Jheronimus Academy of Data Science

Adaptive clinical trial design in a Bayesian framework

Maury Kros

July 2021

Main supervisor

Maurits Kaptein

Second supervisor

Xynthia Kavelaars

Abstract

Conventionally clinical trials are based on fixed samples. The counterpart to fixed sample trials are adaptive trials where efficacy of a treatment is evaluated periodically during interim analyses as data accrues over time. Depending on the observations at each interim analysis, the clinical trial can stop or continue. This involves testing an hypothesis multiple times and thus introduces the problem of multiplicity. That is, the inflation of the type I error rate. The extent to which this occurs depends on the number of interim analyses conducted and design parameters of a trial. In this paper we present a Bayesian framework in which the efficacy of a treatment can be established for binary outcomes in a adaptive manner. For this Bayesian framework we provide how decision rules are implemented and frequentist operating characteristic can be achieved. We compare our Bayesian framework versus a fixed sample size test for proportion to evaluate it's performance. Subsequently, we demonstrate that the the problem of multiplicity persists in our Bayesian model and how we can control it. Specifically, we demonstrate how decision rules and the patient scheme can be altered to control the type I error rate inflation. We discuss the limitations of our work and suggest points for further research.

1 Introduction

Clinical trials aim to establish the efficacy of a treatment. Conventionally, a trial has a fixed sample size which is established during the design phase. Determining the sample size is often done within the null hypothesis significance testing (NHST) framework. In this framework we consider the population parameter of interest to be fixed (e.g. the response rate of a treatment). We test a null hypotheses versus an alternative hypothesis in order to establish whether a treatment is efficacious. The hypothesis are usually constructed in a way such that rejection of the null hypothesis means concluding that the treatment is efficacious. In this framework the null hypothesis can be rejected when it should not be rejected (type I error) or we fail to reject it when we should (type II error). We often control the rate at which we make these errors. In a fixed sample size trial the hypothesis is tested once at the end once all observations have been collected. So, the rate at which we make a type I or type II error should be interpreted as the rate at which we draw incorrect conclusions, should we repeat the experiment over and over again. The type I error rate and type II error rate are controlled by specifying a significance level and power.

In a trial with a fixed sample size the sample size is function of the significance level, power and effect size. The significance level and power are chosen while the effect size is estimated. The sample size can be calculated as these three quantities are chosen. If a trial is started conclusions can only be drawn after trial stops. The trial can always terminate due to practical objections, but then we no longer draw conclusions with the same power. Adaptive trials, contrary to fixed sample size trials, allow researchers to adapt key aspects of the trial after interim analyses as data accrues over time. This flexibility can potentially reduce used resources, improve the chance that results will be scientifically or clinically relevant in comparison to fixed sample size trials [Thorlund et al., 2018] and allow faster product registration [Gaydos et al., 2009]. Furthermore, it provides researchers

a tool to revise assumptions made in the planning phase in case during interim analyses serious deviations from these assumptions are revealed [Bauer et al., 2016]. This can be the difference between continuing or halting a trial.

In this work adaptive trials are considered within the Bayesian framework. One advantage is that this approach allows for incorporation of prior information such that past clinical research can be incorporated in a clinical trial [Shi et al., 2019]. Furthermore, it is a natural approach for sequential data as incoming data can be incorporated in decision making via the posterior distribution at any time. Also, in the Bayesian framework it is possible to test competing hypotheses via the use of Bayes factors [Schönbrodt et al., 2017].

Of particular interest is how to develop an adaptive trial in the Bayesian framework such that the trial has desirable frequentist operating characteristics. These characteristics are the type 1 error rate and power (or 1 - type II error rate) and the bias. The design parameters of a trial, such as group size and minimal number of patients, affect these frequentist operating characteristics. While in the Bayesian framework, decision making is based on posterior probability or Bayes factors, frequentist operating characteristics, are still important. Recall that the type 1 error rate and type 2 error rate represent the proportion of experiments where we draw an incorrect conclusion, had we repeated the experiment over and over again. Generally, researchers are more familiar with frequentist statistics. This in turn makes research more interpretable and makes publication more likely [Sanborn et al., 2014]. Another argument, particularly relevant practitioners, is that desirable frequentist operating characteristics are required by regulatory bodies such as the US food and drug administration (FDA) and the European Medicines Agency (EMA) for drug development and subsequent registration [Thorlund et al., 2018].

Most of the research done so far on adaptive clinical trials in the Bayesian framework focuses on

the frequentist operating characteristics and in particular how to control these [Zhu and Yu, 2017, Shi et al., 2019, Yin et al., 2017]. The main problem is that in an adaptive trial in the Bayesian framework, a hypothesis is continuously tested over time which can inflate the Type I error rate. This is often referred to as the problem of multiplicity.

The goal of this paper is to provide researchers with insights on how to develop an adaptive clinical trial in the Bayesian framework. To that end we address three points. We show 1) how trial designs in a Bayesian framework perform versus a trial based on a fixed sample size calculation. Then, we show 2) how an adaptive trial is affected by the problem of multiplicity. That is, the inflation of the type I error rate as multiple interim analyses are conducted. Finally, we show that 3) the inflation of the type 1 error rate can be controlled by changing design parameters of the trial. Specifically, we demonstrate that the type 1 error rate can be controlled by altering the decision rule and increasing the sample size.

In this work a decision procedure is presented for univariate decision making for single binary outcomes. Tot this end a *Beta-Binomial* model is deployed. This model allows a straightforward interpretation in the sense that the efficacy of a treatment can be directly derived from the binary outcomes of the respective treatment. Furthermore, the Beta and binomial distribution form conjugate priors which is convenient from a computational point of view since a closed expression for the posterior distribution can be obtained, allowing for straightforward sampling.

This paper addresses points 1), 2) and 3) in the remaining sections. First, in section 2, a description of the Bayesian framework is provided. Subsequently, section 3 provides the description of the implementation of the framework in code and the results of the simulation. In the final section limitations and suggestions for further research are discussed.

2 Methodology

In this section the difference between a fixed design and an adaptive trial is addressed. We then shift to a Bayesian adaptive design. We present the specific implementation of a Bayesian model for binary responses in an adaptive clinical design. In doing so, we further address important concepts of adaptive design in the Bayesian framework.

2.1 Adaptive design

Usually in a clinical trial a desirable type I error rate α and power β are determined a priori. Subsequently, an effect size to be detected is chosen and the resulting sample size n can be calculated for the statistical test corresponding to that particular trial. This is what happens in a clinical trial with a fixed sample size. The effect size is a test specific measure that specifies how sure you are that the null hypothesis is false. The problem however is that the effect size is not known in advance and it has to be estimated. If we overestimate the effect size, we no longer draw our conclusions with the prespecified type I error rate α and power β. If we underestimate the treatment effect, we might use more resources then needed.

In an adaptive trial data accrues over time as cohorts of patients are tested. In practice, there is a maximal number of patients that can be used and thus there also is a maximal number of times an interim analyses can be conducted and a maximum cohort size. While there is a practical limit for the number of interim analysis that can be conducted, this does not have to be fixed in advance. The effective number of patients used is not known in advance. This is because after each cohort an interim analysis is performed. Depending on the trial, the trial can stop for either futility of superiority, depending on the decision rules in place. Or, if neither can be concluded, the trial continues with the next cohort.

The ability to change design parameters of a clinical trial is one of the appeals of an adaptive design. In the introduction some advantages were already highlighted. The ability to change an ongoing trial not only allows you to adjust incorrect assumptions but also allow termination of trial in case of ethical and/or economic objection. For example, if an underwhelming response in the first patients is observed the trial can be terminated since it is not ethical to expose patients to a sub-optimal treatment. Or, during research on a rare disease where treatment is expensive we might be able to stop the trial early after a few cohorts if the response is overwhelmingly positive, limiting the resources used. So, we potentially have to treat less patients to establish results as opposed to a fixed sample size trial. Another benefit is that we no long have to estimate an effect size. The advantages that improved testing can offer from a business and societal perspective are outlined appendix A.

Since in an adaptive trial a hypothesis is tested continuously over time at each interim analysis, these trials suffer from the issue of multiplicity [Armitage et al., 1969]. That is, the inflation of the overall Type I error rate α as more tests are conducted. This occurs since at each time the null hypothesis is tested there is a probability that the null hypothesis is incorrectly rejected. Suppose that we conduct a test at the $\alpha = 0.05$ significance level and say that there is no effect (i.e. we should not reject the null hypothesis). The probability of making a type I error is 0.05. Now, if we conduct two tests, the probability of making a type I error is the probability of making a type I error in either of the two tests. Or, 1 minus the probability that we do not make a type I error in both of the tests: $1 - (1 - 0.05)^2 = 0.0975 > 0.05$. So, the type I error rate increase as multiple hypothesis are tested. Or in the context of clinical trials, the type I error rate inflates as multiple interim analyses are conducted.

This problem is not unique to clinical trials but occurs in general when hypothesis are tested

continuously. Various remedies to solve this problem have been proposed from a frequentist point of view. These solutions rely on the maintaining more stringent type I error rates during interim analyses which are determined before the trial starts [Pocock, 1977, O'Brien and Fleming, 1979] or based on the type I error rates used at previous interim analyses [Kim and Demets, 1987].

The issue of multiplicity in a Bayesian framework is a widely researched and ongoing topic. The Bayesian approach offers a different way to test hypothesis that is based on the posterior distribution. Bayesian quantities associated with this hypothesis testing framework are not affected by optional stopping and looking at the data multiple times is permissible [Rouder, 2014, Wagenmakers et al., 2010, Edwards et al., 1963]. For a clinical researcher, frequentist operating characteristics are still important in a Bayesian adaptive design since these are required by regulatory bodies as mentioned earlier. More recently, various solutions for multiplicity have been proposed in a Bayesian sequential design [Shi et al., 2019, Yin et al., 2017]. In the section 2.2.3 we will discuss how frequentist operating characteristics are derived in a Bayesian framework.

Transition from a fixed design to an adaptive design introduces new design parameters such as the number of cohorts and cohortsize. How these additional parameters are chosen affects the outcome of the trial and it is not at all obvious which choices are best, if there are best choices at all. New questions arise such as how many patients must the first cohort contain? Or, how often can we conduct interim analysis while still maintaining the appropriate overall type 1 error rate? The latter question is how the issue of multiplicity can manifest itself in trial design.

2.2 Bayesian sequential design

2.2.1 Beta-Binomial model

In an adaptive clinical trial with binary outcomes the aim is to assess the response rate θ of a treatment, that is the success probability of the treatment. The proposed framework is based on Bayes rule where the posterior distribution is proportional to product of the likelihood function $P(data|\theta)$ and the prior distribution $P(\theta)$:

$$P(\theta|data) \propto P(\theta)P(data|\theta). \tag{1}$$

The posterior distribution $P(\theta|data)$ reflects the belief of response rate θ as data accrues over time. The prior distribution $P(\theta)$ reflects the belief of the response rate when no data has been collected and follows a Beta distribution with parameters α_0 and β_0:

$$P(\theta) = \frac{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{B(\alpha_0,\beta_0)}, \quad \text{where} \quad B(\alpha_0,\beta_0) = \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0+\beta_0)}. \tag{2}$$

Here Γ is the gamma function. The likelihood $P(data|\theta)$ follows a Binomial distribution with parameters n and θ:

$$P(data|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{3}$$

A Beta prior is a conjugate prior for a Binomial likelihood. If a prior combined with a likelihood distribution results in a posterior distribution that has the same distribution (possibly with different parameters) as this prior, the prior is said to be a conjugate prior to the used likelihood distribution. So here the posterior probability is again a Beta distribution (see equation 2)with parameters:

$$\alpha = \alpha_0 + \sum_{j=1}^{n} x_j \quad \text{and} \quad \beta = \beta_0 + (n - \sum_{j=1}^{n} x_j). \tag{4}$$

Where α_0 and β_0 are the parameters of the prior distribution, $\sum_{j=1}^{n} x_j$ is the number of positive responses observed in n patients and $j = 1, 2, \ldots n$ is simply the index of the patient. Note that x_j is a binary outcome so $x_j \in \{0, 1\}$. In a Bayesian framework the parameters α_0 and β_0 of the prior distribution are often called hyperparameters.

Now, the translation to the clinical setting is straightforward. Rather than n patients, a treatment is sequentially tested on a finite number of k cohorts of patients with cohortsize n_k. Let us denote the number of positive response in a cohort by y_k. After the i-th cohort (where $i \in \{1, 2, ..., k\}$) the posterior distribution is a Beta distribution with parameters:

$$\alpha = \alpha_0 + \sum_{k=1}^{i} y_k \quad \text{and} \quad \beta = \beta_0 + (n_k \cdot i - \sum_{k=1}^{i} y_k). \tag{5}$$

Intuitively, α at each interim analysis is the sum of the number of positive responses over all patients observed so far plus the α_0 chosen for the prior. Similarly, β at each interim analysis is the number of negative responses over all patients observed so far plus the β_0 chosen for the prior. Equations 4 and 5 are also provided to highlight that the prior distribution of i-th cohort is the posterior distribution of the $(i-1)$-th cohort. In figure 1 this can be observed as well. The exception is off course $i = 1$ since before the first cohort is observed a prior must be chosen. Also, these equations introduce various design parameters. Namely, the number of cohorts k, the cohortsize n_k, the minimum number of patients $n_{min} = n_k$ and maximum number of patients $n_{max} = k \cdot n_k$.

2.2.2 Decision rules

In a clinical trial the beliefs of the efficacy of the treatment change as the number of treated cohorts of patients increases. Here, superiority of the treatment is concluded if there is a strong conviction that the treatment has positive effect. For example, we might conclude superiority if we observe that the mass of the posterior distribution above $\theta = 0.3$ is greater than 0.95. More formally, superiority is concluded if enough mass of the posterior distribution, above some cutoff value $\phi \in (0, 1)$, exceeds a predefined threshold p_{cut}:

$$P(\theta > \phi \ |\text{data}) > p_{cut} \tag{6}$$

In section 2.2.1 it is discussed that $P(\theta|\text{data})$ has a closed form expression. Hence, $P(\theta > \phi \ |\text{data}) = \int_{\phi}^{1} P(\theta|data)d\theta$ can be computed directly for the Beta distribution where the it's parameters are found using equation 5. Essentially ϕ represents the lower limit for the response rate which is deemed acceptable to conclude that a treatment is effective. p_{cut} in turn represents the probability that the value of θ is at least this lower limit ϕ . Both ϕ and p_{cut} are additional design parameters of a clinical trial.

The decision rule in equation 6 is checked at each interim analysis, so after each cohort of patient is observed. The design parameters ϕ and p_{cut} are fixed in advance. With this decision rule early stopping can only occur for superiority. Either all k cohorts are treated and superiority cannot be concluded or after the i-th cohort superiority is concluded (again, $i \in \{1, 2, ..., k\}$). Once again we refer to figure 1 for an illustration of the applied decision rule to make our formal description more concrete.

2.2.3 Operating characteristics

Clinical trials should still provide their results with appropriate frequentist operating characteristics as mentioned earlier. These are the type I error rate, type II error (or power) rate and bias. Here, we outline conceptually what these entail and in section 3.1 we discuss how these are implemented.

Simulation of a trial is done by running a trial with the same design parameters n_{sim} (these runs are often called iterations). Since in simulation the true response rate θ that generates the population is known, we also know whether we should or should not conclude superiority. A type I error occurs when we incorrectly conclude superiority. A type II error occurs when we fail to conclude superiority, even tho we should have. Then, the type I error rate is the fraction of simulation runs in which superiority is incorrectly concluded. The type II error rate is the fraction of the simulation runs in which we failed to concluded superiority when indeed the treatment was efficacious. Then, the power is 1 minus the type II error rate. That is, the fraction of runs in which superiority is correctly concluded.

After each run of the trial we can end up with a different posterior distribution from which the response rate θ_r can be derived directly. The bias of a simulation is the difference between the true response rate θ that generated the population and the response rate θ_r found after a run of trial, averaged over all runs.

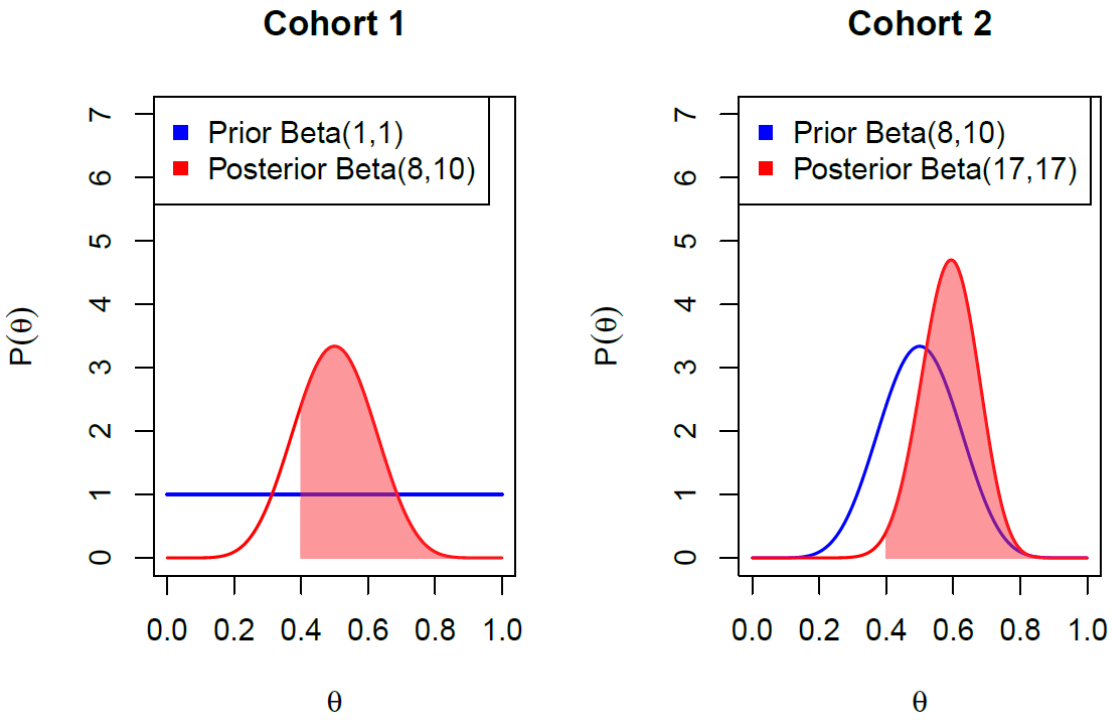

Figure 1: Evaluation of the posterior distribution after the first and second cohort with size $n_k = 16$. **Left:** It can be seen that initially a non-informative prior is used and as data from the first cohort is collected, the posterior distribution is found using equation 5. Then, we apply our decision rule and we chose $\phi = 0.40$ and $p_{cut} = 0.80$. In this case $P(\theta > \phi \,|\text{data}) = 0.64$, which is the shaded area in red. **Right:** After the second cohort of patients are observed we again apply the decision rule. We find $P(\theta > \phi \,|\text{data}) = 0.88$ and stop early concluding superiority. In this example we see visually that indeed the prior at the second interim analysis is the posterior at the first interim analysis. Also, we now see visually that the decision rule entails that we stop early if a enough mass of the posterior distribution exceeds a certain value of ϕ (0.40 in this case).

3 Numerical evaluation

In this section we evaluate the Bayesian design introduced in section 2.2. First, we look at the a fixed sample size calculation to determine group sizes and discuss how operating characteristics are found during simulation. Then we show how the framework is implemented. In the simulation results we focus on various Bayesian designs and demonstrate the effect of conducting multiple interim analyses. The section concludes with investigating how design parameters can be altered to combat the multiplicity problem introduced by conducting multiple interim analyses.

3.1 Simulation setup

3.1.1 Sample size calculation

We are interested in whether the response rate θ of a treatment exceeds a constant value c. In our sample we observe binary outcomes and the proportion of desirable outcomes is the response rate in our sample. From a frequentist point of view, we are interested in testing $H_0 : \theta = c$ versus $H_a : \theta > c$. The effect size in our one-sided test this is a function of the hypothesized response rate c and the true response rate θ_{true}. We apply the method used by [Cohen, 1992] for a test of proportion to which we refer the reader for the full details. Here, we simply present fixed sample sizes n_{fixed} given a selected significance level, power, hypothesized response rate c and true response rate θ_{true}.

We determine sample sizes for 3 scenarios and we shall only vary θ_{true}. The significance level and power are 0.95 and 0.80 respectively. The hypothesized value of the response rate is $c = 0.40$ which means that a treatment is deemed successfully if the associated response rate is greater than 0.40. The resulting couples of the true response rate and resulting required sample size $(\theta_{true}, n_{fixed})$ are $(0.45, 605)$, $(0.50, 156)$ and $(0.55, 69)$.

3.1.2 Operating characteristics

In section 2.2.3 we introduced the operating characteristics conceptually, here we discuss how they are implemented in the simulation. Error rates are assessed by simulating a trial with fixed design parameters n_{sim} times. For the type I error rate, patients are drawn from a population for which the response rate is fixed at the borderline value for which the treatment should not be approved. Similarly, for the type II error rate this is repeated fixing the response rate of the population at plausible values for which the treatment should be approved. We shall denote the rate for simulating the type I error rate and type II error rate θ_I and θ_{II} respectively, these are also design parameters. These are not to be confused type I error rate α and type II error rate β, which are the result of simulations. This approach is consistent with industry standards [LeBlond, 2010]. So actually, operating characteristics of a trial with specific design parameters are found by simulating n_{sim} runs of that trial. Half of these runs are used to find the type I error rate and the other half is used for the type II error rate.

The bias of the simulation of a trial is the difference between the response rate that underlies the population from which the patients are sampled and the expected value of the posterior distribution. Since the posterior distribution follows a Beta distribution the latter is easily found. The bias of a clinical trial is then defined as the average bias over n_{sim} simulations:

$$\text{Bias } = \frac{1}{n_{sim}} \sum_{r=1}^{n_{sim}} \theta_r - \frac{\alpha_r}{\alpha_r + \beta_r}. \tag{7}$$

Where $\theta_r = \theta_I$ for the iterations used for simulating the type I error rate. For iterations used for simulating the type II error $\theta_r = \theta_{II}$. Also, α_r and β_r are the parameters of the posterior distribution displayed in equation 5 (at the r-th run of n_{sim}).

3.1.3 Implementation of the framework

For all simulations we use a non-informative prior so that all response rates $\theta \in (0, 1)$ are equally likely. This results in picking a Beta prior with hyperparameters $\alpha_0 = 1$ and $\beta_0 = 1$. Each simulation consists of $n_{sim} = 2000$ iterations. The decision rule (see equation 6) is implemented with $\phi = 0.40$ and $p_{cut} = 0.95$, meaning that a treatment if considered superior if the mass of the posterior distribution above 0.40 exceeds 0.95. A fixed sample size calculation forms the basis for the number of patients n_{max} used in each simulation. Starting from a 3 fixed sample sizes, multiple simulations are performed where the number of interim analysis is incrementally increases. The results are displayed in table 1. Subsequently, in table 2 we consider a selection of designs from table 1 where various design parameters are altered.

3.2 Results

3.2.1 Problem of multiplicity

We use as starting points sample sizes from a fixed sample size calculation and observe how our Bayesian design performs as we start from the fixed same sample sizes and increase the number of interim analyses. These sample sizes serve as a point of reference, we want the sample size to have a statistical basis, rather than a pragmatic basis.

Our Bayesian design models uncertainty about the the response rate in a population. However, if we want to acquire frequentist operating characteristics we have to assume for a moment that there is indeed a true response rate underlying the population. As described in section 2.2.3 we have to determine a rate for simulating the type I error rate and type II error rate (i.e. $\theta_I,\ \theta_{II}$). We pick the response rate for simulating the type I error equal to the hypothesized response rate c, so $\theta_I = c$. We pick the response rate for simulating the type II error to be equal to the estimated

true population response rate, $\theta_{II} = \theta_{true}$.

In table 1 the simulation results are displayed for Bayesian designs which take as a starting point the fixed sample size based on different estimates of the true response rate θ_{true}. In all three scenarios we see that if there are no interim analyses (D1, D7, D13) and the decision rule is only evaluated after all patients are observed the type I error rate α and power $1 - \beta$ are maintained at an acceptable level. Both the type I error rate and the power deviate less than 0.01 from 0.05 and 0.80 respectively. The problem of multiplicity is clearly observed in each scenario. That is, the type I error rate inflates as we conduct more interim analyses. Designs D6 and D12 demonstrate how severely the type I inflates in the extreme case, when we evaluate our decision rule after every patient. The type I error rate is multiplied by a factor ~ 8 and ~ 7 respectively. With the exception of these two extreme cases, the bias is almost always negligible. In the last column we see that the average number of patients $\tilde{n}$, or expected number of patients, drops as the number of interim analyses increases.

To better understand why this problem of multiplicity occurs let us recall our decision rule in equation 6: $P(\theta > \phi \ |\text{data}) > p_{cut}$. During a simulation we repeatedly draw a group of patients from a population each iteration. With each interim analysis, we check whether the mass in the posterior distribution above ϕ exceeds p_{cut}. For the type I error rate, the response rate is $\theta_I = 0.40$ and p_{cut} rarely exceeds 0.95. However, now that we check multiple times, we risk looking at the exact moment when p_{cut} exceeds 0.95 and we make a type I error. Similarly, for assessing the type II error rate at each iterations patients are drawn from a population with response rate θ_{II}. Now, this is an advantage in the sense that we increase the power. Since the response rate underlying the population is in fact greater than 0.40 multiple interim analyses now allow us to conclude superiority and stop early. This is also observed in table 1. For example, designs D8-D12 all achieve a greater

power than design D7.

3.2.2 Combating the problem of multiplicity

Inflation of the type I error rate is undesirable. In the context of clinical trials, a type I error means concluding incorrectly that a treatment has positive effect. The challenge is to achieve a desirable type 1 error rate while still maintaining an acceptable power, bias and expected number of patients. It could be that the increased type I error rate is minor and that no changes have to be made. For example, if we compare designs D8 and D7 we see that the former has an increased type 1 error rate but 0.072 might still be acceptable. In particular, the trade off between a higher type I error rate for a lower expected number of patients might be acceptable or even desirable. However, as the number of interim analyses increase the inflation quickly becomes unacceptable and change of design parameters are needed. In table 2 we present altered designs from table 1 to show how changing various design parameters can combat multiplicity.

We concluded the last subsection discussing why inflation of the type I error rate occurs. The most natural idea is to increase the value of p_{cut}. If we increase value of p_{cut} we require stronger evidence to conclude superiority. Then we are less likely to conclude superiority when we should not but also when we should. In this sense, we are trading of (excess) power for an improved type I error rate. This is observed in design D16 in table 2 where a more stringent value for p_{cut} is chosen so that the type 1 error rate is maintained, at the expense of a lowered power and increased expected number of patients $\tilde{n}$.

Increasing the number patients n_k observed at each cohort does not bring us much. This is demonstrated by design D18 in table 2. The problem is that this increased cohortsize barely changes the probability of incorrectly concluding superiority. However, it does provide increased

power which in turn can be traded of for a better type I error rate via the increase of p_{cut}. This is observed in table 2 for design D17. If we compare it to the original design D10 in table 1 we see that ultimately if we want to maintain a type 1 error rate of ≤ 0.05 and power ≥ 0.80 conducting interim anlyses comes as the expense of an increased expected number of patients $\tilde{n}$.

So far we have seen that in our Bayesian designs interim analyses increase the type 1 error rate and controlling this comes at the expense of decreased power or increased expected number of patients $\tilde{n}$. So far, a clinical trial based on a fixed sample seems very appealing. However, we stress that these are based on estimates of effect size. For our Bayesian designs we don't have to estimate an effect size. If the effect size is greater than we estimated Bayesian designs can achieve the desirable frequentist operating characteristics and have a lower expected number of patients. This is demonstrated in table 2 by the designs D19. Design D19 shows that if $R_{II} = 0.475$ as opposed to 0.45 we can maintain the type I error rate by increasing p_{cut}. The expected number of patients $\tilde{n}$ drops from 605 to 446 compared to the fixed sample size.

	Design parameters			Operating characteristics			
	k	n_k	n_{max}	α	$1-\beta$	Bias	$\tilde{n}$
Scenario 1: $\theta_{II} = 0.45$							
D1	0	605	605	0.057	0.815	-0.00033	605
D2	1	303	606	0.083	0.832	-0.00130	517
D3	2	202	606	0.090	0.824	-0.00079	483
D4	3	152	608	0.105	0.854	-0.00178	457
D5	4	122	610	0.146	0.891	-0.00348	439
D6	602	1	605	0.419	0.944	-0.03842	270
Scenario 2: $\theta_{II} = 0.50$							
D7	0	156	156	0.048	0.819	-0.00096	156
D8	1	78	156	0.072	0.830	-0.00073	133
D9	2	52	156	0.111	0.880	-0.00416	121
D10	3	39	156	0.128	0.869	-0.00531	116
D11	4	31	155	0.135	0.865	-0.00503	113
D12	155	1	156	0.343	0.920	-0.03737	80
Scenario 3: $\theta_{II} = 0.55$							
D13	0	69	69	0.045	0.794	-0.00044	69
D14	1	35	70	0.099	0.876	-0.00430	58
D15	2	23	69	0.093	0.844	-0.00491	56

Table 1: Effect of conducting multiple interim analyses for different Bayesian designs based on $n_{sim} = 2000$ iterations. The design parameters that change per simulation are in the left-hand side of the table. These are the number of interim analyses k, patients per cohort n_k, and the maximal number of patients observed n_{max}. Note that the total number of analyses is $k+1$ since the decision rule is also evaluated at the end. The operating characteristics are displayed on the right-hand side of the table. The average number of patients $\tilde{n}$ is rounded up to the nearest integer. The design parameters that are not displayed are the same for every simulation. These are the remaining design parameters that we introduced in section 2, $\theta_I = 0.40$, $\phi = 0.40$ and $p_{cut} = 0.95$. The D number is just a label for further reference.

		Design parameters					Operating characteristics			
Label	Original	k	n_k	n_{max}	R_{II}	p_{cut}	α	$1-\beta$	Bias	$\tilde{n}$
D16	D8	1	78	156	0.5	**0.965**	0.045	0.796	-0.00170	137
D17	D10	3	**52**	**208**	0.5	**0.985**	0.040	0.821	-0.00147	168
D18	D10	3	**52**	**208**	0.5	0.95	0.129	0.937	-0.00463	148
D19	D4	3	152	608	**0.475**	**0.98**	0.041	0.964	-0.00036	446

Table 2: In this table we present variations of the designs presented in table 1. The D number in the "Label" column is again the label of the design. The D number in the "Original" column refers to the label on which that design is based, if applicable. For clarity, the design parameters that are changed are typed bold.

4 Discussion

In this work we have deployed a model for decision making in a Bayesian framework with binary outcomes. We address the two main results.

First, it is observed that clinical trials based on our Bayesian designs provide similar operating characteristics compared to clinical trials based on a fixed sample size calculation in case no interim analyses are conducted. That is, our Bayesian designs have a similar type I error rate and power. If everything in a Bayesian designs is kept fixed, the type I error rate increases with the number of interim analyses conducted and can quickly exceed an acceptable level.

Secondly, if the inflation of the type I error rate exceeds an acceptable level, design parameters can be altered in order to maintain an acceptable type I error rate. In particular, altering the decision rule possibly in combination with increasing the cohortsize can control the type I error inflation such that the type I error rate is maintained at a prespecified value. The decision rule is altered such that it is harder to conclude superiority by increasing the required mass above a cutoff value of the posterior distribution at each interim analysis. In doing so, both the type I error rate and power decrease. Essentially excess power is traded off for a better type I error rate.

We find that clinical trials based on our Bayesian designs with interim analyses can have type I error rates similar to a trial based on a fixed sample size calculation if the appropriate decision rule is implemented. An advantage we have seen that our Bayesian design yields a high power, which in turn can be traded off to control the type I error rate. In some cases, we also find that our Bayesian design needs significantly fewer patients in order to draw the right conclusion. Also, practical benefit is that our Bayesian design does not require an a priori estimate of the effect size.

For future work, the design that is presented in this paper can be extended. We suggest two possible directions for future research. First, the way in which the type I error rate is controlled

can be changed. In this work, the decision rule can be made more strict so that the type I error rated is maintained at a prespecified value. If a design shows an unacceptable type I error rate, we simply change the decision rule until the type I error rate is again low enough. An alternative approach is to calculate the appropriate boundaries for the decision rule at each interim analysis such that that the overall type I error rate is maintained at the nominal value (see for example, [Shi et al., 2019, Lewis et al., 2007, Lewis and Berry, 1994]).

Secondly, in our designs we could only stop for superiority. The designs presented in this work can be extended with the possibility for stopping for futility. This can be implemented by changing the used decision rules or by adding additional decision rules. Early stopping is often deployed in clinical trials ([Connor et al., 2013, Ryan et al., 2020]). It can be desirable in case a treatment shows a low response rate, especially when accumulation of patients is difficult, expensive or both.

References

[Armitage et al., 1969] Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2):235–244.

[Bauer et al., 2016] Bauer, P., Bretz, F., Dragalin, V., König, F., and Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, 35(3):325–347.

[Cohen, 1992] Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, 1(3):98–101.

[Connor et al., 2013] Connor, J. T., Elm, J. J., Broglio, K. R., Esett, Investigators, A.-I., et al. (2013). Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. *Journal of clinical epidemiology*, 66(8):S130–S137.

[Edwards et al., 1963] Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review*, 70(3):193.

[Gaydos et al., 2009] Gaydos, B., Anderson, K. M., Berry, D., Burnham, N., Chuang-Stein, C., Dudinak, J., Fardipour, P., Gallo, P., Givens, S., Lewis, R., et al. (2009). Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal*, 43(5):539–556.

[Kim and Demets, 1987] Kim, K. and Demets, D. L. (1987). Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika*, 74(1):149–154.

[LeBlond, 2010] LeBlond, D. (2010). Fda bayesian statistics guidance for medical device clinical trials-application to process validation. *Journal of Validation technology*, 16(4):24.

[Lewis and Berry, 1994] Lewis, R. J. and Berry, D. A. (1994). Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89(428):1528–1534.

[Lewis et al., 2007] Lewis, R. J., Lipsky, A. M., and Berry, D. A. (2007). Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: a frequentist evaluation. *Clinical Trials*, 4(1):5–14.

[O'Brien and Fleming, 1979] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556.

[Pocock, 1977] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

[Rouder, 2014] Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, 21(2):301–308.

[Ryan et al., 2020] Ryan, E. G., Brock, K., Gates, S., and Slade, D. (2020). Do we need to adjust for interim analyses in a bayesian adaptive trial design? *BMC Medical Research Methodology*, 20(1):1–9.

[Sanborn et al., 2014] Sanborn, A. N., Hills, T. T., Dougherty, M. R., Thomas, R. P., Erica, C. Y., and Sprenger, A. M. (2014). Reply to rouder (2014): Good frequentist properties raise confidence. *Psychonomic bulletin & review*, 21(2):309–311.

[Schönbrodt et al., 2017] Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological methods*, 22(2):322.

[Shi et al., 2019] Shi, H., Yin, G., et al. (2019). Control of type i error rates in bayesian sequential designs. *Bayesian Analysis*, 14(2):399–425.

[Thorlund et al., 2018] Thorlund, K., Haggstrom, J., Park, J. J., and Mills, E. J. (2018). Key design considerations for adaptive clinical trials: a primer for clinicians. *bmj*, 360.

[Wagenmakers et al., 2010] Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology*, 60(3):158–189.

[Yin et al., 2017] Yin, G., Lam, C. K., and Shi, H. (2017). Bayesian randomized clinical trials: From fixed to adaptive design. *Contemporary clinical trials*, 59:77–86.

[Zhu and Yu, 2017] Zhu, H. and Yu, Q. (2017). A bayesian sequential design using alpha spending function to control type i error. *Statistical methods in medical research*, 26(5):2184–2196.

Appendix A: Business and societal impact

In this appendix we discuss the societal and business impact of this paper. This paper is part of acquiring a master degree at the Jheronimus Academy of Data Science (JADS). To embed data science in society and a business context is in line with the vision and mission of JADS[1]. Given the theoretical nature of the paper these aspects are not elaborately discussed within the paper hence these are addressed separately.

In this paper a model for decision making in a Bayesian framework with binary outcomes is proposed and therefore we restrict ourselves to the domain of clinical trials. Note that in general this model can be applied in any domain where a binary outcome is observed but we cannot consider all those domains here. As a concrete example, an online marketeer that is interested in clickthrough rate of banner ads can deploy our model. He now has groups of online visitors of his website instead of cohorts of patients and is interested in the click-through rate of a banner ad instead of the response rate of a treatment. However, for an online marketeer, our research may have different implications from both a business and societal perspective.

So, the method of randomized controlled trials is not solely used in the domain of healthcare. In the past years various branches of industry have used systematic experimentation in order to acquire knowledge and making business decisions. The results from this paper can be meaningful in other cases as well.

Business impact

For the business impact of our research, we consider the maturity scale. The maturity scale describes the stages of product development. The scale consists of the following levels:

[1] https://www.jads.nl/about-jads/

1. Incremental model improvement

2. Cost decrease (new solution for existing business)

3. Revenue increase (new solution for existing business)

4. New revenue generation / new business development

In the context of clinical trails, our research provides a contribution to existing knowledge on clinical trials in a Bayesian framework. Specifically, we demonstrated how conducting multiple interim analysis in an adaptive trial can lead to undesirable operating characteristics and how we can control these by considering the trial design. Desirable operating characteristics are important to governing bodies such as the FDA and EMA [2], which in turn approve or disapprove a medicine. Bringing a new medicine (or improved version of existing medicine) focuses on item 3 and 4. An example is the cholestrol-lowering drug *Pravigard Pac*[3]. This drug was developed by Bristol-Myers Squibb and subsequently approved in 2003 by the FDA based on a Bayesian analysis of it's efficacy.

Furthermore, we have seen that our Bayesian designs provides the expected number of patients that have to be observed. A research might find that our Bayesian design needs to enroll fewer patients than the alternative trial design he or she has as hand. Enrolling fewer patients in a trial decreases cost, allow for more rapid innovation and can potentially facilitate earlier market entry. All these are benefits are of great value for a pharmaceutical company. Off course, this is case dependent and therefore our research can potentially focuses on item 2 on the maturity scale.

[2]US Food and Drug Administration, European Medicines Agency
[3]Berry, D. A. (2006). *Bayesian clinical trials.* Nature reviews Drug discovery, 5(1), 27-36.

Societal impact

Here we consider the societal impact from a somewhat broader point of view. By that we mean that we discuss the societal impact that research on clinical trails intend to make since claiming that a single paper in itself has this societal impact is somewhat bold.

Adaptive clinical trials can[4] be completed faster than non-adaptive trials. This in turn can speed up drug development and can speed up the release of a drug to the market which can be beneficial for the general health within a population. For example, this can be particularly useful for a novel treatment to a disease affecting a large portion of the (world) population. The COVID-19 outbreak death toll is approaching 4 million[5] since it outbreak in December 2019. In such an extreme case, if a treatment becomes available one week earlier, it can potentially save thousands of lives. Also, tying in with the benefit mentioned earlier, if we can achieve results while using fewer patients, we expose fewer patients to a inferior treatment if that treatment ultimately does turn out to be inferior.

[4]Thorlund, K., Haggstrom, J., Park, J. J., & Mills, E. J. (2018). Key design considerations for adaptive clinical trials: a primer for clinicians. *bmj*, 360.

[5]https://covid19.who.int/