The Bayesian optimal two-stage design for clinical phase II trials based on Bayes factors

Riko Kelter *
Institute of Medical Statistics and Computational Biology
Faculty of Medicine, University of Cologne

Samuel Pawel
Epidemiology, Biostatistics and Prevention Institute
Center for Reproducible Science
University of Zurich

December 1, 2025

Abstract

Sequential trial design is an important statistical approach to increase the efficiency of clinical trials. Bayesian sequential trial design relies primarily on conducting a Monte Carlo simulation under the hypotheses of interest and investigating the resulting design characteristics via Monte Carlo estimates. This approach has several drawbacks, namely that replicating the calibration of a Bayesian design requires repeating a possibly complex Monte Carlo simulation. Furthermore, Monte Carlo standard errors are required to judge the reliability of the simulation. All of this is due to a lack of closed-form or numerical approaches to calibrate a Bayesian design which uses Bayes factors. In this paper, we propose the Bayesian optimal two-stage design for clinical phase II trials based on Bayes factors. The optimal two-stage Bayes factor design is a sequential clinical trial design that is built on the idea of trinomial tree branching, a method we propose to correct the resulting design characteristics for introducing a single interim analysis. We build upon this idea to invent a calibration algorithm which yields the optimal Bayesian design that minimizes the expected sample size under the null hypothesis. Examples show that our design recovers Simon's two-stage optimal design as a special case, improves upon non-sequential Bayesian design based on Bayes factors, and can be calibrated quickly, as it makes use only of standard numerical techniques instead of time-consuming Monte Carlo simulations. Furthermore, the design allows to ensure a minimum probability on compelling evidence in favour of the null hypothesis, which is not possible with other designs. As the idea of trinomial tree branching is neither dependent on the endpoint, nor on the use of Bayes factors, the design can therefore be generalized to other settings, too.
Keywords: sequential clinical trial design, Bayesian statistics, phase II trial, Bayes factors

1 Introduction

Two-stage designs in clinical trials are primarily used to assess the efficacy of a treatment, especially in early-phase clinical trials where a smaller sample size is typically used (Kieser,

*Correspondence concerning this article should be addressed to ... Draft version 1.0, 28/11/2025. This paper has not been peer reviewed. Please do not copy or cite without author's permission. Data and R code to reproduce our results are openly available at https://osf.io/3dbtv/overview?view_only=9f9d87a6748c446ead76de99c1e830b7. We declare no conflict of interest.

2020; Wassmer & Brannath, 2016). In these trials, a small group of patients is enrolled and their outcome is evaluated in a first step. After the first stage, an interim analysis is performed which determines whether the trial proceeds to the second stage or is stopped. Only if the interim analysis suggests sufficient evidence of efficacy (of a novel drug or treament), a second stage with additional patients is initiated. There are various benefits of two-stage clinical trial designs, in particular, for single-arm phase IIa trials where the goal is to demonstrate a proof of concept for a novel drug or treament. Among these benefits are:

- *Ethical considerations*: If the treatment shows no promising efficacy in the first stage, the trial can be stopped for futility. This avoids further exposure of patients to potentially ineffective treatment (Wassmer & Brannath, 2016).
- *Efficiency considerations*: The second stage is only initiated if there is enough statistical evidence of efficacy, which saves resources and time. (Pallmann et al., 2018).

Bayesian approaches to two-stage designs (and, in general, sequential clinical trial designs) are a fast expanding field of research (Fayers, Ashby, & Parmar, 2005; Ferguson, 2021; Ferreira et al., 2021; Kelter & Schnurr, 2024; Stallard, Todd, Ryan, & Gates, 2020; H. Zhou, Lee, & Yuan, 2017; Y. Zhou, Lin, & Lee, 2021). While Bayesian approaches have certain advantages such as the ability to incorporate historical information via the prior distribution, Bayesian two-stage or sequential designs often require a complicated calibration via Monte Carlo simulation studies (Chevret, 2012). Calibration refers to guaranteeing a prespecified type-I-error rate and power which must be interpretable from a frequentist point of view (Grieve, 2022). This fact is due to health authorities' requirements (U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, 2019; U.S. Food and Drug Administration Center for Drug Evaluation and Research (CDER), 2019) and complicates the design of Bayesian two-stage and sequential designs. Although Monte Carlo studies allow to simulate trial data under the null and alternative hypothesis H_0 and H_1 and study the resulting type-I-error and power – depending on the test criteria used – the reproducibility of Monte Carlo studies is inherently troubled by various aspects, for details see Morris, White, and Crowther (2019), Kelter (2023a), Seibold, Charlton, Boulesteix, and Hoffmann (2021), Boulesteix, Groenwold, et al. (2020) and Boulesteix, Hoffmann, Charlton, and Seibold (2020). Still, calibration of Bayesian two-stage and sequential designs by means of a Monte Carlo simulation is the dominant approach in practice. (Berry, 2011; Giovagnoli, 2021; Grieve, 2022).[1] Based on the literature, there is clear lack of non-simulation based approaches

[1] Regarding the need to carry out possibly complex simulations to calibrate a Bayesian design, (Chevret, 2012, p. 1010) notes that a particular *"challenge in using Bayesian adaptive trials is the lack of user-friendly software for study design and analysis. The practical application of these methods has only become feasible since the early 1990s because of computational advances in Markov chain Monte Carlo methods (...)."*

to calibrate Bayesian two-stage or sequential trials. This would improve both the reproducibility of the statistical methodology as well as the ease of application of such trial designs.

Furthermore, in such trials, the most popular test criteria is the posterior probability of the hypotheses under consideration (T. Zhou & Ji, 2023). Then, the posterior probabilities $P(H_0 \mid y)$ and $P(H_1 \mid y)$ given the data are used to test the null hypothesis H_0 versus the alternative H_1. However, other options exist: Bayes factors have gained increasing attention recently in the context of two-stage assessment of hypotheses (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Stefan, Lengersdorff, & Wagenmakers, 2022; Van de Schoot et al., 2021) as well as in preclinical research (Kelter, 2023b; Pourmohamad & Wang, 2023). The Bayes factor $\mathrm{BF}_{01}(y) := \frac{f(y|H_0)}{f(y|H_1)}$ conceptually is the predictive updating factor from prior to posterior odds:

$$\underbrace{\frac{P(H_0 \mid y)}{P(H_1 \mid y)}}_{\text{Posterior odds}} = \underbrace{\frac{f(y \mid H_0)}{f(y \mid H_1)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{Prior odds}} \tag{1}$$

Bayes factors have become increasingly popular in psychology and other areas (Van de Schoot et al., 2021), most probably because they have certain advantages over posterior probabilities.

First, while posterior probabilities are strongly influenced by the associated prior probabilities $P(H_0)$ and $P(H_1)$ of the hypotheses of interest, Bayes factors are independent of the prior odds of the hypotheses under consideration. Bayes factors are solely influenced by the prior distribution on the parameter itself. As a consequence, even if historical data are used which favour one hypothesis a priori before the trial is conducted, the Bayes factor is not influenced by that prior probability. In contrast, the posterior probability is strongly drawn towards the a priori favoured hypothesis in such a case.

Second, Bayes factors are easy to interpret. For example, a Bayes factor of size ten implies that data are ten times more likely under the null hypothesis H_0 than under the alternative H_1. This allows for simple communication with clinicians. One could argue that the same holds for posterior probabilities, as probabilities are simple to communicate. However, the posterior probability can be influenced strongly by the prior probabilities. In that case, although the posterior probability scale is easy to communicate, it is often difficult to gauge the influence of the prior probability on the resulting posterior probability.

Third, Bayes factors can be used as a frequentist test statistic and the long-term properties of Bayes factors provide a Bayes-frequentist compromise (Grieve, 2022; Kelter & Schnurr, 2024; Stallard et al., 2020; Van de Schoot et al., 2021). We turn to this perspective in the following Section 2 in Equation (3), our main requirement to calibrate a Bayesian two-stage design. Still, we already stress here that our proposed design can be extended to other testing approaches including posterior probabilities.

Fourth, it is well known that sample size planning and the resulting power under H_1 and

type-I-error rate under H_0 can (and often does) vary for different approaches to testing, such as posterior probabilities and Bayes factors (Kelter, 2020b, 2021; Linde, Tendeiro, Selker, Wagenmakers, & van Ravenzwaaij, 2020; Makowski, Ben-Shachar, Chen, & Lüdecke, 2019). Thus, it might be possible that designs which make use of the Bayes factor can improve upon certain operating characteristics of an analogue design which uses posterior probabilities.

Now, an important distinction in Bayesian trial design is between design and analysis priors. While the design prior is selected primarily to incorporate possibly available information and a priori knowledge about the phenomenon under study, the analysis prior is the prior used to calculate the Bayes factor once the data are collected (Berry, 2011; O'Hagan, 2019; Schönbrodt et al., 2017; Stefan et al., 2022). Thus, an analysis prior often needs to be much more objective than a design prior. During the design of a trial, the key requirement is to obtain a calibrated design, which we will explain in further detail in Section 2.

1.1 Outline

Although Bayes factors have some appealing properties as outlined above, few clinical trial designs exist which make use of them (Giovagnoli, 2021; H. Zhou et al., 2017). In this paper, we develop a Bayesian two-stage clinical trial design for phase IIa trials with binary endpoints. In these trials, we focus on the hypotheses

$$H_0 : p \leq p_0 \text{ versus } H_1 : p > p_0 \tag{2}$$

for some $p_0 \in (0, 1)$, so we focus on a proof of concept trial for a novel drug or treatment. Our trial design makes use of Bayes factors to test the above hypotheses, but can conceptually be generalized to any testing approach – including posterior probabilities and others, see Kelter (2020a). Therefore, we introduce a general graphical structure of our Bayesian two-stage design based on a trinomial tree and provide theoretical results which allows for almost instant computation of the Bayesian power and type-I-error rate of our design. Thus, our approach renders complicated Monte Carlo simulations to calibrate the design entirely obsolete, which is even more appealing as our design can be generalized to other testing approaches, as mentioned above.

The structure of our paper is as follows: In the next Section 2 we introduce details about our design and the operating characteristics we focus on when speaking about calibration. Section 3 outlines the key problems which must be addressed when allowing for a single interim analysis in the design – thereby modifying it into a two-stage design. Section 4 provides solutions to these main problems and Section 5 then introduces a calibration algorithm based on these solutions. Section 6 provides illustrating examples and shows how the design works in practical settings and Section 7 closes the paper with a discussion for future work and a conclusion.

2 The Bayesian two-stage design for phase II clinical trials with binary endpoints

The idea of the two-stage design is as follows: We search for two natural numbers n_1 and n_2 with $n_1 < n_2$ for which the two requirements

$$P(\text{BF}_{01}^{n_2}(y) < k \mid H_0) \leq \alpha \qquad \text{and} \qquad P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) \geq 1 - \beta_{n_2} \tag{3}$$

are met for some α and β_{n_2} in $(0, 1)$. The first denotes the Bayesian type-I-error rate, while the second resembles a Bayesian analogue of frequentist power. For example, for $k = 1/10$ – the usual threshold for strong evidence according to the scale of Jeffreys (1939) – when $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0) \leq \alpha$ holds, the probability to obtain a Bayes factor indicating strong evidence against the null hypothesis (that is, in favour of the alternative H_1) is bounded by α under H_0.[2] Thus, for $\alpha := 0.05$ the Bayesian type-I-error rate is controlled at 5% then. Likewise, if $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) > 1 - \beta_{n_2}$ for, say $\beta_{n_2} := 0.2$ and $k = 1/10$, the probability to obtain strong evidence in favour of H_1, when H_1 is true, is 80%.[3] A critical goal in Bayesian design of experiments and sample size calculation therefore is to provide a solution to the inequalities in Equation (3) for some α and β_{n_2} in $(0, 1)$.

In Equation (3), the superscript n_2 in $\text{BF}_{01}^{n_2}(y)$ indicates that the Bayes factor is based on a sample of size n_2. The sample size n_2 is the sample size at the end of the trial after n_2 patients have been enrolled and the outcomes have been observed.

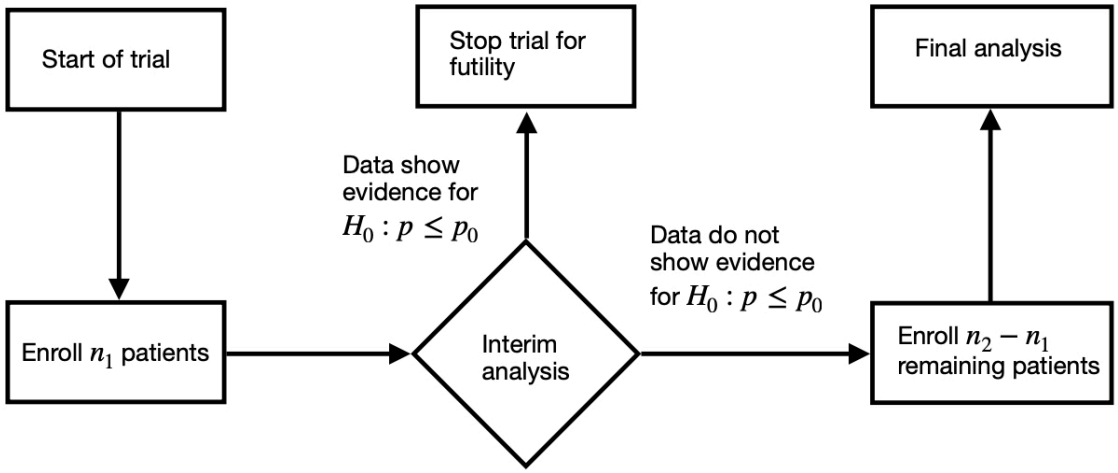


Figure 1: Overview of the Bayesian two-stage design for phase IIa clinical trials

[2] This holds because $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0) = P(1/\text{BF}_{01}^{n_2}(y) > 1/k \mid H_0) = P(\text{BF}_{10}^{n_2}(y) > 10 \mid H_0) \leq \alpha$ and $\text{BF}_{10}^{n_2}(y) > 10$ amounts to the Bayes factor in favour of H_1 indicating (at least) strong evidence for H_1. Thus, $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0)$ resembles an intuitive Bayesian type-I-error rate.

[3] This holds because $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) = P(\text{BF}_{10}^{n_2}(y) > 1/k \mid H_1) = P(\text{BF}_{10}^{n_2}(y) > 10 \mid H_1)$.

The structure of the Bayesian single-arm two-stage design with binary endpoints is shown in Figure 1. The two-stage design makes a single interim analysis after the outcomes of n_1 patients have been enrolled. The principal goal is to stop the trial for futility at this interim analysis, if there is sufficient evidence for the null hypothesis $H_0 : p \leq p_0$ that the drug or novel treatment is ineffective. Here, p_0 denotes a success probability, and if the true success probability of the drug is smaller or equal to p_0, we deem the drug ineffective (possibly compared to a standard treatment of care with efficacy p_0). Thus, the hypotheses of interest are the usual ones considered during a single-arm phase IIa clinical trial – a proof of concept trial – where we test

$$H_0 : p \leq p_0 \text{ versus } H_1 : p > p_0 \tag{4}$$

for some $p_0 \in (0, 1)$, which corresponds to the efficacy of the standard of care. The primary endpoint is therefore supposed to be binary and measure success or failure of the novel drug or treatment.

As Figure 1 shows, there is no option to stop the trial for efficacy – in that case, the trial is continued until reaching the final sample size n_2, which is a common choice in phase II trials (Berry, 2011; Lee & Liu, 2008) – but it can be terminated early due to futility after observing n_1 outcomes. After observing n_1 outcomes, the interim analysis is performed to possibly stop the trial early for futility. As a consequence, the observed data y in $\text{BF}_{01}^{n_2}(y)$ consist of the first $y^1 := (y_1, ..., y_{n_1})$ observations, and the remaining $y^2 = (y_{n_1+1}, ..., y_{n_2})$ observations, where $y_i = 0$ denotes a failure and $y_i = 1$ a success for $i = 1, ..., n_2$.

2.1 Operating characteristics

First and foremost, the two-stage design needs to meet the requirements in Equation (3) regarding Bayesian power and type-I-error rate. We will, see, however, that it is possible to make these error rates fully frequentist under a specific design prior choice, which we will discuss later in detail. Next, to the conditions in Equation (3), there are two further operating characteristics which primarily are involved with the interim analysis. In the interim analysis, two criteria or operating characteristics of the trial design can be considered. First, suppose H_0 indeed holds. If that's the case, then it seems natural to require a minimum probability

$$P(\text{BF}_{01}^{n_1}(y^1) > k_f | H_0) > f \tag{5}$$

for some $f \in (0, 1)$, which can be interpreted as guaranteeing a minimum probability f to stop for futility when the drug is not effective and H_0 holds. For example, if $f = 0.6$, we would stop at the interim analysis with 60% probability when the drug is ineffective and $H_0 : p \leq p_0$ holds. Note that we explicitly use the threshold k_f instead of k, because one could use different

evidence thresholds for the interim analysis to stop for futility and for the final analysis at the end of the trial, compare Equation (3).

We sum up the above and clarify the goal of the Bayesian two-stage design now.

- We search for two sample sizes $n_1, n_2 \in \mathbb{N}$ with $n_1 < n_2$, for which Equation (3) holds for prespecified values of $\alpha \in (0, 1)$ and $\beta_{n_2} \in (0, 1)$, and for which Equation (5) holds for a prespecified $f \in (0, 1)$.

3 Trinomial tree branching and the problems with power and error rates

In this section, we introduce our approach to calculate the operating characteristics in Equation (3) for the Bayesian two-stage design described in the previous section. Therefore, we assume a fixed sample size n_2 at the final analysis and strive for a correct calculation of the Bayesian power $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ in Equation (3) first.

Therefore, we provide the structure of the design graphically in form of a trinomial tree with its different branches. Then, we analyze two core problems which appear when introducing the option to stop the trial for futility after a single interim analysis at n_1. These two problems are:

- The (Bayesian) power $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ is too large and must be adjusted due to the introduced interim analysis.

- The same holds for the (Bayesian) type-I-error rate $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0)$, which is, as it turns out, too conservative and even improves when being adjusted correctly, due to the possibility to stop the trial for futility at the interim analysis.[4]

In the following Section 4, we then turn to solving both of these issues. This allows to calculate the power and type-I-error rate for the Bayesian two-stage design correctly for a fixed sample size n_2. In the subsequent Section 5, we then turn to the task of calibrating the Bayesian two-stage design based on our solution. That is, we turn to the issue of providing values n_1 and n_2 for which Equation (3) and Equation (5) hold. In this step, we make use of our solution for the correct power and type-I-error rate calculations for a fixed sample size n_2 derived in Section 4.

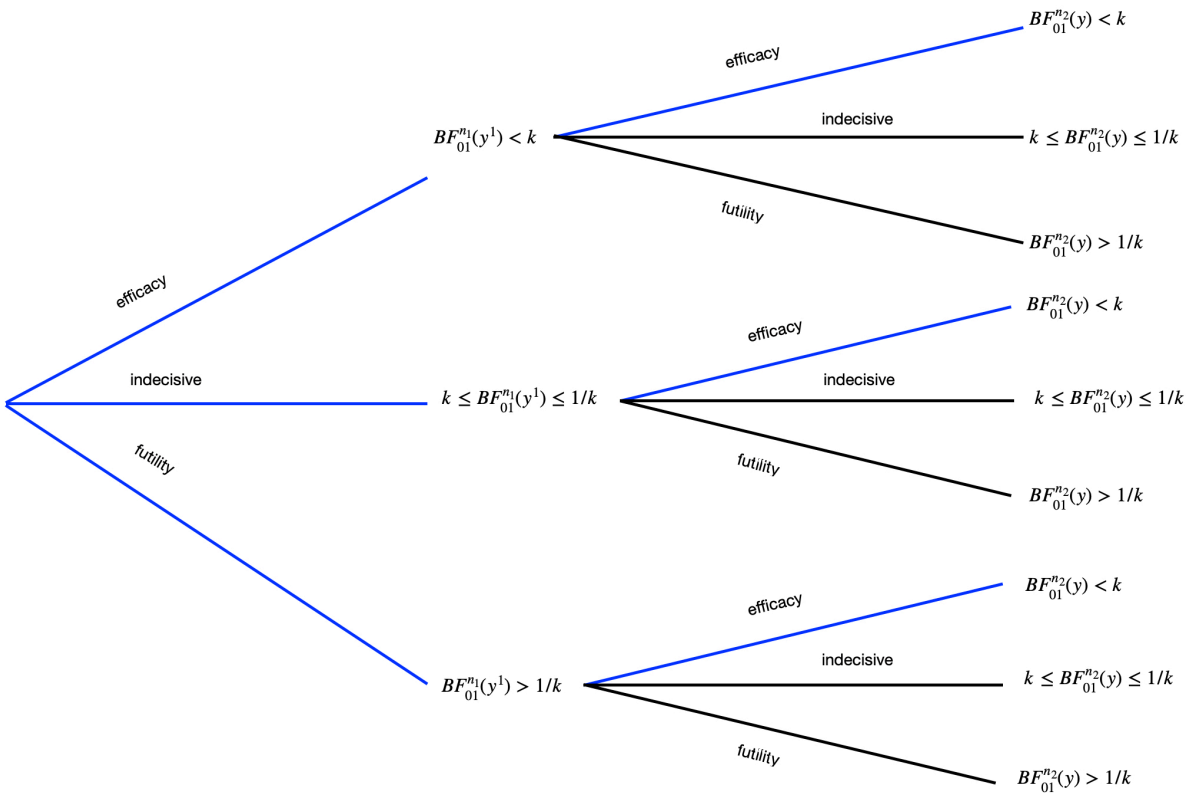


Figure 2: A trinomial tree illustrating the three outcomes the Bayes factor (or a statistical test, in general) can yield during the interim analysis and the final analysis of the Bayesian two-stage design.

3.1 Graphical structure of the design: The trinomial tree

Figure 2 shows a trinomial tree which accounts for the three possibilities to

- obtain a convincing Bayes factor $\mathrm{BF}_{01}^{n_1}(y^1) < k$, which shows evidence for efficacy of the drug, that is, for $H_1 : p > p_0$
- obtain an indecisive Bayes factor $k \leq \mathrm{BF}_{01}^{n_1}(y^1) \leq 1/k$ which shows neither convincing evidence for efficacy or futility, that is, neither for $H_0 : p \leq_0$ or $H_1 : p > p_0$
- obtain an unconvincing Bayes factor $\mathrm{BF}_{01}^{n_1}(y^1) > 1/k$ which shows evidence for inefficacy of the drug, that is, for $H_0 : p \leq p_0$, and therefore for stopping the trial due to futility.

[4] Although surprising at first, this fact makes intuitively sense, because allowing to stop for futility at n_1 implies that under H_0, the Bayes factor cannot be swayed around anymore to show evidence for H_1. Thus, after n_1 patients the Bayes factor could signal evidence for H_0, but after enrolling the remaining $n_2 - n_1$ patients show evidence for H_1, resulting in a false-positive. If an interim analysis allows to stop at n_1 now, the type-I-error rate decreases, as these cases cannot happen anymore. A crucial requirement here is that we do not allow to stop for efficiency, which would influence the type-I-error rate, too.

These three possibilities are possible both at the interim and at the final analysis. The superscripts n_1 at the interim Bayes factors in the middle of Figure 2 indicate that we use only the first batch of sample size n_1 to compute the latter. The final Bayes factors on the right-hand side of the trinomial tree in Figure 2 make use of all data, that is, of the full sample of sample size n_2. All probabilities – that is, the labels 'efficacy', 'indecisive' and 'convincing' in Figure 2 – are conditional on H_i, so different branches of the trinomial tree have different interpretations when $H_i = H_0$ and when $H_i = H_1$. For example, suppose $H_i = H_1$. Then, the Bayesian power in Equation (3) is comprised of three branches in total: These are the three blue branches which eventually end in $\text{BF}_{01}^{n_2}(y) < k$ at the right-hand side of Figure 2. While it seems obvious to calculate power as the sum of the probabilities associated with the three blue branches, there is a caveat: When proceeding like that, the resulting (Bayesian) power is overestimating the true power which results for the two-stage design. We turn to this first problem now.

3.2 Problem I - Calculating the correct power

We first inspect the Bayesian power $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) > 1 - \beta_{n_2}$ based on Figure 3, which shows a more detailed graphical structure of the design and is used to explain the first problem in what follows. Based on Figure 3, we expand $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ as

$$
\begin{aligned}
P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) = & \underbrace{P(\text{BF}_{01}^{n_2}(y) < k \mid \text{BF}_{01}^{n_1}(y^1) < k, H_1)}_{\text{Efficacy at final analysis given efficacy at interim}} \cdot \underbrace{P(\text{BF}_{01}^{n_1}(y^1) < k \mid H_1)}_{\text{Efficacy at interim analysis}} \\
& + \underbrace{P(\text{BF}_{01}^{n_2}(y) < k \mid k \leq \text{BF}_{01}^{n_1}(y^1) < 1/k, H_1)}_{\text{Efficacy at final analysis given indecisive result at interim}} \cdot \underbrace{P(k \leq \text{BF}_{01}^{n_1}(y^1) < 1/k \mid H_1)}_{\text{Indecisive result at interim analysis}} \\
& + \underbrace{P(\text{BF}_{01}^{n_2}(y) < k \mid \text{BF}_{01}^{n_1}(y^1) > 1/k, H_1)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)}_{\text{Futility at interim analysis}} \qquad (6)
\end{aligned}
$$

where the right-hand side equals the probabilities calculated via the path rule in Figure 3 for the three paths which lead to a $\text{BF}_{01}^{n_2}(y) < k$ after n_2 patients have been observed. These are precisely the paths which are shown in blue in Figure 2, and which end with a Bayes factor indicating efficacy of the novel drug irrespective of what the Bayes factor $\text{BF}_{01}^{n_1}(y^1)$ indicated at the interim analysis.

Now, suppose we find a value n_2 for which $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) > 1 - \beta_{n_2}$ for the chosen β_{n_2} now. Then this unconditional probability $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ consists of the three probabilities in the right-hand side of Equation (6). However, as we can stop the trial after n_1 observed outcomes due to futility at interim, it is possible that the last of these three

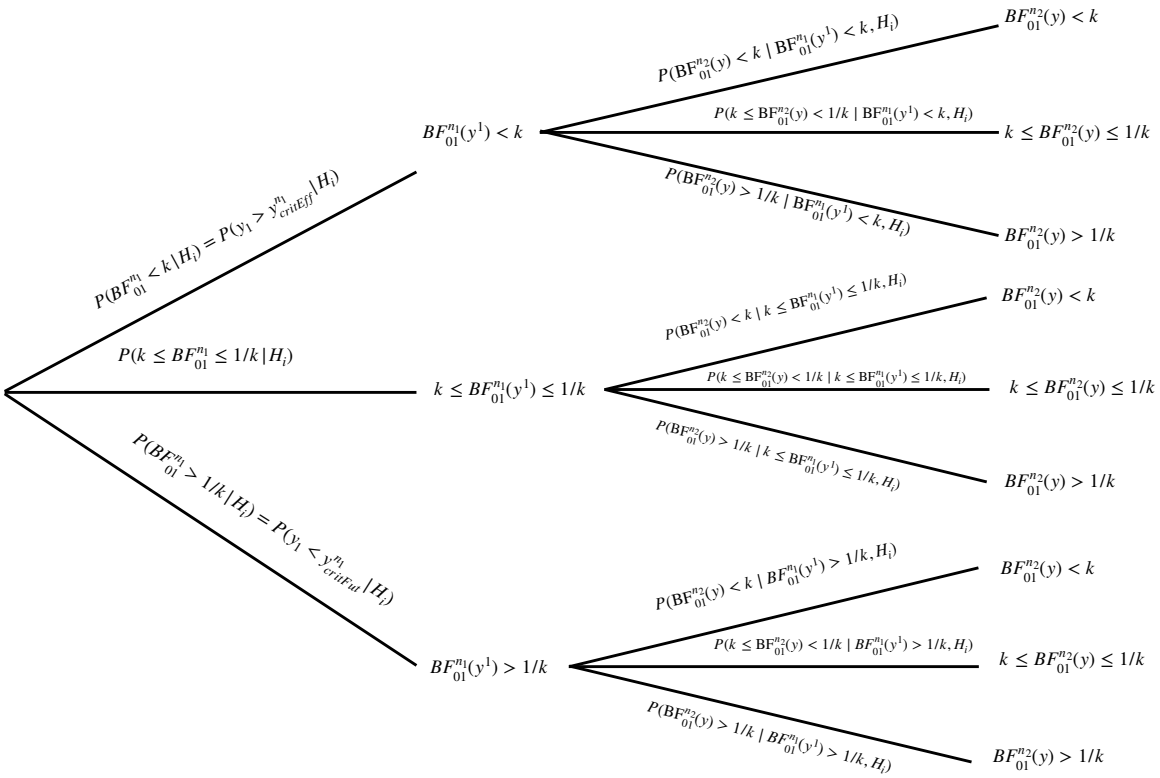


Figure 3: A trinomial tree illustrating the three outcomes the Bayes factor (or a statistical test, in general) can yield during the interim analysis and the final analysis of the Bayesian two-stage design. In addition to Figure 2, the probabilities of each branch are added.

probabilities, – the lowest blue branch in Figure 2 – namely

$$\underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid \mathrm{BF}_{01}^{n_1}(y^1) > 1/k, H_1)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\mathrm{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)}_{\text{Futility at interim analysis}} \tag{7}$$

cannot contribute to this power. Conceptually, in this case the Bayes factor provides convincing evidence for H_0 at the interim analysis, but if the trial would continue it would sway around and indicate convincing evidence for H_1 at the end of the trial, thus contributing to the total power $P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_1)$. However, as we allow the option to stop the trial for futility, once $\mathrm{BF}_{01}^{n_1}(y^1) > 1/k$ holds at the interim analysis the trial is stopped for futility. Thus, the branch with the probability in Equation (7) is 'pruned' in a metaphoric sense. Therefore, when calculating unconditional power $P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_1)$, we must *subtract* the partial power in Equation (7) from the Bayesian power $P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_1)$ to obtain the true power of the resulting two-stage design, compare Equation (6). This adjustment accounts for the possibility that we stop the trial at the interim analysis due to futility. The adjusted Bayesian

power therefore reduces to:

$$
\begin{aligned}
P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_1) = {} & \underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid \mathrm{BF}_{01}^{n_1}(y^1) < k, H_1)}_{\text{Efficacy at final analysis given efficacy at interim}} \cdot \underbrace{P(\mathrm{BF}_{01}^{n_1}(y^1) < k \mid H_1)}_{\text{Efficacy at interim analysis}} \\
& + \underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid k \leq \mathrm{BF}_{01}^{n_1}(y^1) < 1/k, H_1)}_{\text{Efficacy at final analysis given indecisive result at interim}} \cdot \underbrace{P(k \leq \mathrm{BF}_{01}^{n_1}(y^1) < 1/k \mid H_1)}_{\text{Indecisive result at interim analysis}} \\
& + \cancel{\underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid \mathrm{BF}_{01}^{n_1}(y^1) > 1/k, H_1)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\mathrm{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)}_{\text{Futility at interim analysis}}} \qquad (8)
\end{aligned}
$$

3.3 Problem II - Calculating the type-I-error rate

Likewise, suppose now that H_0 holds. In that case, it follows analogue from Equation (7) with H_1 replaced by H_0 that stopping for futility implies that no type-I-error can occur once we stop for futility. In that case, the Bayes factor indicates that we stop the trial for futility after n_1 observed outcomes, but if the trial would continue the Bayes factor could eventually sway around and indicate evidence for H_1, contributing to the type-I-error rate. Thus, when computing the Bayesian type-I-error rate unconditionally as

$$P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_0) \leq \alpha$$

compare Equation (3), we *overestimate* the true Bayesian type-I-error rate. As a consequence, when allowing to stop the trial for futility at the interim analysis, we must *subtract* the probability in Equation (7) with H_1 replaced by H_0 from the Bayesian type-I-error rate $P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_0)$ to obtain the correct Bayesian false-positive rate. The latter therefore reduces to:

$$
\begin{aligned}
P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_0) = {} & \underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid \mathrm{BF}_{01}^{n_1}(y^1) < k, H_0)}_{\text{Efficacy at final analysis given efficacy at interim}} \cdot \underbrace{P(\mathrm{BF}_{01}^{n_1}(y^1) < k \mid H_0)}_{\text{Efficacy at interim analysis}} \\
& + \underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid k \leq \mathrm{BF}_{01}^{n_1}(y^1) < 1/k, H_0)}_{\text{Efficacy at final analysis given indecisive result at interim}} \cdot \underbrace{P(k \leq \mathrm{BF}_{01}^{n_1}(y^1) < 1/k \mid H_0)}_{\text{Indecisive result at interim analysis}} \\
& + \cancel{\underbrace{P(\mathrm{BF}_{01}^{n_2}(y) < k \mid \mathrm{BF}_{01}^{n_1}(y^1) > 1/k, H_0)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\mathrm{BF}_{01}^{n_1}(y^1) > 1/k \mid H_0)}_{\text{Futility at interim analysis}}} \qquad (9)
\end{aligned}
$$

Otherwise, we overestimate the latter due to the option to stop the trial for futility after n_1 observed outcomes.[5]

Summing up, the price paid in statistical terms for introducing an interim analysis in the Bayesian two-stage design is (1) an adjustment which is needed to correct the Bayesian power based on sample size n_2 and (2) an adjustment which is needed to correct the Bayesian type-

[5] See also footnote 4. If we solely base the type-I-error control on the probability in Equation (3), we guarantee that the latter is fully controlled. However, we are too conservative in the sense that we overestimate the true type-I-error rate if we do not adjust it for introducing the option to stop the trial for futility, compare Equation (9).

I-error rate based on sample size n_2. In the following section, we turn to a solution of both of these issues. This allows to correctly calculate the Bayesian power and type-I-error rate of the Bayesian two-stage design in Figure 1 for a fixed sample size n_2, thereby meeting the criteria in Equation (3).

4 Solving both problems

Now, from a computational point of view it is straightforward to obtain the unadjusted Bayesian power $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ or type-I-error rate $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0)$ for a fixed sample size n_2. The principal approach to obtain such a power is detailed in Pawel and Held (2024), and details for a single-arm phase IIa trial are provided in Kelter and Pawel (2025).

Thus, subtracting the probability in Equation (7) from the unconditional (Bayesian) power requires first to compute that probability. Then, one can calculate the unadjusted Bayesian power and subtract the probability in Equation (7), see Equation (8). The probability in Equation (7) can be rewritten as

$$\underbrace{P(\text{BF}_{01}^{n_2}(y) < k \mid \text{BF}_{01}^{n_1}(y^1) > 1/k, H_1)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)}_{\text{Futility at interim analysis}} \tag{10}$$

$$= \underbrace{P(\text{BF}_{01}^{n_2}(y) < k, \text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)}_{\text{Efficacy at final analysis and futility at interim = futility-erased partial power}} \tag{11}$$

which we call henceforth *futility-erased partial power*. The following result provides a convenient numerical way to calculate $P(\text{BF}_{01}^{n_2}(y) < k, \text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)$:

Theorem 1. *Let n_1 and n_2 be the sample sizes of the interim and final analysis in the Bayesian two-stage design. Let the data $Y_i \mid \theta \sim \text{Bin}(n_i, \theta)$ for $i \in \{1, 2\}$ and assume the success probability θ has a truncated Beta prior, $\theta \sim \text{Beta}(a, b)_{[l,u]}$. Denote by $f(y_s^2, y_s^2 | a_d, b_d, l, u)$ the prior-predictive density under the hypothesis H_i, for $i = 0, 1$, where $H_0 : p \leq p_0$ and $H_1 : p > p_0$ for some $p_0 \in (0, 1)$. Denote $y_{critEff}^{n_2}$ as the critical value so that for $y_s^2 > y_{critEff}^{n_2}$ we have $\text{BF}_{01}^{n_2}(y) < k \mid H_i$ based on the prior-predictive. Denote $y_{critFut}^{n_1}$ as the critical value so that for $y_s^1 \leq y_{critFut}^{n_1}$ we have $\text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_i$, where y_s^1 denote the number of successes in the first batch $y^1 := (y_1, ..., y_{n_1})$ of patients and y_s^2 the number of successes in the second batch $y^2 := (y_{n_1+1}, ..., n_2)$ of patients to be possibly enrolled in the second stage of the trial. Then, the following equality holds:*

$$P(\text{BF}_{01}^{n_2} < k, \text{BF}_{01}^{n_1} > 1/k \mid H_i) = \sum_{i=0}^{\lfloor y_{critFut}^{n_1} \rfloor} \sum_{j=\lceil y_{critEff}^{n_2} \rceil - y_i}^{n_2 - n_1} f(y_s^1 = y_i, y_s^2 = y_j | a_d, b_d, l, u) \tag{12}$$

Proof. See the Appendix. □

	0	1	2	$3 = \lfloor y^{n_1}_{critFut} \rfloor$	...	$n_1 = 10$
0						
1						
2						
3						
...						
22				x		
23			x	x		
24		x	x	x		
25	x	x	x	x		
26	x	x	x	x		
27	x	x	x	x		
28	x	x	x	x		
29	x	x	x	x		
$n_2 - n_1 = 30$	x	x	x	x		

Table 1: Illustration of Theorem 1: Columns show the possible number of successes in the first batch of data $y^1 = (y_1, ..., y_{10})$ until the interim analysis is performed. Rows show the number of successes in the second batch of data $y^2 = (y_{11}, ..., y_{30})$. The illustration uses $n_1 = 10$ and $n_2 = 40$, so after $n_1 = 10$ patients the interim analysis is carried out.

Now, the denominator in Theorem 1 can be obtained immediately via the Bayes factor sample size planning with root-finding as outlined by Kelter and Pawel (2025). It is equal to the prior-predictive probability of $y_1 < y^{n_1}_{critFut}$ under H_i for the fixed sample size n_1.

The numerator is more complicated and is a sum over the joint marginal probability mass function. However, as both data in the first batch of size n_1 and in the second batch of size n_2 is binomially distributed, the left-hand side probability in Theorem 1 essentially reduces it to a discrete sum.

Now, suppose we obtain for $n_1 = 10$ the critical value $y^{n_1}_{critFut} = 3$ via root-finding. Suppose further that for $n_2 = 40$ we obtain $y^{n_2}_{critEff} = 25$, also via root-finding as detailed in Kelter and Pawel (2025). This implies that if we achieve $0, 1, 2$ or 3 successes at the interim analysis, we would need $25, 24, 23$ or 22 successes in the final analysis to sway the Bayes factor around to arrive at $\text{BF}^{n_2}_{01} < k$.

Table 1 illustrates what Theorem 1 expresses: In the first sum, we iterate over the first $i = 0$ to $\lfloor y^{n_1}_{critFut} \rfloor$ number of successes in the first batch of data for which $\text{BF}^{n_1}_{01} > 1/k$ holds, so the Bayes factor would stop for futility. In the example above, we used $\lfloor y^{n_1}_{critFut} \rfloor = 3$ so we iterate from zero to three in the first sum, which are the columns of Table 1. Then, the first sum iterates over $y_i = 0, ..., 3$ successes in this first batch of data. Now, if $y^1_s = 0$, we have zero successes in the first batch of data and we need at least 25 successes in the second batch of data $y^2 = (y_{11}, ..., y_{40})$ to sway the Bayes factor around to yield $\text{BF}^{n_2}_{01}(y^2) < k$ in the final analysis. Thus, for $y^1_s = 0$ in the first column of Table 1 the crosses mark the rows from 25 to

30 successes in the second batch of data, which are exactly the predictive probabilities we need to sum over for a fixed $y_s^1 = 0$ to yield $\text{BF}_{01}^{n_2}(y^2) < k$ in the final analysis. The second sum in the right hand-side of Equation (12) therefore sums over the prior-predictive probabilities $f(y_s^1 = y_i, y_s^2 = y_j | a_d, b_d, l, u)$ under H_i for the values $y_j = \lceil y_{critEff}^{n_2} \rceil - y_i = 25 - y_i$ to $n_2 - n_1 = 30$, and the first sum over $y_i = 0$ to $y_i = \lfloor 3 \rfloor$.

If $y_i = 1$ (second column in Table 1) we obtain a single success in the first batch y^1 of data and the second sum in Equation (12) iterates over $y_j = \lceil y_{critEff}^{n_2} \rceil - y_i = 25 - 1 = 24$ to $n_2 - n_1 = 30$ successes, so we obtain 25 to 31 successes in total then to sway the Bayes factor eventually around to $\text{BF}_{01}^{n_2}(y^2) < k$ in the final analysis. Thus, the rows 24 to 30 are marked for column 1 in Table 1.

Table 1 thus shows graphically what Theorem 1 means: We iterate over the number of successes from 0 to $\lfloor y_{critFut}^{n_1} \rfloor$, and for each value we iterate over the needed number of successes in the second batch of patient data of size $n_2 - n_1$ to eventually sway the Bayes factor around in the final analysis, arriving at evidence in favour of H_1. This way, we compute the desired sum of marginal probabilities which equate to the probability $P(\text{BF}_{01}^{n_2} < k | \text{BF}_{01}^{n_1} > 1/k, H_i)$.

Equation (12) has one important advantage: It makes use only of the prior-predictive densities. As we do not condition on the observed data at interim, and take a prediction perspective which uses solely the prior before carrying out the trial, all elements in Equation (12) can be computed easily. As shown in the proof of Theorem 1, the prior-predictive can be computed via beta functions and regularized incomplete beta functions. The critical values $y_{critFut}^{n_1}$ and $y_{critEff}^{n_2}$ can be computed via the root-finding algorithm detailed in Kelter and Pawel (2025) numerically. Thus, Equation (12) can be computed in a relatively straightforward manner.

Taking stock, this allows to compute power and type-I-error rates which adjust for the introduction of a single interim analysis in the two-stage design based on the trinomial tree branching.

4.1 Computing the correct power for fixed sample sizes n_1 and n_2

Now, the previous line of reasoning has led to the solution to obtain the correct power for a fixed sample size n_2, which accounts for the possibility to stop the trial for futility at the interim analysis. We must obtain the probability in Equation (7) and subtract it from the power for sample size n_2, and we proceed in four steps:

1. Compute the unadjusted probability to stop due to futility at interim $P(\text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_1)$ via root-finding based on n_1. The root-finding approach then yields a critical value $y_{critFut}^{n_1}$ for a fixed n_1.

2. Compute the left-hand probability in Equation (6), the unadjusted power $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1)$ at sample size n_2 – for the final analysis – via root-finding. The root-finding

approach yields a critical value $y^{n_2}_{critEff}$ for a fixed n_2, based on which the Bayes factor would signal convincing evidence for H_1 at the end of the trial.

3. Compute Equation (12) for the values $y^{n_1}_{critFut}$ and $y^{n_2}_{critEff}$ obtained in the previous steps.

4. Compute the *futility-erased partial power* in Equation (7) and subtract it from the unadjusted power $P(\mathrm{BF}^{n_2}_{01}(y) < k \mid H_1)$ computed in step two.

Performing these four steps then yields an adjusted Bayesian power which accounts for the loss in power due to allowing to stop the trial early for futility.

4.2 Computing the correct type-I-error rate for fixed sample sizes n_1 and n_2

Computing the correct type-I-error rate for fixed sample sizes n_1 and n_2 proceeds likewise. The only difference is that Equation (7) changes to

$$\underbrace{P(\mathrm{BF}^{n_2}_{01}(y) < k \mid \mathrm{BF}^{n_1}_{01}(y^1) > 1/k, H_0)}_{\text{Efficacy at final analysis given futility at interim}} \cdot \underbrace{P(\mathrm{BF}^{n_1}_{01}(y^1) > 1/k \mid H_0)}_{\text{Futility at interim analysis}} \tag{13}$$

so H_1 is replaced by H_0. Essentially, we make use of Theorem 1 with $H_i = H_0$, so we perform the four steps in the last subsection with H_1 replaced by H_0. In step four, we could rephrase the *futility-erased partial power* in Equation (6) as *futility-erased partial type-I-error* then. Thus, Equation (13) is called *futility-erased partial type-I-error* henceforth.

5 Calibrating the two-stage design

Now, in this section we turn to the calibration of our design. The last section illustrated how to calculate the correct power and type-I-error rate of the Bayesian two-stage design when a single interim analysis is carried out. However, in these considerations, n_1 and n_2 were fixed. In practice, we do however specify the boundaries α and $1 - \beta_{n_2}$ in Equation (3), and want a calibration algorithm to reveal which sample sizes n_1 and n_2 we should use based on the specified boundaries for type-I-error rate and power.

Two outcomes are possible: If we subtract the futility-erased partial power from the Bayesian power under H_1, our power may drop below $1 - \beta_{n_2}$, compare Equation (3). In that case, we must increase n_2 to account for the subtraction of the futility-erased partial power.

On the other hand, under H_0, erasing the futility-erased partial type-I-error can only decrease the type-I-error rate. If two sample sizes n_1 and n_2 therefore are not calibrated with regard to the type-I-error rate in Equation (3), it may happen that the design becomes fully calibrated once

the futility-erased partial power is taken into consideration and subtracted from the original type-I-error rate obtained via root-finding for n_2.

5.1 Calibration algorithm

The algorithm to calibrate the Bayesian two-stage design is therefore relatively simple:

1. Start with a sample size n_2 for which the requirement on Bayesian power in Equation (3) holds strictly. Such a sample size can be found via the root-finding approach outlined in Pawel and Held (2024) and Kelter and Pawel (2025).

2. Fix a smallest sample size $i \in \mathbb{N}$ and set $n_1 := i$ for the interim analysis. Iterate from $n_1 := i$ to $n_1 := n_2 - 1$. For each value, calculate the Bayesian power and type-I-error rates in Equation (3) and adjust these by subtracting the futility-erased partial power and futility-erased partial type-I-error. This step is crucial to obtain valid power and type-I-error rates, which account for multiple testing in the sequential design.

3. If the Bayesian power drops below $1 - \beta_{n_2}$ for the chosen β_{n_2} (e.g. $1 - \beta_{n_2} = 0.80$) after reducing the power by the futility-erased partial power, or if the Bayesian type-I-error rate is larger than α for the chosen α, or optionally, if the condition in Equation (5) is not met, increase n_1 and repeat the last step. If $n_1 = n_2 - 1$, increase n_2 and repeat the last step.

The later the interim analysis is carried out, that is, the larger n_1 is relative to n_2, the less probable it becomes that the Bayes factor signals evidence for H_0 when H_1 is true. As a consequence, the probability to stop early for futility and sway the Bayes factor around in the second stage becomes smaller and smaller for increasing n_1. Therefore, the futility-erased partial power decreases when increasing n_1. Increasing n_2 ensures that from the consistency of the Bayes factor, the power requirement in Equation (3) will eventually be met.[6]

Under H_0, the points above ensure that it becomes more and more difficult, and eventually impossible, to sway the Bayes factor around, when n_1 is increased. Thus, the futility-erased partial type-I-error decreases when increasing n_1. This ensures that increasing n_1 for a fixed n_2 calibrates the Bayesian type-I-error rate. If for the fixed n_2 and n_1 no solution can be obtained, increasing n_2 assures that the Bayesian power requirement will eventually be met, and increasing n_1 for that larger n_2 then calibrates the Bayesian type-I-error rate. Together, these aspects imply a calibration in terms of Equation (3).

[6]The algorithm thus increases n_1 only to the point until the futility-erased partial power is small enough to still yield a power larger than $1 - \beta_{n_2}$

What about Equation (5)? Under H_0, increasing n_1 for a fixed n_2 implies that the probability to stop early for futility at the interim analysis increases, so the condition

$$P(\text{BF}_{01}^{n_1}(y^1) > 1/k|H_0) > f$$

will eventually be met for large enough n_1, too. If no value n_1 can be found for which the latter holds, n_2 can be increased, thereby increasing the Bayesian power, decreasing the type-I-error rate, and allowing larger values of n_1 in turn. This ensures that eventually a large enough value n_1 is found for which the probability above passes the threshold f under H_0, again due to the consistency of the posterior distribution (Kleijn, 2022; van der Vaart, 1998). In some cases it might happen that the Bayesian power is larger than required, or the Bayesian type-I-error smaller than required, or even both, to satisfy the futility condition in the last display above. This is due to the discreteness of the binomially distributed data, a phenomenon which is well documented in the literature, see Kelter and Pawel (2025). We return to this problem and how to solve it in the first example in the following section. In general, it remains difficult to say how the three design characteristics relate to another. We provide results for a variety of realistic settings in Section 6 below and return to this aspect.

5.2 Possible issues with the calibration algorithm

Suppose n_1 is quite small relative to the evidence threshold k for the Bayes factor. Then, it might happen that for that n_1, there is no $y_{critFut}^{n_1} \in \{0, ..., n_1\}$ so that for $y_1 < y_{critFut}^{n_1}$ the condition $\text{BF}_{01}^{n_1}(y^1) > 1/k$ holds under the chosen design prior. As a consequence, the lowest blue branch in Figure 2 has probability zero, that is,

$$P(\text{BF}_{01}^{n_1}(y^1) > 1/k|H_i) = 0 \tag{14}$$

holds. In that case, the probability in Equation (5) can exceed no threshold $f > 0$ to stop for futility with at least probability f. As a consequence, the design cannot be calibrated for that choice of n_1, if the additional requirement in Equation (5) is desired for the resulting sequential design, that is, if a minimum probability f is warranted to obtain compelling evidence under H_0. Thus, the threshold k based on which one stops for futility must be chosen not too conservative, so that there exists a minimal value for n_1 based on which the condition in Equation (5) can hold. In all settings considered and reported in Section 6 below, this phenomenon never occurred. We found that setting the threshold at $k = 3$ – resembling moderate evidence – always yields a calibrated design in our examples, while using the more strict $k = 10$ – resembling strong evidence – is sometimes a too conservative choice and yields no solution to calibrate a design.

5.3 The price of introducing an interim analysis

Also, suppose that the conditions in Equation (3) hold for a value n_2. Then, the futility-erased partial power in Equation (7) shows that the power of the two-stage design with an interim analysis is always smaller than the power of the design without an interim analysis under the following conditions:

Theorem 2. *Let n_1 and n_2 be the interim analysis and final analysis sample sizes in the Bayesian two-stage design, k be the evidence threshold and denote by $\text{BF}_{01}^{n_1}(y^1)$ and $\text{BF}_{01}^{n_2}(y)$ the Bayes factors for the interim and final analysis, where $y^1 := (y_1, ..., y_{n_1})$ is the interim and $y := (y_1, ..., y_{n_1}, y_{n_1+1}, ..., y_{n_2})$ the full trial data. If*

$$P(\text{BF}_{01}^{n_2}(y) < k \mid \text{BF}_{01}^{n_1}(y^1) > 1/k, H_i) \cdot P(\text{BF}_{01}^{n_1}(y^1) > 1/k \mid H_i) \neq 0 \tag{15}$$

holds for $i = 1$, the Bayesian power in Equation (3) of the two-stage design for a fixed sample size n_2 is always smaller than the power of the same trial design without an interim analysis at sample size n_1, for any $n_1 < n_2$. If Equation (15) holds for $i = 0$, the Bayesian type-I-error rate in Equation (3) of the two-stage design for a fixed sample size n_2 is always smaller than the Bayesian type-I-error rate of the same trial design without an interim analysis at sample size n_1, for any $n_1 < n_2$.

Proof. Follows from Equation (3), Equation (6) and the composition of the Bayesian power as shown in Equation (8) for the case of $i = 1$, and from Equation (3), Equation (6) and Equation (9) for the case of $i = 0$. □

The result above clarifies the statistical price paid when introducing an interim analysis into a Bayesian design with Bayes factor testing. The power will almost always be smaller, while the Bayesian type-I-error rate will almost always improve. The condition for these facts to hold is, in other words, that the lowest blue branch in Figure 2 does not vanish by having probability zero.[7] Importantly, Theorem 2 provides a convenient option to check whether an optimal calibration as outlined in the next subsection is possible.

5.4 Optimal calibration algorithm

We close this section by providing an optimal calibration algorithm in addition to the standard calibration algorithm detailed in Section 5.1. In phase II trials, when H_0 holds, one would ideally stop at the interim analysis at sample size n_1 with probability one. On the other hand, if H_1 holds, and no stopping for efficacy is allowed – which is a usual assumption in phase

[7] If this is possible, because n_1 is too small to accumulate evidence $1/k$ in favour of H_0 via $\text{BF}_{01}^{n_1}(y^1) > 1/k$, the possibility of an interim analysis is not given anymore and the two-stage design reduces to a single-stage design.

II trials, see Lee and Liu (2008) – it would be ideal if the trial always is carried out until n_2 patients are enrolled. Two popular strategies which go back to Simon (1989) are to minimize the maximum number of patients which is enrolled, that is, n_2, or to minimize the expected number of patients under H_0, that is, $E[N|H_0]$. Expected sample size under H_0 is calculated as follows:

$$E[N|H_0] = n_1 \cdot \underbrace{P(BF_{01}^{n_1} > 1/k|H_0)}_{\text{stop the trial early for futility at } n_1} + n_2 \cdot \underbrace{(1 - P(BF_{01}^{n_1} > 1/k|H_0))}_{\text{continue the trial until } n_2} \tag{16}$$

Expected sample size under H_1 is calculated likewise as follows:

$$E[N|H_1] = n_1 \cdot \underbrace{P(BF_{01}^{n_1} > 1/k|H_1)}_{\text{stop the trial early for futility at } n_1} + n_2 \cdot \underbrace{(1 - P(BF_{01}^{n_1} > 1/k|H_1))}_{\text{continue the trial until } n_2}$$

When minimizing $E[N|H_0]$, Equation (16) shows that this implies that the probability to stop for futility under H_0 is maximized, while the probability to continue the trial until n_2 under H_0 is minimized. This is the approach taken in two-stage phase II optimal design of Simon (1989), and we provide an analogue optimal Bayesian sequential Bayes factor design by minimizing $E[N|H_0]$ under the constraints that Equation (3) holds.

Definition 1 (Optimal two-stage Bayes factor design)**.** *Let $\alpha \in (0, 1)$ and $\beta_{n_2} \in (0, 1)$ be given and denote by n_{min} the minimum sample size at which the trial is allowed to stop for futility and by n_{max} the maximum trial size of the final analysis. The optimal two-stage Bayes factor design with interim sample size n_1 and full sample size n_2, is the design for which n_1 and n_2 are the solution to the minimization problem*

$$\begin{aligned} \min_{n_1, n_2} \quad & E[N|H_0] \\ \textit{s.t.} \quad & P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_0) \leq \alpha \\ & \textit{and} \qquad P(\mathrm{BF}_{01}^{n_2}(y) < k \mid H_1) \geq 1 - \beta_{n_2} \\ & \textit{and} \qquad n_{min} \leq n_1 < n_2 \leq n_{max} \end{aligned} \tag{17}$$

Optionally, we can provide an optimal Bayesian sequential Bayes factor design by additionally require Equation (5) to hold. The optimal calibration algorithm is relatively simple: Provide a minimum value n_{min} for n_1 and a maximum value n_{max} for n_2 based on subject-domain knowledge about how many patients can realistically be enrolled in a given timeframe. Often, this sets an upper bound n_{max} on n_2. The minimum value n_{min} for n_1 can be set as small as wanted, e.g. $n_1 = 4$ or $n_1 = 5$. In this range, compute the design characteristics in Equation (3) (and optionally in Equation (5)), and filter those values of n_1 and n_2 which are calibrated by running the calibration algorithm in Section 5.1. Next, compute $E[N|H_0]$ for these

calibrated designs and return the design which minimizes the latter.

Before turning to the examples, we return briefly to Theorem 2. Based on Theorem 2, we know that a sequential design's power can only decrease compared to the power the same design without an interim analysis. For that same design without an interim analysis, we can compute the power and type-I-error rate almost instantaneously via root-finding (Kelter & Pawel, 2025). Thus, when a value n_{max} is provided, we can first check if the power of the non-sequential design is above $1-\beta_{n_2}$. If not, Theorem 2 assures that the two-stage design cannot be calibrated and we must increase n_{max}. This can speed up the calibration algorithm even further.

6 Examples

In this section, we provide the results of the sequential two-stage design for a variety of practically relevant settings for a phase-II-a trial with binary endpoint. We stress that the results here are not simulation results but can be obtained in only seconds by applying the calibration algorithm outlined in Section 5. Also, no Monte Carlo standard error is needed.

Regarding the relevant design characteristics, we provide the (Bayesian and frequentist) power, expected sample size under H_0, type-I-error rates (Bayesian and frequentist) and the probability to obtain compelling evidence under H_0.

6.1 An example showing why oscillations cause problems

We start with an example which returns to the fact that oscillations are a well-known phenomenon for the beta-binomial model. Thus, we test $H_0 : p \leq 0.1$ against $H_1 : p > 0.1$, and aim for $\alpha = 0.05$ and $\beta_{n_2} = 0.2$, so we cap the type-I-error rate at 5% and require at least 80% power. We do not add the requirement of a minimum probability f for compelling evidence under H_0 right now. For the optimal Bayesian designs, we search for sample sizes which minimize $E[N|H_0]$ in the range of $n_1 = 5$ to $n_2 = 40$, so $n_{min} = 5$ and $n_{max} = 40$. Also, we always use flat analysis priors in the final analysis, while we investigate the design's operating characteristics under several design priors used during the planning stage of the trial.

Table 2 shows the results for Simon's optimal and minimax designs[8] and for the optimal sequential Bayes factor design using frequentist or Bayesian power and the evidence thresholds $k = 1/3$ and $k_f = 3$ for evidence and futility. Thus, at the interim analysis we check whether $\text{BF}_{01}^{n_1}(y^1) > k_f$ to stop for futility, and at the final analysis we check $\text{BF}_{01}^{n_2}(y) < k$ to test whether $H_1 : p > p_0$ at the full sample size n_2 at the end of the trial holds. Note that the

[8]These can either be computed via the online calculator of the UNC Lineberger Comprehensive Cancer Center at http://www.cancer.unc.edu/biostatistics/program/ivanova/SimonsTwoStageDesign.aspx or via the R software package `clinfun`, available at https://cran.r-project.org/web/packages/clinfun/index.html.

Design	n_1	n_2	type-I-error	power	$E[N\|H_0]$	PET(p_0)
Simon's 2-stage phase II design						
minimax design	15	25	0.0328	0.8017	19.51	0.5490
optimal design	10	29	0.0471	0.8051	15.01	0.7361
Optimal sequential Bayes factor design						PCE(p_0)
frequentist power, $k = 1/3$, $k_f = 3$	10	29	0.0471	0.8051	15.01	0.7361
Bayesian power, $k = 1/3$, $k_f = 3$						
with $\text{Beta}_{[0.1,1]}(1, 1)$ prior	5	15	0.0480	0.8107	9.10	0.5905
with $\text{Beta}_{[0.1,1]}(7, 15)$ prior	11	36	0.0470	0.8017	18.57	0.6974
with $\text{Beta}_{[0.1,1]}(11.29, 25)$ prior	12	28	0.0477	0.8020	17.46	0.6590
with $\text{Beta}_{[0.1,1]}(22, 50)$ prior	10	36	0.0432	0.8038	16.86	0.7361

Table 2: Design characteristics for Simon's two-stage design and different Bayesian sequential designs for Example 1. n_1 and n_2 denote the interim and full sample sizes, the type-I-error is always computed via a point prior at p_0 for all Bayesian designs, while the power is computed as frequentist power with a point prior at $p_1 = 0.3$, or as Bayesian power with a truncated $\text{Beta}_{[0.1,1]}(a_d, b_d)$ prior, where $a_d = b_d = 1$ denotes the flat prior. Larger values of a_d and b_d imply a more informative prior, see Figure 5. Bayesian designs are calibrated as optimal so that they minimize the expected sample size $E[N|H_0]$ under H_0. k denotes the threshold used in Bayes factor designs for type-I-error rates and power in Equation (3), and k_f denotes the threshold used for the additional requirement for compelling evidence under H_0 in Equation (5). PET(p_0) denotes the probability of early termination for Simon's designs, which is the probability to stop the trial early for futility. PCE(p_0) denotes the probability of compelling evidence in favour of H_0, which is computed with a point prior at p_0, and based on the Bayes factor threshold k_f for futility.

Bayesian design calculates the type-I-error rate strictly frequentist using a point prior at $p_0 = 0.1$, while Bayesian power is also strictly frequentist, using a point prior at $p_1 = 0.3$. Thus, the Bayesian design has a fully frequentist interpretation here. Figure 4 shows how oscillations occur for different values of n_1. Thus, it might happen that the power drops below the desired threshold of 80%, e.g. from 0.87 at $n_1 = 8$ to 0.765 at $n_1 = 9$ in Figure 4. A possible solution to this phenomenon is to require the power to stay above the threshold β_{n_2} (and likewise, to require the type-I-error rate to stay below α) for the next 10 natural numbers, see also the discussion in Kelter and Pawel (2025). This is the strategy adopted here and which in all examples we carried out yielded reliable results.[9] We recommend plotting the design's power and type-I-error rate for all interim analyses between n_{min} and n_{max} like in Figure 4 to check whether the resulting power remains calibrated.[10]

Table 2 provides several insights: First, the optimal sequential Bayes factor design with frequentist power and $k = 1/3$ and $k_f = 3$ recovers Simon's two-stage optimal design. All

[9]For details, see the `bfpwr` package on CRAN: https://cran.r-project.org/web/packages/bfpwr/index.html.

[10]This is also relevant in case a trial is stopped due to too many severe adverse events. Then, recalculation of the power and type-I-error rate is possible by plots such as Figure 4 even when a deviation from the trial protocol became necessary.

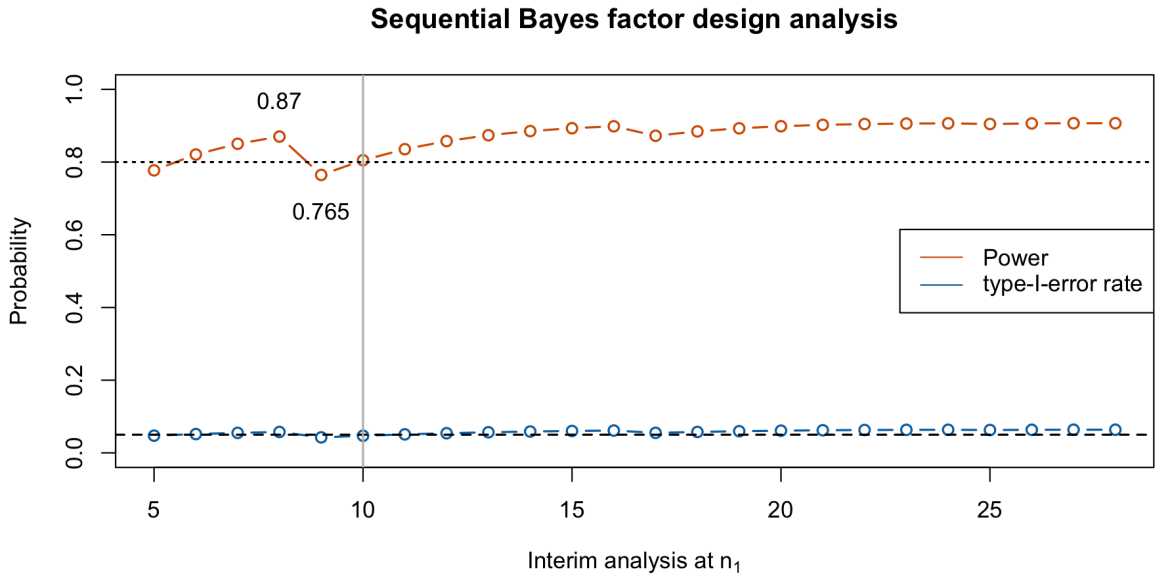


Figure 4: Design characteristics for the first example, showing how oscillations occur due to the discreteness of the binomially distributed data. Vertical line shows the value $n_1 = 10$ which is the optimal value that, together with $n_2 = 29$ yields a calibrated sequential Bayes factor design based on frequentist power computed at $p_1 = 0.3$, and which is optimal in the sense that it minimizes the expected sample size $E[N|H_0]$.

design characteristics are identical, although the concept of probability of compelling evidence under H_0 – that is, PCE(p_0) in Table 2 – which is computed via a point prior at $p_0 = 0.1$, is stronger than the concept of probability of early termination under H_0 – that is, PET(p_0) in Table 2. PCE(p_0) is the probability of compelling evidence under H_0 and implies that we actually have moderate evidence in favour of $H_0 : p \leq 0.1$, while PET(p_0) only implies that we have not sufficient evidence in favour of $H_1 : p > 0.1$. Absence of evidence is, however, no evidence of absence (Altman & Bland, 1995).

Second, shifting to Bayesian power and inspecting the optimal sequential Bayes factor design characteristics demonstrates the influence of the truncated Beta prior on H_1 to calculate the power. For a flat $\text{Beta}_{[0.1,1]}(1, 1)$ prior, larger probabilities like $p = 0.5$ or $p = 0.8$ have the same a priori probability than smaller ones such as $p = 0.25$. As a consequence, the sample sizes n_1 and n_2 for which a calibrated design can be found become smaller. This is, in particular, because the choice of n_1 and n_2 is primarily driven by the requirement on power, compare Figure 4. Table 2 shows that the sample sizes decrease to $n_1 = 5$ and $n_2 = 15$, and the design is still calibrated under a Bayesian notion of power, while the type-I-error-rate is still calibrated from a frequentist point of view (that is, under a point prior on $p_0 = 0.1$). Due to the smaller values $n_1 = 5$ and $n_2 = 15$ compared to $n_1 = 10$ and $n_2 = 29$ under a strictly frequentist power on $p_1 = 0.3$, the resulting expected sample size under H_0 (and H_1) reduces, too. Therefore,

the expected sample size under H_0 beats the one of Simon's two-stage optimal design above, but the price is the shift to a Bayesian concept of power. Also, the probability of compelling evidence under H_0 decreases a little, because n_1 is smaller now and it becomes more difficult for the Bayes factor to accumulate moderate evidence in favour of H_0 after only $n_1 = 5$ instead of $n_1 = 10$ patients.

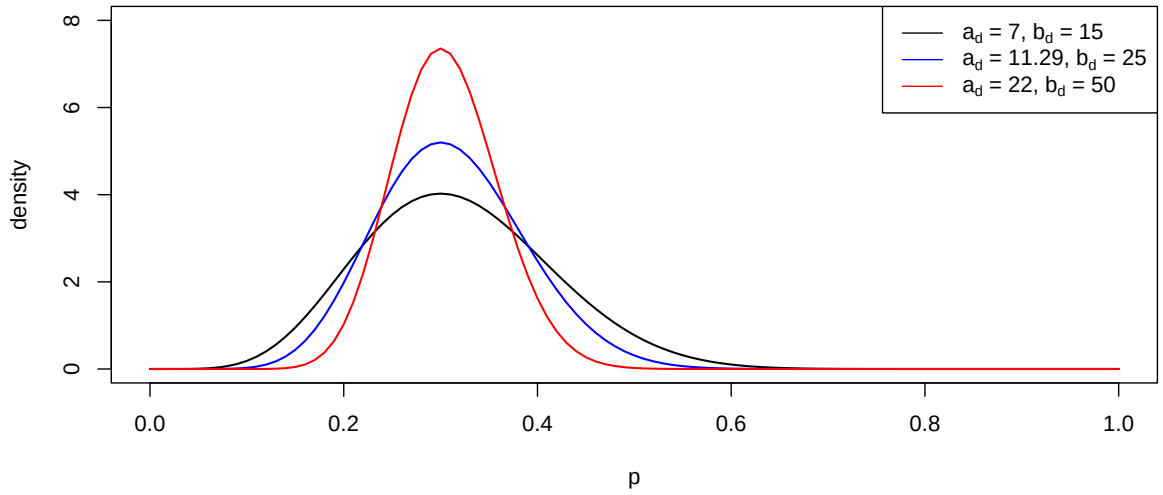


Figure 5: Informative $\text{Beta}_{[0.1,1]}(a_d, b_d)$ priors used in the first example. A prior with larger sum of $a_d + b_d$ corresponds to a more informative prior, where a_d can be interpreted as already observed failures and b_d as already observed successes.

Note that a flat prior on $[0.1, 1]$ might be unrealistic, because extremely large probabilities such as $p = 0.9$ are thought of equally probable a priori as small ones like $p = 0.15$. As a consequence, it might be more realistic to choose a prior centered at the frequentist point prior value $p_1 = 0.3$. We proceed to this option now and investigate how the design characteristics change. Figure 5 shows Beta priors centered at $p = 0.3$ for increasing informativity, and Table 2 includes the optimal sequential Bayes factor design characteristics for these three priors. Two aspects are noteworthy: First, the more informative the prior, the smaller the expected sample size becomes. This is to be expected, because the prior concentrates more closely around $p = 0.3$ and puts less probability mass on smaller values close to $p_0 = 0.1$. Second, the PCE(p_0) is not influenced by the prior on H_1, and differences are solely attributed to the changes in n_1 and n_2.[11]

[11] As a sidenote, when adding the requirement of $\text{PCE}(p_0) > f$ for $f = 0.7$ to the sequential Bayes factor design with the flat Beta prior on H_1, it also recovers Simon's optimal two-stage design as a special case in this example.

6.2 Single-arm phase II trial for nonsmall cell lung cancer

In this subsection, we adopt an example of a clinical trial discussed in Kelter and Schnurr (2024). A single-arm phase IIA study to demonstrate the efficacy of a novel combination therapy as front-line treatment in patients with advanced nonsmall cell lung cancer is considered and the primary endpoint is binary. The primary objective of the study was to assess the efficacy of the novel treatment, which involved the combination of a vascular endothelial growth factor antibody plus an epidermal growth factor receptor tyrosine kinase inhibitor. The primary endpoint is the clinical response rate, that is, the rate of complete response and partial response combined, to the new treatment. The hypotheses of interest are

$$H_0 : p \leq p_0 \text{ versus } H_1 : p > p_1$$

The current standard treatment yields a response rate of $\approx 20\%$, so that $p_0 = 0.2$. The target response rate of the new regimen is 40%, so $p_1 = 0.4$. We stress that somewhat unrealistic, the parameter interval $(0.2, 0.4]$ is somehow excluded from the hypotheses considered, a choice sometimes seen in phase IIA studies, compare Lee and Liu (2008). While such a distance between H_0 and H_1 allows for an easier separation between both hypotheses via statistical testing, we refrain from doing so. We provide the optimal sequential Bayes factor design under truncated Beta design priors centered on $p_1 = 0.4$ with varying informativity, so we select 0.4 for the mode of the design prior and opt for a slightly informative $\text{Beta}_{[0.2,1]}(10.33, 15)$ prior, and a strongly informative $\text{Beta}_{[0.2,1]}(13.66, 20)$ prior. Also, we provide results under a flat Beta prior, and using frequentist power. Like always, type-I-error rates are calculated as frequentist error rates with a point prior at p_0. For the analysis priors, we select flat truncated Beta priors both under H_0 and H_1, which is the usual choice.

Now, in Kelter and Schnurr (2024), a sequential design was used and calibrated via a Monte Carlo simulation to attain the boundaries $\alpha \leq 0.1$ and $\beta \leq 0.1$ for the Bayesian type-I-error rate and power. With the novel idea of trinomial tree branching and the derived results, no simulation is necessary to compute the optimal sequential Bayes factor design now. We also adopt the choice $\alpha \leq 0.1$ and $\beta \leq 0.1$. Now, instead of conducting Monte Carlo simulations, we use trinomial tree branching, correct the required probabilities, and carry out our calibration algorithm to yield a design for which $P(\text{BF}_{01}^{n_2}(y) < k \mid H_0) \leq \alpha$ and $P(\text{BF}_{01}^{n_2}(y) < k \mid H_1) \geq 1 - \beta_{n_2}$ for $\alpha = 0.1$ and $\beta_{n_2} = 0.2$, compare Equation (3).

In addition, we add the requirement $P(\text{BF}_{01}^{n_1}(y^1) > k_f | H_0) > f$, compare Equation (5), in some cases in the following example. Thus, we can add the requirement that we obtain compelling evidence under H_0, e.g. for $f = 0.6$ we have 60% probability to reach a Bayes factor that passes our threshold for evidence. For notational clarity, we denote this threshold as k_f (for futility), because we could choose a different value k_f to stop for futility at n_1 than to stop

for evidence based on threshold k at n_2. To obtain an optimal sequential Bayes factor design with out approach, we analyzed the metrics of the resulting designs in the range of $n_{min} = 5$ to $n_{max} = 60$, while for some cases (with very informative priors and high requirements on power, e.g. 90%) we had to increase n_{max} up to $n_{max} = 150$ in the most extreme case. However, this happened solely under the strongly informative prior and a evidence threshold for strong evidence according to Jeffreys (1939).

Table 3 on the next page provides the resulting characteristics of the different designs: The optimal calibration algorithm for the sequential Bayes factor design recovers Simon's optimal design, compare the values for the Bayesian design with frequentist power, $k = 1/3$, $k_f = 3$. Thus, in this example, under a moderate threshold k for evidence and for futility k_f, Simon's optimal design is a special case of our Bayesian design when power is computed under a frequentist point prior.

Second, when adding the requirement of compelling evidence to the optimal sequential Bayes factor designs, the sample size of the interim analysis often increases. For example, for Bayesian power with $k = 1/3$ and $k_f = 3$, the sample sizes increase from $n_1 = 27$, $n_1 = 28$ and $n_1 = 24$ to $n_1 = 30$ when adding $f = 0.6$ as the minimum probability for $\text{PCE}(p_0)$. One possible reason is that to accumulate 60% probability to reach a Bayes factor of at least 3 in favour of H_0, the interim sample size n_1 must be large enough to express that much evidence statistically. When adding such a requirement, the expected sample size thus increases slightly, as a consequence. Still, there is no option to add this additional layer of evidence on e.g. Simon's optimal design. When the trial is stopped there for futility, there is not sufficient evidence to reject H_0. However, absence of evidence is no evidence of absence, so it remains unclear whether H_0 holds true in that case.[12] For the Bayesian design with that additional requirement, once the trial is stopped at $n_1 = 30$, we can be sure that if H_0 is true, we arrive with 60% at a Bayes factor $\text{BF}_{01} > k_f = 3$. Thus, we have a 60% probability to find moderate evidence for H_0 during the interim analysis, if H_0 holds. The result $\text{BF}_{01} > k_f = 3$ implies at least moderate evidence in favour of H_0, and absence of evidence becomes presence of evidence (for H_0). We can be moderately certain that H_0 is true when we stop for futility for such a design. If we had chosen $k_f = 10$, we could even be strongly certain. This highlights an important aspect of the proposed design: Speaking of stopping early for futility is not appropriate. More appropriate is to speak of stopping for compelling evidence in favour of the null hypothesis, which shows that it is possible to accept H_0 even after the interim analysis at n_1.

Third, Table 3 shows that frequentist power calculations could be regarded as much too optimistic, in particular, when inspecting the result under different Beta priors for the Bayesian designs. It remains entirely unclear how effective the drug is, and indeed, efficacy probabilities

[12] Likewise, a large p-value close to 1 does not imply that H_0 is true.

Design	n_1	n_2	type-I-error	power	$E[N\|H_0]$	PET(p_0)
Simon's 2-stage phase II						
minimax design	19	36	0.0861	0.9024	28.26	0.4551
optimal design	17	37	0.0948	0.9033	26.02	0.5489
Optimal sequential Bayes factor						PCE(p_0)
frequentist power						
$k = 1/3$, $k_f = 3$	17	37	0.0948	0.9033	26.02	0.5489
$k = 1/3$, $k_f = 3$, $f = 0.6$	30	36	0.0886	0.9091	32.36	0.6070
$k = 1/10$, $k_f = 3$	21	51	0.0340	0.9021	33.42	0.5860
$k = 1/10$, $k_f = 3$, $f = 0.6$	30	50	0.0302	0.9011	37.86	0.6070
Bayesian power						
$k = 1/3$, $k_f = 3$						
$\text{Beta}_{[0.2,1]}(1, 1)$ prior	27	54	0.0988	0.9003	39.46	0.5387
$\text{Beta}_{[0.2,1]}(10.33, 15)$ prior	28	67	0.0981	0.9006	47.48	0.5005
$\text{Beta}_{[0.2,1]}(13.66, 20)$ prior	24	58	0.0958	0.9001	42.36	0.4599
$k = 1/10$, $k_f = 3$						
$\text{Beta}_{[0.2,1]}(1, 1)$ prior	42	100	0.0325	0.9002	69.21	0.5309
$\text{Beta}_{[0.2,1]}(10.33, 15)$ prior	65	100	0.0339	0.9001	79.93	0.5735
$\text{Beta}_{[0.2,1]}(13.66, 20)$ prior	40	88	0.0347	0.9005	59.53	0.5931
$k = 1/3$, $k_f = 3$, $f = 0.6$						
$\text{Beta}_{[0.2,1]}(1, 1)$ prior	30	58	0.0928	0.9008	41.00	0.6070
$\text{Beta}_{[0.2,1]}(10.33, 15)$ prior	30	76	0.0923	0.9008	48.08	0.6070
$\text{Beta}_{[0.2,1]}(13.66, 20)$ prior	30	63	0.0978	0.9056	42.97	0.6070
$k = 1/10$, $k_f = 3$, $f = 0.6$						
$\text{Beta}_{[0.2,1]}(1, 1)$ prior	30	130	0.0279	0.9016	69.30	0.6070
$\text{Beta}_{[0.2,1]}(10.33, 15)$ prior	30	147	0.0263	0.9011	75.98	0.6070
$\text{Beta}_{[0.2,1]}(13.66, 20)$ prior	30	108	0.0273	0.9001	60.66	0.6070

Table 3: Design characteristics for Simon's two-stage design and different Bayesian sequential designs for Example 1. n_1 and n_2 denote the interim and full sample sizes, the type-I-error is always computed via a point prior at p_0 for all Bayesian designs, while the power is computed as frequentist power with a point prior at $p_1 = 0.4$, or as Bayesian power with a truncated $\text{Beta}_{[0.2,1]}(a_d, b_d)$ prior, where $a_d = b_d = 1$ denotes the flat prior. Larger values of a_d and b_d imply a more informative prior, see Figure 5. Bayesian designs are calibrated as optimal so that they minimize the expected sample size $E[N|H_0]$ under H_0. k denotes the threshold used in Bayes factor designs for type-I-error rates and power in Equation (3), and k_f denotes the threshold used for the additional requirement for compelling evidence under H_0 in Equation (5). PET(p_0) denotes the probability of early termination for Simon's designs, which is the probability to stop the trial early for futility. PCE(p_0) denotes the probability of compelling evidence in favour of H_0, which is computed with a point prior at p_0, and based on the Bayes factor threshold k_f for futility.

in the range between $p = 0.2$ and $p = 0.4$ are possibly very likely. Frequentist power was computed under $p_1 = 0.4$, however. As a consequence Bayesian power with $k = 1/3$ and $k_f = 3$ could be more realistic, because a flat Beta design prior distributes the probability mass

under $H_1 : p > 0.2$ evenly. Thus, when shifting from frequentist power to Bayesian power with $k = 1/3$ and $k_f = 3$, and a flat $\text{Beta}_{[0.2,1]}(1, 1)$ prior, the sample sizes n_1 and n_2 increase to $n_1 = 27$ and $n_2 = 54$, and the expected sample size under H_0 increases to 39.46. Adding the requirement of compelling evidence under H_0 with 60% on top, these sample sizes increase even slightly further. Now, a flat prior could be criticised as a very large probability such as $p = 0.9$ gets the same prior probability as a small one like $p = 0.25$, but in most cases one would be quite uncertain about high efficacy probabilities in a phase II study. Thus, the Beta priors centered at $p_1 = 0.4$ could be more realistic. The informativity of the $\text{Beta}_{[0.2,1]}(10.33, 15)$ prior is equal to having observed 10.33 successes and 15 failures already (Kruschke & Liddell, 2018), which is not too much for a phase II study. The $\text{Beta}_{[0.2,1]}(13.66, 20)$ prior is more informative, and implies that ≈ 14 failures and 20 successes have already been recorded. The bulk of prior probability mass locates therefore around $p_1 = 0.4$, which is the mode of both priors. Table 3 shows that for such more realistic priors, the sample sizes n_1 and n_2 as well as $E[N|H_0]$ are, in general, slightly larger compared to frequentist power calculations. Two aspects are worth mentioning:

- The results in Table 3 show that shifting from a flat prior to a slightly informative one does not necessarily reduce the required sample sizes. This is due to the fact, that probabilities close to 1 are downweighed strongly even for slightly informative priors, compare Figure 5. Only if a certain amount of informativity is reached, small success probabilities close to p_0 are downweighed enough so that the sample size starts decreasing again. For example, for $k = 1/3$ and $k_f = 3$, $E[N|H_0]$ progresses from 39.46 for the flat prior, over 47.48 for the slightly informative prior, to 42.36 for the strongly informative prior. A similar pattern is visible for $k = 1/10$ and $k_f = 3$ and all other scenarios.

- Shifting from $k = 1/3$ to $k = 1/10$ yields a much larger n_2 and a larger n_1 in most cases, because more observations are necessary for the Bayes factor to express strong evidence compared to only moderate evidence. As a consequence, for different thresholds $k \neq 1/k_f$ such as $k = 1/10$ and $k_f = 3$, the type-I-error rates and power become imbalanced in the sense that the type-I-error rate is not exhausting the 10% upper limit anymore (it is more close to $\approx 3\%$ for $k = 1/10$. Thus, the sample size is driven primarily by the power requirement in these cases. We therefore recommend similar choices for k and k_f to exhaust the design's restrictions in Equation (3).

In summary, the results shown in Table 3 show how flexible the optimal sequential Bayes factor design is. It provides a fully Bayesian interpretation to Simon's two-stage optimal design. Still, if a more flexible power calculation with the option to incorporate prior information is warranted, shifting to Bayesian notions of power might be helpful. The price is an increased sample size at the benefit of more realism. Adding a threshold for a minimum probability for

compelling evidence under H_0 comes at a price, too, but adds a layer of evidence to the analysis, in particular, when the trial is stopped at n_1.

As a last note, the second Bayesian design with frequentist power and $k = 1/3$, $k_f = 3$ and $f = 0.6$ in Table 3 can be interpreted as a generalized Simon's optimal design with a Bayesian validity: It provides frequentist type-I-error rates and power, but a Bayesian notion of compelling evidence for H_0. Thus, it is like a Simon's optimal design where additionally the probability threshold to find compelling evidence under H_0 can be specified.

6.3 Comparison with non-sequential Bayes factor design analysis

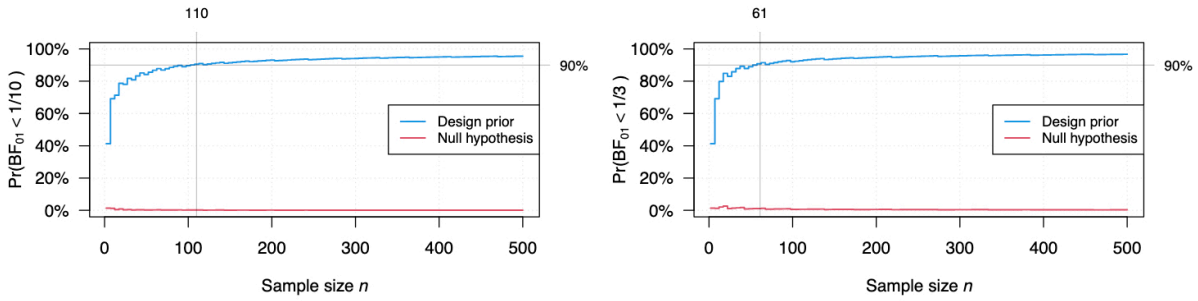
In this subsection, we provide a quick demonstration of what gains can be expected when shifting from a non-sequential Bayes factor design to our novel sequential design. Thus, we compare a Bayesian power analysis without an interim analysis with the proposed optimal sequential Bayes factor design with a single interim analysis at n_1. The general approach to carry out a Bayes factor design analysis without any interim analyses has been detailed by Kelter and Pawel (2025). Figure 6 shows the results for the second example above, where in the phase II trial the hypotheses $H_0 : p \leq 0.2$ and $H_1 : p > 0.2$ were under consideration.

Figure 6a provides the sample sizes $n = 110$ for $k = 1/10$ and $n = 61$ for $k = 1/3$ under a flat Beta design prior on $[0.2, 1]$ under H_1. As there is no interim analysis, this is comparable to the full sample size n_2 in the optimal sequential Bayes factor designs in Table 3. The thresholds for power and type-I-error rates are identically set to 90% and 10%. Table 3 shows that the sequential design reduces $n = 110$ to $n_2 = 100$ with the option to stop early at $n_1 = 42$. The expected sample size is $E[N|H_0] = 69.21$, about 40 patients less than the fixed $n = 110$ in the non-sequential Bayes factor design analysis, which always requires $n = 110$ patients.

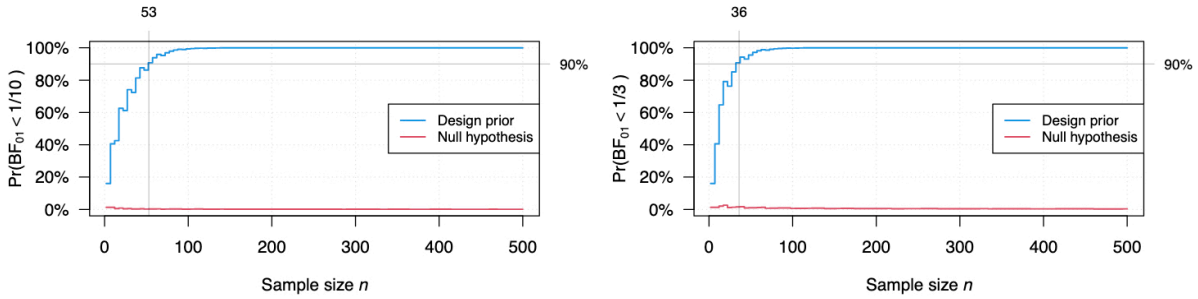
The right panel of Figure 6a shows that when using the evidence threshold $k = 1/3$ for moderate evidence, the required maximum sample size is $n = 61$. Table 3 shows that this sample size reduces to $n = 54$ when using the two-stage Bayes factor design, and the expected sample size becomes $E[N|H_0] = 39.46$ under the null hypothesis. Thus, a reduction of the required number of patients of about a third is possible when shifting from the non-sequential to the proposed two-stage design.

Figure 6b provides analogue results under a point prior at $p_1 = 0.4$, so power is calculated in a fully frequentist sense. The resulting sample sizes $n = 53$ for $k = 1/10$ and $n = 36$ for $k = 1/3$ can be compared to the ones in Table 3 again. For $k = 1/3$, the results in Figure 6b can be compared to the ones for $k = 1/3$ and $k_f = 3$ in Table 3, which shows that the expected sample size in the sequential design is $E[N|H_0] = 26.02$, which is the one of Simon's minimax design.[13] Thus, shifting to the sequential design reduces the required sample sizes significantly

[13]This is reasonable, because the approach in Kelter and Pawel (2025) searches for the smallest sample size which fulfills the requirements in Equation (3), except without any interim analysis.



(a) Evidence threshold $k = 1/10$ (left) and $k = 1/3$ right



(b) Evidence threshold $k = 1/10$ (left) and $k = 1/3$ (right)

Figure 6: Bayesian power and type-I-error rate for the single-arm phase II proof of concept trial with binary endpoint for $k = 1/3$ (left) and $k = 1/10$ (right). Top plots show the Bayesian power $P(\mathrm{BF}_{01}(y) < k \mid H_1)$ under H_1 as the solid blue line, and the Bayesian type-I-error rate $P(\mathrm{BF}_{01}(y) < k \mid H_0)$ under H_0 as the solid red line. A flat $\mathrm{Beta}_{[0.2,1]}(1, 1)$ prior was used. The bottom plots show the same under a frequentist point prior on $p_1 = 0.4$ under H_1. Flat analysis priors are assumed in both cases.

about a third. For $k = 1/10$, the optimal sequential Bayes factor design with $k = 1/10$ and $k_f = 3$ yields a design which has an expected sample size of $E[N|H_0] = 33.42$, about 20 patients less than in the non-sequential design, which needs $n = 54$ patients according to the left panel in Figure 6b.

Expected sample size under H_0 is reduced similarly in other scenarios, when shifting to the sequential Bayes factor design based on our idea of trinomial tree branching.

7 Discussion

Sequential trial design is an important statistical approach to increase the efficiency of clinical trials. Bayesian sequential trial design relies primarily on conducting a Monte Carlo simulation under the hypotheses H_0 and H_1 of interest and the resulting design characteristics are then investigated by means of Monte Carlo estimates. This approach has several drawbacks, namely that calibration of a Bayesian design requires repeating a possibly complex Monte Carlo simulation. This holds both for the process of finding a calibrated design itself, and also for replicating

a calibrated design to check the validity of the proposed design (which is relevant for regulatory agencies). In addition to the time needed for these two points, Monte Carlo standard errors are required to judge the reliability of the simulation.

All of this is due to the fact that closed-form or numerical approaches to calibrate a Bayesian design are essentially missing from the literature. In this paper, we therefore proposed the Bayesian optimal two-stage design for clinical phase II trials based on Bayes factors. Therefore, we developed the approach of trinomial tree branching, a method to visualize and illustrate the challenges which occur during power and type-I-error-rate calculations when introducing a single interim analysis into a static design, thereby making it sequential. Next, we developed theoretical results which allowed to correct the resulting design characteristics such as power and type-I-error rate for introducing an interim analysis.

We build upon this idea to invent a simple calibration algorithm and an optimal calibration algorithm, which yields the optimal Bayesian design that minimizes the expected sample size under the null hypothesis H_0. This is inspired from the well-established Simon's two-stage optimal design for phase II trials with a binary endpoint. Illustrative examples showed that our design recovers Simon's two-stage optimal design as a special case, improves upon non-sequential Bayesian design based on Bayes factors, and can be calibrated very fast with computation times usually below a minute on a regular desktop computer. This is, in turn, due to the almost instant computation of the power and type-I-error rate of our design as developed in Section 3. The last aspect, in particular, makes it a highly attractive choice for application in practice. Comparisons of multiple calibrated designs like in Table 2 and Table 3 – with possibly different priors, power concepts (Bayesian or frequentist), evidence thresholds or futility thresholds – are easily carried out within seconds. Furthermore, the design allows to guarantee a minimum probability on compelling evidence in favour of the null hypothesis, which is that easily possible with other designs.

We close this discussion with a few points which are relevant and some venues for future work:

- The core idea of trinomial tree branching that underpins the calibration algorithm is neither dependent on the binary endpoint, nor on the use of Bayes factors, nor the choice of priors. As a consequence, the design can principally be generalized to other settings with a different endpoint, a different testing approach (e.g. posterior probability, posterior odds), and different priors. This provides a profound branch of future work, namely to develop Bayesian optimal two-stage designs for other settings. For example, a natural extension would be to shift from a single-arm phase II trial to a two-arm phase II trial with binary endpoints (in treatment and control group). Three key preliminaries are then given as follows:
  - First, when shifting to different endpoints, the Bayes factor design analysis provided

by Kelter and Pawel (2025) must be modified accordingly. This enables to compute the power and type-I-error rate for a fixed sample size, compare Section 5. Pawel and Held (2025) provide results how to do this for a variety of endpoints, so this should not present too much of a hurdle for extending the current design to other settings.

- Second, Theorem 1 makes use of the marginal probability mass function

$$f(y_s^1, y_s^2 | a_d, b_d, l, u)$$

given the design prior on H_1. The second preliminary to extend the proposed design therefore is to be able to compute this marginal probability mass (or density, for continuous endpoint(s)) function at least numerically.

- Third, the test criteria of choice must be computable analytically or at least numerically. So, if for example the case of a two-arm phase II trial with binary endpoints is considered, a different Bayes factor is required. However, Bayes factors are available in the literature for most practically relevant settings by now (Du, Edwards, & Zhang, 2019; Ly, Verhagen, & Wagenmakers, 2016; Morey & Rouder, 2011; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019). For other test criteria like posterior probabilities or posterior odds, these can also be computed at least numerically in most cases.

- We did not explicitly incorporate the idea to assert a minimum probability for convincing evidence in favour of H_1 when running the interim analysis at n_1. Also, we did conceptually not allow the trial to be stopped for efficacy at n_1. However, a possible generalization of our design might include such an option.

- We provided an optimal calibration algorithm which makes use of the early ideas of Simon (1989) and which minimizes the expected sample size $E[N|H_0]$ under the null hypothesis H_0, compare the discussion in Section 5. Still, future work could consider different calibration routines such as minimax calibration or even minimin calibration (when carrying out the interim analysis as early as possible is desired). When taking into account Equation (5), even calibration algorithms would be possible which maximize the probability of compelling evidence for the null hypothesis under other constraints like an

upper boundary $n \in \mathbb{N}$ for $E[N|H_0]$:

$$\begin{aligned} \max_{n_1,n_2} \quad & P(\mathrm{BF}_{01}^{n_1}(y^1) > 1/k|H_0) \\ \text{s.t.} \quad & E[N|H_0] \leq n \\ & n_{min} \leq n_1 < n_2 \leq n_{max} \end{aligned} \tag{18}$$

The above shows that there exist plenty of possibilities to extend and generalize the proposed Bayesian optimal two-stage design for clinical phase II trials based on Bayes factors. Given the ease of application and the versatility of our proposed sequential design, we hope that it will be a useful addition to the toolkit of medical statisticians faced with the task of designing a single-arm phase II trial with binary endpoint.

Appendix

Proofs

Proof of Theorem 1.

$$\begin{aligned} P(\mathrm{BF}_{01}^{n_2} < k, \mathrm{BF}_{01}^{n_1} > 1/k \mid H_i) &= P(y_s^2 > y_{critEff}^{n_2} - y_s^1, y_s^1 \leq y_{critFut}^{n_1} | H_i) \\ &= \int_{\{y_s^1 \leq y_{critFut}^{n_1}, y_s^2 > y_{critEff}^{n_2} - y_s^1\}} 1 dP_{H_i} \\ &= \int_{\{y_s^1 \leq y_{critFut}^{n_1}, y_s^2 > y_{critEff}^{n_2} - y_s^1\}} f(y_s^1, y_s^2|H_i) dy \\ &= \int_{\{y_s^1 \leq y_{critFut}^{n_1}, y_s^2 > y_{critEff}^{n_2} - y_s^1\}} f(y_s^1, y_s^2|a_d, b_d, l, u) dy \\ &= \sum_{\{y_s^1 \leq y_{critFut}^{n_1}, y_s^2 > y_{critEff}^{n_2} - y_s^1\}} f(y_s^1, y_s^2|a_d, b_d, l, u) dy \\ &= \sum_{i=0}^{\lfloor y_{critFut}^{n_1} \rfloor} \sum_{j=\lceil y_{critEff}^{n_2} \rceil - y_i}^{n_2-n_1} f(y_s^1 = y_i, y_s^2 = y_j|a_d, b_d, l, u) dy \end{aligned} \tag{19}$$

In the above, we first make use of $\mathrm{BF}_{01}^{n_2}(y) < k$ if and only if $y_s^2 > y_{critEff}^{n_2}$, where y_s^1 and y_s^2 denote the number of successes observed in y^1 and y^2 at sample size n_1 and n_2. Thus, the Bayes factor $\mathrm{BF}_{01}^{n_2}(y)$ based on the full data of sample size n_2 shows evidence k against H_0 if and only if we observe at least $y_{critEff}^{n_2}$ successes in the full trial data y. If we have already

observed y_s^1 successes in the first batch of interim data $y^1 := (y_1, ..., n_1)$, then we need only $y_{critEff}^{n_2} - y_s^1$ successes to yield a Bayes factor $\mathrm{BF}_{01}^{n_2}(y) < k$ in the final analysis based on the full data $y := (y^1, y^2) = (y_1, ..., y_{n_1}, y_{n_1+1}, ..., y_{n_2})$. An analogue holds for $\mathrm{BF}_{01}^{n_1}(y) > 1/k$.

Next, we make use of

$$P(y_s^2 > y_{critEff}^{n_2} - y_s^1, y_s^1 \leq y_{critFut}^{n_1} | H_i) = \int_{\{y_s^2 > y_{critEff}^{n_2} - y_s^1, y_s^1 \leq y_{critFut}^{n_1}\}} 1 dP_{H_i}$$

where P_{H_i} denotes the conditional probability measure $P(\cdot|H_i)$ on the hypothesis H_i. Then, we apply rewriting the Radon-Nikodym-density $dP_{H_i}/dy = f(y_s^1, y_s^2|H_i) = f(y_s^1, y_s^2|a_d, b_d, l, u)$ as we use truncated beta design priors $\theta \sim \mathrm{Beta}(a, b)_{[l,u]}$, for details see below. Also, we make use of the fact that the integral reduces to a discrete sum, which can be expressed as a summation over two variables i and j. In these sums, we need to use floor and ceiling functions, as we need to sum over all $y_s^1 \leq y_{critFut}^{n_1}$ (so that we need to use a floor function on $y_{critFut}^{n_1}$), and over all $y_s^2 > y_{critEff}^{n_2} - y_s^1$ (so we need to use a ceiling function on $y_{critEff}^{n_2}$). Importantly, in the inner sum we can replace $\lceil y_{critEff}^{n_2} \rceil - y_s^1$ with $\lceil y_{critEff}^{n_2} \rceil - y_i$ then, as y_i counts the number of successes $y_i = 0, ..., \lfloor y_{critFut}^{n_1} \rfloor$ in the first batch of data in the outer sum.

To make use of the above representation of the marginal probability mass function, it suffices to note that if $Y_i \mid \theta \sim \mathrm{Bin}(n_i, \theta)$ for $i \in \{1, 2\}$ and $\theta \sim \mathrm{Beta}(a, b)_{[l,u]}$, then the marginal probability mass function $f(y_s^1, y_s^2|a_d, b_d, l, u)$ is given as

$$\begin{aligned}
f(y_s^1,&y_s^2|a_d, b_d, l, u) = \int_l^u f(y_s^1 \mid \theta) f(y_s^2 \mid \theta) f(\theta|a_d, b_d) \mathrm{d}\theta \\
&= \int_l^u \binom{n_1}{y_s^1} \theta^{y_s^1} (1-\theta)^{n_1 - y_s^1} \binom{n_2}{y_s^2} \theta^{y_s^2} (1-\theta)^{n_2 - y_s^2} \frac{\theta^{a_d - 1}(1-\theta)^{b_d - 1}}{\mathrm{B}(a_d, b_d)\{I_u(a_d, b_d) - I_l(a_d, b_d)\}} \mathrm{d}\theta \\
&= \binom{n_1}{y_s^1} \binom{n_2}{y_s^2} \frac{\mathrm{B}(a_d + y_s^1 + y_s^2, b_d + n_1 + n_2 - y_s^1 - y_s^2)}{\mathrm{B}(a_d, b_d)\{I_u(a_d, b_d) - I_l(a_d, b_d)\}} \\
&\times \int_l^u \frac{\theta^{a_d + y_s^1 + y_s^2 - 1}(1-\theta)^{b_d + n_1 + n_2 - y_s^1 - y_s^2 - 1}}{\mathrm{B}(a_d + y_s^1 + y_s^2, b_d + n_1 + n_2 - y_s^1 - y_s^2)} \mathrm{d}\theta \\
&= \binom{n_1}{y_s^1} \binom{n_2}{y_s^2} \frac{\mathrm{B}(a_d + y_s^1 + y_s^2, b_d + n_1 + n_2 - y_s^1 - y_s^2)}{\mathrm{B}(a_d, b_d)\{I_u(a_d, b_d) - I_l(a_d, b_d)\}} \\
&\quad \times \{I_u(a_d + y_s^1 + y_s^2, b_d + n_1 + n_2 - y_s^1 - y_s^2) - I_l(a_d + y_s^1 + y_s^2, b_d + n_1 + n_2 - y_s^1 - y_s^2)\}
\end{aligned}$$

so calculating the marginal probability mass function $f(y_s^1, y_s^2|a_d, b_d, l, u)$ is possible with standard numerical integration. □

References

Altman, D. G., & Bland, J. M. (1995, aug). Statistics notes: Absence of

evidence is not evidence of absence. *BMJ*, *311*(7003), 485. Retrieved from https://www.bmj.com/content/311/7003/485https://www.bmj.com/content/311/7003/485.abstract doi: 10.1136/BMJ.311.7003.485

Berry, S. M. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: CRC Press.

Boulesteix, A. L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., ... Sauerbrei, W. (2020, dec). Introduction to statistical simulations in health research. *BMJ Open*, *10*(12), e039921. Retrieved from https://bmjopen.bmj.com/content/10/12/e039921https://bmjopen.bmj.com/content/10/12/e039921.abstract doi: 10.1136/BMJOPEN-2020-039921

Boulesteix, A. L., Hoffmann, S., Charlton, A., & Seibold, H. (2020, oct). A replication crisis in methodological research? *Significance*, *17*(5), 18–21. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/1740-9713.01444https://onlinelibrary.wiley.com/doi/abs/10.1111/1740-9713.01444https://rss.onlinelibrary.wiley.com/doi/10.1111/1740-9713.01444 doi: 10.1111/1740-9713.01444

Chevret, S. (2012). Bayesian adaptive clinical trials: A dream for statisticians only? *Statistics in Medicine*, *31*(11-12), 1002–1013. doi: 10.1002/sim.4363

Du, H., Edwards, M. C., & Zhang, Z. (2019, oct). Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions. *Behavior Research Methods*, *51*(5), 1998–2021. Retrieved from https://asu.pure.elsevier.com/en/publications/bayes-factor-in-one-sample-tests-of-means-with-a-sensitivity-anal doi: 10.3758/s13428-019-01262-w

Fayers, P. M., Ashby, D., & Parmar, M. K. (2005, aug). Monitoring: Bayesian Data Monitoring in Clinical Trials. *Tutorials in Biostatistics, Statistical Methods in Clinical Studies*, *1*, 335–352. doi: 10.1002/0470023678.CH3B

Ferguson, J. (2021, sep). Bayesian interpretation of p values in clinical trials. *BMJ Evidence-Based Medicine*, *0*, bmjebm–2020–111603. doi: 10.1136/BMJEBM-2020-111603

Ferreira, D., Ludes, P. O., Diemunsch, P., Noll, E., Torp, K. D., & Meyer, N. (2021, feb). Bayesian predictive probabilities: a good way to monitor clinical trials. *British Journal of Anaesthesia*, *126*(2), 550–555. doi: 10.1016/J.BJA.2020.08.062

Giovagnoli, A. (2021). The Bayesian Design of Adaptive Clinical Trials. *International Journal of Environmental Research and Public Health*, *18*(2), 1–15. doi: 10.3390/IJERPH18020530

Grieve, A. P. (2022). *Hybrid frequentist/Bayesian power and Bayesian power in planning and clinical trials*. Boca Raton, FL: Chapman & Hall, CRC Press.

Jeffreys, H. (1939). *Theory of Probability* (1st ed.). Oxford: The Clarendon Press.

Kelter, R. (2020a). Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, *20*(88). doi: https://doi.org/10.1186/s12874-020-00968-2

Kelter, R. (2020b). Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 101–119. doi: 10.1080/15366367.2019.1689761

Kelter, R. (2021). Bayesian Hodges-Lehmann tests for statistical equivalence in the two-sample setting: Power analysis, type I error rates and equivalence boundary selection in biomedical research. *BMC Medical Research Methodology*, *21*(171). doi: 10.1186/s12874-021-01341-7

Kelter, R. (2023a). The Bayesian simulation study (BASIS) framework for simulation studies in statistical and methodological research. *Biometrical Journal*, 2200095. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/bimj.202200095 doi: 10.1002/BIMJ.202200095

Kelter, R. (2023b). Reducing the false discovery rate of preclinical animal research with Bayesian statistical decision criteria. *Statistical Methods in Medical Research*, *32*(10), 1880–1901. Retrieved from https://journals.sagepub.com/doi/full/10.1177/09622802231184636 doi: 10.1177/09622802231184636/ASSET/IMAGES/LARGE/10.1177_09622802231184636-FIG6.JPEG

Kelter, R., & Pawel, S. (2025). Bayesian Power and Sample Size Calculations for Bayes Factors in the Binomial Setting. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2502.02914v1

Kelter, R., & Schnurr, A. (2024). The Bayesian Group-Sequential Predictive Evidence Value Design for Phase II Clinical Trials with Binary Endpoints. *Statistics in Biosciences*((online first)), 1–37. Retrieved from https://link.springer.com/article/10.1007/s12561-024-09430-z doi: 10.1007/s12561-024-09430-z

Kieser, M. (2020). *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*. Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-030-49528-2 doi: 10.1007/978-3-030-49528-2

Kleijn, B. (2022). *The frequentist theory of Bayesian statistics*. Amsterdam: Springer.

Kruschke, J. K., & Liddell, T. (2018). The Bayesian New Statistics : Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*, 178–206. doi: 10.3758/s13423-016-1221-4

Lee, J., & Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, *5*(2), 93–106. Retrieved from https://journals.sagepub.com/

doi/10.1177/1740774508089279 doi: 10.1177/1740774508089279

Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E.-J., & van Ravenzwaaij, D. (2020). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. *psyarxiv preprint, https://psyarxiv.com/bh8vu*.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. doi: 10.1016/j.jmp.2016.01.003

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*, 2767. doi: 10.3389/fpsyg.2019.02767

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406–419. doi: 10.1037/a0024377

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/SIM.8086

O'Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, *73*(sup1), 69–81. Retrieved from https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1518265 doi: 10.1080/00031305.2018.1518265

Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., ... Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, *16*(1), 29. Retrieved from /pmc/articles/PMC5830330//pmc/articles/PMC5830330/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5830330/ doi: 10.1186/S12916-018-1017-7

Pawel, S., & Held, L. (2024). Closed-Form Power and Sample Size Calculations for Bayes Factors. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2406.19940v2

Pawel, S., & Held, L. (2025, apr). Closed-Form Power and Sample Size Calculations for Bayes Factors. *The American Statistician*, 1–15. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/00031305.2025.2467919 doi: 10.1080/00031305.2025.2467919

Pourmohamad, T., & Wang, C. (2023). Sequential Bayes Factors for Sample Size Reduction in Preclinical Experiments with Binary Outcomes. *Statistics in Biopharmaceutical Research*, *15*(4), 706–715. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/19466315.2022.2123386 doi: 10.1080/19466315.2022.2123386

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Retrieved from http://dx.doi.org/10.1016/j.jmp.2012.08.001 doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods.*, *22*(2), 322–339. Retrieved from doi:10.1037/met0000061

Seibold, H., Charlton, A., Boulesteix, A. L., & Hoffmann, S. (2021). *Statisticians, roll up your sleeves! There's a crisis to be solved* (Vol. 18) (No. 4). John Wiley and Sons Inc. doi: 10.1111/1740-9713.01554

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*, *10*(1), 1–10. doi: 10.1016/0197-2456(89)90015-9

Stallard, N., Todd, S., Ryan, E. G., & Gates, S. (2020). Comparison of Bayesian and frequentist group-sequential clinical trial designs. *BMC Medical Research Methodology*, *20*(1), 1–14. Retrieved from https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0892-8 doi: 10.1186/S12874-019-0892-8/FIGURES/4

Stefan, A. M., Lengersdorff, L. L., & Wagenmakers, E.-J. (2022). A Two-Stage Bayesian Sequential Assessment of Exploratory Hypotheses. *Collabra: Psychology*, *8*(1). Retrieved from /collabra/article/8/1/40350/194713/A-Two-Stage-Bayesian-Sequential-Assessment-ofhttps://online.ucpress.edu/collabra/article/8/1/40350/194713/A-Two-Stage-Bayesian-Sequential-Assessment-of doi: 10.1525/COLLABRA.40350

U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research. (2019). Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. *Web archive, accessed 01/02/2022*. Retrieved from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

U.S. Food and Drug Administration Center for Drug Evaluation and Research (CDER). (2019). Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. *https://www.fda.gov/media/78495/download (accessed 01/08/2021)*.

Van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., ... Yau, C. (2021, jan). Bayesian statistics and modelling. *Nature Reviews Methods Primers 2021 1:1*, *1*(1), 1–26. Retrieved from https://www.nature.com/articles/

s43586-020-00001-2 doi: 10.1038/s43586-020-00001-2

Van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(71), 1–12. doi: 10.1186/s12874-019-0699-7

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham: Springer International Publishing Switzerland. Retrieved from http://link.springer.com/10.1007/978-3-319-32562-0 doi: 10.1007/978-3-319-32562-0

Zhou, H., Lee, J. J., & Yuan, Y. (2017). BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in Medicine*, *36*(21), 3302–3314. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/sim.7338https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7338https://onlinelibrary.wiley.com/doi/10.1002/sim.7338 doi: 10.1002/SIM.7338

Zhou, T., & Ji, Y. (2023). On Bayesian Sequential Clinical Trial Designs. *The New England Journal of Statistics in Data Science*, *0*, 1–16. Retrieved from https://nejsds.nestat.org/journal/NEJSDS/article/25/texthttps://nejsds.nestat.org/journal/NEJSDS/article/25 doi: 10.51387/23-NEJSDS24

Zhou, Y., Lin, R., & Lee, J. J. (2021). The use of local and nonlocal priors in Bayesian test-based monitoring for single-arm phase II clinical trials. *Pharmaceutical Statistics*, *20*(6), 1183–1199. Retrieved from https://pubmed.ncbi.nlm.nih.gov/34008317/ doi: 10.1002/PST.2139