Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09430-z

ORIGINAL PAPER

The Bayesian Group-Sequential Predictive Evidence Value Design for Phase II Clinical Trials with Binary Endpoints

Riko Kelter[1] · Alexander Schnurr[1]

Received: 14 August 2023 / Revised: 19 March 2024 / Accepted: 25 March 2024

Abstract

In clinical research, the initial efficacy of a new agent is typically assessed in a phase IIA study. Bayesian group-sequential designs are often based on predictive probability of trial success. In this paper, the novel Bayesian group-sequential predictive evidence value design is introduced, and we prove that the predictive probability approach is a special case of it. A comparison with Simon's two-stage and competing Bayesian designs based on phase IIA cancer trials is provided. Results show that the novel design can improve operating characteristics such as the false-positive rate, probability of early stopping for futility and expected sample size of the trial. Given these advantages, the predictive evidence value design constitutes an important addition to the biostatistician's toolbelt when planning a phase IIA trial the Bayesian way, in particular, when small sample sizes and a large probability for early termination under the null hypothesis are desired.

Keywords Phase II trial · Binary endpoints · Group-sequential design · Stopping for futility · Bayesian statistics

1 Introduction

Bayesian group-sequential phase II designs for clinical trials have received increasing attention in the last years [4, 6, 7, 9, 30]. According to a recent review of [13], Bayesian adaptive designs have attracted keen interest in various disciplines, from a theoretical and practical viewpoint. This is partially due to the increased flexibility which Bayesian analysis offers over frequentist designs [38],

✉ Riko Kelter
riko.kelter@uni-siegen.de

Alexander Schnurr
schnurr@mathematik.uni-siegen.de

[1] Department of Mathematics, University of Siegen, Walter-Flex-Street 2, 57076 Siegen, Northrhine-Westphalia, Germany

Published online: 29 April 2024

and partially due to the availability of software solutions which simplify the application for practitioners [1, 32].

After preliminary information is obtained about the safety profile and dose of a new drug in a phase I trial, the next step consists of determining whether the drug has sufficient efficacy to justify further development [1]. In phase IIA studies, the primary endpoint often is binary and measures response or no response, respectively failure or no failure. For example, in cancer trials the clinical response can be defined as complete or partial response, measured based on tumor volume shrinkage, for details see the RECIST criteria in solid tumors [5, 46]. While the definition of response or no response varies in different contexts, phase IIA studies share the idea that the initial efficacy assessment is often designed as an open-label single-arm study which recruits between 40 to 100 patients in a multi-stage setting [1, 33]. The general idea behind multi-stage designs is to stop the trial early if no efficacy can be found based on an interim data analysis. The trial is thus monitored after recruiting new patients and is possibly stopped for futility or efficacy depending on the observed data. The approach of multi-stage designs goes back to [12], who proposed Gehan's design for cancer drug development, and other approaches are Simon's two-stage designs [41]. Both of these are two-stage designs, that is, a first stage of patient recruiting is observed which ensures a minimum sample size, and then a second stage of data accrual follows depending on the interim analysis results of the available data. One possible benefit of two-stage designs is that if the treatment shows no efficacy in the first stage, the trial can be stopped early for futility to avoid wasting time and resources. However, there are also sequential parallel comparison designs which are not designed to stop early.

Traditional frequentist two-stage designs such as Simon's optimal design were constructed to minimize the expected or maximum sample size under the null hypothesis that the treatment is ineffective. Let $p \in [0,1]$ denote the unknown probability of response to the treatment, henceforth called the response rate, and p_0 is a predefined threshold for judging the efficacy of the new drug. If $p \leq p_0$, the null hypothesis H_0 is true and the drug is considered ineffective for practical purposes. As a consequence, the trial can be stopped for futility. Based on the result obtained in the first stage with n enrolled patients – out of which x show a response – the optimal design specifies when to stop the trial for futility (when X is small enough, that is, $X \leq r_1$ for some positive r_1) while simultaneously controlling the type I error α and type II error β at a prespecified level. The values of n and x in turn depend on the required restrictions for α and β. Also, an operating characteristic which is usually of interest is the probability of early termination (PET), which is $\text{PET}(p_0) = P(\text{Early termination}|H_0) = P(X \leq r_1|H_0)$. Another operating characteristic of relevance is the expected sample size $\mathbb{E}[N|p_0]$ under H_0 and $\mathbb{E}[N|p_1]$ under H_1, where N denotes the random variable which measures the number of patients enrolled in the trial. These are upper bounds on the required sample size of the trial: When $p < p_0$, fewer patients will be required to stop for futility compared to when $p = p_0$. When $p > p_1$, fewer patients will be required to stop early for efficacy compared to when $p = p_1$. Among all designs which fulfill a prespecified type I error rate α and type II error rate β, Simon's optimal two-stage design minimizes $\mathbb{E}[N|p_0]$

and Simon's minimax two-stage design minimizes $N_{\max}$, the maximum number of patients that can be enrolled in the trial.

1.1 Setting

In this paper, we focus on the hypothesis testing framework of a phase IIA clinical trial which is designed to test

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

for some $p_0 \in (0, 1)$ where p_0 represents a prespecified response rate of the current standard treatment. In practice, p_1 denotes a desired target response rate of a new treatment under consideration, where $p_1 > p_0$ [36]. Thus, (a lower boundary of) the power is calculated for p_1. We assume that the trial is designed to fulfill the following requirements:

$$P(\text{Accept new treatment}|H_0) \leq \alpha \tag{1}$$

$$P(\text{Reject new treatment}|H_1) \leq \beta \tag{2}$$

for some prespecified false-positive and false-negative rates α and β. When the inequalities (1) and (2) hold, we speak of a *calibrated* design.[1] Furthermore, we assume that given the probabilities p_0,p_1, the following trial operating characteristics are of interest: (1) the probability of early termination (PET) under H_0 and H_1; (2) the expected sample size $\mathbb{E}[N|p_0]$ and $\mathbb{E}[N|p_1]$ of the trial under H_0 and H_1. We denote by PET(p_0) the probability to stop early for futility when H_0 holds, and by PET(p_1) the probability to stop early for efficacy when H_1 holds. Next to (1) and (2), interest lies in robustness to deviations from the study protocol. The latter includes false-positive control at the required level when a different number of interim analyses is carried out as previously planned. With regard to (2) it is of particular interest how many patients on average are required under H_1 until one can state with certainty that the drug works, if the trial is conducted as planned until the end.

1.2 Outlook

In this paper, we introduce a novel response-adaptive design for clinical trials with binary endpoints based on Bayesian evidence values. Therefore, the next section first outlines the predictive probability approach. The following section then outlines the theory of Bayesian evidence values. Bayesian evidence values have recently

[1] We stress that due to computational reasons, Bayesian designs sometimes do not fully exhaust the boundary (1), as the tradeoff between runtime to calibrate a given design and resulting false-positive rate must be considered in practice. This is similar to frequentist tests such as Fisher's exact test and the χ^2-test for contingency tables, where the former is conservative and the latter fully exhausts the false-positive rate due to the asymptotic χ^2-distribution of its test statistic.

been proposed as a unified approach for Bayesian hypothesis testing and parameter estimation.

The section afterwards shows that the predictive probability approach is a special case of using Bayesian evidence values for stopping the trial early for futility (or efficacy). Theoretical results are provided which clarify the relationship between the predictive probability and Bayesian evidence value approach. After that, we introduce the design which makes use of Bayesian evidence values, henceforth called the predictive evidence value (PEV) design.

The subsequent section then compares the PEV design to existing approaches, including the PP design and Simon's two-stage design. Therefore, two illustrative examples of phase IIA studies are detailed.

The following section investigates the robustness of the PEV design to deviations from the trial protocol, including running a different number of interim analyses than planned, and unplanned early stopping of the trial. Furthermore, a systematic comparison with competing trial designs is provided.

A discussion and outlook for future work concludes the article.

2 Predictive Probability Approach for Binary Endpoints

In this section, the standard Bayesian group-sequential design based on predictive probability is outlined for binary endpoints. Continuous monitoring of trial results in a two-stage design in phase II trials with stopping for futility or efficacy is widely used, see [4, 8, 45] and [16] for examples.

The null hypothesis $H_0 : p \leq p_0$ is tested against the alternative $H_1 : p > p_1$, where $p_0, p_1 \in [0, 1]$, $p_0 \leq p_1$ and p_0 is a predefined threshold for determining the minimum clinically important effect [20]. For simplicity, assume a Beta prior $p \sim \mathcal{B}(a_0, b_0)$ is selected for the response rate p, which offers a broad range of flexibility in terms of modeling the prior beliefs about p.

Let $N_{\max}$ be the maximum number of patients which is possibly recruited during the study, and let X be the random variable which measures the number of responses in the current n enrolled patients, where $n \leq N_{\max}$. A reasonable assumption is that X follows a binomial distribution with parameters n and p, $X \sim \mathrm{Bin}(n, p)$. The $\mathcal{B}(a_0, b_0)$ distribution is a conjugate prior for the binomial likelihood, and thus the posterior $P_{p|X}$ is also Beta-distributed [17]:

$$p|X = x \sim \mathcal{B}(a_0 + x, b_0 + n - x)$$

The idea of the predictive probability approach consists of analyzing the interim data to project whether the trial will result in a conclusion that the drug or treatment is effective or ineffective. When n patients have been enrolled in the trial out of which $X = x$ show a response, there remain $m = N_{\max} - n$ patients which can be enrolled in the trial. Denote by Y the number of responses in the remaining $m = N_{\max} - n$ patients. If out of these remaining m exactly i respond to the treatment, and the conditional probability $P_{p|X,Y}(p > p_0 | X = x, Y = i)$ is larger than a prespecified threshold θ_T, say, $\theta_T = 0.95$, this will be interpreted as the drug being effective. Efficacy is thus declared when the posterior probability fulfills the constraint

$$P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T \tag{3}$$

for some threshold $\theta_T \in [0, 1]$. However, as the number Y of responses in the remaining $m = N_{\max} - n$ patients which can be enrolled in the trial is uncertain, this uncertainty must be modeled, too. Marginalizing out p of the binomial likelihood yields the posterior predictive distribution which is Beta-Binomial, $Y \sim$ Beta-Binom$(m, a_0 + x, b_0 + n - x)$. Additionally, from the conjugacy of the beta prior we have the posterior $P_{p|X,Y}(X = x, Y = i) \sim \mathcal{B}(a_0 + x + i, b_0 + N_{\max} - x - i)$, and the expected predictive probability of trial success – henceforth abbreviated PP – can now be calculated by weighting the posterior probability of trial success $P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T$ when observing $X = x$ and $Y = i$ with the prior predictive probability $P_{Y|X}(Y = i|X = x)$ of observing $Y = i$ responses in the remaining $m = N_{\max} - n$ patients, after $X = x$ responses have been observed in the current n patients:

$$\begin{aligned} \text{PP} &= \mathbb{E}\left[\mathbb{1}_{P_{p|X,Y}(p>p_0|X,Y)>\theta_T}|x\right] = \int_{\mathcal{Y}} \mathbb{1}_{P_{p|X,Y}(p>p_0|X,Y)>\theta_T} dP_{Y|X=x} \\ &= \sum_{i=0}^{m} P_{Y|X=x}(i) \cdot \mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} \end{aligned} \tag{4}$$

where

$$\mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} := \begin{cases} 1, & \text{if } P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T \\ 0, & \text{if } P_{p|X,Y}(p > p_0|X = x, Y = i) \leq \theta_T \end{cases}$$

is an indicator which measures whether the evidence against $H_0 : p \leq p_0$ is large enough – that is, $P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T$ – conditional on $X = x$ and $Y = i$ or not. The predictive probability PP thus quantifies the expected predictive probability of trial success. Figure 1 visualizes the PP design.

To employ the approach in practice, futility and efficacy thresholds θ_L and θ_U out of $[0, 1]$ must be fixed, so that the value of PP can be compared to these thresholds based on available interim data $X = x$. Then, if PP $< \theta_L$ or PP $> \theta_U$, the trial can be stopped early for futility or efficacy. Algorithm 1 shows the PP group-sequential design, see also [1]. Note that in practice, $\theta_U = 1.0$ is often preferred because if the drug is effective one does not want to stop the trial. However, $\theta_L > 0$ is important to stop the trial in case the drug or treatment is not effective to avoid a waste of resources.

Algorithm 1 Phase IIA predictive probability (PP) design

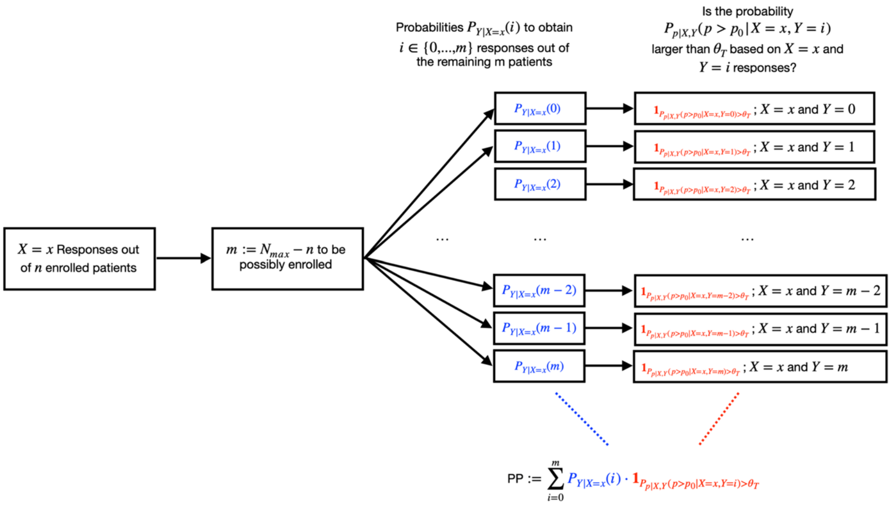

Probabilities $P_{Y|X=x}(i)$ to obtain $i \in \{0,\ldots,m\}$ responses out of the remaining m patients
Is the probability $P_{p|X,Y}(p > p_0 | X = x, Y = i)$ larger than θ_T based on $X = x$ and $Y = i$ responses?
$X = x$ Responses out of n enrolled patients
$m := N_{max} - n$ to be possibly enrolled
$P_{Y|X=x}(0)$
$P_{Y|X=x}(1)$
$P_{Y|X=x}(2)$
$P_{Y|X=x}(m-2)$
$P_{Y|X=x}(m-1)$
$P_{Y|X=x}(m)$
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=0)>\theta_T}$; $X = x$ and $Y = 0$
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=1)>\theta_T}$; $X = x$ and $Y = 1$
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=2)>\theta_T}$; $X = x$ and $Y = 2$
...
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=m-2)>\theta_T}$; $X = x$ and $Y = m - 2$
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=m-1)>\theta_T}$; $X = x$ and $Y = m - 1$
$\mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=m)>\theta_T}$; $X = x$ and $Y = m$
PP $:= \sum_{i=0}^{m} P_{Y|X=x}(i) \cdot \mathbf{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T}$

Fig. 1 Structure of the predictive probability (PP) design: The probability to obtain $Y = i$ successes is weighted with the probability of success $P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T$ for each $i = 0, m$. This weighted sum is the predictive probability of trial success, should the trial be continued until the maximum trial size $N_{\max}$

```
**Require:** $n < N_{\max}$ recruited patients and $X = x$ observed responses
    ○ If PP $< \theta_L$, stop the trial and reject the alternative hypothesis $H_1 : p > p_0$
    ○ If PP $> \theta_U$, stop the trial and reject the null hypothesis $H_0 : p \leq p_0$
    Otherwise recruit the next patient until reaching $N_{\max}$ patients
```

3 The Bayesian Evidence Value (BEV)

The last section outlined the Bayesian group-sequential Phase IIA predictive probability (PP) design. In this section, the theory of Bayesian evidence values is briefly outlined and illustrated with an example. The next section then proposes a novel Bayesian group-sequential design based on Bayesian evidence values.

The Bayesian evidence value (BEV) was recently proposed as a unification of Bayesian hypothesis testing and parameter estimation, which generalizes the Full Bayesian Significance Test (FBST). Details on the FBST can be found in [35] and [21–23], while the BEV was proposed by [24]. The BEV can be computed in any standard parametric statistical model, where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a (possibly vector-valued) parameter of interest, $p(y|\theta)$ is the likelihood and $p(\theta)$ is the density of the prior distribution $\mathbb{P}_\vartheta$ for the parameter θ, and $y \in \mathcal{Y}$ denote the observed sample data, $\mathcal{Y}$ being the sample space.

3.1 Statistical Information, Surprise and the Bayesian Evidence Interval

A natural measure from a Bayesian perspective to quantify the surprise in the observed data $Y = \boldsymbol{y}$ is the Bayesian surprise function which compares the posterior density and a suitable reference function at a given parameter value $\theta \in \Theta$:

Definition 1 (*Bayesian surprise function*) Let $(\Theta, \mathcal{G}, P_\vartheta)$ be the prior model, $\mathscr{P}$ on $(\mathcal{Y}, \mathcal{B})$ be the statistical model and $(\Theta, \mathcal{G}, \{P_{\vartheta|Y} : y \in \mathcal{Y}\})$ be the posterior model. Let μ be a σ-finite dominating measure on $\mathscr{P}$, and denote by $p(\theta|\boldsymbol{y}) := dP_{\vartheta|Y}(\theta)/d\mu$ the corresponding Radon-Nikodým μ-density of the posterior distribution $P_{\vartheta|Y}$. Then, the Bayesian surprise function $s : \Theta \times \mathcal{Y} \to [0, \infty)$ is defined as

$$s(\theta) := \frac{r(\theta)}{p(\theta|\boldsymbol{y})} \tag{5}$$

where $r : \Theta \to [0, \infty)$ is called the reference function.

The inverse of surprise is defined as the Bayesian information as follows:

Definition 2 (*Bayesian information function*) In the setting of Definition 1, the Bayesian information function $I : \Theta \times \mathcal{Y} \to [0, \infty)$ is defined as

$$I(\theta) := \frac{p(\theta|\boldsymbol{y})}{r(\theta)} \tag{6}$$

If $r(\theta) :\equiv 1$, the surprise is smallest for the maximum a posteriori parameter value θ_{MAP}. Equivalently, the information provided by the maximum a posteriori value is largest. A common choice for the reference function $r(\theta)$ is the prior density $p(\theta) := dP_\vartheta(\theta)/d\mu$ [35]. Then, the Bayesian information function quantifies the ratio between prior and posterior density. Importantly, the definition of information as given in Definition 2 can be derived as the probabilistic explication of information from only few very general axioms, see [14], and is motivated by connections to information theory [29, 40]. The Bayesian evidence interval is based on the information function I as follows:

Definition 3 (*Bayesian Evidence Interval*) In the setting of Definition 1, let $I(\theta) := p(\theta|\boldsymbol{y})/r(\theta)$ be the Bayesian information function for a given reference function $r : \Theta \to [0, \infty)$, $\theta \mapsto r(\theta)$. The Bayesian evidence interval $\text{EI}_r(v)$ with reference function $r(\theta)$ to level v is defined as

$$\text{EI}_r(v) := \left\{ \theta \in \Theta \,\middle|\, \frac{p(\theta|\boldsymbol{y})}{r(\theta)} \geq v \right\}. \tag{7}$$

[24] showed that commonly used Bayesian interval estimates are special cases of the EI, and that the EI thus provides an encompassing generalization of various Bayesian interval estimates. For $r(\theta) := p(\theta)$ and $v := k$, the $\text{EI}_r(v)$ evidence interval recovers the support interval as a special case, which was proposed by [47] and

includes the parameter values which have been corroborated by a factor of at least k. That is, all $\theta \in \Theta$ are included which fulfill $p(\theta|y)/p(\theta) \geq k$. Also, for $r(\theta) := 1$ and $\nu := \nu_{\alpha\%}$, the $\mathrm{EI}_r(\nu)$ evidence interval recovers the standard Bayesian $\alpha\%$-HPD interval as a special case if the posterior distribution is symmetric where $\nu_{\alpha\%}$ is the $\alpha\%$-quantile of the posterior distribution $P_{\vartheta|Y}$.

3.2 The Bayesian Evidence Value

It is well known that Bayesian hypothesis tests and parameter estimation can yield contradictory results [47]. Although this is seldom the case, the duality between frequentist Neyman-Pearson tests and the corresponding confidence intervals removes this separation between testing and estimation for frequentists [2], and the Bayesian evidence value was introduced to close this gap. The Bayesian evidence value incorporates the Bayesian evidence interval and provides a theory which unifies Bayesian hypothesis testing and parameter estimation.

Definition 4 (*Bayesian Evidence Value*) Let $H_0 := \Theta_0$ and $H_1 := \Theta \setminus \Theta_0$ be a null and alternative hypothesis with $\Theta_0 \in \Theta$. For a given Bayesian evidence interval $\mathrm{EI}_r(\nu)$ with reference function $r(\theta)$ to level ν, the Bayesian Evidence Value (BEV) $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_0)$ for the null hypothesis H_0 is defined as:

$$\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_0) := \int_{\mathrm{EI}_r(\nu)\cap\Theta_0} p(\theta|\mathbf{y})d\theta \tag{8}$$

The corresponding BEV $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_1)$ for the alternative hypothesis H_1 is defined as:

$$\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_1) := \int_{\mathrm{EI}_r(\nu)\cap\Theta_1} p(\theta|\mathbf{y})d\theta \tag{9}$$

The BEV $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}$ is inspired by the general approach to consider a (small) interval hypothesis instead of a point-null hypothesis, which was first proposed by [18] from a frequentist perspective. Furthermore, the BEV provides a generalization of the FBST which champions the e-value as a Bayesian version of frequentist p-values [35]. As shown by [3], e-values asymptotically recover frequentist p-values under Bernstein-von-Mises regularity conditions, and [24, Theorem 2] showed that the BEV $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_0)$ includes the e-value of the FBST as a special case. Thus, BEVs are, under certain regularity conditions, asymptotically, valid frequentist p-values. The test based on $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_0)$ is also called the Full Bayesian Evidence Test (FBET), or simply Bayesian evidence test. Also, the FBET obtains a widely used decision rule for interval hypothesis testing based on the region of practical equivalence (ROPE) [27, 28] as a special case, see [24]. Now, the BEV depends on three quantities: (i) the choice of the hypothesis $H_0 \subset \Theta$, (ii) the reference function $r(\theta)$ which is used for calculation of the Bayesian evidence interval $\mathrm{EI}_r(\nu)$ and (iii) the evidence threshold ν that is used for deciding which values are included in the Bayesian evidence interval $\mathrm{EI}_r(\nu)$.

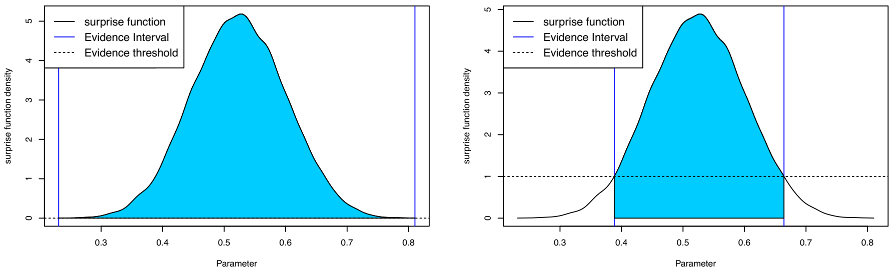


Fig. 2 Visualization of Bayesian evidence values $Ev_{EI_r(\nu)}(H_1)$ in the illustrative example based on $X = 4$ responses in $n = 10$ patients, a vague $\mathcal{B}(1.1, 1.1)$ prior and $H_1 : p \in (p_0, 1]$ for $p_0 = 0.2$. Left: Evidence threshold $\nu := 0$ and flat reference function $r(p) := 1$; Right: Evidence threshold $\nu := 1$ and reference function $r(p)$ selected as the prior density of the $\mathcal{B}(1.1, 1.1)$ prior distribution

Figure 2 shows two examples of the BEV. The left panel shows $X = 4$ responses out of 10 patients in the illustrative example, and the probability mass colored in blue equals the BEV in favor of H_1. The right panel shows the same situation but uses a positive evidence threshold $\nu = 1$ instead of $\nu = 0$, and as a consequence, less probability mass counts as evidence in favor of H_1. The BEV is implemented in the R package `fbst`, which is available on https://cran.r-project.org/web/packages/brada/index.html and detailed in [21].

4 The Predictive Evidence Value (PEV) Design

The last section outlined the theory of Bayesian evidence values and the Full Bayesian Evidence Test. Returning to the context of group-sequential Bayesian trial designs based on predictive probability PP, this section now introduces a modified design based on Bayesian evidence values.

We return to the PP approach, and again consider the competing hypotheses $H_0 : p \leq p_0$ and $H_1 : p > p_1$. The novel Bayesian group-sequential design based on Bayesian evidence values modifies PP as follows into PP_e:

$$\begin{aligned}\text{PP}_e &= \mathbb{E}\left[\mathbb{1}_{\text{Ev}_{\text{EI}_r(\nu)}(H_1)>\theta_T} | x\right] = \int_{\mathcal{Y}} \mathbb{1}_{\text{Ev}_{\text{EI}_r(\nu)}(H_1)>\theta_T} dP_{Y|X=x} \\ &= \sum_{i=0}^{m} P_{Y|X=x}(i) \cdot \mathbb{1}_{\text{Ev}_{\text{EI}_r(\nu)}(H_1)>\theta_T}\end{aligned} \tag{10}$$

where

$$\mathbb{1}_{\text{Ev}_{\text{EI}_r(\nu)}(H_1)>\theta_T} := \begin{cases} 1, \text{ if } \text{Ev}_{\text{EI}_r(\nu)}(H_1) > \theta_T \\ 0, \text{ if } \text{Ev}_{\text{EI}_r(\nu)}(H_1) \leq \theta_T \end{cases}$$

Note that the indicator $\mathbb{1}_{\text{Ev}_{\text{EI}_r(\nu)}(H_1)>\theta_T}$ depends on the value $Y = i$ as well as $X = x$ as the evidence interval $\text{EI}_r(\nu)$ depends on the observed data. Depending on the value of i, the evidence interval thus looks as follows:

$$\mathrm{EI}_r(\nu) := \left\{ \theta \in \Theta \middle| \frac{p(\theta|y')}{r(\theta)} \geq \nu \right\} \tag{11}$$

where $y' := \{X = x, Y = i\}$ is the observed data of x responses in the n enrolled patients and i responses in the remaining $m := N_{\max} - n$ patients to be possibly recruited. Now, PP_e differs from the basic predictive probability approach as follows:

1. The reference function r and the evidence threshold $\nu \geq 0$ influence the result
2. The posterior probability $P_{p|X,Y}(p > p_0|X = x, Y = i) > \theta_T$ condition for effectivity is replaced by the predictive evidence value condition $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_1) > \theta_T$ for trial success, that is, effectivity of the treatment.

PP_e thus is a weighted average of the Bayesian evidence expressed by $\mathrm{Ev}_{\mathrm{EI}_r(\nu)}(H_1)$ in favor of the alternative hypothesis of efficacy and the probability of observing $Y = i$ responses in the remaining $m = N_{\max} - n$ patients when currently n patients are enrolled in the trial. Algorithm 2 shows the phase IIA predictive evidence value (PEV) design.

Algorithm 2 Phase IIA predictive evidence value (PEV) design

```
Require: $n < N_{\max}$ recruited patients and $X = x$ observed responses
  ∘ If $\mathrm{PP}_e < \theta_L$, stop the trial and reject the alternative hypothesis $H_1 : p > p_1$
  ∘ If $\mathrm{PP}_e > \theta_U$, stop the trial and reject the null hypothesis $H_0 : p \leq p_0$
  Otherwise recruit the next batch of patients until reaching $N_{\max}$ patients
```

5 Relationships Between Both Designs

The last section introduced the PEV design. This section presents new results which demonstrate that the PP design is a special case of the PEV design. Theorem 1 establishes this fact:

Theorem 1 If $\nu := 0$ and $r(p) := 1$, then the predictive evidence value design and predictive probability design are equivalent.

Proof See Appendix A.

Theorem 1 shows that for any $N_{\max}$, θ_T, θ_L, and any number and time points of interim analyses, the PP design is a special case of the PEV design. The following Corollary states that due to Theorem 1, the operating characteristics of the PP and PEV design coincide under identical priors and when a flat reference function $r(p) := 1$ with evidence threshold $\nu := 0$ is used in the PEV design:

Corollary 1 *Let* $\nu := 0$ *and* $r(p) := 1$ *and denote* α_{PP} *and* α_{PP_e} *and* β_{PP} *and* β_{PP_e} *as the false-positive and false-negative rates under* $H_0 : p \leq p_0$ *for the predictive probability and predictive evidence value designs. Then,*

$$\alpha_{\text{PP}} = \alpha_{\text{PP}_e} \qquad \text{and} \qquad \beta_{\text{PP}} = \beta_{\text{PP}_e} \tag{12}$$

Proof See Appendix A.

Note that Corollary 1 does not require to specify how a Bayesian false-positive error is defined.

No matter how one specifies a false-positive error that contributes to α_{PP} (or α_{PP_e}), Corollary 1 guarantees that these false-positive error rates will coincide whenever a flat reference function and evidence threshold $\nu := 0$ are used. The same holds for the associated false-negative error rates β_{PP} and β_{PP_e}.

A consequence of Theorem 1 is that the above property also translates to other operating characteristics such as the probability of early termination or the expected sample size until early stopping:

Corollary 2 *Under the conditions of Theorem 1, the operating characteristics of the PP and PEV designs are identical. The latter include the probability of early stopping (PET) and the expected sample size until early stopping as well as their associated variances, both under* H_0 *and* H_1.

Proof Follows from Theorem 1 like Corollary 1.

Theorem 2 below now shows under which conditions the false-positive error rate in the PEV design can be reduced so that it is smaller than the one of the PP design.

Theorem 2 *Let* $r(p) :\equiv 1$. *If* $\nu > 0$, *then*

$$\alpha_{\text{PP}_e} \leq \alpha_{\text{PP}} \tag{13}$$

Proof See Appendix A.

6 Calibration of the PEV Design

The last section provided insights about how the false-positive rate of the PP design can be improved by using the PEV design. Theorem 2 yields the key condition which we use in this section to propose a default way to calibrate the PEV design. Therefore, two choices must be made, the choice of the reference function r and the choice of the evidence threshold ν.

6.1 Choice of the Reference Function

The first choice deals with the reference function $r(p)$. Based on the definition of the evidence value, we propose to use a flat reference function $r(p) :\equiv 1$. This has two advantages: First, using $r(p) :\equiv 1$ implies that the evidence interval measures highest-posterior-density regions, because

$$\mathrm{EI}_p(\nu) : = \left\{ \theta \in \Theta \middle| \frac{p(\theta|y)}{r(\theta)} \geq \nu \right\} = \left\{ p \in [0,1] \middle| \frac{p(p|y)}{r(p)} \geq \nu \right\} \overset{r(p):\equiv 1}{=} \left\{ p \in [0,1] \middle| p(p|y) \geq \nu \right\}.$$

Thus, for any positive $\nu > 0$ the evidence interval includes only a highest-posterior-density region and is equivalent to a highest-posterior-density interval. The larger $\nu > 0$, the smaller this interval will be. This motivates how to choose ν as explained below.

6.2 Choice of the Evidence Threshold

Picking $\nu > 0$ seems reasonable to measure evidence in terms of highest-posterior-density regions.

Theorem 2 shows that when $\nu > 0$ holds, the false-positive rate α_{PP_e} can decrease compared to α_{PP}. Now, Theorem 2 can be made applicable to calibrate the PEV design through the following corollary:

Corollary 3 *There exists a value $\xi \in \mathbb{R}_+$, so that setting $\nu := \xi$ implies that*

$$\alpha_{\mathrm{PP}_e} < \alpha_{\mathrm{PP}} \tag{14}$$

Proof See Appendix A.

Corollary 3 shows that we can calibrate the PEV design as follows: Pick a flat reference function $r(p) :\equiv 1$ and increase ν to a large enough positive value. Then, the false-positive rate α_{PP_e} of the PEV design will – for large enough $\nu > 0$ – become smaller than the false-positive rate of the PP design, α_{PP}.

6.3 The Four-Step Calibration

The following four-step calibration algorithm is proposed for the PEV design:

- **Step 1:** Pick values of θ_T and θ_U for which the false-positive rate is slightly above the desired level α.
- **Step 2:** Increase ν until the false-positive rate α_{PP_e} of the design decreases below the required upper threshold α. Store the smallest evidence threshold for which this holds as ν_c.
- **Step 3:** Check the false-negative rate of the calibrated design with $\nu = \nu_c$. If the false-negative rate β_{PP_e} is above the required threshold β, decrease θ_L until the

Table 1 Overview of the design parameters for the PEV design

Interpretation	Parameter	How to choose
Maximum trial size	$N_{\max}$	Take $N_{\max}$ of Simon's minimax design
Sample size of first interim analysis	n_{init}	Take value of Simon's minimax design
Threshold to declare trial success	θ_T	Calibration parameter
Threshold to stop for futility	θ_L	Calibration parameter
Threshold to stop for efficacy	θ_U	Default value $\theta_U :=1$
Number of interim analysis	-	Domain knowledge
Time points of interim analysis	-	Domain knowledge / equal-spaced
Reference function	$r(p)$	Flat reference function $r(p) :\equiv 1$
Evidence threshold	v	$v>0$ to reduce false-positive rate

false-negative rate decreases below β, and store the largest value of θ_L for which this holds as θ_c. If this step fails increase $N_{\max}$ by one batchsize and repeat (or return to Step 1, if the resulting $N_{\max}$ is judged as too large).

- **Step 4:** Analyze the false-positive and false-negative rate of the resulting design with $v=v_c$ and $\theta_L=\theta_c$. If the false-positive rate $\alpha_{\mathrm{PP}_e}>\alpha$ or the false-negative rate $\beta_{\mathrm{PP}_e}>\beta$ increase $N_{\max}$ and return to Step 2 (or return to Step 1, if the resulting $N_{\max}$ is judged as too large).

A few comments are required regarding the above four-step calibration. First, step one is usually simple as picking starting values is often easy when $N_{\max}$, n_{init} and θ_U are fixed as specified in Table 1. Here, we denote by n_{init} the number of enrolled patients after which the first interim analysis is performed.

With respect to step 2, the `calibrate` function of the `brada` ®package – which is outlined in a separate section below—does this automatically.

Step 3 is also automatically done via the `calibrate` function of the `brada` ®package. However, it may happen that for the specified $N_{\max}$ it is not possible to achieve the desired false-positive and false-negative rates. This is also observed in some cases considered by [31] for the PP design, and a simple solution is to increase $N_{\max}$ slightly in those situations. We experienced that whenever the sum of false-positive and false-negative rates $\alpha_{\mathrm{PP}_e}+\beta_{\mathrm{PP}_e}\leq\alpha+\beta$ held in Step 1, this problem does not occur.

The last step ensures that the calibrated design really achieves the desired operating characteristics.

6.4 Runtime and Computational Efficiency

i

The key advantage of the above four-step calibration algorithm is its (1) computational efficiency and (2) differences in the resulting operating characteristics of the design.

The computational efficiency of calibrating the PEV design is understood best when compared with the calibration of the PP design. To calibrate the PP design one usually has to search the (θ_L, θ_T) space with a linear search and find combinations of θ_L and θ_T for which the resulting false-positive rate and false-negative rates result in the desired specifications.

In the original paper of [31], this requires to search a grid $[0.001, ..., 1.000] \times [0.001, ..., 1.000]$ with 1000^2 points. Making use of $\theta_L, \theta_T \geq 0.5$ which is a reasonable assumption still leaves a grid $[0.001, ...0.499] \times [0.501, ...1.000]$ of 249001 points. At each of these points a Monte Carlo simulation is required to study (A) the false-positive rate under p_0 and (B) the false-negative rate under p_1. The Monte Carlo simulations also must include enough repetitions m to achieve a small enough Monte Carlo standard error of the false-positive and -false-negative rates, compare [34]. Suppose $m = 1000$ Monte Carlo repetitions suffice. Then $249001 \cdot 1000 \cdot 2 = 498002000$ trials must be simulated for the PP design. Depending on $N_{\max}$ and the number of interim analysis and first interim analysis time point the runtime of a single repetition varies, and under the assumption that $m = 1000$ Monte Carlo repetitions take ≈ 5 seconds (which is optimistic, even under full parallelization using multiple cores based on the implementation in the `brada` package detailed below), the PP grid-search calibration takes approximately $250000 \cdot 2 \cdot 5 \approx 2490010$ seconds, which is equal to 28.82 *days*. Shifting to a high-performance-computing cluster with, say, 10 nodes each fully parallelized then reduces the runtime to about 3 days, which is still very long. In contrast, the calibration of the PEV design via the four-step calibration can be achieved in usually less than an hour.

Concerning point (2), the trial operating characteristics which result from the PEV calibration algorithm differ compared to the characteristics obtained via calibrating the PP design via a grid-search. The two examples in the next section illustrate these differences in detail.

6.5 Overview of the Design Parameters

In closing this section, we present an overview about the parameters that can be used to calibrate the PEV design in Table 1.

As noted by [31], the calibration parameters θ_T and θ_L have the following effects: Increasing θ_T decreases the false-positive rate and power to stop for efficacy, decreasing θ_T increases the false-positive rate and power to stop for efficacy. In contrast, decreasing θ_L decreases the false-negative rate and power to stop for efficacy, and increasing θ_L increases the false-negative rate and power to stop for efficacy.

Table 1 shows that we adopt $N_{\max}$ and n_{init} from Simon's minimax two-stage design, and set θ_U to 1 by default (as we usually do not stop the trial when the drug works). The number of interim analyses must be chosen from domain knowledge. However, in practice it is often unrealistic to monitor after each patient and a realistic number of interim analysis possibly spans from $1 - 4$. This is due to logistic and administrative reasons.

The key calibration parameters that remain are θ_L, θ_T and ν as the rest of the parameters have sensible default choices. Note that we always use the flat reference function, so in the PEV design essentially adds one calibration parameter compared to the PP design, namely ν.

7 Comparison Between Predictive Probability, Predictive Evidence and Simon's Two-Stage Design

7.1 Competing Designs

We select the predictive probability design, Simon's minimax two-stage design (except for Example 2) and the BOP2 design as competing designs to which we compare the calibrated PEV design. Details on other possible competitors and the BOP2 design are provided in the supplementary material.

7.2 Example 1—A Lung Cancer Trial

First, we use the lung cancer trial example also used by [31]. The primary objective of the study was to assess the efficacy of a combination therapy as front-line treatment in patients with advanced nonsmall cell lung cancer. The study 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 current standard treatment yields a response rate of $\approx 20\%$, so we have $p_0 = 0.2$. The target response rate of the new regimen is 40%, so $p_1 = 0.4$.

First, Simon's two-stage design is applied. Therefore, we follow [31] in specifying $\alpha \leq 0.1$ and $\beta \leq 0.1$ both for the minimax and optimal design.

For the calibrated PP design we use the values $N_{\max} = 36$ which is also the maximum sample size of Simon's two-stage minimax design, perform the first interim analysis after 10 patients and consistently monitor the result after each new patient. We investigate deviations from this unrealistic monitoring plan in a separate section below. Note also that Simon's two-stage minimax design makes the interim analysis after 10 patients, too. We use the $B(0.2, 0.8)$ prior for p that is also used by [31] to allow for a fair comparison of both designs. The thresholds θ_T, θ_L are then taken from the grid-search that is performed by [31] which results in $\theta_T = 0.922$ and $\theta_L = 0.001$. Using these values is simple, finding them is not. Finding these values requires a computationally very expensive grid-search as discussed above.

Calibration of the PEV design proceeds by following the four-step calibration algorithm outlined in the last section. An R package has been created to facilitate application of the PEV and PP designs, the R package `brada`. The abbreviation brada stands for Bayesian response-adaptive design analysis, and currently includes the group-sequential PP and PEV designs. It automatically sets up a cluster to make full use of multicore environments and parallelizes and vectorizes computations automatically. This achieves efficient runtimes in practice and the package allows to fit a trial design with the `brada` function, plot and summarize

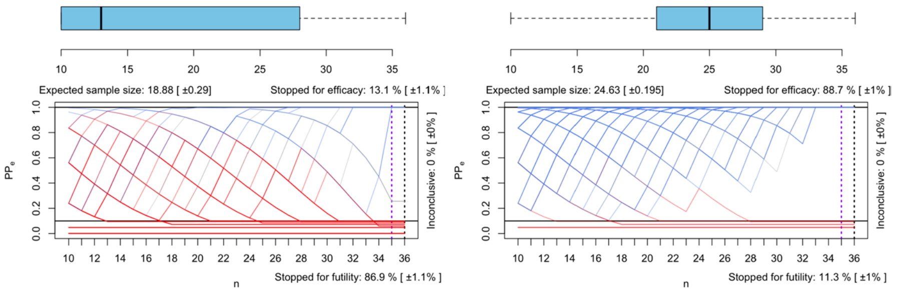


Fig. 3 Results of the uncalibrated PEV design under $H_0 : p \leq 0.2$ (left) and $H_1 : p > 0.2$ (right). Trajectories show simulated trial runs, where blue lines show trials that reach the threshold θ_U and red lines are trials which reach θ_L. The expected sample size under H_0 (left panel) and H_1 (right panel) are shown at the top of the panels, together with the probability to stop for efficacy. At the bottom of each panel the probability to stop early for futility is shown. The boxplot at the top shows the distribution of resulting sample sizes where the trial is terminated

the results with `plot` and `summary` functions, and calibrate the PEV design via the four-step algorithm via the `calibrate` function. Details on the package can be found in the accompanying Quarto file provided at the Open Science Foundation under https://osf.io/zmfyn/?view_only=348067ed1ccc498da7e4a11d949c84df. Further information is also provided in the separate section on the package.

We implement the four-step calibration algorithm as follows with the help of the `brada` ®package:

- **Step 1:** To calibrate the PEV design with the `brada` package we start with liberal thresholds $\theta_T = 0.8$ and $\theta_L = 0.1$. We investigate the false-positive and false-negative rates with a call to the `brada` function of the `brada` R package, call the `plot` function in R for the resulting object and obtain Fig. 3.

The standard output when plotting a `brada` object in the `brada` package is a plot which shows the simulated trial trajectories together with the percentages of how many trials stopped early for futility or efficacy, and a boxplot which shows the distribution of the sample size of the trial. The horizontal black lines in the trajectory plot are the thresholds $\theta_U = 1$ and $\theta_L = 0.1$ where the trial is stopped for efficacy and futility. Note that due to $\theta_U = 1$, the trial is never stopped for efficacy. As a consequence, the percentage of trials which is stopped for efficacy according to the plot is actually the percentage of trials which finishes at $N_{\max}$. Under H_0, this percentage can be interpreted as a false-positive because the trial finishes although it should be stopped for futility. Under H_1, this percentage can be interpreted as the power to reject H_0, because the trial finishes and is not stopped for futility.

The dashed-blue vertical line in the trajectory plots (at the 35th patient) visualize the time of the last interim analysis. If a trajectory has not passed θ_L or θ_U at this point, the advantage of a group-sequential design has vanished because $N_{\max}$ patients were recruited.

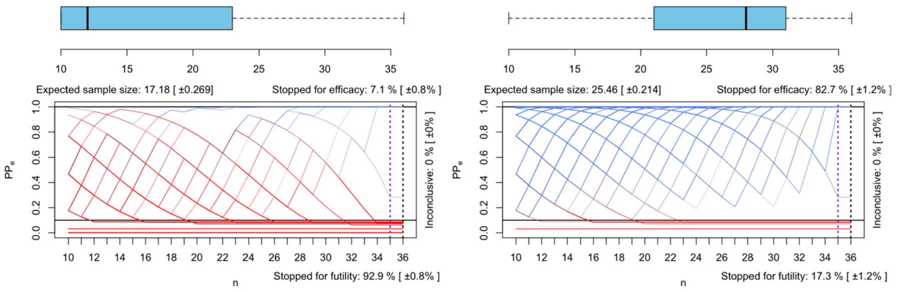


Fig. 4 Results of the ν-calibrated PEV design under $H_0 : p \leq 0.2$ (left) and $H_1 : p > 0.2$ (right)

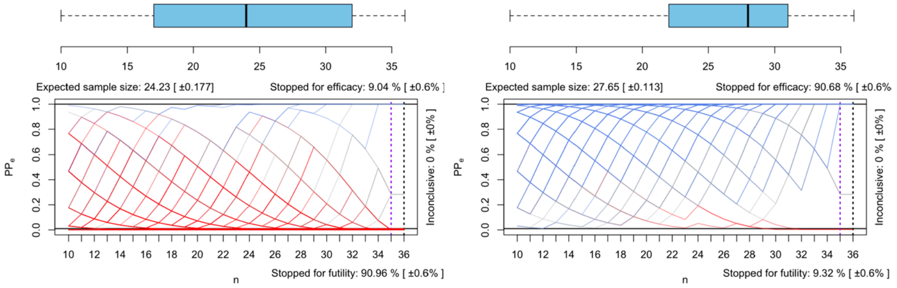


Fig. 5 Results of the fully calibrated PEV design under $H_0 : p \leq 0.2$ (left) and $H_1 : p > 0.2$ (right)

- **Step 2:** Next, we call the `calibrate` function of the `brada` package and are recommended to increase ν from $\nu = 0$ to $\nu = 1.3$.

Figure 4 shows the trial's operating characteristics after this first calibration step. Note that now the false-positive rate has dropped to 7.1%, and the false-negative rate is at 17.3%. Before, we had a false-positive rate of 13.1%—see the left plot in Fig. 3—and a false-negative rate of 11.3%—see the right plot in Fig. 3.

- **Step 3:** Next, we call the `calibrate` function of the `brada` package again to calibrate θ_L and are adviced to decrease θ_L from 0.1 to 0.01.
- **Step 4:** Refitting the design with the `brada` function then yields the fully calibrated design shown in Fig. 5. The result shows that both the false-positive rate and the false-negative rate are well controlled below their boundaries of 10%. Note that to further improve our design we could try the same four-step calibration with a smaller $N_{\max}$ than $N_{\max} = 36$. For example, one could use $N_{\max}$ of Simon's two-stage optimal design.

Table 2 shows a comparison of the designs. All Bayesian solutions use continuous monitoring, and the expected sample size $\mathbb{E}[N|p_0]$ under $H_0 : p \leq 0.2$ is better for the calibrated PEV design than for the calibrated PP solution. Also, the PEV design outperforms Simon's two-stage minimax design with regard to the

Table 2 Comparison of Simon's two-stage minimax design, the calibrated PP design and the calibrated PEV design for the first example; $\mathbb{E}[N|p_1]$ is not reported by [31] and not available for Simon's two-stage minimax design; operating characteristics are simulation-based and obtained with the brada R package for the calibrated PP and PEV design

	$N_{\max}$	First interim analysis	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
Simon's two-stage minimax design								
	36	10	0.086	0.098	0.4600	0.902	28.26	-
Calibrated PP design design								
	36	10	0.088	0.094	0.8600	0.906	27.67	-
Calibrated PEV design								
	36	10	0.0904	0.0932	0.9036	0.908	24.40	27.76
BOP2 design								
	36	10	0.1005	0.1382	0.8995	0.8618	18.10	33.50

average sample size, as it requires ≈ 4 patients less under H_0, and still ≈ 3 patients less than the calibrated PP design. Table 2 also shows that the BOP2 design does not control the false-negative rate (and the false-positive rate only if rounding to two digits). This is, however, to be expected because the BOP2 design maximizes the power and does not assert an upper false-negative rate. Although it requires the smallest expected sample size under H_0, it violates the requirement (2) on the false-negative rate which was formulated in advance.

All of the above simulations took ≈ 15 minutes on a desktop computer, while the grid-search to calibrate the PEV design takes much longer. Furthermore, the false-positive and -negative estimates of the calibrated PEV design include the Monte Carlo standard error (MCSE). For example, the MCSE of the false-positive rate is 0.6%, so one can judge the uncertainty in the Monte Carlo estimate [34]. Results of [31] include no MCSEs but we could replicate them using the `brada` package. MCSEs are computed automatically in the `brada` package for all relevant quantities by means of 10000 bootstrap samples, see [25].

Some comments are in order with regard to the plots above. First, although the labels at the top right of each plot say stopped for efficacy, all simulations used $\theta_U = 1$. Thus, stopping early for efficacy is not possible. Stopped for efficacy is, as a consequence, to be interpreted that the trial continues until $N_{\max}$ patients were recruited. Under H_0, this can be interpreted as a false-positive result. Under H_1, it can be interpreted as the power to reject H_0.

Secondly, in Table 2 the expected sample size $\mathbb{E}[N|p_1]$ under H_1 is the expected value of the sample size under H_1 for the BOP2 design, which is not the expected value specifically under p_1. As a consequence, the expected sample size under H_1 is smaller for BOP2 than the expected sample size specifically under p_1, because the average includes also the sample sizes under probabilities $p > p_1$ (where fewer patients are required).

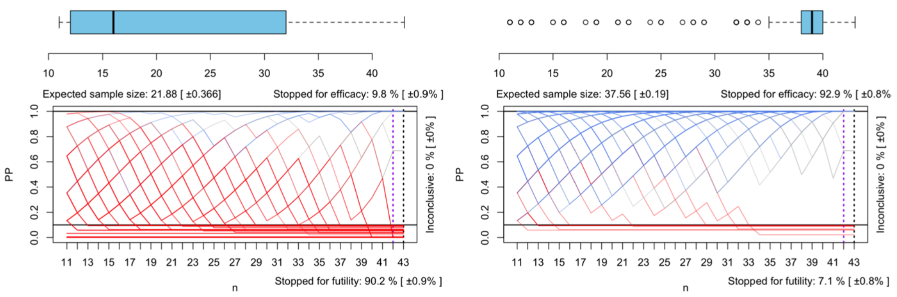


Fig. 6 Results of the uncalibrated PEV design under $H_0 : p \leq 0.6$ (left) and $H_1 : p > 0.8$ (right) for the tongue cancer trial

For the calibrated PEV design, $\mathbb{E}[N|p_1]$ is the expected sample size until PP_e reaches the threshold $\theta_U = 1$. Thus, it represents the average sample size required to state with certainty that the drug works.

It should be noted that these quantities differ for the PEV and BOP2 designs, so using them as a benchmark is only partially justified. In practice, the expected sample size under H_0 is of larger relevance because the success rate of phase II studies is only modest. Stopping a trial early for futility while controlling α and β is thus of primary importance.

7.3 Example 2—A Tongue Cancer Trial

Next, we reproduce the tongue cancer trial example of [31]. There, the primary objective is to assess the efficacy of induction chemotherapy (with paclitaxel, ifosfamide, and carboplatin) followed by radiation in treating young patients with prior untreated squamous cell carcinoma of the tongue. Previous results showed that radiation alone yields a response rate of 60%, so $p_0 := 0.6$. With induction chemotherapy plus radiation, the target response rate is set at 80%, so $p_1 := 0.8$. The constraints for the type I and II error rates are given as follows:

$$\alpha \leq 0.05 \text{ and } \beta \leq 0.20$$

In contrast to the first example, we now take Simon's optimal two-stage design as a competitor. Thus, we use $N_{\max} = 43$ from Simon's optimal two-stage design. The calibrated PP design reported in [31] uses a different initial sample size at which the first interim analysis is made than Simon's two-stage optimal design. That is, the first interim analysis is performed after 10 instead of 11 patients. For the PEV design, we use the 11 patients after which the first interim analysis is performed in Simon's two-stage design.

We proceed with the four-step calibration as follows:

- **Step 1:** To calibrate the PEV design with the `brada` package we start again with liberal thresholds $\theta_T = 0.9$ and $\theta_L = 0.1$. We investigate the false-positive and false-negative rates with a call to the `brada` function of the `brada` R package, call the `plot` function in R for the resulting object and obtain Fig. 6.

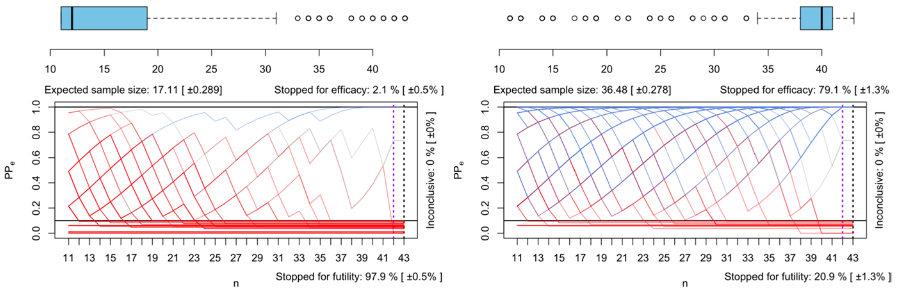


Fig. 7 Results of the ν-calibrated PEV design under $H_0 : p \leq 0.6$ (left) and $H_1 : p > 0.8$ (right) for the tongue cancer trial

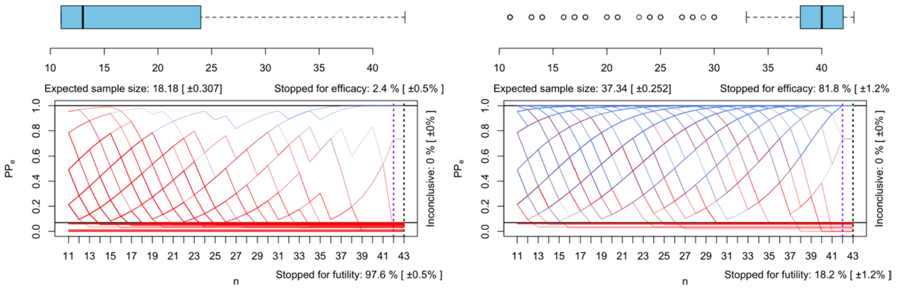


Fig. 8 Results of the fully calibrated PEV design under $H_0 : p \leq 0.6$ (left) and $H_1 : p > 0.8$ (right) for the tongue cancer trial

The design achieves a false-positive rate of 9.8% and a false-negative rate of 7.1%. Thus, the design violates the requirement $\alpha \leq 0.05$. We proceed with the step two of the four-step calibration:

- **Step 2:** We call the `calibrate` function of the `brada` package and are recommended to increase ν from $\nu = 0$ to $\nu = 1.6$.

Figure 7 shows the trial's operating characteristics after this first calibration step. Note that now the false-positive rate has dropped to 2.1%, and the false-negative rate is at 20.9%.

- **Step 3:** Next, we call the `calibrate` function of the `brada` package again to calibrate θ_L and are adviced to decrease θ_L from 0.1 to 0.07.
- **Step 4:** Refitting the design with the `brada` function then yields the fully calibrated design shown in Fig. 8. The result shows that both the false-positive rate and the false-negative rate are well controlled below their boundaries of 5% and 20%. Note that to further improve our design we could try the same four-step calibration with a smaller $N_{\max}$ than $N_{\max} = 43$, e.g. $N_{\max} = 35$ of Simon's two-stage minimax design.

Table 3 Comparison of Simon's two-stage minimax design, the calibrated PP design and the calibrated PEV design for the second example; $\mathbb{E}[N|p_1]$ is not reported by [31] and not available for Simon's two-stage minimax design; operating characteristics are simulation-based and obtained with the brada R package for the calibrated PP and PEV design

$N_{\max}$	First interim analysis	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
Simon's two-stage minimax design							
35	13	0.0499	0.1918	0.6470	0.8082	20.77	-
Simon's two-stage optimal design							
43	11	0.0489	0.1976	0.7037	0.8024	20.48	-
Calibrated PP design							
43	10	0.033	0.134	0.9675	0.8663	27.04	39.67
Calibrated PEV design							
43	10	0.024	0.182	0.9760	0.8180	18.18	37.34
BOP2 design							
43	10	0.0398	0.1697	0.9566	0.8277	17.4	39.20

Table 3 shows a comparison of the calibrated PP design, calibrated PEV design, and Simon's optimal and minimax two-stage designs.

We see that the calibrated PEV design has a larger probability of early termination PET(p_0) under H_0 than all other designs. The expected sample size $\mathbb{E}[N|p_1]$ under H_1 is smallest for the calibrated PEV design. The expected sample size $\mathbb{E}[N|p_0]$ is about one patient larger compared to the BOP2 design, and the PET(p_1) is also about 2% smaller compared to the BOP2 design. In this example, the calibrated PEV design and BOP2 perform comparable. Both designs outperform Simon's two-stage minimax and optimal design and the calibrated PP design.

8 Simulation Study

The last section discussed two examples of phase II studies with binary endpoints in detail and compared several group-sequential designs and their resulting performance characteristics. It was shown that the calibration of the PEV design is straightforward with the help of the `brada` R package and the four-step calibration algorithm. The two detailed examples of the last section showed that the calibrated PEV design yields larger probabilities of early termination and smaller expected sample sizes under H_0 than Simon's two-stage designs and the calibrated PP design. The price paid for this improvement is a slightly higher α and a slightly lower β than for Simon's designs and the PP design. The calibrated PEV design performs comparable or better than the BOP2 design in the two examples.

In this section, we provide additional simulations to investigate the performance of the calibrated PEV design. First, we provide a systematic comparison under a selection of different contexts. Second, we explore how deviations from the sampling plan affect the resulting operating characteristics of the PEV design. Thus, we deviate from continuous monitoring in the two examples of the last section and replace this unrealistic monitoring scheme with $1-4$ interim analyses which is more realistic in clinical research. Finally, we investigate how an unplanned early termination influences the design characteristics.

8.1 Systematic Comparison

Table 4 shows the systematic comparison between the calibrated PEV design, the calibrated PP design, Simon's two-stage minimax design and the BOP2 design. Operating characteristics are simulation-based and obtained with the brada R package for the calibrated PP and PEV designs. We built on the calibrated solutions of [31] for the selected settings (shown in the rows with italics *PP*), and took the values of θ_T and θ_L they found via a two-dimensional grid search. For the calibrated PEV design, we applied the four-step calibration. The latter worked in a single cycle except for the last setting and setting five, where we had to increase $N_{\max}$ once. Note that b in Table 4 denotes the batchsize after which the next interim analysis is performed. Thus, if $b = 1$, we monitor after each patient continuously. This is shown for the PP design in the rows denoted with CPP, and as this is unrealistic in practice, the rows with PP below show the results for a more realistic batchsize. In all settings, we aimed for $1-4$ interim analyses, which seems possible in practice. The batchsize b was then chosen accordingly.

A flat prior was used in all simulations, and two comments are in order regarding the PET(p_1) and $\mathbb{E}[N|p_1]$. As noted previously, all simulations used $\theta_U = 1$, so stopping early for efficacy is not possible. Thus, PET(p_1) is actually the probability of terminating the trial with $N_{\max}$ patients and can be interpreted as the Bayesian power to reject H_0 given that $p = p_1$ is the true success probability. Furthermore, the expected sample size $\mathbb{E}[N|p_1]$ is the expected sample size under H_1 and not p_1 for the BOP2 design. For the other rows *PP*, PP and PEV it is the expected sample size until PP respectively PP$_e$ reaches $\theta_U = 1$. Thus, it can be interpreted as the average number of patients required to state with certainty that the drug works (although the trial is not stopped early then). This interpretation is in particular helpful, because if stopping for efficacy at $\theta_U = 1$ would be allowed, it reflects the sample size at which the trial would be stopped for efficacy under this protocol.

There are a few comments worth mentioning with regard to Table 4:

- *Setting 1:* The PEV design achieves the smallest combined sample sizes and best PETs both under H_0 and H_1.
- *Setting 2:* The BOP2 design's solution does not strictly fulfill the requirement $\beta \leq 0.1$ (first bold entry in Table 4). The same holds for the false-positive rate of the PP design in the last setting in Table 4. The preferred solution in setting 2 thus is the PEV design.
- *Setting 3:* Both the calibrated PP and PEV designs require to increase $N_{\max}$ compared to Simon's two-stage design. Still, the expected sample size of the *PP* solution is best, except when a very small sample size under H_0 is desired. Then, BOP2 is better. However, BOP2 has a substantially smaller PET under H_0, so the calibrated *PP* design seems best. Still, shifting from continuous

Table 4 Comparison of Simon's two-stage minimax design, the calibrated PP design, the calibrated PEV design and the BOP2 design

p_0/p_1	Design	$N_{\max}$	n_{init}	b	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
0.1/0.3										
	Minimax	25	16	–	0.0951	0.0970	0.5147	0.9030	20.4	-
	CPP	25	10	1	0.0988	0.0912	0.9012	0.9088	21.80	16.71
	PP	25	10	5	0.096	0.084	0.9011	0.9160	24.13	18.02
	PEV	25	10	5	0.095	0.094	0.905	0.906	17.58	17.55
	BOP2	25	10	5	0.0894	0.0625	0.720	0.8864	16.8	24.3
0.2/0.4										
	Minimax	36	19	–	0.0861	0.0976	0.4551	0.9024	28.26	-
	CPP	36	10	1	0.0876	0.0905	0.9124	0.9095	29.41	28.75
	PP	36	9	9	0.0905	0.090	0.9124	0.9095	32.74	35.95
	PEV	36	9	9	0.089	0.093	0.911	0.907	27.1	35.59
	BOP2	36	9	9	0.0741	**0.103**	0.771	0.8524	19.8	33.7
0.3/0.5										
	Minimax	39	28	–	0.0943	0.0999	0.3648	0.9001	35.0	-
	CPP	42	10	1	0.0956	0.0833	0.9044	0.9167	33.37	33.99
	PP	42	15	9	0.0957	0.0829	0.9043	0.9171	37.85	39.23
	PEV	42	15	9	0.0890	0.0780	0.9110	0.9220	37.94	39.13
	BOP2	42	15	9	0.0878	0.0592	0.7620	0.8978	27.2	41
0.4/0.6										
	Minimax	41	28	–	0.0951	0.0991	0.5510	0.9009	33.80	–
	CPP	41	10	1	0.0902	0.0909	0.9098	0.9091	31.24	35.26
	PP	41	11	5	0.0904	0.0908	0.9096	0.9092	33.41	37.42
	PEV	41	21	5	0.0790	0.0970	0.9210	0.9030	27.93	36.80
	BOP2	41	21	5	0.0897	0.0660	0.7963	0.8878	27.90	40.00
0.5/0.7										
	Minimax	39	23	–	0.0978	0.0985	0.5000	0.9015	31.0	–
	CPP	39	10	1	0.0967	0.0929	0.9033	0.9071	28.95	24.40
	PP	39	23	4	0.0969	0.0924	0.9031	0.9076	31.65	37.20
	PEV	43	23	4	0.100	0.077	0.9000	0.9230	29.23	38.87
	BOP2	39	23	4	0.0955	0.0541	0.7835	0.8974	28.40	38.40
0.6/0.8										
	Minimax	35	27	–	0.0965	0.0997	0.8161	0.9003	28.5	–
	CPP	36	10	1	0.0901	0.0878	0.9099	0.9122	25.48	32.38
	PP	36	16	10	0.0902	0.0876	0.9098	0.9124	29.80	35.87
	PEV	36	16	10	0.0850	0.0990	0.9150	0.9010	24.10	35.32
	BOP2	36	16	10	0.0890	0.0254	0.6429	0.9084	26.70	35.70
0.7/0.9										
	Minimax	25	16	–	0.0905	0.0980	0.5501	0.9020	20.00	–
	CPP	25	10	1	0.0911	0.0989	0.9089	0.9011	16.32	23.09
	PP	25	10	5	**0.1072**	0.0548	0.8928	0.9452	22.00	37.42
	PEV	30	10	5	0.0630	0.0650	0.9370	0.9350	18.18	29.36

Table 4 (continued)

p_0/p_1	Design	$N_{\max}$	n_{init}	b	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
	BOP2	25	10	5	0.0895	0.0471	0.7681	0.8997	16.80	24.60

Table 5 Runtimes for calibrating the PEV design in the seven simulation settings shown in Table 4; m = minutes, s = seconds

p_0/p_1	0.1/0.3	0.2/0.4	0.3/0.5	0.4/0.6	0.5/0.7	0.6/0.8	0.7/0.9
Runtime	55.80 m	34.01 m	22.72 m	46.01 s	25.07 m	5.55 m	3.29 m

monitoring with $b = 1$ to the more realistic $b = 9$ the PEV design becomes comparable to the PP design.

- *Setting 4:* As in Setting 1, the PEV design achieves the smallest combined sample sizes and best PETs both under H_0 and H_1.
- *Setting 5:* As in Setting 3, the *PP* design with continuous monitoring is best, but for $b = 4$ it is worse than the PEV and BOP2 designs. BOP2 and PEV are comparable, with the PEV design yielding a larger PET under H_0 and H_1 and BOP2 yielding a slightly smaller sample size under H_0.
- *Setting 6:* Again, the PEV design is best in terms of PET under H_0 and H_1 and the required sample sizes under both hypotheses.
- *Setting 7:* The four-step calibration requires to increase $N_{\max}$ by one batchsize to $N_{\max} = 30$. All other designs can be calibrated with $N_{\max} = 25$. Here, BOP2 yields smaller sample sizes and the PEV design again higher PETs under H_0 and H_1.

There are two conclusions: First, the PEV design *always* achieves the highest probability of early termination under the null hypothesis. Although the BOP2 design sometimes requires fewer patients under H_0, the PEV design always yields a larger PET(p_0).

Secondly, the calibrated PEV design performs better than Simon's two-stage minimax design in all settings, and performs better than or comparable to the calibrated PP and BOP2 designs. BOP2 achieves smaller sample sizes in settings 3 and 7, and was outperformed in the other settings by the PEV design.

Table 5 shows that calibration of the PEV design typically takes less than an hour, while in most cases less than half an hour is possible on a regular desktop computer. Note that while Simon's two-stage designs are calibrated almost instantaneously, calibration of the PP design via a 2-dimensional grid-search of θ_L and θ_T requires multiple hours in the best case, while in the worst case it may even take more than a full day on regular desktop machines. This happens, in particular, when the parameters for which the design is calibrated are located in regions that are visited by the search algorithm only at the end of the sequential procedure, compare Sect. 6.4. Note that the runtimes in Table 5 are the times needed from an uncalibrated design to a fully calibrated PEV design, where the calibration algorithm is run multiple times in some simulation settings (e.g. setting 3 and 5).

8.2 Deviations from the Study Protocol

In this section, we investigate deviations from the study protocol. Thus, we reexamine Example 1 and 2 discussed earlier and analyze how the operating characteristics of the calibrated PEV design change when using a different number of interim analyses than specified in the study protocol. We vary between 1, 3, 13 and 26 (that is, continuous monitoring) interim analyses in the first example, and between 1, 2, 4, 8, 16 and 32 (continuous monitoring) interim analyses in the second example and investigate how the false-positive and false-negative rates, the expected sample size and PET under H_0 and H_1 change. We use equally spaced interim analyses after the first one. That means when 2 interim analyses are specified and the first interim analysis is conducted after e.g. 10 patients, and $N_{\max}$ is specified as $N_{\max} = 30$, we use time points 10 and 20 for the first and second interim analysis.

Table 5 shows the results of deviations from the study protocol. Results indicate that the false-positive and false-negative rates and the PET(p_0) and PET(p_1) are robust against changes to deviations from the study protocol. As with all group-sequential designs, the expected sample sizes increase when fewer interim analyses are performed. In practice, a balance between logistic effort and expected sample sizes thus must be made with any group-sequential trial.

Importantly, the PET under H_0 and H_1 is *always* much better than for Simon's two-stage design, even when conducting only a *single* interim analysis. This shows that the large PET of the PEV design under H_0 and H_1 is not due to a large number of interim analyses, but a property of this trial design. If possible, researchers should still aim for a large number of interim analyses to reduce the expected sample sizes of the PEV design under H_0 and H_1, see Table 5.

8.3 Unplanned Early Termination

The performance of a trial design under unplanned early termination is important, because clinical trials sometimes are terminated earlier than planned because of slow accrual or other reasons. For this situation, we investigate the performance of the calibrated PEV designs in Example 1 and 2 when the trials are terminated earlier than specified in the study protocol (Table 6).

In the first example, we suppose that we have to stop the trial after 20 patients. In the second example, we proceed identical and report the resulting false-positive rate and false-negative rates, and the power to stop early for efficacy and futility.

In the first example, 32.76% of the trials are stopped for futility under H_0 and 1.96% under H_1. Thus, the false-negative rate is still controlled at $\beta \leq 0.1$. Only 0.04% of the trials result in a false-positive conclusion under H_0, and 9.8% of the trials are stopped for efficacy under H_1. Thus, in case of an unplanned early termination the error rates are still controlled in the first example, but the power under H_1 drastically decreases, like the PET under H_0.

Table 6 Deviations from the study protocol for the lung cancer and tongue cancer trials in Example 1 and 2

Example 1—Phase IIA lung cancer trial

# interim analyses	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
1	0.0940	0.0860	0.9060	0.9140	33.35	35.86
2	0.0940	0.0864	0.9060	0.9136	30.68	35.43
13	0.0924	0.0904	0.9076	0.9096	25.44	29.01
26	0.0904	0.0932	0.9096	0.9068	24.23	27.65

Example 2—Phase IIA tongue cancer trial

# interim analyses	α	β	PET(p_0)	PET(p_1)	$\mathbb{E}[N\|p_0]$	$\mathbb{E}[N\|p_1]$
2	0.028	0.1650	0.9720	0.8350	23.69	41.13
4	0.027	0.1710	0.9730	0.8290	21.23	40.57
8	0.027	0.1720	0.9730	0.8280	20.28	40.06
16	0.024	0.1770	0.9760	0.8230	18.91	39.20
32	0.024	0.182	0.9760	0.8180	18.18	37.34

In the second example, 71.10% of the trials are stopped for futility under H_0 and 8.0% under H_1. Thus, the false-negative rate is still controlled at $\beta \leq 0.1$. Zero percent of the trials result in a false-positive conclusion under H_0, and zero percent of the trials are stopped for efficacy under H_1. Thus, in case of an unplanned early termination the error rates are still controlled in the second example, but the PET under H_1 is not sufficient to stop early for efficacy. The mean PP_e when stopping at 20 patients under H_0 and H_1 are 0.18 and 0.78 for the first and 0.15 and 0.72 for the second example.

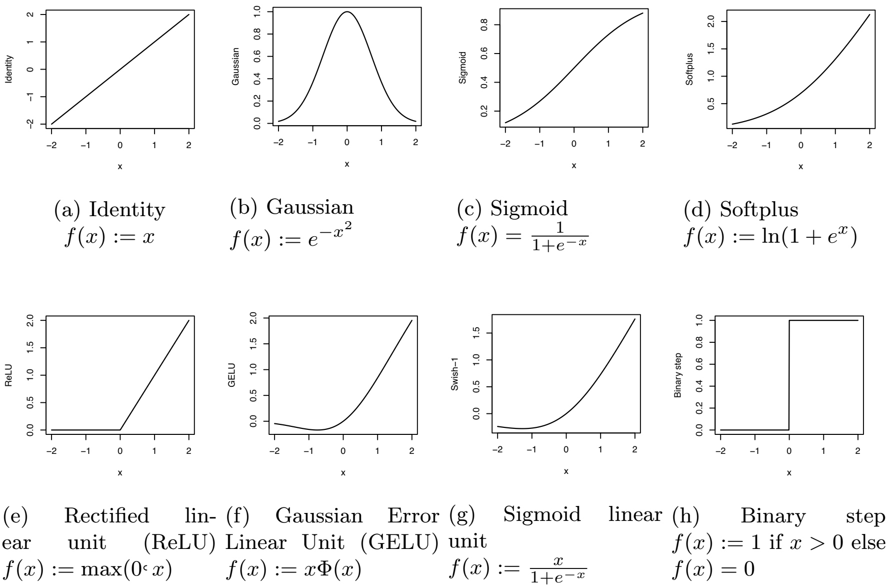
In summary, the most severe problem, an inflation of false-positive or false-negative rates, does not happen when an unplanned early termination of a trial happens with the PEV design.

9 The Brada R Package

All of the above examples, plots and simulations were computed with the accompanying R package brada. Details on the brada R package which implements the PEV design are provided in the supplementary material.

10 Discussion

The previous sections demonstrated the versatility of the proposed predictive evidence value design in real data examples and simulations. In this section, we discuss some limitations and points not covered in detail thus far. Importantly, we address the case of a non-flat reference function now.



(a) Identity $f(x) := x$

(b) Gaussian $f(x) := e^{-x^2}$

(c) Sigmoid $f(x) = \frac{1}{1+e^{-x}}$

(d) Softplus $f(x) := \ln(1 + e^{x})$

(e) Rectified linear unit (ReLU) $f(x) := \max(0‹ x)$

(f) Gaussian Error Linear Unit (GELU) $f(x) := x\Phi(x)$

(g) Sigmoid linear unit $f(x) := \frac{x}{1+e^{-x}}$

(h) Binary step $f(x) := 1$ if $x > 0$ else $f(x) = 0$

Fig. 9 Overview of different activation functions for ANN nodes

All applications in this paper rest on the choice of a flat reference function $r(p) :\equiv 1$. This choice leads to the case where the evidence interval recovers HPD-intervals, which will be unambiguous for most Bayesians. However, selecting a flat reference function is not mandatory. Although the flat reference function is the canonical choice, there is a huge palette of alternatives available. As the success probability p is inside [0, 1], any function $f : [0, 1] \rightarrow \mathbb{R}$, $p \mapsto f(p)$ is a potential candidate for the reference function $r(p)$.

Figure 9 shows some possible alternative choices for the reference function, which are inspired from the literature on artificial neural networks (ANNs). In fact, the functions in Fig. 9 are the most widely used activation functions for artificial neural networks (ANNs). Activation functions substantially influence the performance of an ANN and are classified into ridge, radial and fold functions depending on their mathematical properties. A popular activation function is the rectified linear unit (ReLU) shown in Fig. 9e. The ReLU has emerged for visual feature extraction since the 1960s in hierarchical neural networks [10, 11]. The choice of activation functions in ANNs has a substantial effect on the resulting training dynamics and performance of the resulting neural network. We expect the same for the resulting operating characteristics of the PEV design when shifting from the flat reference function to any of these functions. An appealing property which makes these functions attractive as candidates for non-flat reference functions is that most of them are monotonically increasing. This property should work as a kind of penalty for evidence about large success properties $p > p_0$ on H_1, thus reducing the false-positive rate under H_0 at the price of slightly decreased power under H_1 for large true success probabilities. When modifying the functions in Fig. 9 slightly, e.g. truncating them to the domain [0, 1] with associated image (e.g. the identity in Fig. 9a to $f : [0,1] \to [0,1]$, $x \mapsto f(x) := x$) they seem like reasonable choices for non-flat reference functions. However, we decided against including non-flat reference functions in the examples and simulations because of two primary reasons:

(a) Firstly, *any* of the functions in Fig. 9 can be parameterized into a whole family of possible functions. For example, the ReLU is easily generalized into a parameterized version $f(x) := \xi + c \cdot \max(p_0, 1)$ for $\xi \in \mathbb{R}$, $c \in \mathbb{R}_+$ and the cutoff p_0 between $H_0 : p \leq p_0$ and $H_1 : p > p_0$. The Softplus, GELU and Sigmoid linear unit can be parameterized into parabolically shaped families, while the sigmoid function allows parameterization into a logarithmically shaped family. Without further constraints (e.g. that a reference function should be *proper*, that is, integrate to unity, or be continuous) for each of these parameterized versions, there are infinitely many choices for such a reference function, which troubles application.

(b) Secondly, even after parameterization, formulation of further constraints and identification of canonical – or even possibly unique – choices among these differently shaped parameterized families of functions, there remain ethical problems. For example, some functions in Fig. 9 are monotonically increasing, or even strictly monotonically increasing. Thus, evidence for large success probabilities p close to 1 is increasingly stronger penalized when opting for a strictly monotonically increasing reference function. In contrast, using e.g. the binary step, evidence on $H_1 : p > p_0$ is equally penalized for any $p > p_0$. These aspects are crucial for ethical reasons, because several questions arise here:

- Is it justified to penalize large success probabilities more than small to moderate ones? This could be the case, because a steeper reference function on $H_1 : p > p_0$ should also lead to smaller false-positive rate, so calibration of a design should be possible at the price of risking that large success probabilities are not identified. In some contexts, e.g. oncology, large success probabilities are, however, often unrealistic so such a choice might be justified (depending on the precise context).
- Also, how should the reference function treat evidence on $H_0 : p \leq p_0$? For example, choosing $\xi = 1$ in the generalized ReLU means a flat reference function is chosen on H_0. However, smaller or larger values express favor or skepticism about H_0 being true. A regulatory agency will typically accept a critical stance towards H_1 (the drug being effective), while a critical stance towards H_0 must be justified (e.g. by improved operating characteristics of the design, while still being calibrated in a frequentist sense).

Further theoretical work is required before simulations under these non-flat reference functions can become helpful for practitioners, and align with ethical or regulatory agency requirements.

The two points (a) and (b) clarify why we opted for the canonical case of a flat reference function in this paper. Work on the theoretical side regarding (a) is current work in progress. This should help in providing a sound base for finding answers to the questions formulated in point (b).

However, some aspects are unambiguous even without the availability of further theoretical results: Firstly, we expect that the resulting operating characteristics of the design will be sensitive to the specific choice of non-flat reference function. The shape of the reference function will be crucial here, see (a).

Secondly, another layer of complexity is given by the interplay of prior and reference function, where we expect that they can work synergistic or cancel the effect of each other out, also depending on the specific choices. It has been shown that power gains are typically not possible in Bayesian designs when requiring strict false-positive control [26]. This is another interesting venue of future research now, because it should be possible to use informative priors based on historical data – e.g. power priors, see [15] – which may cause a design to be uncalibrated in the first place. Then, by using a similarly shaped, but slightly right-shifted reference function, the reference function raises the bar for evidence to accumulate for even more optimistic success probabilities than the ones specified in the informative prior. This way, historical data is incorporated and it could become possible to achieve power gains *only in certain regions* of the success probability $p \in [0, 1]$, for example for the region $p \in [p_0, p_0 + 0.2]$ of small to moderate success probabilities under H_1. See also the discussion in [26]. Conceptually speaking, by using an informative prior and an appropriate reference function it may become possible to achieve power gains through the prior at the price of suffering power in other areas of H_1 via the reference function.

Thirdly, the calibration algorithm proposed in this paper will work also for non-flat reference functions. However, we expect that extreme choices of reference functions may lead to cases where a calibration itself becomes impossible for a given $N_{\max}$. Such extreme choices should, however, be less relevant in practice.

11 Conclusion

In clinical research, the initial efficacy assessment of a new agent is typically considered in a phase IIA study which investigates the response rate of patients to the agent under consideration. Bayesian group-sequential designs for phase IIA studies are in practice often calculated based on the predictive probability approach which uses the predictive probability of concluding efficacy or futility based on interim data analyses under the premise that the trial will be conducted to the maximum planned sample size. The predictive probability of trial success is then used to stop the trial early for futility or efficacy.

In this paper, a novel group-sequential design for binary endpoints based on Bayesian evidence values – the PEV design – was proposed and its theoretical properties and operating characteristics were analyzed. It was shown that the predictive probability approach is a special case of the latter, and that the PEV design can improve the operating characteristics of the resulting trial in a variety of cases. The simulation and theoretical results demonstrated that Bayesian evidence values offer another layer of flexibility for error control in Bayesian group-sequential clinical trial designs, and offer the possibility to achieve smaller expected sample sizes and larger probabilities of early termination.

We provided default choices for the reference function (flat) and a four-step algorithm to calibrate the operating characteristics of the PEV design. The four-step algorithm is based on the theoretical results in Corollary 3, which shows how an improvement of the false-positive rate by increasing ν is possible. The developed R package `brada` facilitates calibration of the PEV design, and offers further methods for visualization, monitoring and reporting of a trial. Additionally, the `brada` package is designed for multicore environments and achieves efficient runtimes.

Our results indicate that the PEV design is quite robust to deviations from the sampling protocol and unplanned early termination. However, there are also limitations. First, the calibration is still computationally more difficult than for the competing BOP2 design and Simon's two-stage minimax and optimal designs. Second, there are cases where the BOP2 design achieves smaller sample sizes (e.g. setting 7 in Table 4). However, when a practically feasible number of interim analyses is carried out, the PEV design often outperforms Simon's two-stage design and the BOP2 design. In particular, while the expected sample sizes of the BOP2 design are often slightly smaller, the probability of early termination was always largest for the PEV design in all scenarios considered. This is an appealing feature, because the primary goal of a group-sequential trial design is to yield an early conclusion based on an interim analysis. As the expected sample sizes of the PEV and BOP2 design are often comparable, the PEV design provides an attractive competitor for a Bayesian phase IIA trial with a binary endpoint.

Although the PEV design often performed better than the standard PP design, it is not helpful to conclude that the PEV is "superior" to the standard PP design, as in this paper it was shown that the standard PP design is simply a special case of the PEV design. The latter is appealing because Bayesian group-sequential trials can be tailored to attain the desired frequentist operating characteristics and are gaining in popularity (e.g. the Biontech-Pfizer mRNA vaccine Comirnaty against SARS-Cov-2 used a Bayesian adaptive trial design based on a beta-binomial model similar to the one discussed in this paper. In particular, a $\mathcal{B}(0.700102, 1)$ prior adjusted for surveillance time was used, see page 91 in the EPAR available at https://www.ema.europa.eu/en/documents/assessment-report/comirnaty-epar-public-assessment-report_en.pdf. See also page 74 for details on the approach (e.g. the Bayesian design was calibrated to attain a frequentist $\alpha = 0.025$), and the posterior probability of vaccine efficacy (VE) being larger than 30% had to pass the threshold of 98.60% to declare VE. Note, however, that the Comirnaty trial was a phase II/III design with planned interim analyses at at least 32, 62, 82 and 120 cases (not participants)). This is also reflected in the ongoing interest in Bayesian group-sequential phase II designs with binary endpoints [19, 48].

Future work could extend the results obtained herein to other endpoints, because the derivations in this paper are not special to binary endpoints at all. Theorems 1 and 2 as well as the Corollaries, therefore, should also hold for continuous endpoints. Also, an extension to a two-group phase IIb design with treatment and control group should be straightforward. We expect the advantages in terms of smaller expected sample sizes and larger probabilities of early termination under H_0 will translate to this setting, too.

Appendix A Proofs

Proof of Theorem 1 It suffices to show that PP = PP$_e$ under the stated conditions, as then the group-sequential designs given in Algorithm 1 and 2 will yield identical decisions based on observed data $X = x$, fixed $\theta_L, \theta_U, \theta_T$, n and $N_{\max}$.

Based on Eqs. (4) and (10) the values for PP and PP$_e$ will be identical based on $X = x$ if and only if

$$P_{p|X,Y}(p > p_0|X = x, Y = i) \equiv \mathrm{Ev}_{EI_r(\nu)}(H_1) \tag{A1}$$

holds for $H_1 : p > p_1$ with arbitrary $p_1 > 0$. If Eq. (A1) holds, then

$$\mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} \equiv \mathbb{1}_{\mathrm{Ev}_{EI_r(\nu)}(H_1)>\theta_T} \tag{A2}$$

holds for fixed $\theta_T > 0$, and therefore the values of PP and PP$_e$ will be identical for fixed thresholds θ_L, θ_U, n and maximum sample size $N_{\max}$. Thus, it remains to show that Eq. (A1) holds. Therefore, write the observed data $y' := \{X = x, Y = i\}$, and we have

$$\begin{aligned}
\mathrm{Ev}_{EI_r(\nu)}(H_1) &\overset{(1)}{:=} \int_{EI_r(\nu)\cap H_1} p(p|y')dp \\
&\overset{(2)}{=} \int_{\{p\in[0,1]|s(p)\geq\nu\}\cap\{p\in[0,1]|p>p_0\}} p(p|y')dp \\
&\overset{(3)}{=} \int_{\{p\in[0,1]|p(p|y')\geq 0\}\cap[p_0,1]} p(p|y')dp \\
&\overset{(4)}{=} \int_{[0,1]\cap[p_0,1]} p(p|y')dp \\
&\overset{(5)}{=} \int_{[p_0,1]} p(p|y')dp \\
&\overset{(6)}{=} P_{p|X,Y}(p > p_0|X = x, Y = i)
\end{aligned} \tag{A3}$$

In (1), the definition of $\mathrm{Ev}_{EI_r(\nu)}(H_1)$ was used. In (2), the definition of $\mathrm{EI}_r(\nu)$ and H_1 was used. In (3) the assumptions $r(p) :\equiv 1$ and $\nu := 0$ were used and the fact that under a continuous beta prior the singleton $\{p_0\}$ is a Lebesgue null-set. In (4) to (5) the sets over which the integration is performed were rewritten, and the fact that the support of $p(p|y)$ equals $[0, 1]$ was used, and in (6) the integral in the previous step was identified as the appropriate posterior probability $P_{p|X,Y}(p > p_0|X = x, Y = i)$.

Equation (A3) now states that

$$\mathrm{Ev}_{EI_r(\nu)}(H_1) = P_{p|X,Y}(p > p_0|X = x, Y = i).$$

This was the statement in Eq. (A1). Thereby, Eq. (A2) follows which implies the statement in the theorem.

Proof of Corollary 1 The result follows from Theorem 1, because under $r(p) := 1$ and $v := 0$ the designs yield identical values for PP and PP_e. As a consequence, when $H_0 : p \leq p_0$ holds for a fixed $p_0 \in [0, 1]$, the corresponding type I error rates α_{PP} and α_{PP_e} are identical. The same holds for the corresponding type II error rates β_{PP} and β_{PP_e}, if data follow $H_1 : p > p_1$ for any fixed $p_1 \in [0, 1]$.

Proof of Theorem 2 To show Eq. (13), we show

$$\mathrm{Ev}_{EI_r(v)}(H_1) \leq P_{p|X,Y}(p > p_0|X = x, Y = i) \tag{A4}$$

except for the case when both sides are zero. If Eq. (A4) holds, then

$$\mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} = 0 \Rightarrow \mathbb{1}_{\mathrm{Ev}_{EI_r(v)}(H_1)>\theta_T} = 0 \tag{A5}$$

but

$$\mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} = 1 \not\Rightarrow \mathbb{1}_{\mathrm{Ev}_{EI_r(v)}(H_1)>\theta_T} = 1 \tag{A6}$$

for any fixed $\theta_T > 0$. In particular, this implies that $\mathrm{PP}_e \leq \mathrm{PP}$ because PP respectively PP_e depend on the values $P_{p|X,Y}(p > p_0|X = x, Y = i)$ respectively $\mathrm{Ev}_{EI_r(v)}(H_1)$. There may be cases in which $\mathbb{1}_{P_{p|X,Y}(p>p_0|X=x,Y=i)>\theta_T} = 1$ but $\mathbb{1}_{\mathrm{Ev}_{EI_r(v)}(H_1)>\theta_T} = 0$, and in these situations $\mathrm{PP}_e < \mathrm{PP}$. If a false-positive error is defined based on the threshold $\mathrm{PP} > \theta_U$ respectively $\mathrm{PP}_e > \theta_U$ (stopping early for futility although $H_0 : p \leq p_0$ holds), then $\mathrm{PP}_e < \mathrm{PP}$ implies that a false-positive error can occur based on PP, while it is possible that PP_e does not pass the critical threshold θ_U simultaneously. Note, however, that this must not be the case, which is why no *strict* inequality holds in general.

Equation (A4) remains to show. Therefore, notice that (we substitute the parameter θ for p for notational convenience)

$$\begin{aligned}\mathrm{Ev}_{EI_r(v)}(H_1) &:= \int_{EI_r(v)\cap H_1} p(\theta|y')d\theta \\ &= \int_{\{\theta\in\Theta|\frac{p(\theta|y)}{r(\theta)}\geq v\}\cap H_1} p(\theta|y')d\theta\end{aligned} \tag{A7}$$

$$\leq \int_{\{\theta\in\Theta|p(\theta|y)>0\}\cap H_1} p(\theta|y')d\theta \tag{A8}$$

$$\begin{aligned}
&= \int_{\Theta \cap H_1} p(\theta|y')d\theta \\
&= \int_{H_1} p(\theta|y')d\theta \\
&= \int_{(p_0,1]} p(p|y')dp \\
&= P_{p|X,Y}(p > p_0|X = x, Y = i)
\end{aligned} \tag{A9}$$

where in the second-last equality the parameter $\theta := p$ was re-substituted for notational reasons, and the last equality follows as $p(\cdot|y')$ is the posterior density, where again the observed data are denoted as $y' := \{X = x, Y = i\}$ in the steps before. Note that in Eq. (A8), $r(\theta) := r(p) :\equiv 1$ was used, and we have

$$\left\{\theta \in \Theta \middle| p(\theta|y) \geq \nu\right\} \subseteq \left\{\theta \in \Theta \middle| p(\theta|y) > 0\right\} \tag{A10}$$

as $\nu \geq 0$ by definition.

From Eq. (A9) now follows Eq. (A4) which finishes the proof.

Proof (Corollary 3) Now, if $\nu > 0$, Eq. (A10) becomes

$$\left\{\theta \in \Theta \middle| p(\theta|y) \geq \nu\right\} \subset \left\{\theta \in \Theta \middle| p(\theta|y) > 0\right\} \tag{A11}$$

As a consequence, inequality (A8) can become a strict inequality only if $\nu > 0$. Then (A4) becomes a strict inequality, too. For $\nu \to \infty$ we have

$$\left\{\theta \in \Theta \middle| p(\theta|y) \geq \nu\right\} \to \emptyset$$

so there exists a $\xi \in \mathbb{R}$ large enough, for which setting $\nu := \xi$ implies $\alpha_{\text{PP}_e} < \alpha_{\text{PP}}$, which is Eq. (14) and finishes the proof.

Appendix B Competing Designs

Simon's two-stage designs First, the main competitor is Simon's two-stage minimax and optimal design. Here, we focus mainly on Simon's minimax design. The reason is that in the PEV design, we want to fix $N_{\max}$ to the same value as in Simon's minimax design to avoid having another free parameter to calibrate. Still, there is no problem to compare the PEV design to Simon's optimal two-stage design, compare the second trial in Example 2. However, a reason for the focus on Simon's minimax design here is that in certain situations the PP and PEV design do require a larger $N_{\max}$ than Simon's minimax design.

Predictive probability design The second key competitor is the standard predictive probability approach. Although this is not really a competitor—Theorem 1 showed that the PP design *is*, in fact, a PEV design – we compare the PEV design with calibrated $\nu > 0$ to the calibrated PP design to see whether it is possible to obtain better trial operating characteristics with the PEV design.

Bayesian posterior probabilities Another possible competitor includes designs which are based on Bayesian posterior probabilities without any notion of predictive probability. Thus, given the interim data, the posterior probability of the drug being effective is computed. [31] showed that this approach is less favorable than the predictive probability approach, and we, therefore, do not include it here.

BOP2 design Another competitor that is proposed in the literature is the BOP2 design of [49]. The BOP2 design includes (amongst others) binary endpoints. At each interim, the go/no-go decision in the BOP2 design is made by evaluating a set of posterior probabilities of the events of interest, which is optimized to maximize power or minimize the number of patients under the null hypothesis. Unlike other existing Bayesian designs, the BOP2 design explicitly controls the type I error rate, thereby bridging the gap between Bayesian designs and frequentist designs. For our current purposes, a comparison with the BOP2 design is only partially useful because as stressed in the introduction, we focus on the false-positive *and* false-negative constraints in Eqs. (1) and (2) simultaneously. Therefore, it may be the case that the BOP2 design yields a better power than the PP or PEV design but has no false-negative control. One example of this problem occurs in Table 4. Still, we will report the results of the BOP2 design anyway to get an intuition whether the power of a design can be improved through introduction of the reference function and evidence threshold. Calculation of the BOP2 design is straightforward via https://biostatistics.mdanderson.org/shinyapps/BOP2/.

Sequential Bayes factor designs A last branch of competing designs deals with Bayes factors: Examples are given in [37, 39, 43] and [42]. Although the approach is appealing, there is currently no efficient software implementation. There is the `bfda` package on GitHub, compare https://rawgit.com/nicebread/BFDA/master/package/doc/BFDA_manual.html, but there is still no R package on CRAN. Furthermore, Bayes factors are not without controversy, see for example [44] or [23]. However, future research should analyze group-sequential Bayes factor tests more thoroughly.

The Brada R Package

Application of the PEV design is facilitated by the ®package `brada`. The package was developed to simplify the planning, calibration, monitoring and reporting of a Bayesian phase IIA trial with either the PP or PEV design. All computations in this paper have been carried out with the help of the `brada` package, and the package is available on CRAN under https://cran.r-project.org/web/packages/brada/index.html. The package consists of the main function `brada` to fit a trial, `plot` to produce plots in the style of Fig. 4, and calibrate to calibrate ν and θ_L as detailed in the four-step calibration, compare the previous section. When analyzing the operating characteristics of a design, the brada package internally sets up a multicore cluster to reduce computation time. In practice, the analysis of a PP or PEV design therefore takes little time even on desktop computers. As the calibration of Bayesian designs is simulation-based, the default plots of the brada package always include Monte Carlo errors in square brackets behind each estimate, compare e.g. Figs. 4, 5, 6, 7 and 8.

We have provided a vignette to get started with the package on CRAN at https://cran.r-project.org/web/packages/brada/index.html. Also, the accompanying Quarto replication scripts (see https://osf.io/zmfyn/?view_only=348067ed1ccc498da7e4a11d949c84df) at the Open Science Foundation include all code to reproduce the results in this manuscript and serve as a further source of how to use the package.

Author Contributions RK came up with the original idea, wrote up a first draft, derived the main results and carried out all simulations. AS proofread the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors did not receive any funding for this research.

Data Availability The accompanying Quarto replication scripts (see https://osf.io/zmfyn/?view_only=348067ed1ccc498da7e4a11d949c84df) at the Open Science Foundation include all code to reproduce the results in this manuscript

Code Availability the accompanying Quarto replication scripts (see https://osf.io/zmfyn/?view_only=348067ed1ccc498da7e4a11d949c84df) at the Open Science Foundation include all code to reproduce the results in this manuscript

Declarations

Competing interests The authors declare no conflict of interest.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

References

1. Berry SM (2011) Bayesian adaptive methods for clinical trials. CRC Press, Boca Raton
2. Casella G, Berger RL (2002) Statistical inference. Thomson Learning, Stamford

3. Diniz M, Pereira CAB, Polpo A, Stern JM, Wechsler S (2012) Relationship between Bayesian and frequentist significance indices. Int J Uncertain Quantif 2(2):161–172
4. Dmitrienko A, Wang MD (2006) Bayesian predictive approach to interim monitoring in clinical trials. Stat Med 25(13):2178–2195. https://doi.org/10.1002/SIM.2204
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Verweij J (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45(2):228–247. https://doi.org/10.1016/J.EJCA.2008.10.026
6. Fayers PM, Ashby D, Parmar MK (2005) Monitoring: Bayesian data monitoring in clinical trials. Tutor Biostat Stat Methods Clin Stud 1:335–352. https://doi.org/10.1002/0470023678.CH3B
7. Ferguson J (2021) Bayesian interpretation of p values in clinical trials. BMJ Evid-Based Med. https://doi.org/10.1136/BMJEBM-2020-111603
8. Ferreira D, Ludes PO, Diemunsch P, Noll E, Torp KD, Meyer N (2021) Bayesian predictive probabilities: a good way to monitor clinical trials. Br J Anaesth 126(2):550–555. https://doi.org/10.1016/J.BJA.2020.08.062
9. Freedman LS, Spiegelhalter DJ, Parmar MK (1994) The what, why and how of Bayesian clinical trials monitoring. Stat Med 13(13–14):1371–1383. https://doi.org/10.1002/SIM.4780131312
10. Fukushima K (1969) Visual feature extraction by a multilayered network of analog threshold elements. IEEE Trans Syst Sci Cybern 5(4):322–333. https://doi.org/10.1109/TSSC.1969.300225
11. Fukushima K (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36(4):193–202. https://doi.org/10.1007/BF00344251
12. Gehan EA (1961) The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis 13(4):346–353. https://doi.org/10.1016/0021-9681(61)90060-1
13. Giovagnoli A (2021) The Bayesian design of adaptive clinical trials. Int J Environ Res Public Health 18(2):1–15. https://doi.org/10.3390/IJERPH18020530
14. Good I (1966) A derivation of the probabilistic explication of information. J R Stat Soc Ser B (Methodol) 28(3):578–581
15. Gravestock I, Held L (2017) Adaptive power priors with empirical Bayes for clinical trials. Pharm Stat 16(5):349–360. https://doi.org/10.1002/PST.1814
16. Guo B, Liu S (2020) An optimal Bayesian predictive probability design for phase II clinical trials with simple and complicated endpoints. Biom J 62(2):339–349. https://doi.org/10.1002/BIMJ.201900022
17. Held L, Sabanés Bové D (2014) Applied statistical inference. Springer, Berlin
18. Hodges JL, Lehmann EL (1954) Testing the approximate validity of statistical hypotheses. J R Stat Soc: Ser B (Methodol) 16(2):261–268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x
19. Kaizer A, Zabor E, Nie L, Hobbs B (2022) Bayesian and frequentist approaches to sequential monitoring for futility in oncology basket trials: a comparison of Simon's two-stage design and Bayesian predictive probability monitoring with information sharing across baskets. PLoS ONE 17(8):e0272367. https://doi.org/10.1371/JOURNAL.PONE.0272367
20. 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 Med Res Methodol 21:171. https://doi.org/10.1186/s12874-021-01341-7
21. Kelter R (2021) fbst: an R package for the Full Bayesian Significance Test for testing a sharp null hypothesis against its alternative via the e value. Behav Res Methods. https://doi.org/10.3758/s13428-021-01613-6
22. Kelter R (2021) How to choose between different Bayesian posterior indices for hypothesis testing in practice. Multivar Behav Res. https://doi.org/10.1080/00273171.2021.1967716
23. Kelter R (2021) On the measure-theoretic premises of Bayes factor and full Bayesian significance tests: a critical reevaluation. Comput Brain Behav. https://doi.org/10.1007/s42113-021-00110-5
24. Kelter R (2022) The evidence interval and the Bayesian evidence value—on a unified theory for Bayesian hypothesis testing and interval estimation. Br J Math Stat Psychol 75(3):550–592. https://doi.org/10.1111/bmsp.12267
25. Koehler E, Brown E, Haneuse SJ (2009) On the assessment of Monte Carlo error in simulation-based statistical analyses. Am Stat 63(2):155. https://doi.org/10.1198/TAST.2009.0030
26. Kopp-Schneider A, Calderazzo S, Wiesenfarth M (2020) Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biom J 62(2):361–374. https://doi.org/10.1002/BIMJ.201800395

27. Kruschke JK (2018) Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci 1(2):270–280. https://doi.org/10.1177/2515245918771304
28. Kruschke JK, Liddell T (2018) The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25:178–206. https://doi.org/10.3758/s13423-016-1221-4
29. Kullback S (1959) Information theory and statistics. Wiley, New York
30. Lee J, Chu CT (2012) Bayesian clinical trials in action. Stat Med 31(25):2955–2972. https://doi.org/10.1002/SIM.5404
31. Lee J, Liu DD (2008) A predictive probability design for phase II cancer clinical trials. Clin Trials 5(2):93–106. https://doi.org/10.1177/1740774508089279
32. Lesaffre E, Baio G, Boulanger B (2020) Bayesian methods in pharmaceutical research. Chapman & Hall/CRC, Boca Raton
33. Matthews JN (2006) Introduction to randomized controlled clinical trials, 2nd edn. CRC Press, Boca Raton
34. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38(11):2074–2102. https://doi.org/10.1002/SIM.8086
35. Pereira CADB, Stern JM (2020) The e-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program. São Paulo J Math Sci. https://doi.org/10.1007/s40863-020-00171-7
36. Qin F, Wu J, Chen F, Wei Y, Zhao Y, Jiang Z, Yu H (2020) Optimal, minimax and admissible two-stage design for phase II oncology clinical trials. BMC Med Res Methodol 20(1):1. https://doi.org/10.1186/S12874-020-01017-8
37. Rosner GL (2020) Bayesian adaptive designs in drug development. In: Lesaffre E, Baio G, Boulanger B (eds) Bayesian methods in pharmaceutical research. CRC Press, Boca Raton, pp 161–184
38. Ryan EG, Bruce J, Metcalfe AJ, Stallard N, Lamb SE, Viele K, Gates S (2019) Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. BMC Med Res Methodol 19(1):1–10. https://doi.org/10.1186/S12874-019-0739-3/TABLES/4
39. Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, Perugini M (2017) Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. Psychol Methods 22(2):322–339. https://doi.org/10.1037/met0000061
40. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
41. Simon R (1989) Optimal two-stage designs for phase II clinical trials. Controll Clin Trials 10(1):1–10. https://doi.org/10.1016/0197-2456(89)90015-9
42. Stefan AM, Lengersdorff LL, Wagenmakers EJ (2022) A two-stage Bayesian sequential assessment of exploratory hypotheses. Collabra Psychol 8(1):1. https://doi.org/10.1525/COLLABRA.40350
43. Stefan AM, Schönbrodt FD, Evans NJ, Wagenmakers EJ (2022) Efficiency in sequential testing: comparing the sequential probability ratio test and the sequential Bayes factor test. Behav Res Methods 1:1–18. https://doi.org/10.3758/S13428-021-01754-8/TABLES/1
44. Tendeiro JN, Kiers HA (2019) A review of issues about null hypothesis Bayesian testing. Psychol Methods 24(6):774–795. https://doi.org/10.1037/met0000221
45. Thall PF, Wathen JK (2007) Practical Bayesian adaptive randomization in clinical trials. Eur J Cancer (Oxford, England: 1990) 43(5):859. https://doi.org/10.1016/J.EJCA.2007.01.006
46. Therasse P, Eisenhauer EA, Verweij J (2006) RECIST revisited: a review of validation studies on tumour assessment. Eur J Cancer 42(8):1031–1039. https://doi.org/10.1016/J.EJCA.2006.01.026
47. Wagenmakers EJ, Gronau QF, Dablander F, Etz A (2020) The support interval. Erkenntnis. https://doi.org/10.1007/s10670-019-00209-z
48. Zabor EC, Kaizer AM, Pennell NA, Hobbs BP (2022) Optimal predictive probability designs for randomized biomarker-guided oncology trials. Front Oncol. https://doi.org/10.3389/FONC.2022.955056
49. Zhou H, Lee JJ, Yuan Y (2017) BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. Stat Med 36(21):3302–3314. https://doi.org/10.1002/SIM.7338