Received: 25 June 2022 | Revised: 30 January 2023 | Accepted: 7 April 2023

DOI: 10.1002/bimj.202200183

Biometrical Journal

RESEARCH ARTICLE

Exact sample size determination for a single Poisson random sample

Susanna Gentile | **Valeria Sambucini**

Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy

Correspondence
Valeria Sambucini, Department of Statistical Sciences, Sapienza University of Rome, P.le Aldo Moro 5, 00185, Rome, Italy.
Email: valeria.sambucini@uniroma1.it

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Classical power analysis for sample size determination is typically performed in clinical trials. A "hybrid" classical Bayesian or a "fully Bayesian" approach can be alternatively used in order to add flexibility to the design assumptions needed at the planning stage of the study and to explicitly incorporate prior information in the procedure. In this paper, we exploit and compare these approaches to obtain the optimal sample size of a single-arm trial based on Poisson data. We adopt exact methods to establish the rejection of the null hypothesis within a frequentist or a Bayesian perspective and suggest the use of a conservative criterion for sample size determination that accounts for the not strictly monotonic behavior of the power function in the presence of discrete data. A Shiny web app in R has been developed to provide a user-friendly interface to easily compute the optimal sample size according to the proposed criteria and to assure the reproducibility of the results.

KEYWORDS

analysis and design prior distributions, exact sample size determination, fully Bayesian approach, hybrid classical Bayesian approach, Poisson data

1 | INTRODUCTION

The sample size determination (SSD) is a crucial element of the planning phase of any clinical trial, and it is usually conducted through a pre-experimental power analysis. In general terms, the purpose is to select the smallest sample size that allows to achieve a desired level for the probability of correctly rejecting the null hypothesis, H_0.

The traditional procedure consists in selecting a fixed *design value* for the parameter of interest under the alternative hypothesis and in using it to evaluate the conditional probability of rejecting H_0 computed under a frequentist framework. The prespecified design value plays a key role and should be carefully selected as a clinically relevant effect size that we wish to detect with high probability. An obvious drawback of the method is that the determination of the sample size may be strongly affected by the choice of this point alternative. To account for the uncertainty about the design effect size, a so-called "hybrid" classical Bayesian approach can be adopted (see Spiegelhalter et al., 2004) by averaging the traditional power function using a prior distribution for the unknown parameter. In the literature, this approach has been implemented to obtain an unconditional probability of the success of the trial (see, for instance, Carroll, 2013;

Chuang-Stein, 2006; O'Hagan et al., 2005; D. G. Chen & Ho, 2017). In line with other works (see Ciarleglio & Arendt, 2019; Lan & Wittes, 2012; Sambucini, 2017), we suggest to use a prior distribution for θ, named the *design prior distribution*, that formally introduces uncertainty of the guessed design value under the condition that it belongs to the alternative hypothesis. In this way, the average of the classical power more closely resembles the concept of frequentist power, since it is computed under the assumption that the alternative hypothesis is true and, therefore, can be interpreted as a probability of making the correct decision.

Alternatively, it is possible to adopt a "fully Bayesian" approach that exploits only Bayesian concepts both at the analysis and at the design stage of the trial. Since the early 2000s, there has been growing interest within the pharmaceutical industry in the use of Bayesian procedures at various stages of the research. The U.S. Food and Drug Administration (FDA) *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (*2010*)* recognizes the merits of Bayesian methods and provides detailed recommendations for their use in FDA-regulated clinical trials. The recent status of Bayesian statistics in terms of regulatory acceptance by the FDA is addressed in Campbell (2020). A prior distribution, called the *analysis prior distribution*, is introduced to formalize pre-experimental information available on the unknown parameter and is used to compute the posterior distribution. The null hypothesis is rejected if the posterior probability assigned to the alternative hypothesis is sufficiently high. Then, the sample size is determined by ensuring a large level for the probability of rejecting H_0, under the assumption that the alternative is true. Analogously to the procedures based on a frequentist final analysis, this optimistic assumption can be realized either by setting a fixed design value suitably selected under the alternative hypothesis or by eliciting a design prior distribution to account for uncertainty in the guessed parameter value. Wang and Gelfand (2002) provided an exhaustive formulation of this approach for SSD, known as the *two-priors approach*, that was subsequently adopted by several authors (see, among others, De Santis, 2006; Matano & Sambucini, 2016; Psioda & Ibrahim, 2019; Sahu & Smith, 2006). Applications of both the hybrid classical Bayesian and the fully Bayesian approaches when the focus is on the mean of a normal distribution or a single binomial proportion are provided by Gubbiotti and De Santis (2008) and Sambucini (2017), respectively.

In this paper, we address the problem of SSD based on power analysis when the focus is on a single Poisson rate. In clinical trials, for instance, in certain chronic diseases, such as angina pectoris and epilepsy, symptoms manifest acutely in an episodic nature and it is assumed that the number of attacks for a given patient within a specified time interval follows a Poisson distribution (Layard & Arvesen, 1978). Examples of single-arm noncontrolled studies based on count data are provided by Stein et al. (2009) and Wasserman et al. (2011), where the number of acute serious bacterial infections is used as primary endpoint to test the effectiveness of immunoglobulin replacement therapies in patients with primary immunodeficiency disease. In this context, the commonly used SSD procedures are based on the application of classical power calculations with normal approximations. An exact approach to determine the minimum sample size for estimating a Poisson parameter, such that the prespecified levels of relative precision and confidence are guaranteed, has been provided in Z. Chen and Chen (2016). Bayesian approaches have been also used. For instance, Stamey et al. (2004) used the average coverage criterion for a single Poisson rate with underreported data. These authors exploited simulation-based methods based on a Bayesian symmetric interval estimator. Their approach has been extended by Stamey et al. (2006), who exploited the highest posterior density interval for the single Poisson rate and also implemented the average length criterion and a posterior variance criterion. Interval-based Bayesian criteria have been also proposed in Zaslavsky (2010), where prior data are used to express uncertainty about the Poisson event rate through an empirical procedure. Moreover, Zaslavsky (2012) determined the sample size for the estimation of a Poisson rate using the concept of conditional power and reformulating the Type I and II error probabilities in terms of confidence and credible limits. Furthermore, for phase II single-arm trials, Hand et al. (2016) extended the standard scheme to conduct a two-stage design to the case of a Poisson-distributed response variable and constructed a Bayesian predictive two-stage design.

We consider a single-arm trial where the primary endpoint is the number of events occurring in a given time period and follows a Poisson distribution. To the best of our knowledge, for this experimental situation, implementations of exact SSD methods based on power analysis exploiting the classical Bayesian hybrid and the fully Bayesian approaches has not been presented in the literature. We illustrate how to derive these criteria that allow to add flexibility to the classical procedure. Furthermore, to take into account the discrete nature of the data and the consequent sawtooth behavior of the power functions as the sample size varies, we suggest adopting a conservative criterion to have the conditions of interest satisfied for all sample sizes greater than the one selected (see Chernick & Liu, 2002; Sambucini, 2017). Finally, let us notice that we exploit exact distributions, instead of relying on asymptotic normality, and we perform the required computations by enumerating the probabilities of all the possible results. Thus, no simulation tools are involved to obtain numerical results.

The rest of the paper is organized as follows. In Section 2, we formalize the problem setting and describe the exact SSD criteria for single-arm trials based on Poisson data, when both "not desired" and "positive" events are considered. In

Section 3, we provide suggestions on the choice of the prior distributions involved either to formalize prior information or to represent design expectations. Numerical results are presented and discussed in Section 4, where a case study example is also presented. An interactive Shiny web application has been developed to facilitate the implementation of the proposed criteria, and in Section 5 we provide the link to access. Finally, a brief discussion is provided in Section 6.

2 | SSD METHODS FOR POISSON DATA

Let us consider a clinical trial aimed at evaluating the efficacy of a new experimental treatment through a single-arm study. The primary endpoint X is the number of occurrences of an event of interest over a prefixed interval of time. Thus, for each of the n patients enrolled in the study, we observe the random variable X_i that follows a Poisson distribution of rate $\theta > 0$. Moreover, without loss in generality, we assume that the event of interest has a negative connotation so that the focus is on testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$. The target event rate θ_0 is a fixed value that should ideally represent the true rate for the standard of care therapy and is typically selected by exploiting historical data.

The aim of this section is to derive exact criteria to determine the optimal value of n based on prestudy power analyses. The first approach considered is the classical one, which requires an initial guess about the true value of the unknown parameter. Then, in order to add flexibility to the procedure, we describe how to implement criteria based on a hybrid classical Bayesian approach and a fully Bayesian approach.

2.1 | Classical approach

We aim at using exact sample size procedures, and, therefore, we focus on the frequentist rejection region of H_0 based on the sufficient statistic $S_n = \sum_{i=1}^{n} X_i$, whose sampling distribution is

$$f(s_n|\theta) = \text{pois}(s_n; n\theta), \qquad \text{for } s_n \in 0, 1, 2, \dots,$$

where $\text{pois}(\cdot; \lambda)$ denotes the probability mass function of a Poisson variable of parameter λ.

The null hypothesis is rejected at the fixed significance level α, if $S_n \leq k_F$, where the critical value k_F is such that

$$k_F = \max \left\{ u \in \{0, 1, 2, \dots\} : \sum_{i=0}^{u} \text{pois}(i;\, n\theta_0) \leq \alpha \right\}.$$

In fact, due to the discreteness of the test statistic, it is not possible to hit the nominal level α exactly and the standard exact procedure ensures that the actual Type I error rate is smaller than or equal to the fixed α (see Ryan, 2013).

The traditional frequentist SSD criterion is based on the conditional probability of being able to make a correct decision in favor of the alternative hypothesis. A suitable design value of θ, θ^D, which represents a clinically relevant value important to detect, is suitably fixed under the alternative hypothesis. Then, the focus is on the so-called *conditional frequentist power*, defined as the probability of correctly rejecting H_0 when the actual θ is equal to θ^D,

$$\eta_F^C(n) = \mathbb{P}_{f(\cdot|\theta^D)}(S_n \leq k_F) = \sum_{i=0}^{k_F} \text{pois}(i;\, n\theta^D). \tag{1}$$

The typical criterion selects the optimal sample size as the smallest n such that $\eta_F^C(n) > \gamma$, where γ is a desired prefixed threshold. Since it is not possible to obtain a closed-form formula for the sample size, we need to use a numerical procedure that searches upwards from a minimum value of n until the condition is satisfied.

However, this way of proceeding does not take into account the not-monotonically increasing behavior of $\eta_F^C(n)$ as a function of n. The conditional frequentist power in (1), in fact, tends to increase showing a sawtooth behavior that typically occurs when dealing with discrete data. A discussion about this behavior in the presence of a binary endpoint is provided in Chernick and Liu (2002) and Sambucini (2017). To provide an example, in Figure 1, we plot $\eta_F^C(n)$ with respect to n when $\theta_0 = 2$, $\theta^D = 1.6$, and $\alpha = 0.05$. To analyze in detail the reasons for this sawtooth shape for count data, for each possible value of n between 5 and 85, in Table 1 we report the corresponding critical value k_F and the value of the conditional frequentist power. We can note that the oscillations depend on how k_F increases with the sample size. The critical value, in fact, depends on the distribution of S_n when the null hypothesis is true. Thus, when we increase the sample size by

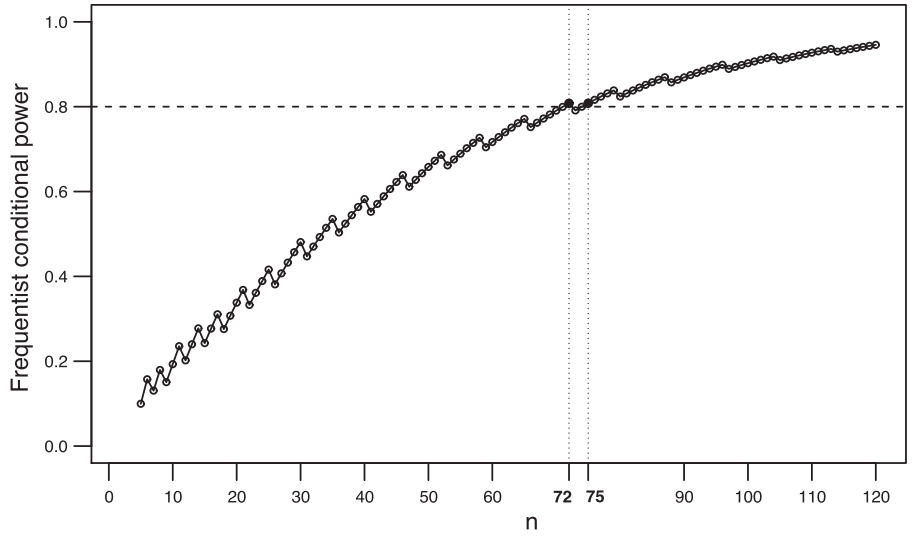

FIGURE 1 Behavior of $\eta_F^C(n)$ as a function of n, when $\theta_0 = 2$, $\theta^D = 1.6$, and $\alpha = 0.05$.

TABLE 1 Critical value and conditional frequentist power as n varies, when $\theta_0 = 2$, $\theta^D = 1.6$, and $\alpha = 0.05$.

n	k_F	$\eta_F^C(n)$	n	k_F	$\eta_F^C(n)$	n	k_F	$\eta_F^C(n)$	n	k_F	$\eta_F^C(n)$
5	4	0.0996	24	36	0.3890	43	70	0.5887	62	105	0.7398
6	6	0.1574	25	38	0.4160	44	72	0.6060	63	107	0.7507
7	7	0.1307	26	39	0.3812	45	74	0.6227	64	109	0.7612
8	9	0.1794	27	41	0.4072	46	76	0.6388	65	111	0.7713
9	10	0.1507	28	43	0.4325	47	77	0.6114	66	112	0.7519
10	12	0.1931	29	45	0.4570	48	79	0.6275	67	114	0.7621
11	14	0.2354	30	47	0.4808	49	81	0.6430	68	116	0.7720
12	15	0.2021	31	48	0.4471	50	83	0.6580	69	118	0.7816
13	17	0.2400	32	50	0.4702	51	85	0.6725	70	120	0.7907
14	19	0.2774	33	52	0.4927	52	87	0.6864	71	122	0.7995
15	20	0.2426	34	54	0.5144	53	88	0.6616	72	124	0.808
16	22	0.2770	35	56	0.5355	54	90	0.6756	73	125	0.7912
17	24	0.3107	36	57	0.5035	55	92	0.6891	74	127	0.7998
18	25	0.2758	37	59	0.5242	56	94	0.7021	75	129	0.8082
19	27	0.3072	38	61	0.5442	57	96	0.7147	76	131	0.8162
20	29	0.3380	39	63	0.5635	58	98	0.7267	77	133	0.8239
21	31	0.3681	40	65	0.5822	59	99	0.7045	78	135	0.8313
22	32	0.3327	41	66	0.5523	60	101	0.7167	79	137	0.8384
23	34	0.3612	42	68	0.5708	61	103	0.7284	80	138	0.8239

one, if k_F increases by 2 units, that is the expected number of events under H_0, then the power gets larger. Otherwise, the power decreases. We use alternate blocks of colors to underline when these fluctuations happen, and we can note that they are more frequent for smaller sample sizes.

In line with other works, in order to account for this fluctuating trend, we suggest to adopt a more conservative SSD criterion by selecting the smallest sample size for which $\eta_F^C(n)$ exceeds the desired level γ at all larger sample sizes, that is

$$n_F^C = \min\left\{n^* \in \mathbb{N} : \eta_F^C(n) \geq \gamma,\ \forall n \geq n^*\right\}.$$

Also in this case, it is necessary to proceed numerically by evaluating the condition of interest for decreasing values of the sample size, starting from a very high value, until reaching the optimal one. According to this conservative criterion, the

optimal sample size of the illustrative example is $n_F^C = 75$, while the lower value 72 allows to achieve the desired level of power without accounting for the fluctuating behavior.

2.2 | Hybrid classical Bayesian approach

The SSD criteria based on the classical approach suffer from local optimality because the selected sample size is strongly affected by the choice of the design value θ^D, whose uncertainty is not addressed. To overcome this limit, we suggest to use a hybrid classical and Bayesian procedure that allows to model uncertainty on the design value using a prior distribution according to the Bayesian approach, while still analyzing the data in a frequentist framework. Specifically, we elicit the so-called *design prior distribution*, $\pi^D(\theta)$, and then we obtain the *predictive frequentist power* by averaging the traditional frequentist power with respect to this prior. It can be proved that this is equivalent to compute the probability of rejecting H_0 with respect to the prior predictive distribution of the data (see, for instance, Spiegelhalter et al., 2004; Gubbiotti and De Santis, 2008; Sambucini, 2017).

In the case of Poisson data, it is useful to introduce a gamma design prior distribution,

$$\pi^D(\theta) = \text{gamma}\left(\theta; \alpha^D, \beta^D\right) = \frac{\left(\beta^D\right)^{\alpha^D}}{\Gamma(\alpha^D)}\, \theta^{\alpha^D - 1}\, \mathrm{e}^{-\beta^D \theta},$$

where $\text{gamma}(\cdot|a, b)$ denotes the probability density function of a Gamma distribution with shape parameter $a > 0$ and rate parameter $b > 0$. Then, the corresponding prior predictive distribution of S_n is

$$\begin{aligned} m^D(s_n) &= \int_0^\infty f(s_n|\theta)\pi^D(\theta)d\theta \\ &= \text{bin-neg}\left(s_n; \alpha^D, \frac{\beta^D}{\beta^D + n}\right) \qquad \text{for } s_n \in 0, 1, 2, \dots, \end{aligned} \tag{2}$$

where $\text{bin-neg}(\cdot; m, p)$ is the probability mass function of a Negative Binomial distribution with parameters m and p. Therefore, the predictive frequentist power is given by

$$\eta_F^P(n) = \mathbb{P}_{m^D(\cdot)}(S_n \le k_F) = \sum_{i=0}^{r} \text{bin-neg}\left(i; \alpha^D, \frac{\beta^D}{\beta^D + n}\right), \tag{3}$$

where $\mathbb{P}_{m^D(\cdot)}$ denotes the probability measure associated with prior predictive distribution of S_n. Since the marginal distribution $m^D(s_n)$ is discrete, also $\eta_F^P(n)$ presents a sawtooth behavior as a function of n. Therefore, we recommend to select the optimal n according to the conservative criterion described above, that is

$$n_F^P = \min\left\{n^* \in \mathbb{N} : \eta_F^P(n) \ge \gamma,\ \forall n \ge n^*\right\},$$

for a fixed threshold γ. Note that, when using this approach, it is essential to restrict the design prior distribution, which plays the key role of weight function, to assume that the alternative hypothesis is true. This ensures that the predictive frequentist power is conceptually analogous to the conditional one in that it provides a measure of the probability of correctly rejecting H_0.

2.3 | Bayesian approach

When planning a clinical trial, we may wish to take into account prior knowledge available and derived from various sources, such as historical data, pilot studies, expert opinions, etc. In this case, it is necessary to assume that a Bayesian analysis will be performed at the end of the trial. The prior information can be represented through a prior distribution, $\pi^A(\theta)$, that in this context is typically called *analysis prior distribution*. Differently from the design prior, it plays the role

of the usual prior distribution introduced to formalize pre-experimental knowledge in a Bayesian analysis and used to compute the posterior distribution.

In our specific case, we rely on conjugate analysis and assume that θ follows a Gamma prior distribution, $\pi^A(\theta) = \text{gamma}(\alpha^A, \beta^A)$. Given the observed result s_n, the posterior distribution is still a Gamma density with updated parameters, $\pi^A(\theta|s_n) = \text{gamma}(\alpha^A + s_n, \beta^A + n)$. In a pre-experimental setting and under a Bayesian framework, we need to establish a condition to reject the null hypothesis on the basis of the random result S_n. Typically, the null hypothesis is rejected if its posterior probability is less than or equal to a threshold ε, chosen as a small value. Equivalently, we can claim that we reject the null hypothesis if the posterior probability that θ belongs to the alternative one is sufficiently high, that is if

$$\begin{aligned}\mathbb{P}_{\pi^A(\cdot|S_n)}(\theta < \theta_0) &= \int_0^{\theta_0} \text{gamma}(\theta; \alpha^A + S_n, \beta^A + n)d\theta \\ &= F_{\text{Gamma}}(\theta_0; \alpha^A + S_n, \beta^A + n) > 1 - \varepsilon, \end{aligned} \tag{4}$$

where $\mathbb{P}_{\pi^A(\cdot|S_n)}$ is the probability measure associated with the posterior distribution of θ and $F_{\text{Gamma}}(\cdot; a, b)$ is the cumulative distribution function of a Gamma variable of parameters a and b. For a fixed value of n, the posterior probability in (4) increases as S_n decreases and we can find an integer k_B such that

$$F_{\text{Gamma}}(\theta_0; \alpha^A + k_B, \beta^A + n) > 1 - \varepsilon \qquad \text{and} \qquad F_{\text{Gamma}}(\theta_0; \alpha^A + k_B + 1, \beta^A + n) \leq 1 - \varepsilon.$$

Therefore, the Bayesian rule consists in rejecting H_0 if the experimental result is such that $S_n \leq k_B$, where

$$k_B = \max\left\{u \in \{0, 1, 2, ...\} : F_{\text{Gamma}}(\theta_0; \alpha^A + u, \beta^A + n) > 1 - \varepsilon\right\}.$$

In other words, k_B is the largest total number of events out of n patients that leads to reject H_0. Hence, when a Bayesian final analysis is planned, the power of the study can be obtained by computing the probability that $S_n \leq k_B$, under the assumption that the alternative hypothesis is true. It is possible to exploit a conditional approach by specifying a suitable design value θ^D for the parameter of interest under H_1. Then, we obtain the *conditional Bayesian power* as the probability of rejecting H_0 computed with respect to the sampling distribution of S_n conditional on θ^D,

$$\eta_B^C(n) = \mathbb{P}_{f(\cdot|\theta^D)}(S_n \leq k_B) = \sum_{i=0}^{k_B} \text{pois}\left(i; n\theta^D\right). \tag{5}$$

Alternatively, to avoid local optimality, we introduce a design prior distribution that models uncertainty on the guessed value θ^D. In this latter case, the prior predictive distribution of S_n provided in (2) is used to obtain the *predictive Bayesian power*

$$\eta_B^P(n) = \mathbb{P}_{m^D(\cdot)}(S_n \leq k_B) = \sum_{i=0}^{k_B} \text{bin-neg}\left(i; \alpha^D, \frac{\beta^D}{\beta^D + n}\right). \tag{6}$$

Since both the sampling and the marginal distributions of S_n are discrete, we suggest to select the optimal sample size according to the conservative criterion described above to account for the sawtooth behavior of the Bayesian power functions. Given the threshold of interest γ, the criteria are

$$n_B^C = \min\left\{n^* \in \mathbb{N} : \eta_B^C(n) \geq \gamma,\ \forall n \geq n^*\right\} \qquad \text{and} \qquad n_B^P = \min\left\{n^* \in \mathbb{N} : \eta_B^P(n) \geq \gamma,\ \forall n \geq n^*\right\}.$$

2.4 | Reversal of hypotheses

The majority of the clinical studies based on Poisson outcomes refer to the count of "negative" events, and examples include the number of falls for patients with Parkinson's disease, of asthma attacks in children, of relapses, of scan lesions in multiple sclerosis, or of seizures in epilepsy. Thus, without loss in generality, we have considered the count of an

event that represents a not-desired outcome for patients, focusing on the null hypothesis $H_0 : \theta \geq \theta_0$ to select the optimal sample size.

The proposed SSD criteria can be similarly derived when the hypotheses are reversed because a "positive" event is considered. In this case, the interest is on testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ and, when the final analysis is conducted under a frequentist framework, H_0 is rejected at level α if $S_n \geq \widetilde{k}_F$, where

$$\widetilde{k}_F = \min \left\{ u \in \{0, 1, 2, ...\} : \sum_{i=u}^{\infty} \text{pois}(i, n\theta_0) \leq \alpha \right\}.$$

Instead, if we plan to conduct a Bayesian final analysis, we introduce a Gamma analysis prior distribution for θ, $\pi^A(\theta) = \text{gamma}(\alpha^A, \beta^A)$, to represent prior knowledge, as previously described, and use it to obtain the posterior distribution of the parameter of interest. Then, H_0 is rejected if $S_n \geq \widetilde{k}_B$, where

$$\widetilde{k}_B = \min \left\{ u \in \{0, 1, 2, ...\} : 1 - F_{\text{Gamma}}(\theta_0; \alpha^A + u, \beta^A + n) > 1 - \varepsilon \right\},$$

where ε is a prespecified small threshold.

Hence, the conditional and predictive powers under the two inferential approaches are provided by

$$\eta_J^C(n) = \mathbb{P}_{f(\cdot|\theta^D)}(S_n \geq \widetilde{k}_J | n\theta^D) = 1 - \sum_{i=0}^{\widetilde{k}_J - 1} \text{pois}(i; n\theta^D)$$

and

$$\eta_J^P(n) = \mathbb{P}_{m^D(\cdot)}(S_n \geq \widetilde{k}_J) = 1 - \sum_{i=0}^{\widetilde{k}_J - 1} \text{bin-neg}\left(i; \alpha^D, \frac{\beta^D}{\beta^D + n} \right),$$

for $J = F, B$, where θ^D is the fixed design value larger than θ_0. Here, the design prior distribution, $\pi^D(\theta) = \text{gamma}(\theta; \alpha^D, \beta^D)$, addresses uncertainty on θ^D by assigning a negligible prior probability to values of θ smaller than θ_0. Given the power functions, by using the conservative criterion which accounts for their sawtooth behavior as n varies, the optimal sample sizes are selected as the ones ensuring that the power will not drop below the desired threshold for any larger sample.

3 | CHOICE OF THE PRIOR DISTRIBUTIONS

A fundamental element that characterizes and distinguishes the different approaches is the possible use of two different Gamma prior distributions, the analysis and the design priors, respectively, introduced to formalize prior information and to represent design expectations. To elicit these prior distributions, we resort to a way of proceeding often used to elicit beta prior densities when testing binary data. The idea is to express the hyperparameters of the priors in terms of (i) a *measure of central location* and (ii) a parameter that can be interpreted as the *prior sample size*.

With the purpose of illustrating the procedure, we focus on $\pi^D(\theta)$ whose hyperparameters can be fixed as

$$\alpha^D = n^D\theta^D + 1 \quad \text{and} \quad \beta^D = n^D,$$

in order to obtain a Gamma prior distribution with prior mode at θ^D and prior sample size n^D. This latter parameter reflects the dispersion of the distribution around its mode: the larger it is, the more concentrated the Gamma prior is. Hence, we can set the prior mode equal to the clinically relevant design value that we would select under the conditional approach and use n^D to express the desired degree of uncertainty around it. It is worth noting that, if $n^D \to \infty$, the prior variance $(n^D\theta^D + 1)/(n^D)^2$ goes to 0 so that the prior density corresponds to the degenerate distribution at θ^D. In this case, no uncertainty is introduced on the design value and, as a consequence, the conditional and predictive approaches lead

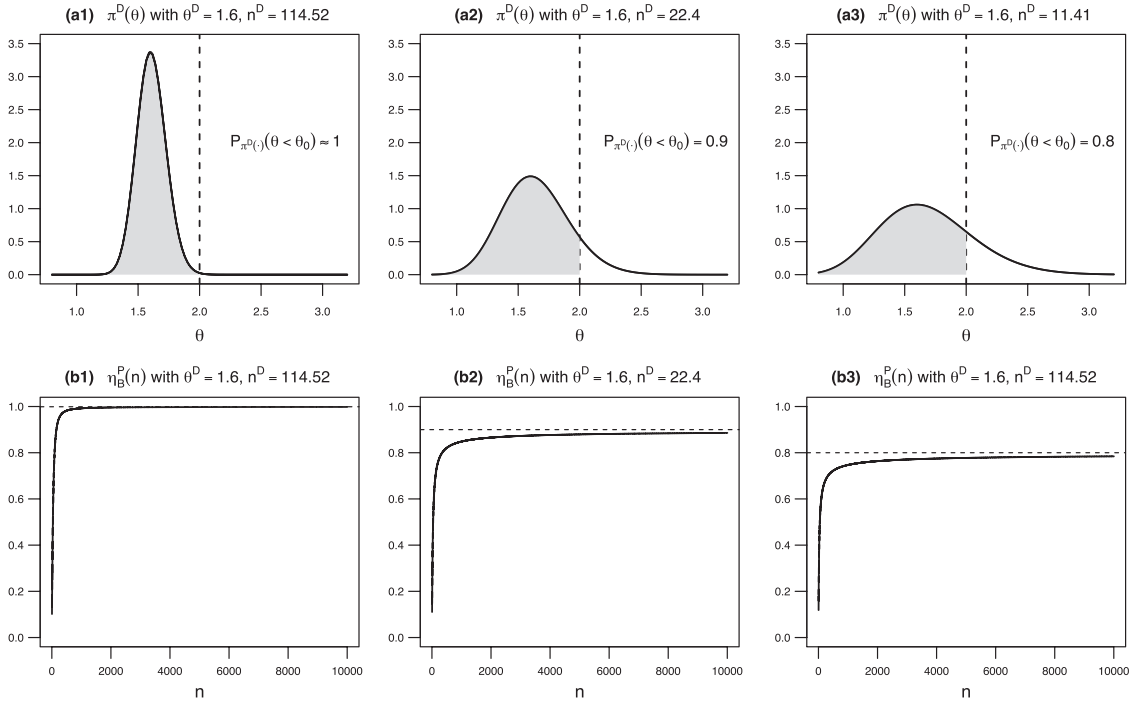

FIGURE 2 Design prior distributions that assign different probabilities to the alternative hypothesis (graphs (a1), (a2), and (a3)). Corresponding behavior of the predictive frequentist powers as a function of n, when $\alpha = 0.05$ (graphs (b1), (b2), and (b3)).

to the same optimal sample size, being $\eta_F^P(n) = \eta_F^C(n)$ and $\eta_B^P(n) = \eta_B^C(n)$. Of course, it is also possible to choose the prior mean as the measure of centrality of interest: in this case, the hyperparameters are fixed as $\alpha^D = n^D\theta^D$ and $\beta^D = n^D$.

We have already stressed that the design prior distribution should describe a design scenario that supports values of θ under H_1. In addition to the need to deal with a probability of making a correct decision, there is a further reason for this, related to computational aspects due to the limiting behavior of the predictive power functions. In fact, it is possible to prove that the limit of both $\eta_F^P(n)$ and $\eta_B^P(n)$ as $n \to \infty$ is given by $\mathbb{P}_{\pi^D(\cdot)}(\theta < \theta_0)$, which is the probability that the design prior assigns to the alternative hypothesis (see Eaton et al., 2013). To show this result empirically, we set $\theta_0 = 2$ and consider three different design prior distributions for θ with the same mode $\theta^D = 1.6$, but different dispersion. More specifically, the probability assigned to the alternative hypothesis by these prior densities is about equal to 1, 0.9, and 0.8, respectively, as it is possible to appreciate in Figure 2 (graphs (a1), (a2), and (a3)). In the same figure (graphs (b1), (b2), and (b3)), we show the corresponding behavior of predictive frequentist powers as a function of n when $\alpha = 0.05$. As expected, in all the cases considered, the limit that $\eta_F^P(n)$ approaches as n increases is equal to $\mathbb{P}_{\pi^D(\cdot)}(\theta < \theta_0)$. An analogous behavior can be shown for the Bayesian predictive power. Therefore, if the design prior does not assign a negligible probability to values of θ under H_0, the frequentist and Bayesian predictive powers do not approach 1 as the sample size goes to infinity, and we need to bound the threshold γ below the limiting value to obtain a finite sample size that satisfies the criterion. The usefulness of avoiding these situations has also been stressed by Lan and Wittes (2012) and Ciarleglio and Arendt (2019).

As regard the choice of the analysis prior distribution, its support and shape depend on the pretrial knowledge we want to incorporate in the final statistical analysis. Differently from the design prior, that must always be a proper distribution in order to have a proper predictive distribution of the data, $\pi^A(\theta)$ can be chosen as an improper noninformative prior. For instance, the choice $\alpha^A = 1$ and $\beta^A = 0$ leads to an improper uniform distribution or we can set $\alpha^A = 1/2$ and $\beta^A = 0$ to obtain the Jeffreys prior. Alternatively, we can employ standard procedures to formalize information from historical data to elicit expert knowledge or to construct archetypal prior distributions to express prior skepticism or enthusiasm about the new treatment efficacy. In these latter cases, it can be useful to express α^A and β^A in terms of prior mode and prior sample size as described for the design prior density.

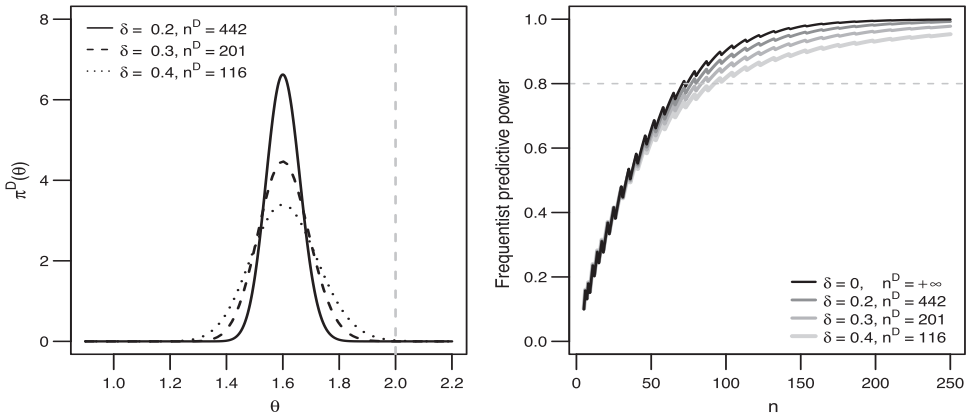

FIGURE 3 Left panel: Possible choices of the design prior distribution using method (i) for different values of δ, when $\theta_0 = 2$ and $\theta^D = 1.6$. Right panel: Behaviors of the corresponding $\eta_F^P(n)$ and of $\eta_F^C(n)$ (case $\delta = 0$) as a function of n, when $\alpha = 0.05$.

4 | NUMERICAL RESULTS

In this section, we illustrate the main features of the different SSD criteria based on the conservative selection method that takes into account the sawtooth behavior of the power functions. A case study example is also provided. All computations are performed using the R programming language.

4.1 | Behavior of the power functions and optimal sample sizes

First, we focus on the hybrid classical Bayesian approach and, therefore, on the power function in (3). As previously stressed, selecting a proper design prior is critical in determining the predictive frequentist power. We consider two possible strategies to ensure that $\pi^D(\theta)$ assigns a negligible probability to the null hypothesis by exploiting the hyperparameters selection procedure described in Section 3. Specifically, given the prior mode θ^D, we set n^D so that

(i) $\pi^D(\theta)$ assigns a 0.999 probability to a symmetric interval $(\theta^D - \delta, \theta^D + \delta)$, where δ is a nonnegative real number such that $\theta^D + \delta \leq \theta_0$;
(ii) $\pi^D(\theta)$ assigns a 0.999 probability to the alternative hypothesis.

In the left panel of Figure 3, we show three possible design priors obtained by using method (i) for different values of δ, when $\theta_0 = 2$ and $\theta^D = 1.6$. The behavior of the corresponding $\eta_F^P(n)$ as a function of n is represented in the right panel of the figure. The limiting case $\delta \to 0$, that is, $n^D \to +\infty$, which corresponds to a degenerate design prior to θ^D, is also considered. As previously remarked, in this case $\eta_F^P(n)$ is equivalent to the conditional frequentist power $\eta_C^F(n)$. Clearly, the uncertainty around θ^D increases with δ, resulting in a smaller n^D and a more dispersed prior distribution. Accordingly, the predictive frequentist power decreases, so that larger sample sizes are needed to achieve the desired level γ. For instance, when $\gamma = 0.8$, the resulting optimal sample sizes are 75, 77, 84, and 93, for δ equal to 0, 0.2, 0.3, and 0.4, respectively.

In Figure 4, we show three possible design prior distributions, and the corresponding behaviors of $\eta_F^P(n)$ as a function of n, obtained by selecting n^D with method (ii) for different θ^D, when $\theta_0 = 2$. In this case, for a fixed θ^D, n^D is selected as the smallest value such that $\mathbb{P}_{\pi^{D(\cdot)}}(\theta < \theta_0)$ is higher than or equal to 0.999. As a result, the larger the difference between the prior mode θ^D and θ_0, the lower both the n^D and the degree of concentration of the design prior distribution required to satisfy the condition regarding $\mathbb{P}_{\pi^{D(\cdot)}}(\theta < \theta_0)$. As θ^D approaches θ_0, the prior design scenario considered is less optimistic and the predictive frequentist power decreases, yielding to larger optimal sample sizes. More specifically, when $\gamma = 0.8$, the optimal values of n_F^P are 61, 93, and 168 for θ^D equal to 1.7, 1.6, and 1.5, respectively. Additional results about the optimal sample sizes according to the predictive frequentist criterion are provided in the Supporting Information to this article.

We now focus on the exact calculations of the sample size based on the Bayesian power functions in (5) and (6). In the right panel of Figure 5, we show the behavior of $\eta_B^C(n)$ as a function of n for different choices of $\pi^A(\theta)$, when $\theta_0 = 2$, $\theta^D = 1.6$, and $\varepsilon = 0.05$. More specifically, we consider three weakly informative analysis prior distributions that express

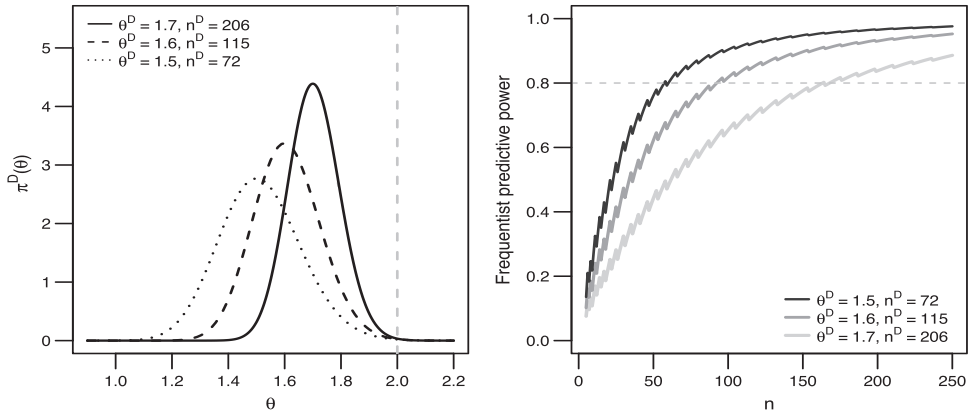

FIGURE 4 Left panel: Possible choices of the design prior distribution using method (ii) for different values of θ^D when $\theta_0 = 2$. Right panel: Behavior of the corresponding $\eta_F^P(n)$ as a function of n, when $\alpha = 0.05$.

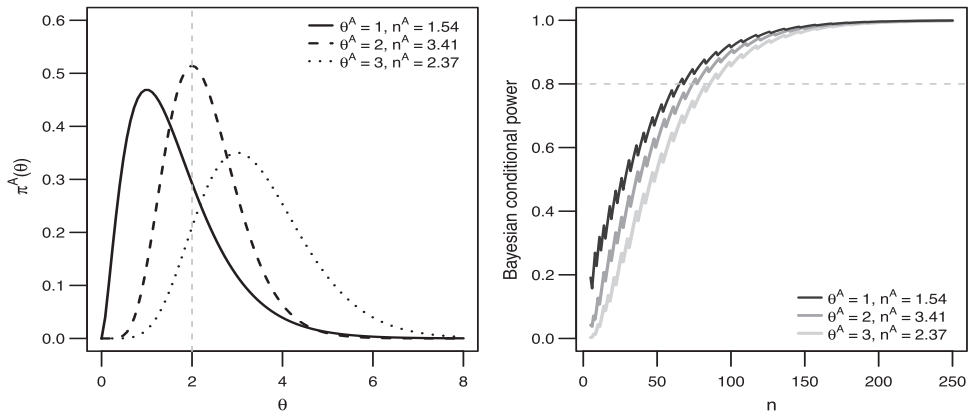

FIGURE 5 Left panel: Possible choices of the analysis prior distribution when $\theta_0 = 2$. Right panel: Behaviors of the corresponding $\eta_B^C(n)$ as a function of n, when $\varepsilon = 0.05$.

different prior beliefs towards the treatment efficacy and are shown in the left panel of Figure 5. They are obtained by fixing the prior mode θ^A and by determining the prior sample size n^A to ensure that the prior probability assigned to the alternative hypothesis, $\mathbb{P}_{\pi^A(\cdot)}(\theta < \theta_0)$, is equal to a prespecified level. The levels considered are 07, 0.4, and 0.1, for prior modes equal to 1, 2, and 3, respectively. Clearly, the greater θ^A, the stronger the prior skepticism expressed. As expected, $\eta_B^C(n)$ rises more slowly as the skepticism increases and the gap is more evident for small sample sizes because as n increases, the prior information becomes less influential. However, it still impacts the optimal sample sizes n_B^C at the level $\gamma = 0.8$, which are 69, 75, and 86, respectively.

To evaluate how the concentration of the analysis prior distribution around its prior mode affects the selection of the sample size, in Figure 6 we plot the behavior of n_B^C as a function of n^A for different values of θ^A, when $\theta_0 = 2$, $\theta^D = 1.6$, $\varepsilon = 0.05$, and $\gamma = 0.8$. In the left panel of the figure, we consider values of θ^A larger than θ_0 and therefore $\pi^A(\theta)$ expresses increasing skepticism about the treatment effect as n^A increases, whereas in the right panel the analysis prior distributions considered express more enthusiasm as n^A increases since the specified prior modes are smaller than θ_0. In both the panels, the dashed horizontal line corresponds to the optimal sample size obtained when $n^A \to 0$: whatever the value of θ^A is, this value coincides with n_C^F, that is the sample size selected under the conditional frequentist approach since the analysis prior introduces no information. As it is shown in the left panel of Figure 6, when $\theta^A > \theta_0$, n_B^C assumes values greater than n_C^F for $n^A > 0$ and grows monotonically with n^A, as a consequence of the higher degree of skepticism represented by $\pi^A(\theta)$. Moreover, as θ^A increases, n_B^C results to be uniformly larger for any value of n^A. The opposite situation occurs when $\theta^A < \theta_0$ as we can see from the right panel of Figure 6: n_B^C is below n_C^F for $n^A > 0$, decreases as n^A moves away from 0 and assumes smaller values if the prior mode θ^A decreases. More specifically, we can note that, when θ^A approaches 0, n_B^C drops rapidly and reaches 0. This happens when, given the high degree of optimism expressed by the analysis prior distribution, the conditional Bayesian power is uniformly above the desired threshold $\gamma = 0.8$ for

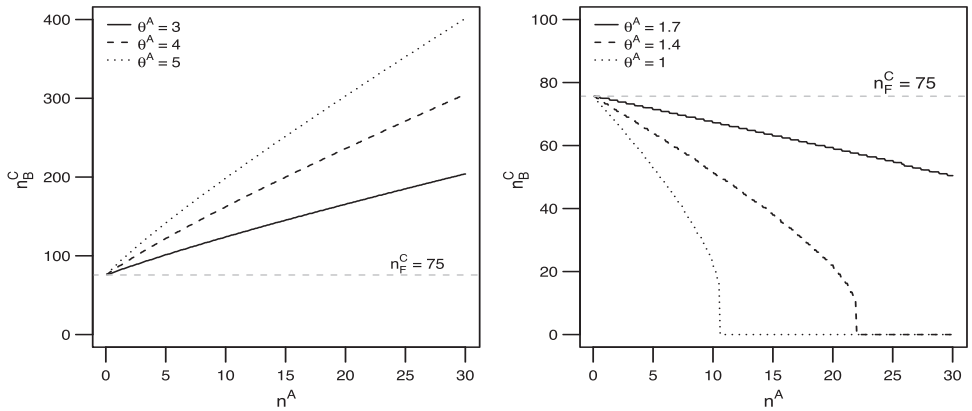

FIGURE 6 Behavior of n_B^C at level $\gamma = 0.8$ as a function of n^A when $\pi^A(\theta)$ expresses skepticism (left panel) or enthusiasm (right panel) towards the treatment efficacy.

any sample size. Additional computations about the predictive Bayesian approach are available in the online Supporting Information. These results confirm the tendencies previously observed and show that the Bayesian predictive power can be interpreted as a generalized power function which contains the other three as specific cases.

4.2 | A case study example

Primary immunodeficiency disorders are a group of heterogeneous conditions due to immune system malfunctions. As a result, patients affected by them may suffer from recurrent bacterial infections. A well-established replacement therapy, especially for primary humoral immunodeficiency, consists of regular administrations of polyclonal immune globulin preparations of human origin, such as the IGIV (i.e., immune globulin intravenous). In a related Guidance for Industry (2008), FDA recommends assessing the efficacy of investigational IGIV products in open-label, historically controlled single-arm trials. More specifically, the trial should provide a statistical demonstration that the annual rate of severe bacterial infections per patient (to avoid seasonal bias) is less than 1. The sample size should ensure at least 80% power with one-sided hypothesis testing based on a Type I error rate equal to 0.01. The FDA also underlines that the estimated rate for patients receiving routine IGIV transfusions is less than 0.5 per year against a historical rate of 4 or more infections per year before treatment. Examples of applications of this protocol may be found in Wasserman et al. (2011), Stein et al. (2009), and Hand et al. (2016), who assume that the number of infections is a Poisson distributed variable.

Following these guidelines, we can size the single-arm trial according to $\eta_F^C(n)$ by setting $\theta_0 = 1$, $\theta^D = 0.5$, $\alpha = 0.01$, and $\gamma = 0.8$. The corresponding optimal sample size is $n_F^C = 34$. Alternatively, since the design value is an estimate based on historical data, we may consider the hybrid classic-Bayesian approach to account for uncertainty on it. By exploiting method (ii), a proper design prior centered in $\theta^D = 0.5$ is $\pi^D(\theta) = \text{gamma}(\theta|17.99, 33.98)$ and the corresponding optimal sample size is $n_F^P = 46$. Moreover, we may consider the fully Bayesian approaches. Let us suppose that there is some skepticism towards the novel preparation efficacy. This attitude can be incorporated in the SSD procedure, for instance, by considering the analysis prior $\pi^A(\theta) = \text{gamma}(\theta|5, 1)$, which is weakly informative ($n^A = 1$) and centered on the historical rate of the nontreated $\theta^A = 4$. We set $\varepsilon = 0.01$ to ensure comparability with the frequentist results. Then, the corresponding optimal sample sizes at level $\gamma = 0.8$ are $n_B^C = 45$ if we consider the conditional approach and $n_B^P = 63$ for the predictive one.

5 | IMPLEMENTATION

To provide a user-friendly and interactive way to apply the methodologies described in the article, we have implemented an R Shiny App that is available at the link https://susanna-gentile.shinyapps.io/Poisson_SSD/.

The app allows to compute the optimal sample sizes according to the four power functions, when the focus is on the count of "negative" ($H_0 : \theta \geq \theta_0$) or "positive" ($H_0 : \theta \leq \theta_0$) events. Moreover, it is possible to use the conservative criterion, which takes into account the saw-toothed behavior of the power functions or the standard one. Furthermore,

the app provides some tools to help selecting the analysis and the design prior distributions properly. In the Supporting Information of the paper, the contents and the possible usages of the app are described in detail.

6 | DISCUSSION

In this paper, we address the problem of exact SSD for single-arm trials based on Poisson data. To overcome the drawbacks of procedures based on classical power analysis, we exploit analogous criteria based on "hybrid classical Bayesian" or "fully Bayesian" approaches. The first approach allows to model uncertainty on the design value of the parameter of interest through the introduction of a *design prior distribution*, which expresses the scenario under which the trial is planned. The second approach, instead, also allows to take into account pre-experimental knowledge about the Poisson rate by suitably specifying an *analysis prior distribution*. This conceptual distinction between the priors used to formalize uncertainty on guessed values of the parameter and to represent extra-experimental opinions or information is now becoming more and more popular in the statistical literature. It represents an essential element of the proposed criteria and some guidelines for choosing the prior distributions involved are discussed.

The SSD criteria derived are based on exact tests, which are preferred because normal approximations do not work well especially when the Poisson rate is small (see Ryan, 2013). As a consequence, dealing with discrete distributions of the data, we obtain that the relationship between the frequentist and Bayesian power functions and the sample size is not strictly monotonic. Thus, in line with other works (see, for instance, Chernick & Liu, 2002), we suggest to adopt a conservative criterion that takes into account the sawtooth behavior of the power functions. The idea is to avoid the paradox of selecting a sample size that meets the required criterion, but the same criterion is no longer fulfilled for a larger sample size. However, it is fair to point out that there is not a unanimous agreement on the appropriateness of such an approach (see, for instance, Ryan, 2013, p. 115). Therefore, to let researchers use the criterion they consider more reasonable, the R Shiny App developed to implement the proposed SSD methods allows to select the optimal sample size by using both the conservative and the standard approach.

ACKNOWLEDGMENTS

The authors would like to thank an associate editor and an anonymous referee for their helpful comments and constructive suggestions that contributed to improve the revised version of this paper.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable; no new data is generated

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Susanna Gentile https://orcid.org/0000-0002-1274-8177
Valeria Sambucini https://orcid.org/0000-0003-2101-480X

REFERENCES

Campbell, G. (2020). FDA Regulatory acceptance of Bayesian statistics. In E. Lesaffre, G. Baio, & B. Boulanger (Eds.), *Bayesian methods in pharmaceutical research* (pp. 41–51). Chapman & Hall.

Carroll, K. J. (2013). Decision making from phase II to phase III and the probability of success: reassured by assurance? *Journal of Biopharmaceutical Statistics*, *23*(5), 1188–1200.

Chen, D. G., & Ho, S. (2017). From statistical power to statistical assurance: It's time for a paradigm change in clinical trial design. *Communications in Statistics-Simulation and Computation*, *46*(10), 7957–7971.

Chen, Z., & Chen, X. (2016). Exact calculation of minimum sample size for estimating a Poisson parameter. *Communications in Statistics-Theory and Methods*, *45*(16), 4692–4715.

Chernick, M. R., & 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*(2), 149–155.

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics*, *5*(4), 305–309.

Ciarleglio, M. M., & Arendt, C. D. (2019). Sample size re-estimation in a superiority clinical trial using a hybrid classical and Bayesian procedure. *Statistical Methods in Medical Research*, *28*(6), 1852–1878.

De Santis, F. (2006). Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association*, *101*(473), 278–291.

Eaton, M. L., Muirhead, R. J., & Soaita, A. I. (2013). On the limiting behavior of the probability of claiming superiority in a Bayesian context. *Bayesian Analysis*, *8*(1), 221–232.

Food and Drug Administration. (2008). *Guidance for industry: Safety, efficacy, and pharmacokinetic studies to support marketing of immune globulin intravenous (human) as replacement therapy for primary humoral immunodeficiency*. https://www.fda.gov/media/124333/download

Food and Drug Administration. (2010). *Guidance for the use of Bayesian statistics in medical device clinical trials*. https://www.fda.gov/media/71512/download

Gubbiotti, S., & De Santis, F. (2008). Classical and Bayesian power functions: Their use in clinical trials. *Biomedical Statistics and Clinical Epidemiology*, *2*(3), 201–211.

Hand, A. L., Scott, J. A., Young, P. D., Stamey, J. D., & Young, D. M. (2016). Bayesian adaptive two-stage design for determining person-time in Phase II clinical trials with Poisson data. *Journal of Applied Statistics*, *43*(9), 1625–1635.

Lan, K. G., & Wittes, J. T. (2012). Some thoughts on sample size: A Bayesian-frequentist hybrid approach. *Clinical Trials*, *9*(5), 561–569.

Layard, M. W. J., & Arvesen, J. N. (1978). Analysis of Poisson data in crossover experimental designs. *Biometrics*, *34*(3), 421–428.

Matano, F., & 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*(6), 517–530.

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*(3), 187–201.

Psioda, M. A., & Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, *20*(3), 400–415.

Ryan, T. P. (2013). *Sample size determination and power*. Wiley.

Sahu, S. K., & Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A*, *169*(2), 235–253.

Sambucini, V. (2017). Bayesian vs frequentist power functions to determine the optimal sample size: Testing one sample binomial proportion using exact methods. In J. P. Tejedor (Ed.), *Bayesian inference* (pp. 77–97), IntechOpen.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Wiley.

Stamey, J. D., Seaman, J. W., Jr, & Young, D. M. (2004). Bayesian sample size determination for estimating a Poisson rate with underreported data. *Communications in Statistics-Simulation and Computation*, *33*(2), 341–354.

Stamey, J. D., Young, D. M., & Bratcher, T. L. (2006). Bayesian sample-size determination for one and two Poisson rate parameters with applications to quality control. *Journal of Applied Statistics*, *33*(6), 583–594.

Stein, M. R., Nelson, R. P., Church, J. A., Wasserman, R. L., Borte, M., Vermylen, C., & Bichler, J. (2009). Safety and efficacy of Privigen, a novel 10 liquid immunoglobulin preparation for intravenous use, in patients with primary immunodeficiencies. *Journal of Clinical Immunology*, *29*(1), 137–144.

Wang, F., & 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*(2), 193–208.

Wasserman, R. L., Melamed, I., Kobrynski, L., Strausbaugh, S. D., Stein, M. R., Sharkhawy, M., Engl, W., Leibl, H., Sobolevsky, L., Gelmont, D., Schiff, R. I., & Grossman, W. J. (2011). Efficacy, safety, and pharmacokinetics of a 10% liquid immune globulin preparation (GAMMAGARD LIQUID, 10%) administered subcutaneously in subjects with primary immunodeficiency disease. *Journal of Clinical Immunology*, *31*(3), 323–331.

Zaslavsky, B. G. (2010). Empirical Bayes models of Poisson clinical trials and sample size determination. *Pharmaceutical Statistics*, *9*(2), 133–141.

Zaslavsky, B. G. (2012). Bayesian sample size estimates for one sample test in clinical trials with dichotomous and countable outcomes. *Statistics in Biopharmaceutical Research*, *4*(1), 76–85.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gentile, S., & Sambucini, V. (2023). Exact sample size determination for a single Poisson random sample. *Biometrical Journal*, 2200183. https://doi.org/10.1002/bimj.202200183