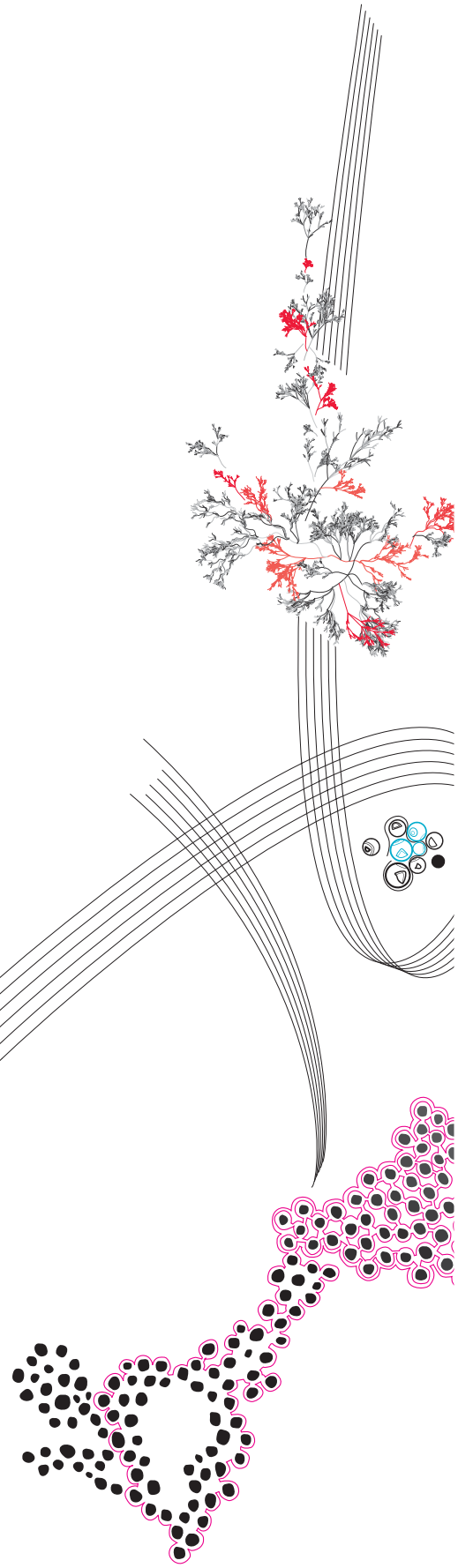

BSc Thesis Applied Mathematics

Computation of the Control Group Size in Randomized Controlled Trials

Irene Gabi van de Zande

Supervisor: R. J. Boucherie

June, 2019

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Preface

Before you lies my Bachelor Thesis, on which I have been working for the past twenty weeks. I would have liked to have had more time for the assignment.

I would like to thank my supervisor Richard Boucherie for his guidance during the process. In addition I would like to thank my friends and family for their support, and the coffee machine in the university library for almost always being able to provide me with my daily dose of caffeine.

Computation of the Control Group Size in Randomized Controlled Trials

Irene Gabi van de Zande*

June, 2019

Abstract

In this paper the size needed for the control group in a randomized controlled trial is investigated. This is done by studying the fundamental Bayesian statistics that are significant for correctly claiming causal inference and by looking into how and when historical information can be used, it is possible that the historical data needs to be adjusted for covariates before it can be of use. Historic borrowing methods are compared by their mean square error (MSE), type I error, and power. A computation of the optimal number of patients needed for the current control group and a ratio including the effective sample size are included. When historical data is available on sufficiently similar previous studies, the current control group size can be reduced accordingly.

Keywords: Randomized Controlled Trial, Bayesian Statistics, Historic Borrowing, Covariate Adjustment.

1 Introduction

A new medical treatment must first get registered as a safe and effective medicine before it is allowed to use it. In order to get registered, information has to be gathered and the treatment has to go through three phases of trial studies before it can be submitted to the European Medicine Agency (EMA) in Europe or the Food and Drug Administration (FDA) in the United States of America.
The safety of the intervention will be tested in the phase I study and it takes several months to complete. The trial usually includes a small number between the 20 and 100 healthy volunteers. Side effects which may occur as dosage level increases are also investigated in this phase. In phase II the efficacy of the intervention is studied and the duration time and number of participants will be increased to somewhere between a few months and two years, and several hundred patients respectively. Phase III, the last phase, is where the randomized controlled trial comes in. This method represents the effectiveness and causal inference by looking at what difference the treatment would make with applying the standard treatment available, which is sometimes no treatment at all. The patients will be allocated into at least two groups, the treatment group and the control group by randomization and afterwards results will be analyzed. Ideally everyone involved with the trial is blinded, as to reduce errors due to biases. The duration of the study and the participants both increase even more. The trial will be conducted on hundreds to thousands of patients and because it is of such a large scale it may also take several years to complete.

*Email: i.g.vandezande@student.utwente.nl

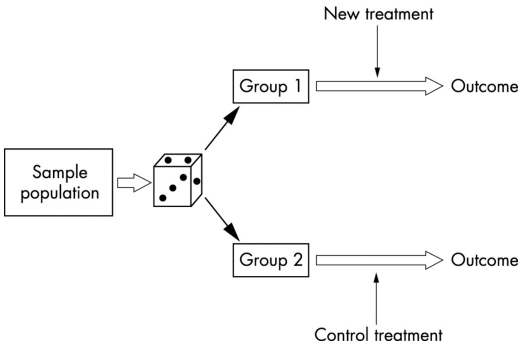

FIGURE 1: Allocation of sample population in a randomised controlled trial

It can happen that patients of a phase III trial continue receiving the intervention till it can be purchased in potential lifesaving cases when the intervention is pending for regulatory submission. [3] But the control group for this study then has not gotten treatment for several years. This might raise some ethical questions, considering that a phase III study takes quite long to complete. If it is recognized as unethical to stop giving patients treatment in possible matters of life and death, one might consider it to be unethical that there even exists a control group for whom it is decided by allocation that they will not receive the treatment.
Ordinarily when there is a fixed amount of patients participating in the trial and there is one treatment group and one control group, the group sizes are equal. But perhaps the size of the control group can be smaller than the treatment group, or include no control group in the current study at all. This could be beneficial for some, considering life threatening cases and when it is already clear that the intervention is of great significance. But when is that the case? Historical data on previous studies might provide sufficient and useable information for a new randomized controlled trial to alter the size of the control group. When it is wanted to make use of the historical data from these previous studies, they also need to be taken into account statistically. This can be done by making use of the Bayesian approach to causal inference instead of the ordinary frequentist approach where data are a repeatable random sample. With Bayesian reasoning data are fixed and the unknown parameters are described as probabilities. The aim of this paper is to provide an answer to the question of how the size of the control group of a randomized controlled trial should be computed. First further knowledge on the required statistics and causal inference are discussed in section 2. Afterwards in section 3 more information is provided on how historical information can be used to alter the size of the current control group in a trial while still being able to determine statistically correct the causal effect of the treatment relative to the control group.

2 Statistics of Randomized Controlled Trials

2.1 Mathematical Model of a Randomized Controlled Trial

A statistical model is provided for a randomized controlled trial, based on Holland's article [5]. The people included in the target population U for the new intervention are $u \in U$. The sample population $\tilde{U} \subset U$ are patients from the target population who are participating in the trial and $|\tilde{U}| = N$.
A patient can be exposed to a specific cause $S(u)$, or have certain properties or characteristics $A(u)$. Whether a patient has a characteristic is known upfront, but to what cause

the patient will be exposed needs to be determined. Exposure to treatment (or control) is denoted as $S = t$ (or c). The group of the sample population receiving treatment (or control) is denoted as $\tilde{U}_t$ (or $\tilde{U}_c$) where $|\tilde{U}_t| = N_t$ (and $|\tilde{U}_c| = N_c$).
Exposure to a cause must happen at some specific time or within a specific period of time. This means we have a state of pre-exposure and of post-exposure. The measure of effect of a cause Y is post-exposure. These values of post-exposure are potentially affected by the treatment t or control c. Which implies that causes have effects. $Y_t(u) = Y(t, u)$ is the value response after exposure to t, and $Y_c(u) = Y(c, u)$ is the value response after exposure to c. Thus if $S(u) = t$, $Y_t(u)$ is observed, and if $S(u) = c$, $Y_c(u)$ is observed.

The observed response on a patient u is $Y_{S(u)}(u)$, and the observed response variable is Y_S. Thus the observed data for each patient is (S, Y_S). When patients of a sample population are exchangeable with patients in the rest of the target population, the sample population can be used to estimate population parameters. $E(Y_t)$ is the average value of $Y_t(u)$ over all u in U, and $E(Y_S|S = t)$ is the average value of $Y_t(u)$ over only those u in U that were exposed to t. (S, Y_S) can give information on $E(Y_S|S = t) = E(Y_t|S = t)$. It is important to notice that $E(Y_t)$ is in general not $E(Y_t|S = t)$.

It is impossible to measure the effect of treatment and control on the same patient at the same time, thus one cannot always be 100% sure that there is actual causal inference. But with this model the statistical solution will replace the impossible to observe causal effect of t on one patient u with the estimate of the average causal effect of t over the whole sample population $\tilde{U}$. $Y_t(u) - Y_c(u)$ is the effect of t on u measured by Y and relative to c over the whole sample population. $T = E(Y_t - Y_c) = E(Y_t) - E(Y_c)$ is the average causal effect of t, relative to c. The prima facie causal effect $T_{PF} = E(Y_t|S = t) - E(Y_c|S = c) = E(Y_S|S = t) - E(Y_S|S = c)$ is the regression of Y_S on S. The term prima facie causal effect is used to determine the difference from the true average causal effect T and in general $T \neq T_{PF}$ [11].

In the model we have the variables S, Y_t, and Y_c and in the process of observation we have S, and Y_S. It is crucial for the analysis of causation that the distinction between the measurement process Y that produces the response variable, the two versions of the response variable Y_t and Y_c correspond to exposure to the cause, and the observed response variable Y_s is clear.

2.2 Associational Inference and Causal Inference

The first thing that needs to be stated is that association is not causation [1]. Association is the relation between two or more variables, whereas causation means that the change in one variable directly causes change in another variable. And although everything might have a cause, it does not mean that everything can be a cause. Patients have characteristics $A(u)$, and $Y_a(u)$ can be defined for all $u \in U$ where $A(u) = a$. The same holds for $Y_b(u)$, where for all $u \in U$ $A(u) = b$. A patient can have the characteristic a, b, or neither, but never both a and b. Hence the causal effect $Y_a(u) - Y_b(u)$ cannot be defined for characteristics, for any patient $u \in U$ [5].
For all experiments on causal inference it important that a control for comparison is included, otherwise no causal effect can be determined at all because two causes are required for the definition of an effect. How much can randomized clinical trials tell us about causation? Only when $Y_t(u)$ and $Y_c(u)$ can be defined, the causal effect $Y_t(u) - Y_c(u)$ is possible to determine. But $Y_c(u)$ and $Y_t(u)$ cannot be observed for the same patient u. Hence assumptions need to be made, which also bring uncertainty.

There is also a distinction between medical interventions such as a physical device or a drug. Take for instance a surgical implant such as a screw to repair a broken bone.

Patients with broken arms who did not get a screw implanted in their bone but received the standard treatment which is a simple cast, form the control group. The effect of a device such as a surgical implant is local and thus easier to predict. Now it can be easily seen if there is a causal effect $Y_t(u) - Y_c(u)$. The patients from the control group could in this case even have been from historical studies on which rehabilitation times have been kept in their dossiers. The assumption is then made that the cases of the historical broken arms are sufficiently similar enough to the current target population.
For drugs it is more difficult to make that assumption because the effects are less predictable than those of medical devices. Where for a new device previous trials which have been conducted in a different country, patient registries, previous studies, studies of the device on similar patient populations, and possibly nonclinical studies are potentially reliable sources, this will not be self-evident for a new drug.

2.3 Special cases

There are several statistical special cases that might occur in trials and are thus of interest for this paper. It is not that straightforward to make a correct statistical claim of causal inference.

2.3.1 Temporal stability and causal transience

When there is temporal stability and thus $Y_c(u)$ does not depend on time, and in addition the value of $Y_t(u)$ is not affected by prior exposure to c. Then one can simply measure $Y_t(u)$ and $Y_c(u)$ by sequential exposure to c then t. This means that the sample population can be allocated first to control and afterwards receive the treatment. $T = E(Y_t) - E(Y_c)$ is not affected by Y_c. For instance when t means that a broken arm is casted, and c means that it has not.

2.3.2 Unit Homogeneity

If $Y_t(u_i) = Y_t(u_j)$ and $Y_c(u_i) = Y_c(u_j)$ for all units u_i and u_j, then we assume unit homogeneity. Then the causal effect of t is $Y_t(u_i) - Y_c(u_j)$. This would mean that every patient would have the exact same effect as the others from the treatment, the same holds for the control. The assumption of unit homogeneity is not necessarily valid, but when the sample population is particularly carefully chosen to be almost identical and randomization has been carried out correctly the assumption might hold.

2.3.3 Independence

When the allocation of patients to treatment t and control c has been done such that the determination of which intervention u is exposed to is regarded as statistically independent of all other variables, then $E(Y_t) = E(Y_t|S = t)$ and $E(Y_c) = E(Y_c|S = c)$. Thus when the randomized allocation has been carried out correctly, it is plausible that S is independent of Y_t, Y_c and all other variables over U. Under this independence assumption $T = E(Y_t) - E(Y_c) = E(Y_S|S = t) - E(Y_S|S = c) = T_{PF}$. Now (S, Y_S) can be used to estimate T by taking the difference. Thus if randomization is possible, the average causal effect T can always be estimated.

2.3.4 Constant effect

Assume that the effect of t is the same on every patient, then $T = Y_t(u) - Y_c(u)$ is the causal effect for all patients $u \in U$. Then $Y_t(u) = Y_c(u) + T$ for all $u \in U$, and $E(Y_t|S = t) = T + E(Y_c|S = t)$. Hence $T_{PF} = T + E(Y_c|S = t) - E(Y_c|S = c)$. But $T_{PF} = T$ again only when S is independent and thus the patients have been randomly allocated. Thus only when the independence assumption also holds the true average causal effect can be estimated.

3 Applying Historical Information

3.1 Introducing Bayesian Statistics

When throwing a dice, the probability of throwing a six is expected to be $\frac{1}{6}$. It does not matter how many times each side has been thrown before, because we expect every side to have the same chance to land on. Suppose a friend asks what the chances are of the dice landing on a four. By approaching it from the frequentist point of view, it would still be $\frac{1}{6}$. But from the Bayesian approach and taking into account that this friend is known for his tricks, the chance could be expected to be way smaller and for instance $\frac{1}{50}$. The probability is then not seen as being derived from a long run frequency distribution but as degrees of belief in a proposition.

Suppose you have lost your keys, but there are eight spots where you might have left them. If there are a few spots of which you remember that you have left the keys there before, it would be smart to have a look there first. This is also a Bayesian approach, because what has happened before can be of use. The same is what is wanted to be reached for medical trials by looking into historical data on previous studies. What happened to patients in previous studies can be of great information for a current trial. With the Bayesian approach the uncertainty of an unknown quantity of a parameter θ of interest is represented by probabilities for the values possible. In case of medical trials this parameter would be the response variable $\theta = Y$. The prior distribution $p(\theta)$ then are the prior probabilities which are assigned to the possible values of the specific parameter θ before the trial. This distribution also reflects the knowledge on θ of the designers of the trial and is usually based on comparable relevant previous trials, thus the historical data D_H. $v \in V$ are in total the M patients from the historical trials, there can be several historic populations V_i $1 \leq i \leq H$, where $\sum_i^H V_i = V$. It is important that the patients from the historical studies and the current trial have the same chances of success and are hence exchangeable.
After data from the current trial D_0 are gathered, the prior probabilities can be updated by Bayes' theorem. This is how the posterior distribution $p(\theta|D_0)$ is computed: the posterior probabilities are the updated probabilities for values of the unknown parameter after new data has been observed.

3.2 Methods of Historic Borrowing

Below six methods which can be used for historic borrowing are explained.

3.2.1 Separate

In a separate trial the historical data is ignored and the analysis is done by only looking at the current data. Hence no historic borrowing takes place. The Fisher exact test is used

and there are $|\tilde{U}| = N$ patients in a trial, so ordinarily the group sizes of the treatment and control group are $|\tilde{U}_t| = |\tilde{U}_c| = N/2$.

3.2.2 Pooling

For a pooled trial the history control group is added and pooled together with the current sample population as if they are all included in the current trial, thus $\tilde{U} = \tilde{U} + V$ and $N = N + M$ patients are included in the trial. Because the M patients from the historical control group are automatically allocated to the control group in the current trial, more patients of the current trial can be allocated to the treatment group. Again the Fisher exact test is used.

3.2.3 Single arm

No control group is included for single arm trials and thus $\tilde{U} = \tilde{U}_t$, and $\tilde{U}_c = \emptyset$. Single arm trials are used for instance when it is unethical to have a control group. The null hypothesis is then obtained from the historical information, for example $H_0 : p = 0.65, H_1 : p > 0.65$, where 0.65 is conducted from the historical data and exact binomial test is used.

3.2.4 Test-then-pool

For test-then-pool trials it is first tested whether the historical control data differs significantly from the current control population. $H_0 : p_0 = p_H, H_1 : p_0 \neq p_H$, where p_0: from current control group and p_H from the historical data. This method now splits into either the separate or the pooling approach, depending on whether the null hypothesis will be rejected or not.

3.2.5 Power prior

It is possible that weight gets assigned to the historical data depending on the current data, which is called a power prior. This can happen when there is a lot of historical information available which needs to be downscaled to a decent ratio with the new data D_0. A weight between 1 and 0 is assigned to the historical data D_H, where 1 is equal to the pooling approach and 0 to the separate approach. $\tilde{U} = \tilde{U} + wV$, with $w \in [0, 1]$. Hence the historical data will be pooled with the current to some degree.

3.2.6 Hierarchical modeling

In hierarchical modeling Bayes' theorem is used, by means of the posterior distribution. It is used in clinical trials when there are several parameters that play a role. First parameters of the prior distribution, called hyperparameters, are computed from which distributions then again the hyperpriors will be computed.
These sort of methods are sometimes used in clinical cancer trials when several drugs are combined in one trial [10] and in section 3.5 a clinical trial on a cardiovascular device with hierarchical borrowing on a has been worked out.

3.3 Comparison of Historic Borrowing

To obtain significant results from a trial a low mean square error (MSE) and type I error are wanted, and a high power. In figure 2 three graphs are shown in which the MSE, type I error, and power for a detection of 12% improvement on the treatment group are set

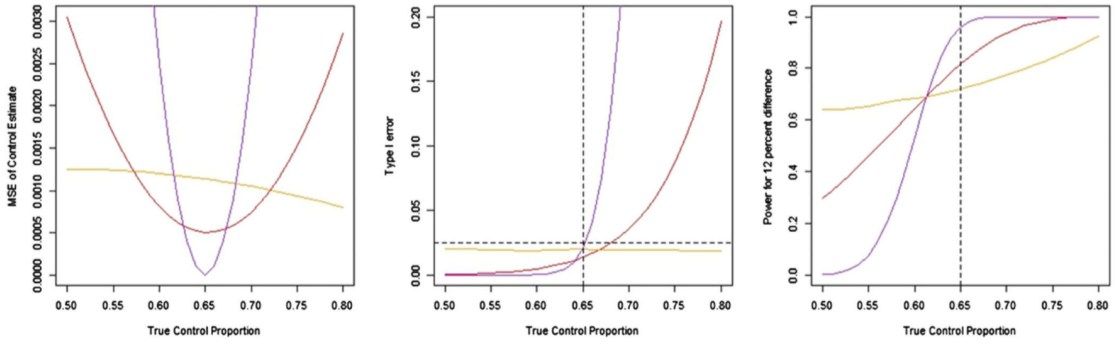

FIGURE 2: Comparison of the MSE, type I error, and power for separate (yellow), single arm trial (purple), and pooled (red) designs[6].

against the current control rate of separate trials, pooled, and single arm trials are put together. It is assumed that there are $N = 200$ patients included and from the historical data $p_H = 0.65$ is set at $\alpha = 0.025$, these values are also shown with the dashed lined.

For the separate trial (yellow) where none of the historical information is taken into account the mean square error does not change much but has a maximum at ... = 0.5. The type I error is also flat and stays a bit below 0.025, because of the design of the trial. But the power however does change over the true control rate and increases as the control group increases.

The single arm trial (purple), which can of course be seen as the exact opposite of the in terms of use of the historical data available, shows different behaviour. Because this trial assumes that the control rate is 0.65 has no mean square error when the current control rate is the same, but increases rapidly as the current rate is further from the historical. Thus when there is drift in the trials, first the results should be calibrated when wanting to use a single arm design. The type I error stays small till the historical rate is reached and than grows uncontrolled. The power is reduced for current rates below the found historical rate, but when the current rate is larger success can be declared earlier.

When the prior information is pooled with the current trial (red), the MSE is lowered around the historical rate. Not as much as the single arm trial and it also has a wider range where the MSE is below the separate trial and thus reasonably small. For the type I error again the pooled trial behaves similar to the single arm trial, but has a wider range around the historical rate where the type I error is still smaller than α and is controlled a bit longer as the slope is less steep. The power of the pooled trial is larger when the current control rate lies around the history control rate compared to a separate trial, but still lower than single armed. As for current control rate smaller than the history rate the power decreases, but more gradually.

When valid data is borrowed there is a region about the historical control rate where the MSE and type I error are lowered, and the power is greater than for a separate trial. In this region borrowing is dominant because apparently the prior information is a correct estimate of the current control rate, which means that using the information can only be of assistance and have no disadvantages. When the current control is smaller the pooled trial has a small type I error but also a reduced power, and is only greater than separate trials around the historical. Thus it becomes a bit harder to claim trial success when they are both lowered. When the current control rate is larger than the historical rate, it is easier to declare success. In general single arm trials perform worse and the larger the region where borrowing is dominant, the more appealing it is to use such a method.

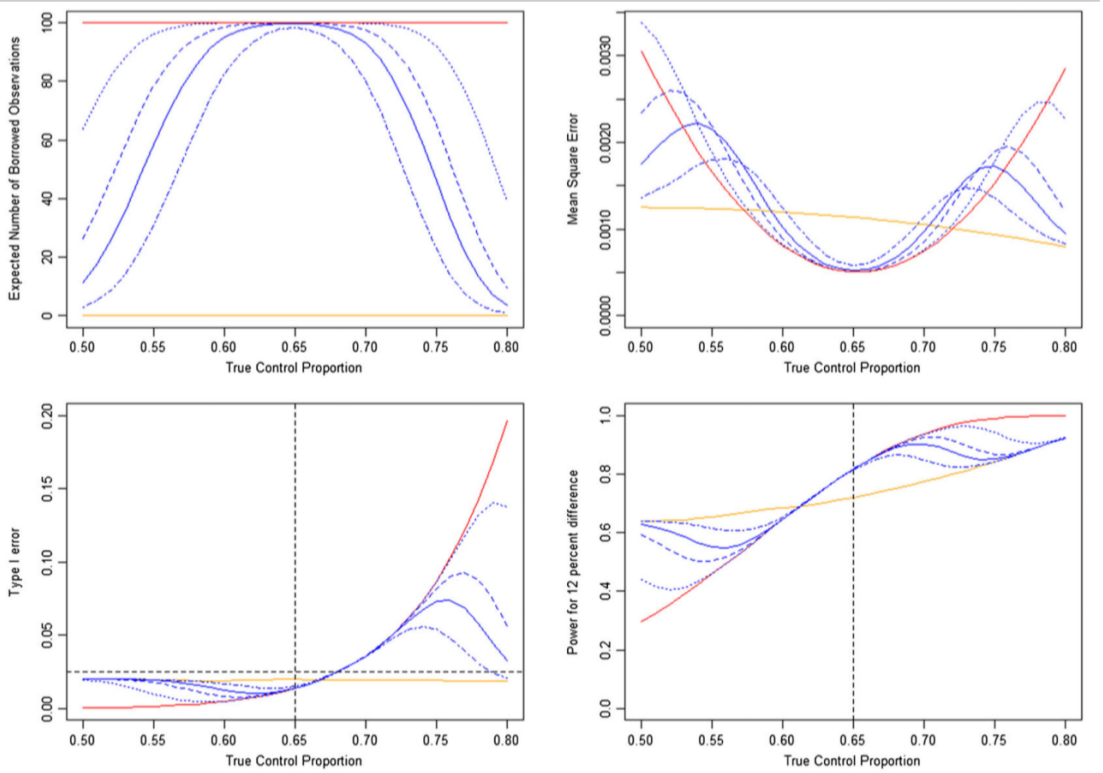

FIGURE 3: Comparison of borrowing, MSE, type I error, and power for separate (yellow), pooled (red) and test-then-pool with sizes of $\alpha = 0.20, 0.10, 0.05$ and 0.01 (blue)[6].

In figure 3 we look again at the MSE, type I error, and power but the single arm trial is left out and test-then-pool trials of sizes $\alpha = 0.20, 010, 0.05$, and 0.01 are included (blue). Because for test-then-pool trials it is either pooled or separate depending on the null hypothesis, the region where information is borrowed can be altered by changing the size of the test. An improvement of this method in comparison to pooled trials is that there is a restriction on the borrowing, which results in dynamic borrowing. If the current control arm is significantly different from the historical control arm data will not be borrowed.
In the graphs of the MSE, type I error, and power it can all be seen that around the historical rate it behaves like a pooled trial, but when the current control rate is either significantly smaller or larger it converges to the separate trial.

In figure 4 it is shown how power priors influence the MSE, type I error, and power. The aim of power priors is to downweight the historical information to some degree. Whereas in figure 3 it could be seen that the test-then-pool moves from pooled behaviour around the historical rate to separate when the current rate differs significantly, it now can be seen that the power priors lie in a range between pooled and separate trials depending on their weight parameter. The region where borrowing trials dominates the separate trial is again evident in the figures.

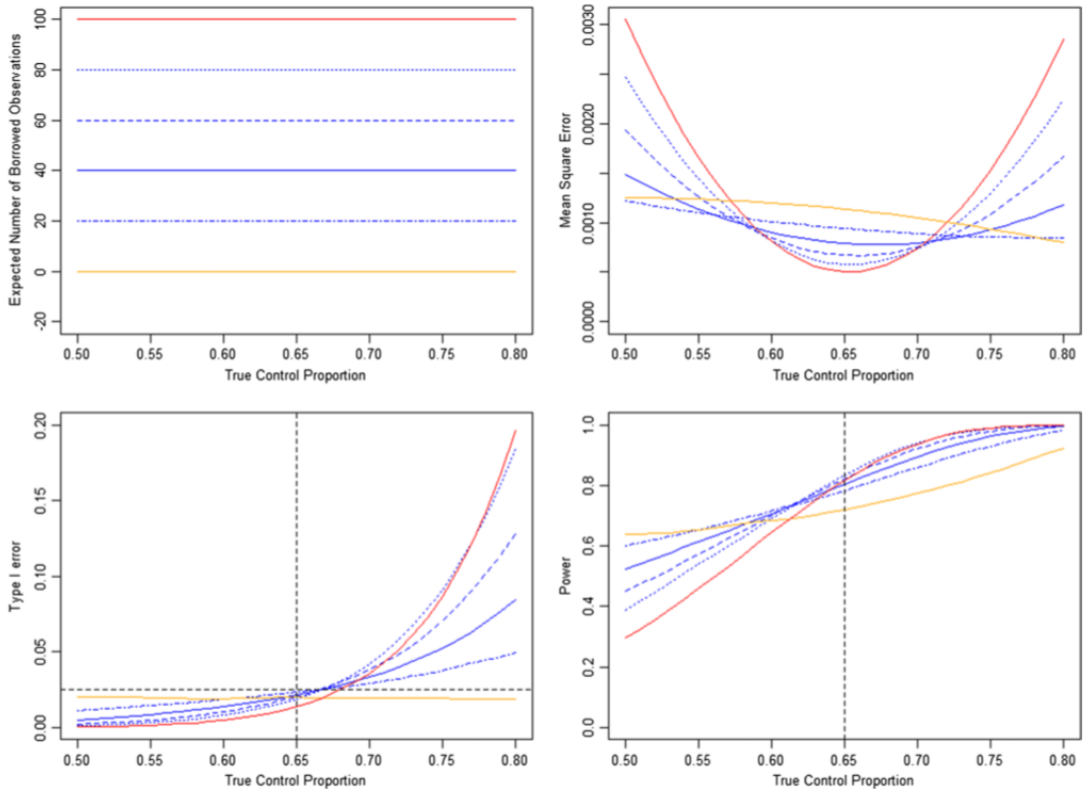

FIGURE 4: Comparison of borrowing, MSE, type I error, and power for separate (yellow), pooled (red) and power priors with weight parameters of $20\%, 40\%, 60\%$ and 80% as shown in the upper left panel[6].

3.4 Optimal Control Group Size Computation

The data on the historical control group can be combined with the data of the current treatment group. A model to compute the required sample size of the current control group in case of historic borrowing is provided.

The response variable Y is still considered, where $y \in Y$ is the observation from the random variable and normally distributed $Y \sim N(\mu_Y, \sigma_Y^2)$ with μ_Y and σ_Y^2 unknown. Let the current and historical populations also have normal random variables for Y, with corresponding unknown means μ_t, μ_c, and μ_H and variances σ_t^2, σ_c^2, and σ_H^2. The observable means of these are denoted as $\bar{y}_t$, $\bar{y}_c$, and $\bar{y}_H$ where $\bar{y}_t = \sum_t y/N_t$, $\bar{y}_c = \sum_c y/N_c$, and $\bar{y}_H = \sum_H y/M$.

The goal of the trial is to obtain an accurate estimate of the relative effectiveness of the treatment; $T = E(Y_t) - E(Y_c) = \mu_t - \mu_c$. With no valid information on historical trials available, $\bar{y}_t - \bar{y}_c$ is the point estimate. Because there can be no historical bias and it has minimal standard error it is the best estimate for this case. But if there is prior information available, $\bar{y}_H$ will not be ignored. The aim is to optimize the $\mu_t - \mu_c$ estimate. But there is a potential bias which we have to take into account in the prior information since we cannot assume the historical data to be 100% reliable, this unknown bias is denoted as $b = \mu_c - \mu_H$. In trials it is unknown how the bias is distributed, but we set the mean to zero because it the sign not known either. However the variance σ_b^2 is set fixed and rather a bit larger than smaller since we are never sure of the certainty of the historical data.

Suppose prior to the historical information μ_H is ignored, μ_H and is uniformly dis-

tributed and has a normal posterior distribution mean $\bar{y}_H$ and variance σ_H^2/M, then $\mu_c = \mu_H + b$. Thus after including the prior information but before including the randomized current controls, μ_c is normally distributed, $\mu_c \sim N(\bar{y}_H + b, \sigma_H^2/M + \sigma_b^2)$, and the prior distribution $p(\mu_c)$ will be adapted by the data from $\bar{y}_c$ and σ_c^2. Thus the posterior distribution of μ_c is normally distributed, $p(\mu_c|\bar{y}_c, \sigma_c^2) \sim N(\bar{y}_k, v_k^2)$ with

$$\bar{y}_k = \frac{(\sigma_H^2/M + \sigma_b^2)\bar{y}_c + (\sigma_c^2/N_c)\bar{y}_H}{(\sigma_c^2/N_c) + (\sigma_H^2/M) + \sigma_b^2}, \tag{1}$$

$$v_k^2 = \frac{1}{(N_c/\sigma_c^2) + \frac{1}{\sigma_b^2} + (\sigma_H^2/M)}. \tag{2}$$

Because if $x|\theta$ has a normal distribution, $x|\theta \sim N(\theta, \sigma^2)$, and the prior is normal, $\theta \sim N(\mu, \tau^2)$. Then the posterior also normal and the Bayes estimator is given by

$$\hat{\theta}(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \text{ where } \sigma^2 = \frac{\sigma_H^2}{M} + \sigma_b^2, \mu = \bar{y}_c, \tau^2 = \frac{\sigma^2}{N_c}, \text{ and } x = \bar{y}_H. \tag{3}$$

We can rewrite $\bar{y}_k = \frac{\bar{y}_c + W\bar{y}_H}{1+W}$, where $W = \frac{\sigma_c^2/N_c}{\sigma_H^2/M + \sigma_b^2}$. This means that the optimal point estimate is indeed a weighted sum of the sample means $\bar{y}_c$ and $\bar{y}_H$. To assess the relative effectiveness we still have to look at the difference between μ_t and μ_c. Assume a uniform prior for μ_t, then its posterior is normal with mean $\bar{y}_t$ and variance σ_t^2/N_t. Thus the posterior of $\mu_t - \mu_c$ is normal with mean $\bar{y}_t - \bar{y}_k$ and variance $\sigma_t^2/N_t + v_k$, with $\bar{y}_k$ and v_k^2 defined before.

As mentioned before, in a traditional randomized controlled trial $N_t = N_c$, which will result in $\bar{y}_t - \bar{y}_c$ having a minimal variance when $\sigma_t = \sigma_c$ holds. But due to the use of the historical information the current control group may small down and $N_t > N_c$. From the formulas above which still apply, a N_c should be chosen such that the variance is minimized. In other words: Under the constraint that $N = N_t + N_c$ is fixed, design a trial s.t. $\mu_t - \mu_c$ is estimated as precisely as possible [9].
$p(\mu_t - \mu_c|\bar{y}_c, \sigma_c^2) \sim N(\bar{y}_t - \bar{y}_k, \sigma_p^2)$ and Then the variance σ_p^2 can be rewritten as

$$\sigma_p^2 = \sigma_t^2/N_t + v_k^2 = \frac{\sigma_t^2}{N - N_c} + \frac{\sigma_c^2(\sigma_h^2/M) + \sigma_b^2}{\sigma_c^2 + N_c(\sigma_H^2/M + \sigma_b^2)}. \tag{4}$$

The optimal value of N_c is obtained by setting the derivative of equation 3.4 equal to zero. This minimizes the variance of $\mu_t - \mu_c$. The optimal value then becomes

$$N_c = \frac{\sigma_c}{\sigma_c + \sigma_t}\Big(N - \frac{\sigma_c\sigma_t}{\sigma_H^2/M + \sigma_b^2}\Big). \tag{5}$$

If the variances are set equal and thus $\sigma_H = \sigma_c = \sigma_t = \sigma$ the formula can be simplified to

$$N_c = \frac{1}{2}\Big(N - \frac{M}{1 + (\sigma_b^2 M/\sigma^2)}\Big), \quad N_t = N - N_c = \frac{1}{2}\Big(N + \frac{M}{1 + (\sigma_b^2 M/\sigma^2)}\Big). \tag{6}$$

These sample sizes N_t and N_c have been computed for a single response variable Y. However, for some trials it is necessary to look at several response variables. For example with a severe stage of breast cancer it will be useful to look at the patient survival time but also the maximum shrinkage of the tumor is interesting when assessing the treatment. When using this method to compute the group sizes, the sizes are determined for each response variable separately. At the trial design weight relative to the importance of the different variables should be assigned such that a weighted mean group size can be computed. For this example it is for the patient more of importance that he or she survives longer than that the tumor shrinks more.

3.5 Hierarchical Borrowing Example

A study of a new cardiovascular device on 200 patients is considered. It is tested how many patients will suffer within thirty days after placement of the device from a major adverse cardiovascular event (MACE), such as a nonfatal stroke, nonfatal myocardial infarction, or cardiovascular death. It concerns hypothetical data on the MACE rate in thirty days for six historical studies and one current study[8]. A synthetic control group is derived from the historical studies and will be used as the control group for the new study. The historical population means are exchangeable with the mean of the synthetic control group that will be computed.

The claim to be shown is that $p_0 < 0.249$, where $p_j, j = 0, ..., 6$ are the rates of how many patients suffer from a MACE within thirty days after the device has been placed. The seven studies in total are single-arm studies and concern no separate control group within the same study. From the historical studies together the prior probability p_0 of the new study was computed. The prior distribution is conditional on the six historical studies and does not include the new study.
The patients who suffer from a MACE event within thirty days after placement have a binomial distribution $s_j \sim Bin(M_j, p_j)$ where $j = 0, ..., H$, $j = 0$ concerns the new study and $j = H = 1, ..., 6$ concern the six historical studies. $logit(p_j) = \mu_j \sim N(\mu_0, \sigma_\mu^2)$, $\mu_0 \sim N(0, 1000)$ is the mean, and $\sigma_\mu^2 \sim \Gamma(0.001, 0.001)$ the precision.

TABLE 1: Hypotheticla data on the 30-day MACE rate for 6 historical (HS) and one new study of a cardiovascular device [8].

Device	Number of patients	Events	Rate
New	200	NA	NA
HS1	135	20	0.15
HS2	260	55	0.21
HS3	1960	325	0.17
HS4	415	60	0.15
HS5	205	43	0.21
HS6	25	5	0.20

From table 2 we can see that the prior probability of the claim $p_0 < 0.249$ is 97.3%. Because the results from the historical studies are really close to each other and thus the variance between studies μ^2 is small, the prior probability is really high and close to the success criterion of 0.975. But this is the case for hierarchical modeling without a weight being given to the historical data. Otherwise the previous studies would suggest that a clinical study of the new device is not required. But a new device cannot be approved by previous studies alone, some current study has to be included. Hence the prior information is downweighted.
A quick solution is to change the hyperprior on the between study precision to $\sigma_\mu^2 \sim \Gamma(10, 10)$, which causes the prior probability to drop to 0.674. This results in a much smaller estimate of the between study precision 220.7 (the posterior mean conditional on historical studies only), thus less prior information will be borrowed when estimating p_0.

TABLE 2: Summary of prior distribution for 30-day MACE rate for the new device [8].

Parameter	Mean	SD	2.5%	97.5%
p_1	0.176	0.034	0.121	0.252
$0.249 - p_1$	0.073	0.034	- 0.003	0.128
$p_1 < 0.249$	0.973	0.163	0.0	1.0

TABLE 3: Summary of prior distribution for 30-day MACE rate for the new device after discounting the prior information [8].

Parameter	Mean	SD	2.5%	97.5%
p_1	0.215	0.155	0.028	0.608
$0.249 - p_1$	0.034	0.155	- 0.359	0.221
$p_1 < 0.249$	0.674	0.469	0.0	1.0

3.6 Effective Sample Size

When considering the same variable θ as before, the effective sample size (ESS) equals the ratio of the variation of θ on the current data D_0 and given that historical data D_H are ignored over the variation when historical studies are taken into account, multiplied by the sample size of the current study N [7].

$$ESS = N\frac{Var(\theta|D_0, D_H ignored)}{Var(\theta|D_0, D_H utilized)} \tag{7}$$

By computing this effective sample size (ESS) we are able to calculate how many extra patients the prior distribution was worth. We use again the cardiovascular device example with a binomial-normal hierarchical distribution from before. Suppose that of the in total 200 patients included in the trial 40 suffer from a MACE within 30 days after placement of the device. The posterior mean and standard deviation for the new MACE rate p_0 are now 0.184 and 0.018686, respectively. But if the diffuse hyperprior distribution $logit(p_0) = \mu_0 \sim N(0, 1000)$ was used, the posterior mean is 0.200 and has a posterior standard deviation of 0.02826. Then $ESS = 200 * (0.02826/0.018686)^2 = 457.4$. This would mean that 257.4 patients were effectively borrowed from the historical studies. Since this is more than the number of patients included in the current study, it is another sign that the prior information should be downweighted in this hierarchical model.
For clinicians who are not experts in statistics this is also a useful instance to check whether the historical data is too informative or not.

Equation 3.6 can be rewritten into the following:

$$\frac{ESS}{N} = \frac{Var(\theta|D_0, D_H ignored)}{Var(\theta|D_0, D_H utilized)} \leq C \leq 2. \tag{8}$$

As mentioned before, it is not wanted to that there are more patients effectively borrowed than participating in the current itself, thus a boundary can be set on this ratio.

3.7 Adjustment of covariates

3.7.1 Historical data: Statistical Analysis

What is wanted is to compute over a fixed sample population of patients the optimal number of patients to randomize to the current treatment or control group. As seen in the previous section, optimal allocation often means that more patients of the current sample population will be allocated to the treatment group then to the control due to the extra information on the control arm that is obtained via historical information.
The historical information is only used when it is considered to be useful, however sometimes first the historical data needs to be adjusted before it can be used. For instance assume that a new weight loss device has been invented, but there have also been studies on previous versions of such a device which are considered to be similar enough. However, the patients involved in the historical studies are not comparable to the current sample population due to their initial weight. If the current sample population is significantly heavier than the historical population, the data needs to be adjusted for this covariate before the prior information can be used.

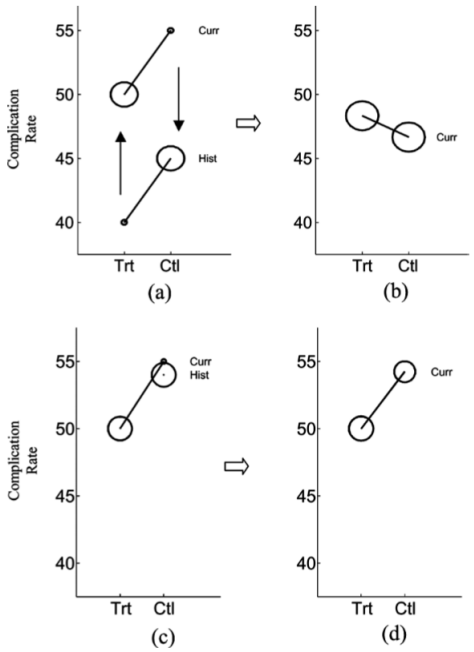

FIGURE 5: Comparison between borrowing within arms with adjustments for health-related covariates (panel (c) then panel (d)) and without (panel (a) then panel (b)) [8].

In figure 5 where the circles are in proportion with the effective sample sizes, the upper two graphs (a) and (b) where information is borrowed without covariate adjustment show that the current control rate is decreased towards the historical control rate, and the historical treatment rate is increased the other way around. In (c) there has been adjustment for health-related covariates first, which results in less separation between the current and historical control rates before the borrowing takes place in (d). Borrowing without adjustment for covariates can make it needlessly harder to make correct statements about inference because the historical information might bias the trial. Without the adjustment of covariates in this case no historical borrowing should take place.

Another example of how data can be adjusted such that it is useful for the current trial is by looking at the variances of and between patient populations. These is a difference between borrowing on treatment-control differences and within different arms. Because there might be a control group in all of the studies, previous and current, a treatment group might be only included in the latest. Although the study-specific treatment effects or study-specific control effects may vary seriously over all the studies, the study-specific differences between treatment and control are possibly quite constant [8]. This may result in a higher precision on the treatment-control difference. To calibrate the studies, the control is in this case used as baseline covariate. When borrowing takes place within the different arms of a study it could require further adjustment for other baseline covariates as explained before.

3.7.2 Design stage: Primary Analysis

The European Union also states several recommendations for adjustment on baseline covariates in clinical trials. Data can be adjusted either at the design stage of a trial, or afterwards when statistical analysis is performed. When designing the trial and at primary analysis, not too many covariates should be included. Adjustment for these covariates usually result in stronger and more precise evidence [4], which means smaller P-values and narrower confidence intervals. For smaller populations even fewer, and allocation to either treatment or control should be done by minimization [2]. This is an allocation method due to which the groups will be more balanced. Patients are assigned to a group which will result in the best balance of the covariates that are included. This way by adjusting the group allocation during the randomization process the covariates are accounted for already in the beginning. All covariates measured after randomization should not be included in the primary analysis of the trial [2]. If the allocation has been correctly randomized and any imbalance is observed, it should be considered a random phenomenon and not a proper reason to include this covariate in the primary analysis of this trial. Although for further research in the field the covariate may suggest to be included.

4 Discussion

It is difficult to state whether there is causal inference in a clinical trial, but when including a control group such as in randomized controlled trials it becomes possible to estimate. The size of the control group, whether it is from the current trial or from historical studies, and what the control group will receive (standard treatment or nothing at all) still needs to be determined.

When there is data on historical studies available, it first needs to be determined whether the historic population is comparable with the sample population of the current trial. If this is not the case, historic borrowing should in general not take place. Except when the historicla data is not directly applicable but there are covariates for which can be adjusted such that the prior information can be used in the clinical trial.
When historic borrowing can take place, it needs to be determined in what manner and how much. Thus what method of historic borrowing should be used?

As could be seen in section 3.3, there is a region around the historical rate where historical borrowing is favourable over a separate trial where none of the prior information is utilized, provided that the historical data provides valid information for the trial. In this

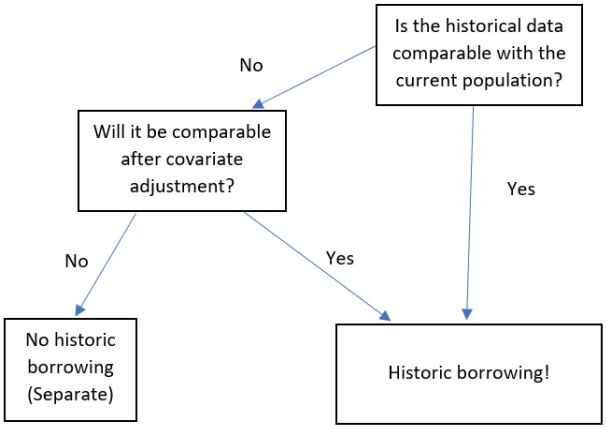

FIGURE 6: Can historical data be borrowed?

region the MSE and type I error are lowered and it has a higher power. But when the current control rate is not significantly similar to the historical rate, the historical data should not be of great influence or have none at all. By using the test-then-pool borrowing method the range of the region where historic borrowing is more beneficial can be altered by changing the size of the test α. If the current control rate lies not in that region, it will tend to the results of a separate trial. For power prior borrowing, where the historical information is downweighted to some degree, the region where borrowing dominates remains.
For all borrowing methods, when the current control rate is smaller than the historical rate the MSE increases whereas the type I error and power decrease. For a current control rate larger than the historical rate the MSE again increases, and the same happens to the type I error and power in opposition to when the current control is smaller. A small MSE and type I error, and a high power are wanted for valuable results.

When a few assumptions are made, a formula for the optimal control group size can be computed, dependent on the variances and total number of patients the current study as the historical studies.

$$N_c = \frac{\sigma_c}{\sigma_c + \sigma_t}\Big(N - \frac{\sigma_c \sigma_t}{\sigma_H^2/M + \sigma_b^2}\Big). \tag{9}$$

This holds under the assumptions that the response variable, prior distribution, and posterior distribution have a normal distribution although this is not entirely certain. Also the bias is unknown but still needs to be taken into account but cannot be accurately estimated.

By using the ESS a constant C can be determined by which the ratio needs to be restricted. Ultimately this is 2, but it could be smaller.

$$\frac{ESS}{N} = \frac{Var(\theta|D_0, D_H ignored)}{Var(\theta|D_0, D_H utilized)} \leq C \leq 2. \tag{10}$$

What could also still happen is that during a clinical trial interim analysis takes place. When there is already sufficient data to reject or accept the null hypothesis with enough certainty, the trial could be stopped. This can lead to faster conclusions and shorter trial time. Looking into this more precisely and determining when

5 Conclusions

By using historical information the size of the current control group of a randomized controlled trial can be altered and mainly decreased. For cases where the historical studies are comparable but the data needs to be adjusted before historic borrowing can take place. There are several methods which can be used to choose how much information will be borrowed.
Depending on how close the current control rate lies to the historical control rate, the MSE, type I error, and power behave differently for each borrowing method. In the region where they are very near each other borrowing is dominating, whereas when they lie further away from each other it is the other way around and the MSE of borrowing methods always increases. When the current control rate is smaller compared to the historical rate, the type I error and power are reduced. When the current rate is greater than the historical, both the type I error and power are raised.
The optimal control group size of the current trial can also be computed, together with a bound on the effective sample size.

References

[1] G. A. Barnard. Causation. *Encyclopedia of Statistical Sciences*, 1:387–389, 1982.

[2] J. M. Bland D. G. Altman. Treatment allocation by minimisation. *BMJ*, 2005.

[3] Food and Drug Administration. Guidance for institutional review boards and clinical investigators. 1999.

[4] Committee for Medicinal Products for Human Use (CHMP). Guideline on adjustment for baseline covariates in clinical trials. 2015.

[5] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[6] B. Neuenschwander et al K. Viele, S. Berry. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat.*, 13(1):41–54, 2014.

[7] D. Malec. A closer look at combining data among a small number of binomial experiments. *Statistics in Medicine*, 20(12):1811–1824, 2001.

[8] Gene Pennello and Laura Thompson. Experience with reviewing bayesian medical device trials. *Journal of Biopharmaceutical Statistics*, 18(1):81–115, 2007.

[9] Stuart J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976.

[10] C. Hamada S. Yada. Application of bayesian hierarchical models for phase i/pahse ii clinical trials in oncology. *Pharm Stat.*, 16(2):114–121, 2017.

[11] P. C. Suppes. A probabilistic theory of causality. 1970.