ ~ Be(0.5,0.5)-mean

Type I error - Power prior, $\pi(\alpha_0)$ ~ Be(0.5,0.5)-mean

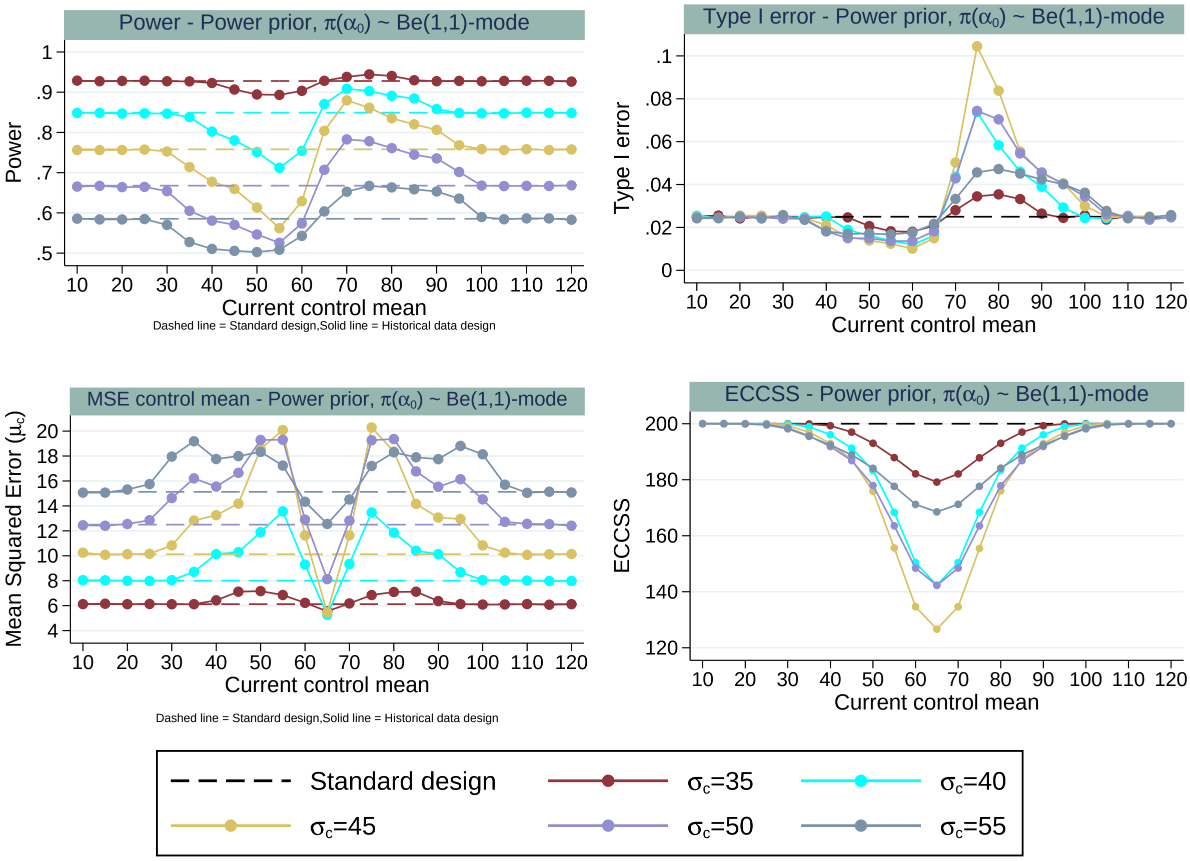
Dashed line = Standard design,Solid line = Historical data design

MSE(μ_c) - Power prior, $\pi(\alpha_0)$ ~ Be(0.5,0.5)-mean

ECCSS - Power prior, $\pi(\alpha_0)$ ~ Be(0.5,0.5)-mean

Dashed line = Standard design,Solid line = Historical data design

Figure I.3: Comparison of the power, type I error, mean squared error and expected current control sample size across different true means and standard deviations in the current trial control arm for the adaptive design using the using the modified power prior with a Beta$(1, 1)$ prior on α_0 taking the mode of the posterior distribution of α_0 as fixed weight and a standard design incorporating no historical data. Example, historical data $\bar{x}_h = 65, \hat{\sigma}_h = 45, n_h = 100$, $n_c = n_t = 200$, $n_{c1} = 100$, $nmin = 20$ and treatment effect $\mu_t = \mu_c + 12$.


Power - Power prior, π(α0) ~ Be(1,1)-mode
Power
Current control mean
Dashed line = Standard design,Solid line = Historical data design
Type I error - Power prior, π(α0) ~ Be(1,1)-mode
Type I error
Current control mean
MSE control mean - Power prior, π(α0) ~ Be(1,1)-mode
Mean Squared Error (μc)
Current control mean
Dashed line = Standard design,Solid line = Historical data design
ECCSS - Power prior, π(α0) ~ Be(1,1)-mode
ECCSS
Current control mean
Standard design
σc=35
σc=40
σc=45
σc=50
σc=55