Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 90-105
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Frequentist and Bayesian Sample Size Determination for Single-Arm Clinical Trials Based on a Binary Response Variable: A Shiny App to Implement Exact Methods

Susanna Gentile, Valeria Sambucini

Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
Email: susanna.gentile@uniroma1.it, valeria.sambucini@uniroma1.it

How to cite this paper: Gentile, S. and Sambucini, V. (2024) Frequentist and Bayesian Sample Size Determination for Single-Arm Clinical Trials Based on a Binary Response Variable: A Shiny App to Implement Exact Methods. *Open Journal of Statistics*, **14**, 90-105.
https://doi.org/10.4236/ojs.2024.141004

Received: November 13, 2023
Accepted: February 26, 2024
Published: February 29, 2024

Abstract

Sample size determination typically relies on a power analysis based on a frequentist conditional approach. This latter can be seen as a particular case of the two-priors approach, which allows to build four distinct power functions to select the optimal sample size. We revise this approach when the focus is on testing a single binomial proportion. We consider exact methods and introduce a conservative criterion to account for the typical non-monotonic behavior of the power functions, when dealing with discrete data. The main purpose of this paper is to present a Shiny App providing a user-friendly, interactive tool to apply these criteria. The app also provides specific tools to elicit the analysis and the design prior distributions, which are the core of the two-priors approach.

Keywords

Binomial Proportion, Frequentist and Bayesian Power Functions, Exact Sample Size Determination, Shiny App, Two-Priors Approach

1. Introduction

Sample Size Determination (SSD) is an essential step in the design of a research study, especially in clinical trials. Let us denote by θ the parameter of interest, which measures the efficacy of a novel treatment, and assume that we are interested in testing $H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$, where Θ_0 and Θ_1 form a partition of the parameter space Θ.

A well-established strategy for SSD exploits the concept of *power function*: the

study is sized to guarantee a large probability of rejecting the null hypothesis H_0, when it is actually false. The decision about the rejection of H_0 can be made under a frequentist framework or by performing a Bayesian analysis. In this latter case, a prior distribution, called the *analysis prior*, is introduced to incorporate in the procedure pre-trial knowledge the researcher wants to take into account, together with pre-experimental evidence if available. Moreover, the conjecture that the alternative hypothesis is true represents an essential element of the methodology. It can be realized by assuming that the true θ is equal to a fixed design value θ^D, suitably selected under H_1 and, in this case, the probability of rejecting H_0 is evaluated by exploiting the sampling distribution of the test statistic conditional on θ^D (*conditional power*). Alternatively, we can introduce uncertainty on the guessed design value by incorporating another prior distribution, called the *design prior*, which assigns negligible probability to values of θ under H_0. In this latter case, the probability of rejecting H_0 is computed by exploiting the prior predictive distribution of the test statistic under the assumption that θ is distributed according to the design prior (*predictive power*). By combining frequentist and Bayesian procedures of analysis, with both the conditional and the predictive approach, we can obtain four power functions that we can use for sample size determination (see [1]). The general idea is to select the minimum sample size necessary to achieve a desired level of power. This methodology, based on the introduction of two distinct prior distributions and thus based on the so-called *two-priors approach*, has been initially formalized by Wang and Gelfand [2]. It is now well known with many implementations presented in the literature (see, among others, [3]-[10]).

In this paper, we consider the problem of SSD based on power analysis when the focus is on single-arm studies based on a single binomial proportion. This design is typically used in Phase II of clinical trials, where the parameter of interest is the probability of response to a novel therapy. Sambucini [1] derived the four power functions described above by using frequentist and Bayesian exact methods at the analysis stage, which are particularly attractive because Phase II sample sizes are usually small. It is interesting to remark that, since we are dealing with discrete data, the power functions show a basically increasing, but not-monotonic, behaviour as a function of the sample size. This "saw-tooth behaviour" requires a modification of the standard criterion to select the optimal sample size, if we are interested in having the condition regarding the power functions fulfilled also for all the sample size values greater than the optimal one ([1] [11]). This modification of the SSD criteria has been also introduced in Gentile and Sambucini [12], where the four power functions have been derived for single-arm trials based on count data. The aim of this paper is to present an R Shiny web application (app) developed to implement the SSD criteria provided in [1]. Some R functions, contained in already existing packages, are available to compute the optimal sample size for a single binomial proportion, but they are based only on the frequentist conditional power and rely on asymptotic ap-

proximations. For instance, the functions pwr.p.test and prop1, implemented in the R packages pwr [13] and pwr2ppl [14], exploit the arcsine transformation of the proportion [15], whereas the function power.prop1.test of the package MKpower uses the normal approximation [16]. Furthermore, some functions allow for exact sample size computation, but do not account for the saw-tooth behaviour of the power function. Examples are the power_binom_test function in the package MESS [17] and the propTestPower function in the package EnvStats [18]. The function power.diagnostic.test of the package MKpower accounts for the saw-tooth behaviour, but it aims to size diagnostic tests for an expected sensitivity or specificity [16]. Functions implementing the exact frequentist conditional power are also available in software specific for sample size determinations. Examples of freeware are G*Power [19] and Lenth's applet [20]. For a more exhaustive list the reader is referred to the textbook by Ryan [15].

In practice, when the interest is focused on a single binomial proportion, many software tools have been developed to implement the standard procedures for SSD, based on power analysis conducted using the frequentist conditional approach. Instead, to our best knowledge, up until now, no software tool has been available to implement exact criteria based on the other three power functions. Thus, we developed an R Shiny App [21] [22] that provides a user-friendly and interactive environment to obtain the optimal sample size according to the criteria based on the two-priors approach and derived in [1]. The app allows the visualization of the behaviour of the four power functions as the sample size increases and lets the user decide whether to take into account or not the saw-tooth behaviour of the power when selecting the sample size. It also contains specific tools to suitably select the analysis and the design prior distributions.

The rest of the paper is organized as follows. In Section 2, we revise the exact procedures, based on the two-priors approach, to select the optimal sample size for a single binomial proportion. Section 3 discusses some strategies to elicit the prior distributions. In Section 4, we present the Shiny App and illustrate its features through an example. Finally, Section 5 contains some concluding remarks.

2. Exact SSD Methods for a Single Binomial Proportion

In this Section, we revise the exact SSD procedures based on four possible power functions, assuming that interest is on one-sample testing problems with a binary response [1].

Specifically, let us suppose that we are interested in testing the proportion of responders to a novel therapy. We consider n patients, each of whom receives the same treatment dosage, and classified as responders or not to the therapy by using a binary variable Y. We assume that we are interested in testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$, where θ denotes the parameter of interest, *i.e.* the true response rate, while θ_0 is a fixed target value that should represent the efficacy rate of the standard of care and is usually estimated through historical data.

2.1. Frequentist Power Functions

Initially we assume that, at the analysis stage, the decision of rejecting the null hypothesis is made under a frequentist framework. The test statistic to use is the number of responders out of the n patients, Y_n, whose sampling distribution of is

$$f_n\left(y_n \mid \theta\right) = \text{bin}\left(y_n; n, \theta\right), \text{ for } y_n = 0, \cdots, n,$$

where $\text{bin}\left(\cdot; n, p\right)$ denotes the probability mass function of a binomial with parameters n and p. Therefore, the frequentist rejection region at level α is $R_{H_0} = \left\{y_n \in \{0,1,\cdots,n\} : y_n \geq r\right\}$, where the critical value r is given by

$$r = \min\left\{k \in \{0,1,\cdots,n\} : \sum_{i=k}^{n} \text{bin}\left(i; n, \theta_0\right) \leq \alpha\right\}. \tag{1}$$

Note that, since the binomial distribution is discrete, the actual Type I error rate does not hit α exactly, but it is always less than or equal to it. In order to exploit the power function for SSD purposes, at the design stage, we need to consider a scenario under which the alternative hypothesis is true. A first possibility is to specify a design value θ^D for θ that belongs to H_1. In this case, we obtain the *frequentist conditional power*

$$\eta_F^C\left(n\right) = \mathbb{P}_{f_n\left(\cdot \mid \theta^D\right)}\left(Y_n \in R_{H_0}\right) = \sum_{y_n = r}^{n} \text{bin}\left(y_n; n, \theta^D\right),$$

where $\mathbb{P}_{f_n\left(\cdot \mid \theta^D\right)}$ denotes the probability measure associated with the sampling distribution of Y_n for $\theta = \theta^D$. Note that $\eta_F^C\left(n\right)$ represents the probability of correctly rejecting H_0 using a frequentist procedure, when θ is equal to θ^D. However, we can add flexibility to the procedure by avoiding the use of a fixed design value. In fact, it is possible to introduce uncertainty on the suitable design value to specify by eliciting a *design prior distribution*, $\pi^D\left(\theta\right)$. This latter is an instrumental tool that allows to model design expectations on θ, under the assumption that the treatment is effective. Consequently, it should be chosen as an informative distribution, as we will discuss further in Section 3. In our specific case, given a beta design prior, $\pi^D\left(\theta\right) = \text{beta}\left(\theta \mid \alpha^D, \beta^D\right)$, the prior predictive distribution of Y_n is

$$m_n^D\left(y_n\right) = \text{beta-bin}\left(y_n; \alpha^D, \beta^D, n\right), \text{ for } y_n = 0, \cdots, n, \tag{2}$$

where $\text{beta-bin}\left(\cdot; p, q, n\right)$ is the probability mass function of a beta-binomial with parameters p, q and n. Therefore, the *frequentist predictive power* is given by

$$\eta_F^P\left(n\right) = \mathbb{P}_{m_n^D(\cdot)}\left(Y_n \in R_{H_0}\right) = \sum_{y_n = r}^{n} \text{beta-bin}\left(y_n; \alpha^D, \beta^D, n\right)$$

where $\mathbb{P}_{m_n^D(\cdot)}$ denotes the probability measure associated with the prior predictive distribution of Y_n in (2) and r is the critical value defined in (1). Note that $\eta_F^P\left(n\right)$ provides the probability of correctly rejecting H_0 using a frequentist procedure, when θ is guessed to belong to the alternative hypothesis, where it is distributed according to $\pi^D\left(\theta\right)$.

2.2. Bayesian Power Functions

Alternatively, it is possible to perform the analysis under a Bayesian framework. This choice allows us to take into account pre-experimental information available on the treatment, for instance, based on historical data or on the subjective opinions of experts. The information is incorporated through the elicitation of another prior distribution on the parameter, the *analysis prior distribution*, $\pi^A(\theta)$. By exploiting conjugate analysis results, we consider a beta density, $\pi^A(\theta) = \text{beta}(\theta; \alpha^A, \beta^A)$, so that the corresponding posterior distribution is

$$\pi_n^A(\theta \mid y_n) = \text{beta}(\theta; \alpha^A + y_n, \beta^A + n - y_n).$$

Under this setup, to build a Bayesian equivalent of a power function, we need to determine the set of values of Y_n that, if observed, would lead to rejecting the null hypothesis. In line with Spiegelhalter *et al.* [23], we name this condition on the random result as the "Bayesian significance" and establish that it consists in rejecting H_0 if the posterior probability of the alternative hypothesis is sufficiently high. Thus, in our specific case, Y_n can be considered "significant" in a Bayesian perspective if

$$\mathbb{P}_{\pi_n^A(\cdot \mid Y_n)}(\theta > \theta_0) > 1 - \varepsilon \tag{3}$$

where $\mathbb{P}_{\pi_n^A(\cdot \mid Y_n)}$ denotes the probability measures associated with the posterior distribution of θ and ε is a probability threshold typically selected as a small value. The condition in (3) is a random object at the design phase because it depends on the future result Y_n. Nevertheless, for a fixed value of n, the probability that it is fulfilled increases with Y_n. Therefore, the Bayesian rule consists in rejecting H_0 if $Y_n \geq \tilde{r}$, where

$$\tilde{r} = \min\left\{k \in \{0,1,\cdots,n\} : \mathbb{P}_{\pi_n^A(\cdot \mid k)}(\theta > \theta_0) > 1 - \varepsilon\right\}. \tag{4}$$

Then, we need to compute the probability that the Bayesian significance condition is fulfilled under the optimistic assumption that the treatment is effective. Once again, we may realize this assumption by using either a conditional or a predictive approach. In the first case, we fix a suitable design value θ^D under H_1 and define the *Bayesian conditional power* as

$$\eta_B^C(n) = \mathbb{P}_{f_n(\cdot \mid \theta^D)}\left[\mathbb{P}_{\pi_n^A(\cdot \mid Y_n)}(\theta > \theta_0) > 1 - \varepsilon\right] = \sum_{y_n = \tilde{r}}^{n} \text{bin}(y_n; n, \theta^D).$$

In the second case, we elicit a design prior distribution for θ, as described above, and obtain the *Bayesian predictive power*, that is

$$\eta_B^P(n) = \mathbb{P}_{m_n^D(\cdot)}\left[\mathbb{P}_{\pi_n^A(\cdot \mid Y_n)}(\theta > \theta_0) > 1 - \varepsilon\right] = \sum_{y_n = \tilde{r}}^{n} \text{beta-bin}(y_n; \alpha^D, \beta^D, n).$$

Clearly, both the power functions $\eta_B^C(n)$ and $\eta_B^P(n)$ provide the probability of correctly rejecting H_0 using a Bayesian procedure, under the assumption that the alternative hypothesis is true. Moreover, it is worth pointing out that the Bayesian predictive power is the one that allows to model both prior knowledge and uncertainty on the design value: it includes the other power functions as

special cases. In fact, if we consider a point-mass design distribution on θ^D, then no design uncertainty is involved, and the predictive approach coincides with the conditional one. On the other hand, If no pre-experimental information is available, a non-informative analysis prior can be elicited and the Bayesian powers coincide with the frequentist one.

2.3. Sample Size Determination Criteria

Whatever the power function chosen, the standard SSD criterion selects the optimal n as the minimum value such that the power exceeds a threshold of interest γ. Hence, the optimal sample sizes are obtained as

$$n_J^I = \min\left\{n \in \mathbb{N} : \eta_J^I(n) \geq \gamma\right\},\ I = C, P,\ J = F, B. \tag{5}$$

where the superscript refers to the approach used at the design stage, while the subscript refers to the approach used at the analysis stage. However, given the saw-tooth shape of the power curves as a function of n, a slightly different and more conservative SSD criterion can be adopted. The idea is to select the smallest sample size such that the condition on the power is fulfilled also for all the sample size values greater than it, that is

$$n_J^I = \min\left\{n^* \in \mathbb{N} : \eta_J^I(n) \geq \gamma, \forall n \geq n^*\right\},\ I = C, P,\ J = F, B. \tag{6}$$

This latter criterion prevents the condition of interest from being satisfied for the selected sample size, but no longer satisfied for some larger values of n.

3. Prior Distributions Selection

This section discusses some strategies to elicit the design and analysis prior distributions, accounting for their different aims. We start focusing on the design prior distribution $\pi^D(\theta)$. The idea is to express the hyperparameters in terms of the prior mode θ^D and the prior sample size n^D by using [1] [24]:

$$\alpha^D = n^D\theta^D + 1 \text{ and } \beta^D = n^D\left(1-\theta^D\right)+1. \tag{7}$$

We can center $\pi^D(\theta)$ on the design value we would consider in the conditional approach and regulate the concentration through the choice of the prior sample size. It is crucial to emphasize that $\pi^D(\theta)$ should be an informative distribution. First, it serves to realize the assumption that θ belongs to the alternative hypothesis. Furthermore, as n approaches infinity, $\eta_B^P(n)$ and $\eta_F^P(n)$ tend to the probability assigned to the alternative hypothesis by $\pi^D(\theta)$, denoted by $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0)$ [25]. Thus, $\mathbb{P}_{\pi^D(\theta)}(\theta > \theta_0)$ should be close to one to ensure that the power tends to 1 as n goes to infinity. We suggest the use of two possible strategies to ensure that the design prior distribution satisfies these features. Once θ^D has been specified, we determine n^D numerically so that:

1) $\pi^D(\theta)$ assigns a probability close to one to the alternative hypothesis;
2) $\pi^D(\theta)$ assigns a probability close to one to a symmetric interval $\left(\theta^D - \delta, \theta^D + \delta\right)$, where δ is a non-negative real number such that $\theta^D - \delta \geq \theta_0$.

Both these procedures are implemented in the Shiny App described in the

next Section. Finally, as n^D tends to infinity, $\pi^D(\theta)$ tends to assign all the probability mass to the prior mode θ^D, resulting in no variability introduced around it. Consequently, the predictive and conditional approaches align.

The elicitation of analysis prior distribution can be based on historical data or on subjective opinions of experts. However, one of the most common ways of proceeding is to choose a non-informative density or a density based on very weak information, to avoid the possibility of underweighting the results of the current experiment in determining the analysis outcome. Thus, for instance, we may rely on the Uniform distribution on the interval $(0,1)$ by specifying $\alpha^A = \beta^A = 1$ or on the non-informative Jeffreys prior by setting $\alpha^A = \beta^A = 0.5$. As an alternative and similarly to the choice of $\pi^D(\theta)$, we can express the hyperparameters of the analysis prior distribution in terms of prior mode, θ^A and prior sample size n^A, where n^A is typically fixed equal to one, or equal to a very low value, in order to obtain a weakly informative prior distribution. This way of proceeding allows to express skepticism, neutrality or optimism about large treatment effects through the choice of the prior mode θ^A. Finally, let us notice that if we introduce no prior information, *i.e.* if n^A is set equal to 0, the Bayesian and the frequentist setup coincide.

4. Shiny App

This Section presents a Shiny App that implements the sample size criteria described in Section 2. It is available at the following link:

https://susanna-gentile.shinyapps.io/SSD_singlearm.

The Shiny package in R enables the creation of interactive web apps directly from R ([21] [22]). Our Shiny App aims to provide an intuitive and user-friendly tool for applying the methodologies discussed in this paper. The main functionalities of the app are:

- To allow computing the optimal sample size according to the four power functions;
- To implement both the standard and the conservative criterion to select the optimal sample size;
- To display, if requested, the power function behaviour as a function of n;
- To enable storing the design parameters and results into a table and to download it as a CSV file;
- To help to select the analysis and the design prior distributions and to visualize them.

Thus, the users can select either the conservative criterion in (6), accounting for the saw-tooth behaviour, or the standard criterion. We suggest using the first criterion. However, we let this choice be at the user's discretion as there is no unanimous agreement on the appropriateness of this methodology [15]. The tools to select the design and analysis prior are organized into two separate panels. As previously stressed, the two distributions have different aims and should be distinguished.

4.1. User Interface Structure

We start by describing the User Interface (UI). The UI changes accordingly to the methodologies chosen to conduct the design and analysis phases, as depicted in Figures 1-4. In the upper part of the UI, users can input the design parameters, split into three groups.

General setting: The inputs include the historical control θ_0, the power level γ, and the maximum sample size. Users can choose whether to use the conservative criterion (the default), the standard one, or both.

Analysis stage: The inputs depend on the planned final analysis. For a frequentist analysis, the app requires the Type-I error probability α (Figure 1). For a Bayesian analysis, the app requires the Bayesian significance level $1-\varepsilon$ and the analysis prior's hyperparameters. Users can exploit the "Analysis prior" panel to select them (Figure 3, panels (a) and (b)).

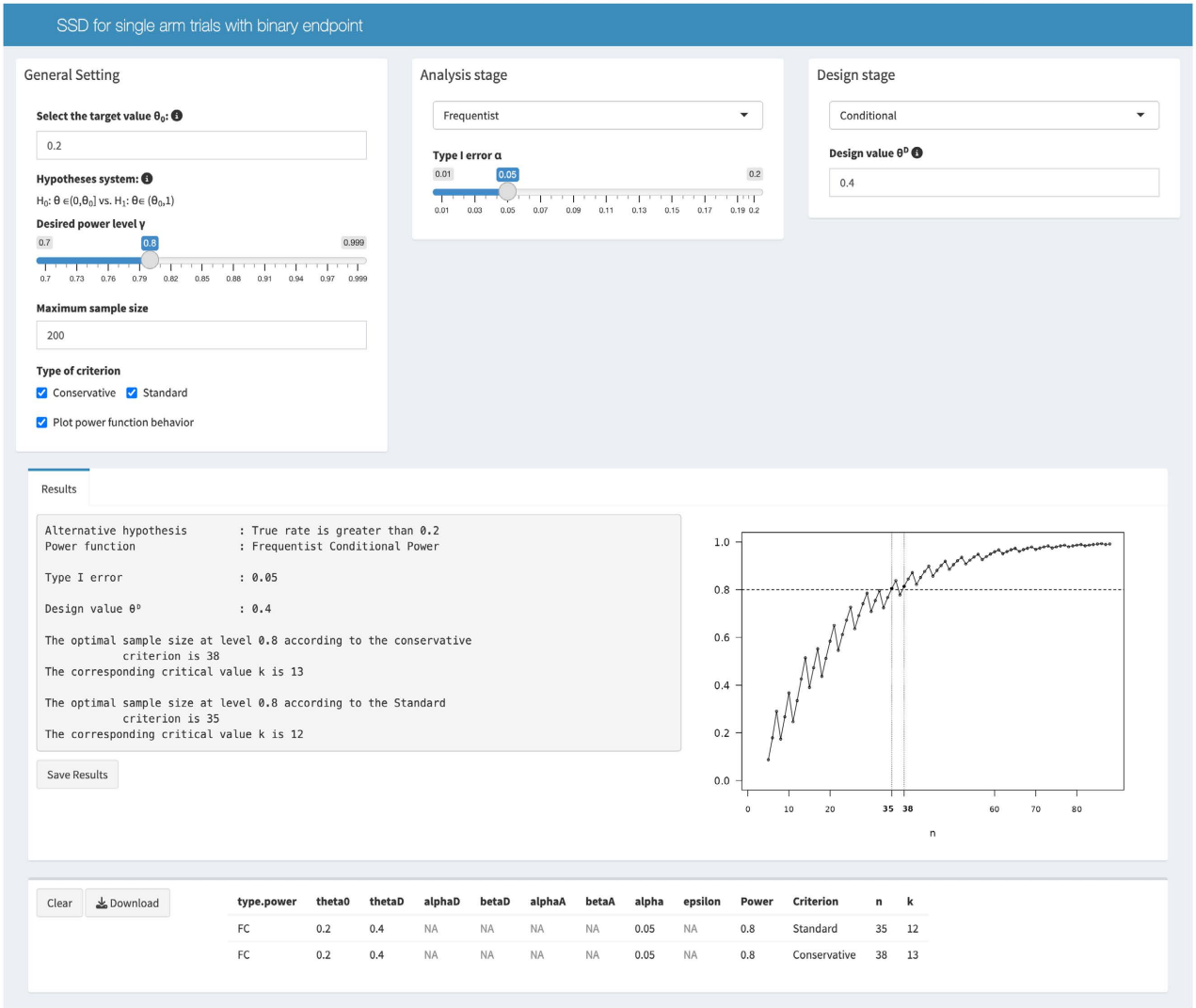
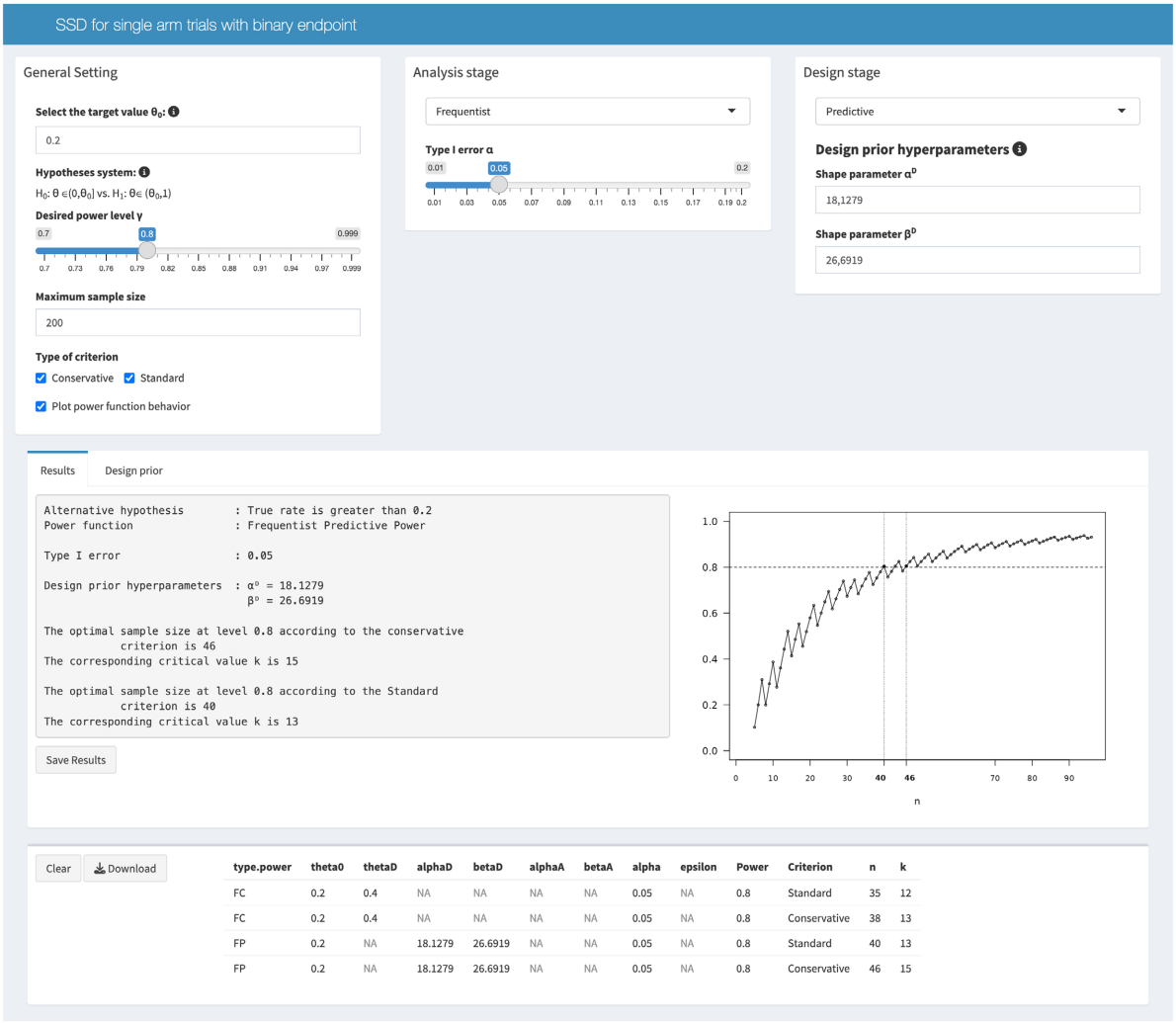
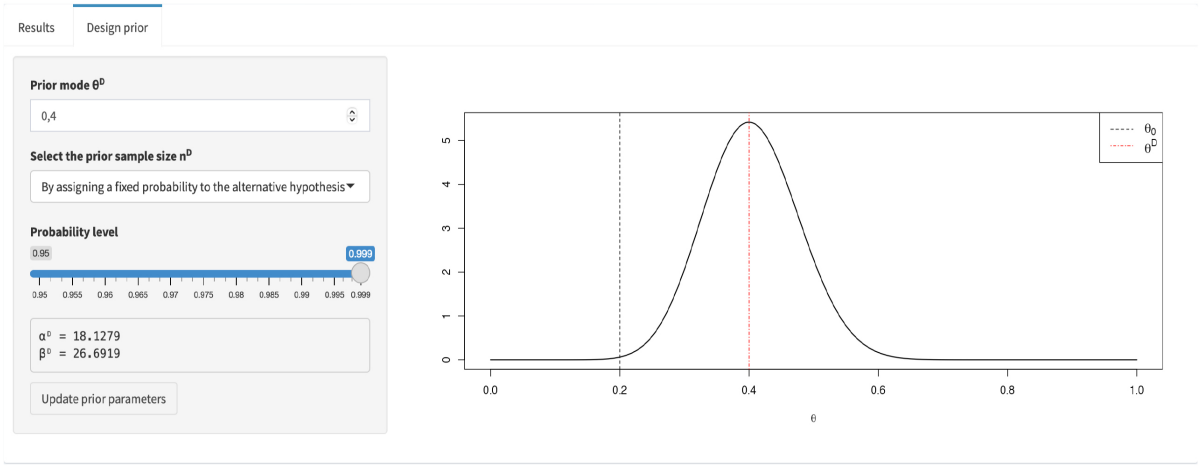


Figure 1. User interface when the aim is to compute n_F^C, when $\theta_0 = 0.2$, $\gamma = 0.8$, $\alpha = 0.05$ and $\theta^D = 0.4$. We consider both the conservative and the standard criterion.

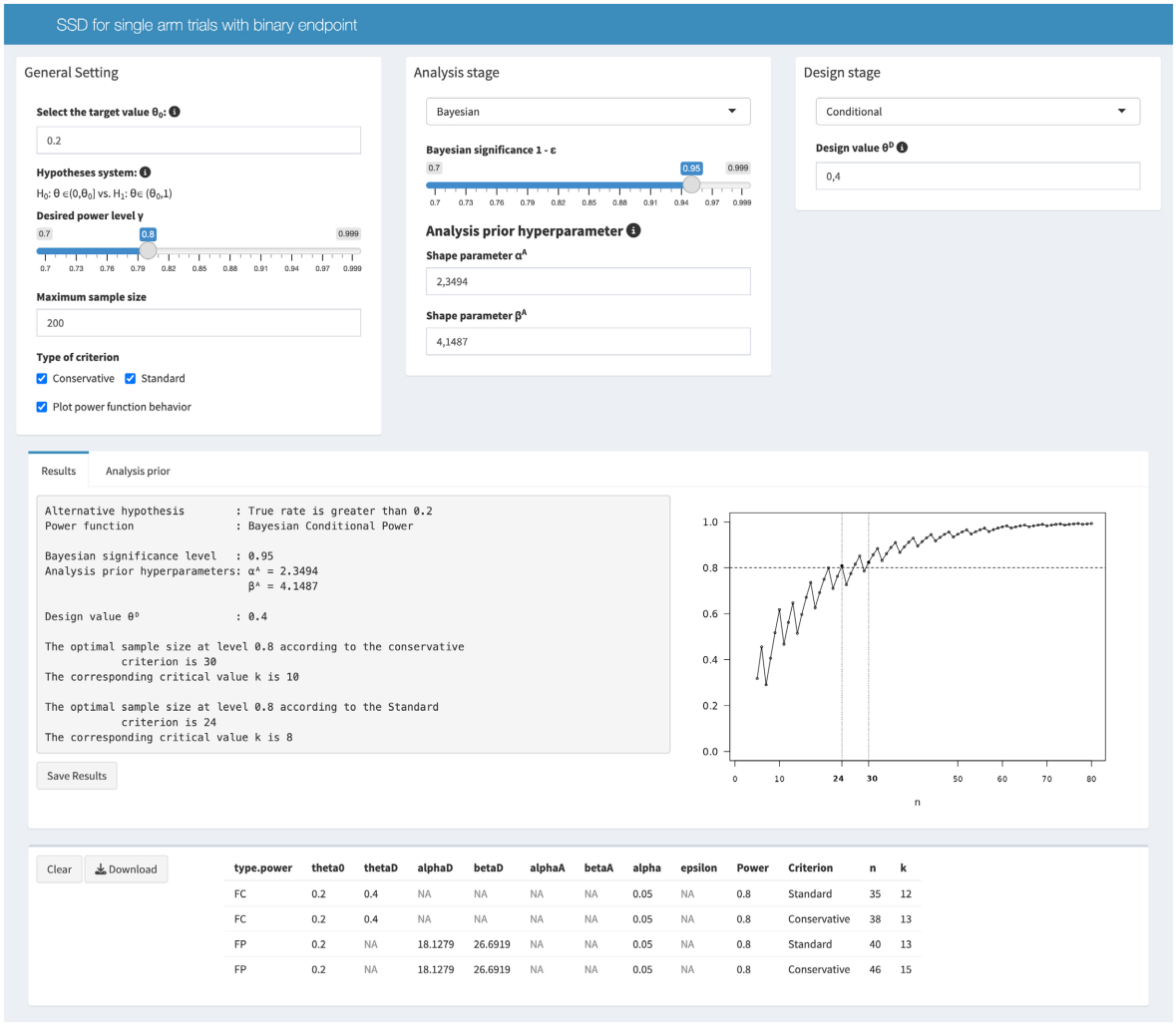


(a)

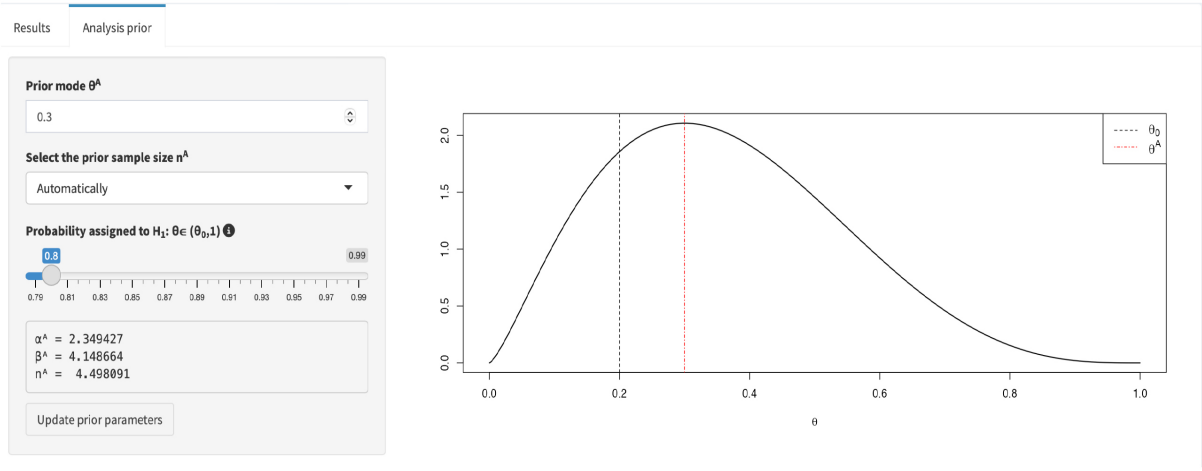


(b)

Figure 2. User interface and design priors' selection panel when the aim is to compute n_F^P, when $\theta_0 = 0.2$, $\gamma = 0.8$, $\alpha = 0.05$ and $\theta^D = 1.5$. n^D is selected so that $\mathbb{P}_{\pi^D(\cdot)}(\theta < \theta_0) = 0.999$. (a) User interface and results for the frequentist predictive power; (b) Design prior selection panel.



(a)



(b)

Figure 3. User interface and analysis priors' selection panel when the aim is to compute n_B^C, when $\theta_0 = 0.2$, $\theta^D = 0.4$, $\gamma = 0.8$ and $\varepsilon = 0.05$. For the analysis prior, we set $\theta^A = 0.3$ and n^A is selected so that $\mathbb{P}_{\pi^A(\cdot)}(\theta < \theta_0) = 0.8$. (a) User interface and results for the Bayesian conditional power; (b) Analysis prior selection panel.

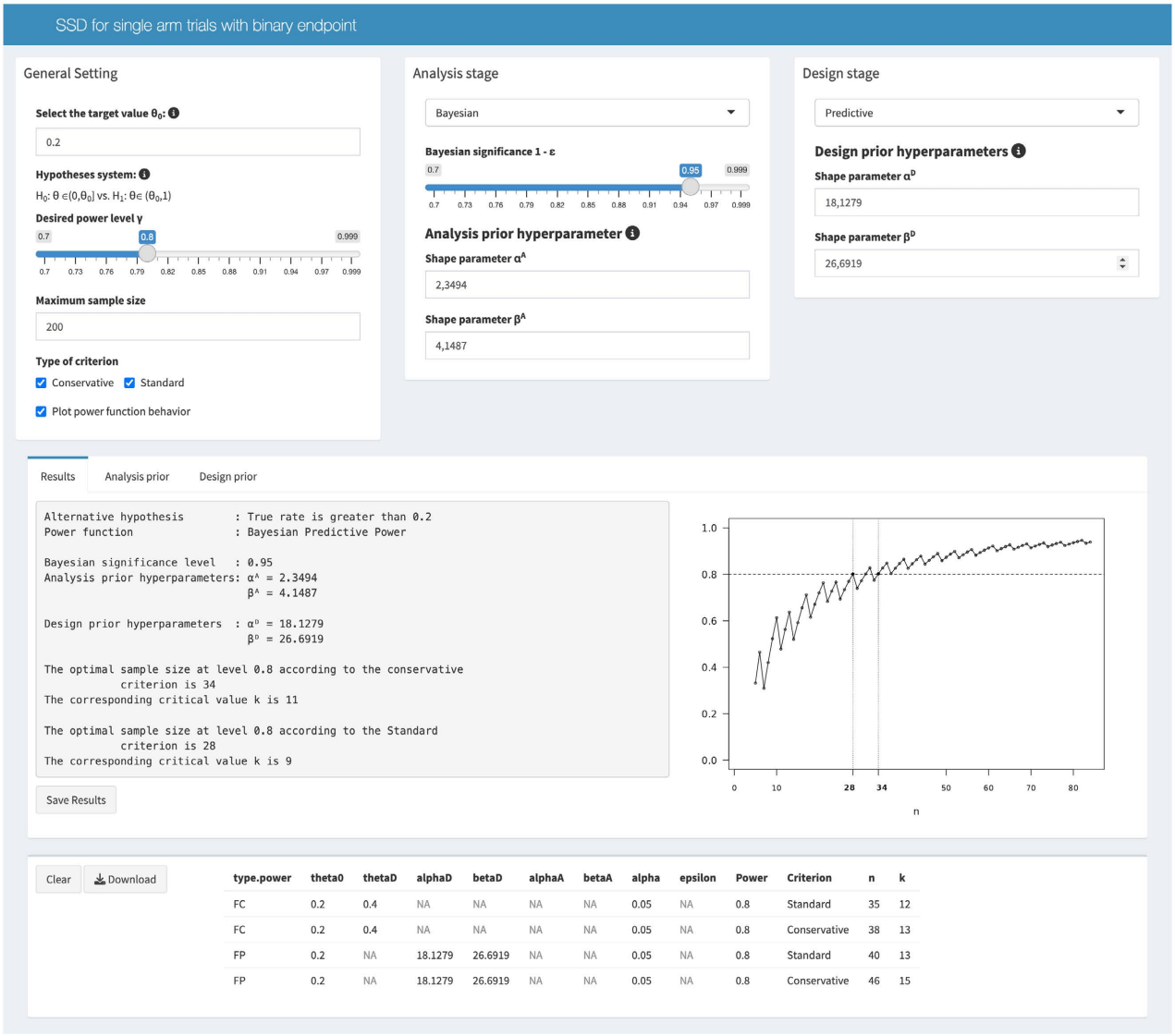


Figure 4. User interface when the aim is to compute n_B^P, when $\theta_0 = 0.2$, $\gamma = 0.8$ and $\varepsilon = 0.05$. The design and the analysis priors are the ones selected previously.

Design stage: The inputs depend on the approach used to realize the optimistic assumption that the experimental treatment is effective. The app requires the design value θ^D for the conditional approach (Figure 1) and the hyperparameters of the design prior distribution for the predictive approach. We recommend using the "Design prior" panel to select the design prior (Figure 2, Panel (b)).

Once all the required inputs have been provided, the "Results" Panel prints on the left a summary of the design parameters, the optimal sample size, and the critical value. On the right, if requested, the power as a function of n is displayed. Users can save the results in a table by clicking "Save results". The table can then be downloaded as a CSV file (Figure 1). The info icons provide some suggestions on the choice of the parameters.

Note that if the user changes the target value parameter θ_0, he will have to insert the other design parameters again. This mechanism prevents the app from crashing if the old input parameters are inconsistent with the new target value and the corresponding hypothesis system. Once the user provides θ_0, the app will check if all the inputs satisfy the boundaries before computing the results and inform the user otherwise.

The maximum sample size input allows the user to specify the maximum sample size available. If the optimal sample size exceeds the maximum, the app prints a message and the power corresponding to the maximum sample size in the "Result" panel. By default, the app still computes the optimal sample size according to the selected criterion, as shown in the plot on the right. However, suppose the optimal sample size is greater than 1000. In that case, the app stops, and the user can decide whether to increase the maximum sample size by modifying the input parameter, as suggested by a printed message. This boundary also prevents the app from crashing if the optimal sample size at the desired level does not exist, which may happen if the prior distributions are not well specified.

4.2. Tools for Selecting the Prior Distributions

The "Design prior" panel appears if the user opts for the predictive approach at the design stage (**Figure 2**, Panel (b)). The panel requires the specification of the prior mode θ^D. Then, the **"Select the prior sample size n^D"** drop-down list allows to select n^D according to three possible strategies:

By assigning a fixed probability to the alternative hypothesis:

the resulting $\pi^D(\theta)$ assigns the selected probability to the alternative hypothesis.

By assigning a fixed probability to an interval:

the resulting $\pi^D(\theta)$ assigns the selected probability to a symmetric interval $[\theta_0 - \delta, \theta_0 + \delta]$.

Manually: the user can select the prior sample size.

As emphasized in the previous section, the design prior should be highly informative and assign a negligible probability to the null hypothesis. The first two methods implement the strategies described in the previous Section and ensure this condition by selecting n^D numerically. However, users can also choose n^D at their discretion. In this case, we encourage users to verify if the probability assigned to the alternative hypothesis is greater than the desired power level γ. To help the user in the choice, this probability corresponding to the inserted n^D is printed under the hyperparameters. Regardless of the selected method, the design distribution is displayed on the right. Finally, the "Update prior parameters" button allows for updating the corresponding inputs in the user interface with the hyperparameters of the chosen distribution.

The "Analysis prior" panel appears when the user decides to conduct the analysis stage under a Bayesian framework (see **Figure 3**, Panel (b)). Firstly, the user needs to specify the prior mode θ^A. Then, the drop-down list **"Select the**

prior sample size n^A" allows selecting the prior sample size n^A "Manually", *i.e.*, at the user's discretion, or "Automatically". In the latter case, we select n^A by fixing the probability of the alternative hypothesis $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0)$. The range of possible probabilities is determined numerically by ensuring that:

1) The hyperparameters α^A and β^A are both greater than 1, so that $\pi^A(\theta)$ admits a mode in θ^A;

2) The prior sample size n^A is less than 100.

As outputs, the panel returns the hyperparameters and a graphical representation of the analysis prior. If the user opts for selecting n^A manually, the probability assigned to the alternative hypothesis is also shown for a check. If the user opts for the automatic selection, the app prints the prior sample size n^A instead. Finally, α^A and β^A can be stored in the corresponding UI values by clicking on "Update prior parameters".

4.3. Illustrative Example

We now illustrate an example of the app utilization. When inserting the inputs, we suggest starting from the target value θ_0 so that the app can automatically check if the other parameters are well selected.

Let us start by considering the frequentist conditional power; the corresponding UI is shown in **Figure 1**. We assume that the aim is to test the null hypothesis that the actual response rate is less than or equal to $\theta_0 = 0.2$ at level $\alpha = 0.05$. We set $\theta^D = 0.4$ as we consider clinically relevant an increase of 0.2. The desired power level is $\gamma = 0.8$. We set the maximum sample size to 200 and consider both the standard and the conservative criterion. The two optimal sample sizes are respectively 35 and 38, due to the saw-tooth behaviour.

Then, we switch to a predictive approach by selecting "Predictive" in the Design Stage window. This choice leads to the User Interface in **Figure 2**. We rely on the "Design prior" panel to select the design prior hyperparameters. More specifically, we consider the same design value $\theta^D = 0.4$ and select n^D using the "By assigning a fixed probability to the alternative hypothesis" method. Since we require that the design prior assigns the $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0) = 0.999$, the resulting beta density $\pi^D(\theta) = \text{beta}(\theta; \alpha^D = 18.13, \beta^D = 26.69)$. The optimal sample size is 40 for the standard criterion and 46 for the conservative one. As expected, n_F^P is greater than n_F^C because we are accounting for the uncertainty around the design value θ^D.

Let us suppose now that there is an optimistic prior opinion toward the treatment efficacy, and the most plausible value for the parameter, according to experts, is $\theta = 0.3$. We switch to a Bayesian analysis framework to incorporate this information. **Figure 3** and **Figure 4** show the Shiny App screenshots for the Conditional and Predictive approaches. We set $\varepsilon = 0.05$ to ensure comparability with the previous results and use the "Analysis Prior" panel to select α^A and β^A. More specifically, we set $\theta^A = 0.3$, while n^A is selected so that $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0) = 0.8$. The resulting analysis prior is

$\pi^A(\theta) = \text{Beta}(\alpha^A = 2.35, \beta^A = 4.15)$, corresponding to a prior sample size $n^A = 4.50$. Considering the conditional approach, as in Figure 3, the optimal sample size according to the conservative criterion is $n_B^C = 30$. Specifically, $n_B^C < n_F^C$ because the selected analysis prior distribution expresses a modest enthusiasm towards treatment efficacy. Similarly, if we adopt a predictive approach to account for the uncertainty around the design value, as in Figure 4, the optimal sample size is $n_B^P = 34$. As expected, the latter is greater than n_B^C due to the predictive approach but smaller than the corresponding frequentist sample size n_F^P.

5. Conclusions

The methodologies based on the two-priors approach allow to exploit four different power functions to determine the optimal sample size. We revise these procedures when the focus is on single-arm studies based on a single binomial proportion. Although there are several R packages and software tools to implement the classical procedures for SSD, based on the frequentist conditional approach, easy-to-use computational tools to implement criteria based on the other three power functions are not yet available. To fill this gap, we developed an interactive and used-friendly Shiny application, whose main functionalities are presented in this paper.

In addition to allowing the calculation of the optimal sample size, the Shiny app allows us to check the behaviour of the four power functions as the sample size varies and let the users choose between the standard and the conservative criterion for SSD, which takes into account the saw-tooth behaviour of the power functions. Moreover, since the distinction between analysis and design priors is an essential element of the implemented procedures, the app provides two separate panels to help the selection of both the prior distributions: different strategies to elicit these priors can be used and the app allows us to visualize the corresponding plot.

Finally, let us notice that in this paper we refer to single-arm designs in phase II of clinical trials. These designs are frequently used to determine whether a new treatment is likely to meet a basic level of efficacy, before comparing it with the standard therapy in larger and randomized phase III trials, and the efficacy is commonly measured as a response rate. However, the SSD procedures implemented in the Shiny App we present can be used to size experiments conducted in fields other than the clinical one, as long as based on a single binomial proportion.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sambucini, V. (2017) Bayesian vs Frequentist Power Functions to Determine the

Optimal Sample Size: Testing One Sample Binomial Proportion Using Exact Methods. In: Tejedor, J.P., Ed., *Bayesian Inference*, IntechOpen, Rijeka, 77-97. https://doi.org/10.5772/intechopen.70168

[2] Wang, F. and Gelfand, A.E. (2002) A Simulation-Based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science*, **17**, 193-208.

[3] Sahu, S.K. and Smith, T.M.F. (2006) A Bayesian Method of Sample Size Determination with Practical Applications. *Journal of the Royal Statistical Society*: *Series A*, **169**, 235-253. https://doi.org/10.1111/j.1467-985X.2006.00408.x

[4] De Santis, F. (2006) Sample Size Determination for Robust Bayesian Analysis. *Journal of the American Statistical Association*, **101**, 278-291. https://doi.org/10.1198/016214505000000510

[5] Brutti, P., De Santis, F. and Gubbiotti, S. (2008) Robust Bayesian Sample Size Determination in Clinical Trials. *Statistics in Medicine*, **27**, 2290-2306. https://doi.org/10.1002/sim.3175

[6] Gubbiotti, S. and De Santis, F. (2008) Classical and Bayesian Power Functions: Their Use in Clinical Trials. *Biomedical Statistics and Clinical Epidemiology*, **2**, 201-211.

[7] Sambucini, V. (2008) A Bayesian Predictive Two-Stage Design for Phase II Clinical Trials. *Statistics in Medicine*, **27**, 1199-1224. https://doi.org/10.1002/sim.3021

[8] Matano, F. and Sambucini, V. (2016) Accounting for Uncertainty in the Historical Response Rate of the Standard Treatment in Single-Arm Two-Stage Designs Based on Bayesian Power Functions. *Pharmaceutical Statistics*, **15**, 517-530. https://doi.org/10.1002/pst.1788

[9] Berchialla, P., Zohar, S. and Baldi, I. (2019) Bayesian Sample Size Determination for Phase IIA Clinical Trials Using Historical Data and Semi-Parametric Prior's Elicitation. *Pharmaceutical Statistics*, **18**, 198-211. https://doi.org/10.1002/pst.1914

[10] Turchetta, A., Moodie, E.E.M., Stephens, D.A. and Lambert, S.D. (2023) Bayesian Sample Size Calculations for Comparing Two Strategies in SMART Studies. *Biometrics*, **3**, 2489-2502. https://doi.org/10.1111/biom.13813

[11] Chernick, M.R. and Liu, C.Y. (2002) The Saw-Toothed Behavior of Power versus Sample Size and Software Solutions: Single Binomial Proportion Using Exact Methods. *The American Statistician*, **56**, 149-155. https://doi.org/10.1198/000313002317572835

[12] Gentile, S. and Sambucini, V. (2023) Exact Sample Size Determination for a Single Poisson Random Sample. *Biometrical Journal*, **25**, Article ID: 2200183. https://doi.org/10.1002/bimj.202200183

[13] Champely, S. (2020) PWR: Basic Functions for Power Analysis. R Package Version 1.3-0. https://CRAN.R-project.org/package=pwr

[14] Aberson, C. (2021) pwr2ppl: Power Analyses for Common Designs (Power to the People). R package version 0.5.0. https://cran.r-project.org/web/packages/pwr2ppl/pwr2ppl.pdf

[15] Ryan, T.P. (2013) Sample Size Determination and Power. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781118439241

[16] Kohl, M. (2020) MKpower: Power Analysis and Sample Size Calculation. R Package Version 0.7. https://cran.rproject.org/web/packages/MKpower/MKpower.pdf

[17] Ekstrm, C. (2022) MESS: Miscellaneous Esoteric Statistical Scripts. R package Version 0.5.9. https://CRAN.R-project.org/package=MESS

[18] Millard, S.P. (2013) EnvStats: An R Package for Environmental Statistics. Springer Science & Business Media, Berlin. https://doi.org/10.1007/978-1-4614-8456-1

[19] Faul, F., Erdfelder, E., Lang, A. and Buchner, A. (2007) G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, **39**, 175-191. https://doi.org/10.3758/BF03193146

[20] Lenth, R. V. (2018) Java Applets for Power and Sample Size. http://homepage.divms.uiowa.edu/~rlenth/Power/

[21] Chang, W., Cheng, J., Allaire, J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. and Borges, B. (2021) Shiny: Web Application Framework for R. R package Version 1.7.1. https://CRAN.R-project.org/package=shiny

[22] Wickham, H. (2021) Mastering Shiny. O'Reilly Media, Inc., Sebastopol.

[23] Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P. (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley & Sons, Hoboken. https://doi.org/10.1002/0470092602

[24] Lesaffre, E. and Lawson, A.B. (2012) Bayesian Biostatistics. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119942412

[25] Eaton, M.L., Muirhead, R.J. and Soaita, A.I. (2013) On the Limiting Behavior of the Probability of Claiming Superiority in a Bayesian Context. *Bayesian Analysis*, **8**, 221-232. https://doi.org/10.1214/13-BA809