Classical and Bayesian power functions: their use in clinical trials

Stefania Gubbiotti, Fulvio De Santis

Department of Statistics, Probability and Applied Statistics,
"Sapienza" University of Rome, Italy

Corresponding Author:
Stefania Gubbiotti, Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università degli Studi di Roma "Sapienza"
Piazzale A. Moro, 5 - 00185 Roma - Italy
E-mail: stefania.gubbiotti@uniroma1.it

Summary
The most widely used method for sample size determination in clinical trials is based on the power function. This function expresses the probability of rejecting a statistical null hypothesis on the quantity of interest, typically the unknown difference between the effects of two alternative treatments. The standard classical power function, which we will refer to hereafter as the *Conditional Frequentist Power function*, does not take into account the following: (a) uncertainty on the design value used for the unknown parameter to compute the power; (b) pre-experimental information on the difference of unknown effects, provided, for instance, by previous clinical studies. By taking into account (a) and (b), several extensions of the power function have been proposed: *the Predictive Frequentist Power function*, the *Conditional* and *Predictive Bayesian Power functions*. We review these methods, their relationships with the standard approach and implications on sample size determination.

KEY WORDS: *power, sample size determination, analysis prior, design prior.*

Introduction

Sample size determination (SSD) is a crucial problem in experimental design. From a technical point of view, the number of units to be enrolled in a trial is defined as *optimal* when it fulfils a pre-specified criterion that guarantees good quality inference. In general, the main purpose of a study is to observe the minimum number of individuals allowing inferential analysis to be conclusive. Nevertheless, in the specific context of clinical trials, we also have to deal with budget problems and, above all, ethical implications. In this paper, we review SSD methods based on power functions and in particular, we compare the features of both frequentist and Bayesian approaches.

Let us suppose that the objective of a study is inference on a parameter of interest, θ, representing for instance a treatment effect or a measure of comparison between alternative therapies. The traditional frequentist SSD criterion suggests choosing the minimum number n that guarantees a given power for performing tests on the mean θ. This method is widely used in applications, although it has two relevant drawbacks. First of all, the optimal sample size depends on a prefixed design value for the alternative hypothesis, yielding local optimality of the selected sample sizes. Secondly, when adopting a frequentist approach, we do not exploit pre-experimental information that could potentially reduce the required number of subjects. Therefore, a solution to these two problems is needed. On the one hand, it is possible to take into account the uncertainty around the design value. This leads to the introduction of a predictive method, which represents a more cautious option resulting in larger optimal sample sizes. On the other hand, pre-experimental information can be incorporated into the analysis by adopting the Bayesian approach. Information known about the unknown parameter θ can be formalized through a prior probability distribution. Besides contributing

to a reduction in the overall sample size, the use of initial information also allows for more flexibility, reflecting the actual knowledge on the phenomenon before performing the experiment. Moreover, the Bayesian approach involves two main advantages, as pointed out in (3). Assigning a prior distribution to the unknown quantities different plausible scenarios can be considered. Technically speaking, this avoids local optimality as we discuss in "Conditional frequentist power" section. In addition, this approach enables one to deal with additional unknown quantities that are not of direct scientific interest (*i.e.,* nuisance parameters), such as the parameters that measure sampling variability (see (3) for further details).

The increasing interest in Bayesian methods as applied to clinical trials has been recently attested to by the FDA *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2006)* (4). This document highlights the advantages of the Bayesian approach and provides the official guidelines for its correct use in clinical practice. It represents a turning point in clinical trial methodology, since it is the first time that the use of Bayesian statistics in clinical trials is not only allowed but even encouraged.

Finally, note that prior knowledge for a clinical trial can be derived from several sources of information, such as opinions of experts, historical data, previous trials, pilot studies, etc. A detailed discussion on the use and incorporation of prior information in planning and analyzing a clinical trial is beyond the scope of this paper; this aspect having already been discussed extensively in the literature (see for instance (1, 5-8)). For further details on Bayesian statistics in clinical trials, and in particular on Bayesian SSD, see among others (9-13).

The outline of the paper is as follows. Firstly, we recall the framework presented in (1) on the normal approximation for survival data and we introduce an example which is further discussed in the following sections. Then, we start from the frequentist approach introducing the classical conditional power. In the following section, we show how to deal with uncertainty on the design value adopting a predictive approach. Finally, we illustrate how prior knowledge can be exploited for SSD criteria, using the Bayesian power function.

Normal likelihood and normal approximation for survival data

Let us recall that we focus on a parameter of interest, θ. For the sake of simplicity in the following exposition, we focus on the normal model, namely we assume $Y_n \sim N(\theta, \frac{\sigma^2}{n})$, where Y_n is an estimator of the parameter of interest, θ; σ^2, assumed to be known; and n, the sample size yet to be determined. As argued in Spiegelhalter et al. (2004) (1) this framework can be adopted not only with normal data but also when a normal approximation applies. The authors show that the assumption that the data relevant to θ are summarised by a normally distributed statistic Y_n can be reasonable in many contexts, for instance with binary data, survival data and count data. However, in these cases n does not represent the total number of enrolled patients, but instead represents the so-called *effective sample size*, *i.e.* the number of events. In order to clarify this notion, let us recall some details on the normal approximation for survival data, which is used in the example illustrated throughout the paper (see again (1)). When dealing with survival data, a standard measure of comparison between two alternative treatments is the hazard ratio $HR = \frac{h_1(t)}{h_2(t)}$, where $h_i(t)$ is the hazard rate under treatment i (*i.e.* the probability of an event occurring in a short time interval, given that the event has not occurred yet at time t). Values of HR range from 0 to ∞ and, assuming proportional hazards, are constant over time. However, to make the normal likelihood assumption more plausible, we transform HR on the logarithmic scale and have $\theta = \log(HR) \in (-\infty, \infty)$ as a parameter of interest. Moreover, let us consider the observed log rank statistic L_n, which is defined as the excess of events under the innovative therapy compared to the number of events under the null hypothesis of no treatment effect. Assuming equal allocation and a balanced follow-up, under the null hypothesis the number of events in the treatment group is $n/2$ and therefore $L_n = O_T - n/2 = (O_T - O_C) / 2$ where O_j is the number of occurred events in arm, $j = T, C$, T denotes the treatment arm and C denotes the control arm. In this setup it has been shown by Tsiatis (2) that in large trials, θ can be estimated using the sta-

tistic $y_n = 4L_n / n = 2 (O_T - O_C) / n$, which is approximately normal with mean θ and variance $4 / n$. This leads us to adopt a normal likelihood with $\sigma = 2$, as in the application illustrated in the following sections. However, note that if an estimate of the variance of L_n is provided, one could proceed by equating it to $n / 4$ in order to obtain the effective number of events n. Obviously, non integer values for n may result, even though the approximate meaning of the effective sample size holds.

Example: SSD for a superiority trial

In order to illustrate the usage and the characteristics of the many power functions, we compare each function's advantages and disadvantages and we illustrate the different approaches for SSD, using as a guideline an example presented in (1).

A randomized controlled trial is designed for testing the effect difference of two competing cancer treatments. The outcome is mortality and the log hazard ratio of death is chosen as a measure to compare two randomized arms of patients. As described in the Introduction, a normal approximation is used for the log hazard ratio θ. We also assume that a positive value θ favours the new treatment, while a negative value supports the standard one. The null hypothesis we want to verify is $H_0 : \theta = 0$, corresponding to no benefit in the new treatment.

Conditional frequentist power

Let us refer to the setting as just described in the previous section. First of all, let us define the **conditional frequentist power** as the probability of rejecting H_0 given θ, defined as

$$\beta_F^C(\theta) = \Phi\left(\frac{\theta\sqrt{n}}{\sigma} + z_\alpha\right) \qquad [1]$$

where Φ is the cumulative distribution function of the standard normal random variable and z_α is the quantile of a standard normal distribution at significance level α; furthermore the superscript C and the subscript F stand respectively for *conditional* and *frequentist*. Notice that $\beta_F^C(\theta)$ is a function of the parameter θ, and that it also depends on the sample size n and z_α, with $\sigma = 2$ as discussed in the previous section.

Recall now that we want to determine the optimal sample size. Hence, our objective is to reach a given power, say 80%, at a pre-specified significance level, conventionally set at $\alpha = 0.05$. In [1] we need to fix a *design value* θ_D that can be interpreted as the target effect difference we aim to detect. In other words, we are assuming that the sampling distribution of future data Y_n is $f(y_n \mid \theta_D) = N(y_n \mid \theta_D, \sigma^2 / n)$. Therefore, we compute the frequentist power conditional to θ_D, an increasing function of n, and from there the optimal sample size is defined as the minimum number of units that guarantees a given power, *i.e.*,

$$n_F^C = \{\min n : \beta_F^C(\theta_D) > \eta\} \qquad [2]$$

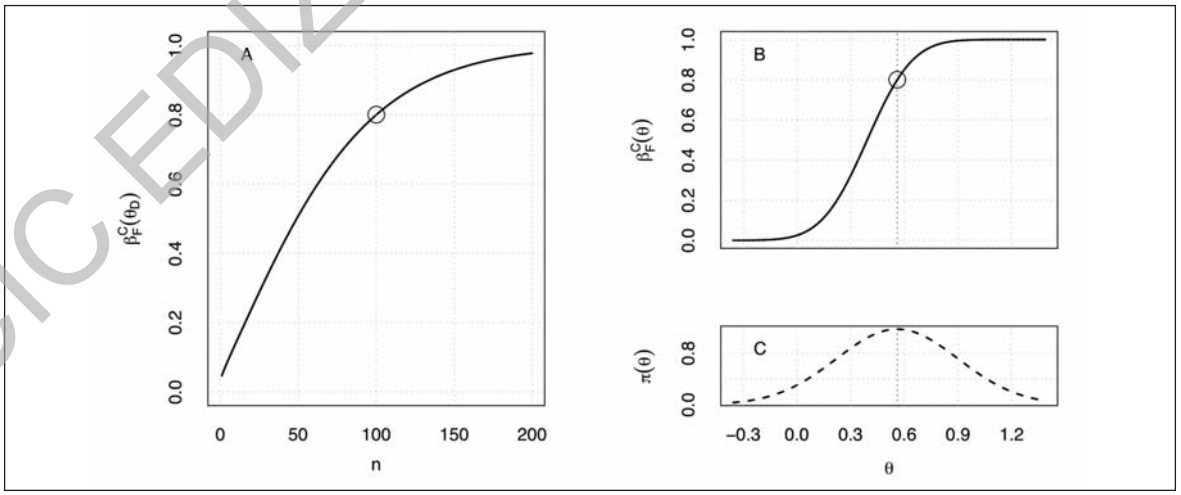


Figure 1. A. Conditional frequentist power $\beta_F^C(\theta_D)$ with respect to n, where the design value is $\theta_D = 0.56$. The optimal sample size is $n = n_F^C = 100$, corresponding to the required 80% power. B. Conditional frequentist power curve $\beta_F^C(\theta)$ with respect to the parameter θ, for a fixed sample size $n = 100$. C. Enthusiastic prior for $\theta : \pi(\theta)$ is a normal density of mean $\theta_D = 0.56$ and variance $\sigma^2 / n_0 = 4 / 34.5$.

where the threshold value η can be set for instance at 80%.

Following (1) in panel A of Figure 1, we set, as an example, $\theta_D = 0.56$ and we then obtain the corresponding optimal sample size, $n_F^C = 100$. In panel B, however, we highlight the dependence of the frequentist power on θ: for $n = 100$, the power corresponding to $\theta_D = 0.56$ is equal to 80% as designed, but $\beta_F^C(\cdot)$ is an increasing function of the design value. For example, if we set $\theta_D = 0.4$, the conditional frequentist power dramatically decreases to 0.52.

In Figure 2 we show how the choice of θ_D affects the optimal sample size. The actual values are reported in the corresponding table: for a slight change in the design parameter (for instance, $\theta_D = 0.5$ instead of the previous 0.56), the resulting sample size is remarkably larger (126 instead of 100). This is the problem of local optimality we referred to in the introductory section: since a small reduction of the design value implies a large decrease of the power level, n_F^C is the optimal solution of [2] for $\theta_D = 0.56$ and it can be considered "acceptable" only for values of θ "close" to θ_D.

In summary, the smaller the effect difference to be detected, the larger the optimal sample size, given the same power. This intuitive relationship between design value and power clearly shows how crucial the choice of θ_D is for sample size determination. It is natural, then, to consider a predictive approach that takes into account uncertainty on the design value, as we illustrate in the next section.

A second thing to note is that the frequentist approach commonly exploits prior information to specify an alternative hypothesis, but it does not formally incorporate pre-experimental knowledge from previous studies or expert opinions into the actual analysis. Conversely, according to a Bayesian perspective, it is actually possible to take into account pre-experimental information with a certain amount of uncertainty, which can be formalized by specifying a prior distribution for θ. For simplicity, we consider as prior distribution $\pi(\theta) = N(\theta \mid \theta_0, \sigma^2 / n_0)$, where n_0 is the so-called *prior sample size*. Adopting this formulation, introduced in (1), we are actually fixing σ both in the likelihood (see Introduction) and in the prior, but it should be remembered that the prior is based on an 'implicit' number of events. In other words, the information contained in the prior distribution is equivalent to that of an hypothetical previous study of (effective) sample size n_0. Also, note that when n_0 tends to 0, the variance increases and in the limit the distribution becomes 'flat', representing a non-informative density distribution (see again (1)).

In the above example, for instance, we can express an enthusiastic opinion about the benefit of the new treatment by eliciting a normal prior density centred on a positive value of θ, for example $\theta_D = 0.56$. Then, assuming a remote chance of negative values for θ, for instance a 5% prior probability that $\theta < 0$, we get $n_0 = 34.5$, with $\sigma = 2$. Hence, superimposing the prior $\pi(\theta)$ on the power curve provides a rough indication on the plausibility of the values of the parameter with respect to the corresponding power (see panel C of Figure 1). As already noted above, this procedure only provides an approximate idea; a more formal method is provided by the Bayesian approach, which allows one to incorporate the prior $\pi(\theta)$ into the power function and, consequently, in the SSD criterion. We deal with this problem in the last section.

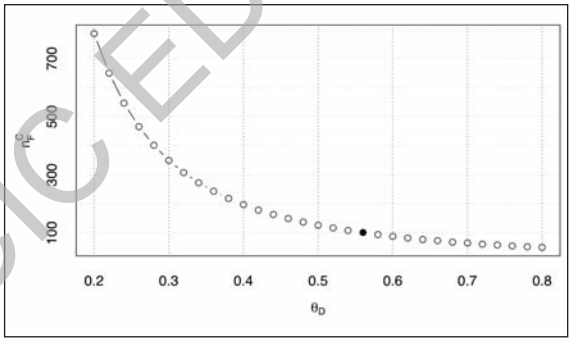


Figure 2. Optimal sample sizes n_F^C for several values of θ_D.

θ_D	0.1	0.2	0.3	0.4	0.5	**0.56**	0.6	0.7	0.8
n_F^C	3140	785	349	197	126	**100**	88	65	50

Predictive frequentist power

As shown in the previous section, conditional frequentist power is strongly related to the chosen design value θ_D, which influences sample size selection. In other words, by increasing (or decreasing) the effect difference to be detected, θ_D, we reach completely different indications on the optimal sam-

ple size for the trial. Hence, in order to avoid local optimality, instead of considering a single design value we take into account the uncertainty on this value in the power function. According to the Bayesian approach, we model uncertainty on θ by specifying a prior probability distribution. Specifically, we elicit the prior distribution π_D for θ – where the subscript D denotes the *design* prior distribution – and then we average the conditional frequentist power of [1] with respect to this prior:

$$\beta_F^P(\pi_D) = \int_\Theta \beta_F^C(\theta)\pi_D(\theta)d\theta, \qquad [3]$$

where the superscript P reminds that it is a *predictive* power function that is the unconditional probability of rejecting H_0. The notation also highlights that the predictive power depends on the prior π_D for θ. Again, we assume that π_D is a normal density of mean θ_D and variance denoted by σ^2 / n_D, where $\sigma = 2$. A technical remark: instead of using [3], $\beta_F^P(\pi_D)$ can be directly computed as the probability of rejecting the null hypothesis (or equivalently, of getting a significant result) with respect to the marginal distribution of the data, $m_{\pi_D}(y_n) = N(y_n \mid \theta_D, \sigma^2(\frac{1}{n_D} + \frac{1}{n}))$, that is, the average of the sampling distribution $f(\cdot;\theta)$, with respect to the prior π_D. Hence, we have the **predictive frequentist power**

$$\beta_F^P(\pi_D) = \Phi\left(\sqrt{\frac{n_D}{n_D + n}}\left(\frac{\theta_D\sqrt{n}}{\sigma} + z_\alpha\right)\right) \qquad [4]$$

and it is straightforward to define the following predictive SSD criterion:

$$n_F^P = \{\min n : \beta_F^P(\pi_D) > \eta\} \qquad [5]$$

where the threshold η is conventionally equal to 0.80. Note that in this case we want the conclusions of the study to be entirely classical, namely we do not intend to incorporate prior information in the final analysis. That is why this approach is also referred to as *hybrid classical-Bayesian* in (1).

Let us return to the example presented in the previous section. We choose the prior $\pi_D(\theta) = N(\theta \mid \theta_D = 0.56, \sigma^2 / n_D = 4 / 34.5)$ to model uncertainty on $\theta_D = 0.56$. For a sample size $n = n_F^C = 100$, while the conditional power $\beta_F^C(\theta_D)$ reaches the required level of 80%, as designed, the predictive power $\beta_F^P(\pi_D)$ declines to 0.66. In this case, in order to obtain the same power level we should increase the number of observations to $n_F^P = 240$. In general, we have $n_F^P > n_F^C$. From panel A of Figure 3, we notice that averaging with respect to the enthusiastic prior π_D slightly raises the power for small values of the sample size; while, as n increases, the predictive power becomes correspondingly lower than the conditional power.

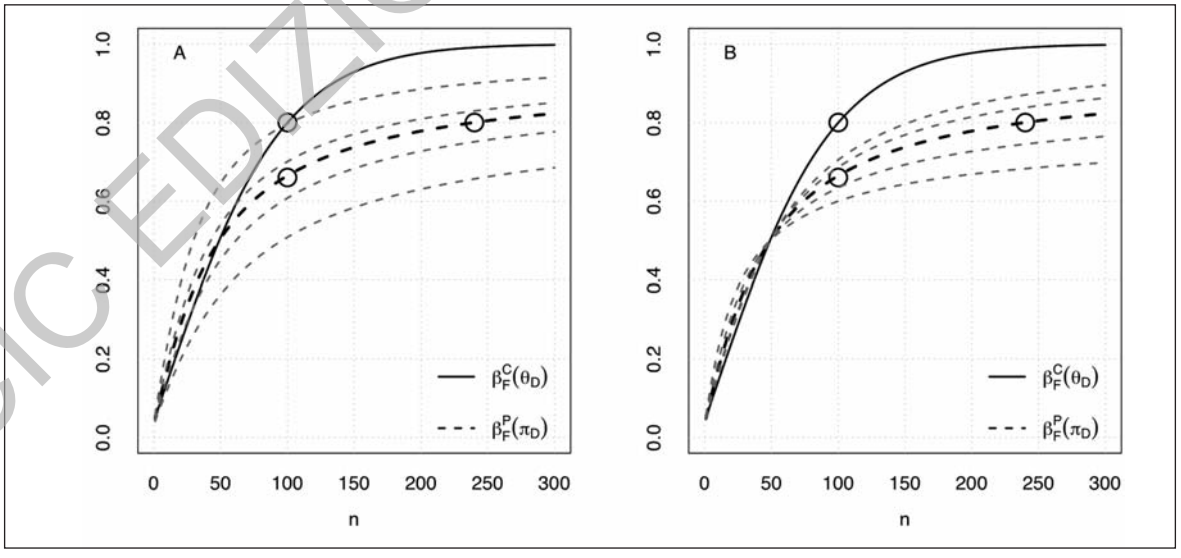


Figure 3. $\beta_F^C(\theta_D)$ (continuous line) $\beta_F^P(\pi_D)$ (dashed line) are plotted with respect to n, with design value $\theta_D = 0.56$ and design prior $\pi_D(\theta) = N(\theta \mid \theta_D = 0.56, \sigma^2 / n_D = 4 / 34.5)$ respectively. The dashed gray lines represent $\beta_F^P(\pi_D)$ for different choices of: A. the design prior mean $\theta_D = 0.4$, $\theta_D = 0.5$, $\theta_D = 0.6$ and $\theta_D = 0.72$, (from the bottom to the top); B. the prior sample size $n_D = 10$, $n_D = 120$, $n_D = 50$ and $n_D = 70$ (from the bottom to the top).

This is even more evident when considering the predictive power curves corresponding to larger prior means, the prior variance held equal. Specifically, notice that we need to increase the design prior mean to $\theta_D = 0.72$, in order to have $\beta_F^P(\pi_D) = 0.80$ in correspondence to $n = n_F^C = 100$. On the contrary, if we shift the design prior mean towards smaller values - expressing less optimistic opinions on the innovative therapy benefit - we obtain lower values for the power.

In panel B of Figure 3, we consider several alternative values for the prior variance, maintaing $\theta_D = 0.56$. As expected, for small values of n_D, the prior variance increases, with reduction of predictive power. In contrast, if we consider larger values of n_D, the prior π_D is more concentrated around its mode, raising the predictive power curve.

From the comparison of [1] and [4], it follows that $\beta_F^C(\theta)$ is a special case of $\beta_F^P(\pi_D)$: as $n_D \to \infty$ the prior π_D tends to concentrate on θ_D and $\beta_F^P(\pi_D)$ tends to $eta_F^P(\theta_D)$. However, as already discussed, for finite n_D we have $\beta_F^P(\pi_D) > \beta_F^C(\theta_D)$, provided that $\beta_F^P(\pi_D) > 0.50$. Finally, notice that if we allow $n_D \to 0$, which implies adopting a non informative flat design prior, from [4] we have $\beta_F^P(\pi_D) = 0.5$, regardless of the sample size. In other words, if we want the predictive SSD criterion in [5] to be conclusive, we need to specify a proper design prior.

Bayesian powers

Suppose now that in planning the experiment there exists available initial information regarding the difference in treatments, for example, the results of a previous trial or a pilot study are known. Suppose also that we want to perform a fully Bayesian analysis, in which prior information is incorporated. For instance, we elicit for θ the prior distribution $\pi_A(\theta) = N(\theta \mid \theta_A, \sigma^2 / n_A)$. Here, the subscript A stands for *analysis* because we assume that this prior is used in the inferential phase, in contrast with the design prior introduced at the beginning of the previous section. In general π_A does not necessarily coincide with π_D. These two priors are conceptually different: the former is used for designing the experiment and it represents the goal of the trial; the latter is the prior distribution actually used for inference and it is based on the available pre-experimental information. In the following subsection, we go into further details and we provide some references for this approach.

Two-priors approach

In general, most of the Bayesian SSD criteria use the same prior distribution for computing both the posterior and the predictive distributions (see, among others, (19) and (20)). However, several authors have argued that two distinct priors should be used, due to the conceptual difference between their roles: on one hand the *design prior* models uncertainty on unknown parameter and it is used to obtain the predictive distribution; on the other hand the *analysis prior* models pre-experimental information and it is used to obtain the posterior distribution. This distinction motivates what we refer to as *two-priors approach*.

The possibility of using different priors for design and estimation was first acknowledged in (21). In that paper, the author justified the apparent inconsistency of this innovative idea providing technical reasons: "using a design prior with variance much larger than believed reasonable is likely to lead to a wasteful experiment", while it is common practice to consider non-informative priors for final inference. After this pioneering paper, this concept has been refined by Etzioni and Kadane in "*Optimal experimental design for another's analysis*" (14). The motivating idea of the article is that the party performing the experiment and the party analysing experimental results are not necessarily the same. Sometimes, even if they have common goals, their priors may be different. This also responds to the point emphasized in (22): despite the experimenters' prior opinion, inference should convince people evaluating medical trials, who ultimately determine whether new treatments are adopted in clinical practice.

In the most recent literature, the use of two priors has been considered in a paper by Wang and Gelfand (16) who provide an exhaustive formulation of this approach that has constituted the paradigm for a set of following works, such as (8, 13, 15, 17, 23). The authors note that it is convenient to choose a relatively non–informative analysis prior – that they call *'fitting'*, since it is used to fit the model once the data

are obtained – because, in general, it is preferable to let the data drive inference. On the other hand, the design prior – *'sampling'* prior, in their terminology – represents the scenario we expect to observe, and in this sense it must be chosen to be informative. Moreover in this way one can play with different scenarios and compare the results: this is what the authors mean by the expression *'what if'* spirit.

In conclusion, we find convincing the idea of the two priors approach that also guarantees a substantive advantage in terms of flexibility and interpretability with respect to alternatives methods. As we argue in the present work, the two–priors approach also constitutes a general framework which includes as special cases both the hybrid classical-Bayesian (described in (1) and already mentioned in the previous section) and the classical approach: this allows an interpretation of the predictive Bayesian power function, which we are soon to introduce within the next few subsections, as a generalized power function.

Conditional Bayesian power

Let us return to the example previously introduced, and let us assume that the previous study provides an optimistic indication about the new treatment in terms of log-hazard ratio, which can be formalized by eliciting a prior distribution of mean $\theta_A = 0.56$ and prior sample size $n_A = 34.5$. According to the Bayesian approach, inference is based on the posterior distribution of θ, given the data Y_n. From standard Bayesian analysis it is well known that

$$\pi_A(\theta \mid Y_n) = N\left(\theta \,\middle|\, \frac{n_A\theta_A + nY_n}{n_A + n}, \frac{\sigma^2}{n_A + n}\right). \quad [6]$$

In order to derive the Bayesian power function, according to (1), we start providing the following definition: without loss of generality we say a Bayesian result is 'significant' if we have a low posterior chance, say $\alpha = 0.05$, that θ is negative. Basic calculations show that this happens if:

$$Y_n > \frac{-\sqrt{n_A + n}\, z_\alpha \sigma - n_A\theta_A}{n}, \quad [7]$$

as shown in (1) (see in particular Section 6.5).

Two alternatives for computing the probability of event [7] are available: we can either use the sampling distribution $f(\cdot;\theta_D)$ or the marginal distribution $m_{\pi_D}(\cdot)$. In the present section, we adopt the former, yielding the **conditional Bayesian power**; the latter will be used in next subsection to compute the **predictive Bayesian power**. Thus, in the first case we have

$$\beta_B^C(\theta_D) = \Phi\left(\frac{\theta_D\sqrt{n}}{\sigma} + \frac{\theta_A n_A}{\sigma\sqrt{n}} + \sqrt{\frac{n_A + n}{n}} z_\alpha\right) \quad [8]$$

which defines the conditional Bayesian power. The corresponding SSD criterion is

$$n_B^C = \{\min n : \beta_F^C(\theta_D) > \eta\}, \quad [9]$$

where the threshold η is equal to 0.80, with no loss of generality.

In Figure 4, we apply criterion [9] and we compare the conditional Bayesian power curve with the conditional frequentist one. The use of the enthusiastic prior π_A increases the power up to 0.93. This results in a smaller optimal sample size $n_B^C = 53$ with respect to $n_F^C = 100$ for the usual threshold $\eta = 0.80$. In panel A, we also compare the impact of different choices for the prior means on the optimal sample size, being the prior sample size fixed to $n_A = 34.5$. As expected, the more enthusiastic the prior mean the higher the power. On the contrary, a prior mean expressing scepticism towards the treatments difference (for instance, $\theta_A = 0.1$) leads to a Bayesian power $\beta_F^C(\theta)$ uniformly lower than $\beta_F^C(\theta_D)$, the conditional value being equal. In panel B, we proceed in the opposite way: we fix $\theta_A = 0.56$ and plot $\beta_F^C(\theta_D)$ for several value of the prior sample size n_A. Note that the conditional frequentist power is a particular case of $\beta_F^C(\theta_D)$ corresponding to $n_A = 0$, *i.e.*, to a flat non-informative prior. Then, considering increasing values of n_A we observe at each step a raise in the Bayesian power curve, reflective of the enthusiastic prior becoming more and more informative. Similar remarks can be drawn from panel C and panel D, where $\beta_B^C(\theta)$ is plotted with respect to θ, for fixed $n = 100$. Again, because this power function is conditional, we can highlight the problem of local optimality, as was already discussed in the second section with respect to $\beta_F^C(\cdot)$: the curve shape implies that small variations of have a strong impact on the corresponding power level.

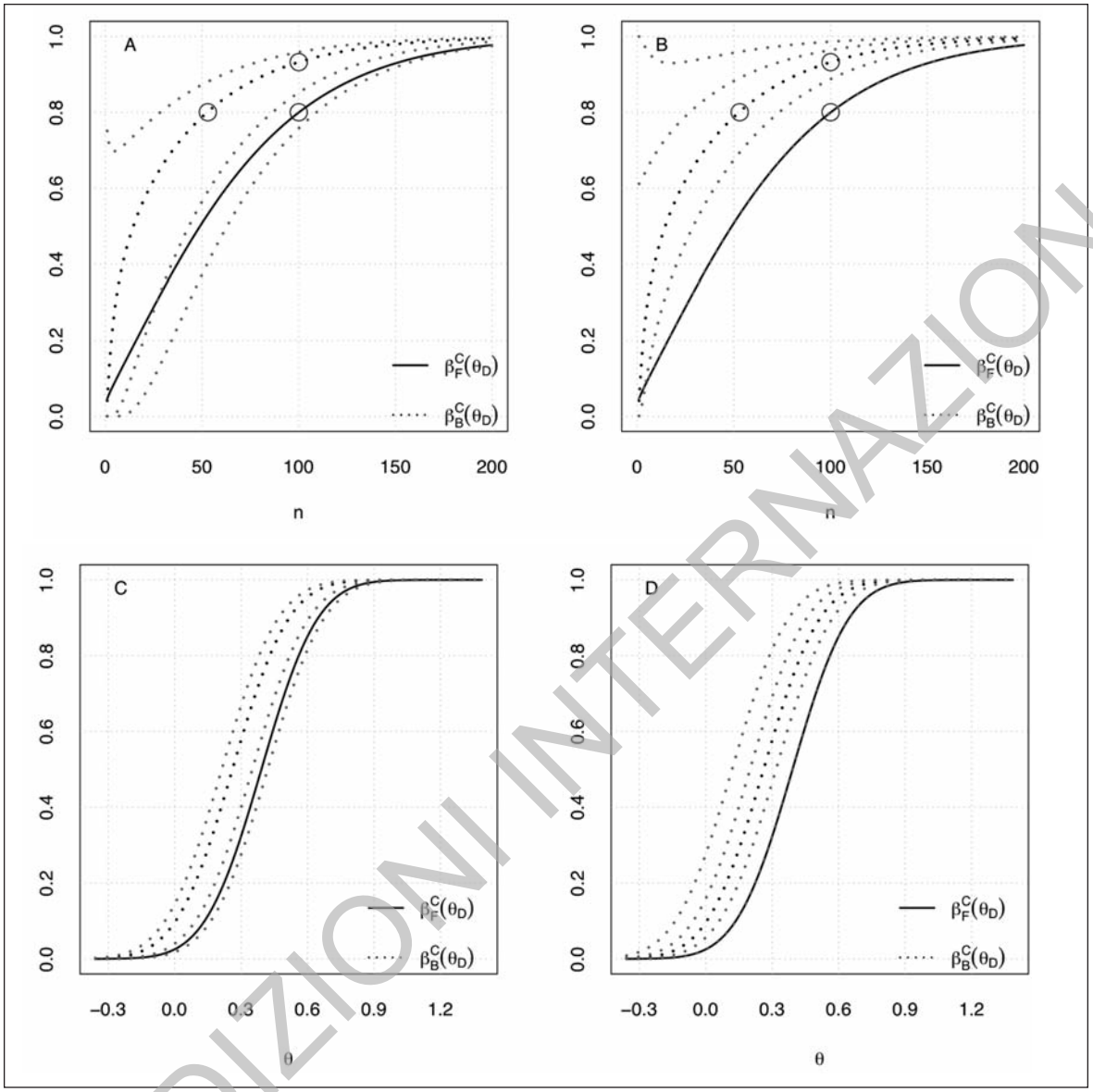


Figure 4. The conditional frequentist (continuous line) and Bayesian (dotted lines) power curves $\beta_F^C(\theta_D)$, with $\theta_D = 0.56$, are plotted A. with respect to n (with $\theta_D = 0.56$) and C. with respect to θ (with $n = 100$), for several values of the analysis prior means $\theta_A = 0.1$, $\theta_A = 0.3$, $\theta_A = 0.56$, $\theta_A = 0.7$ (dotted gray lines from right to left), with fixed prior sample size $n_A = 34.5$; B. with respect to n (with $\theta_D = 0.56$) and D. with respect to θ (with $n = 100$), for several values of the analysis prior sample size $n_A = 0$ (coinciding with $\beta_F^C(0_{\alpha_D})$), $n_A = 20$, $n_A = 34.5$, $n_A = 50$ and $n_A = 70$ (dotted gray lines from right to left), with given prior mean $\theta_A = 0.56$.

Predictive Bayesian power

Let us now consider the second alternative suggested in the previous section: we compute the probability of event [7] with respect to the marginal distribution. Hence, we have the predictive Bayesian power:

$$\beta_B^P(\pi_D) = \Phi\left(\frac{1}{\sigma\sqrt{\frac{1}{n_D}+\frac{1}{n}}}\left(\frac{\sqrt{n_A+n}z_\alpha\sigma + n_A\theta_A + n\theta_D}{n}\right)\right) \quad [10]$$

Note that the expression in [10] can be further simplified in case we assume $\pi_A = \pi_D$, that is

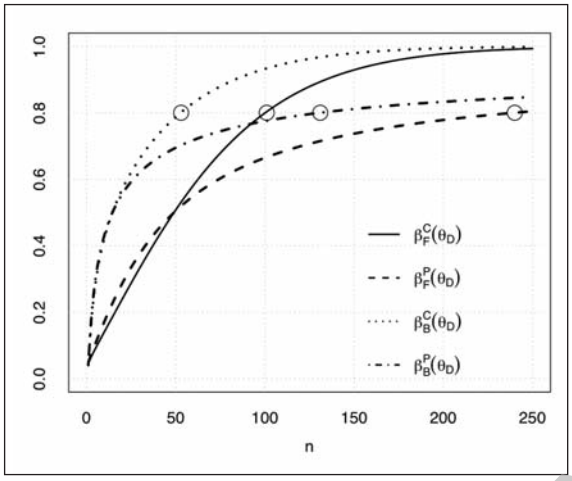
$$\beta_B^P(\pi_D) = \Phi\left(\frac{\theta_A\sqrt{n_A+n}\sqrt{n_A}}{\sigma\sqrt{n}} + \sqrt{\frac{n_A}{n}}z_\alpha\right) \quad [11]$$

where $\theta_A = \theta_D$ and $n_A = n_D$. In (1) the authors only deal with this case; nevertheless, we highlight again that the use of two distinct priors may be considered, as pointed out in the introductory section. Finally, giv-

en a suitable threshold value, the optimal sample size is the minimum number satisfying the following condition:

$$n_B^P = \{\min n : \beta_B^P(\pi_D) > \eta\}. \qquad [12]$$

Again, switching from a conditional power to a predictive one, we actually take into account the uncertainty on the design value and, therefore, we obtain a lower power, because we average the power function with respect to the design prior. Note that this is consistent with the considerations presented in the "Predictive frequentist power" section and it is represented in Figure 5 where the conditional Bayesian power (dotted curve) is compared with the predictive one (dashed-dotted curve).



n_F^C	n_F^P	n_B^C	n_B^P
100	240	53	131

Figure 5. $\beta_F^C(\theta_D)$ (continuous line), $\beta_P^F(\pi_D)$ (dashed line), $\beta_P^B(\theta_D)$ (dotted line) and $\beta_P^B(\pi_D)$ (dashed-dotted line) are plotted with respect to the sample size n. The conditional value is $n_D = 0.56$; the prior parameters are $\theta_D = \theta_A = 0.56$, $n_D = n_A = 34.5$. The resulting optimal sample sizes are reported in the table above.

Furthermore, in Figure 6, we plot the predictive Bayesian power $\beta_F^P(\pi_D)$ with respect to the sample size n. We consider several values for design prior parameters and we reach conclusions that are similar to the ones in "Predictive frequentist power" section: increasing the uncertainty on the design value (*i.e.*, decreasing the prior sample size n_D), the power curve rises (see panel A). The same happens if we choose increasingly larger design prior means.

It is interesting to remark that $\beta_P^B(\pi_D)$ can actually be considered a generalized power function that includes the other power functions as special cases (as summarized in Table 1). In $\beta_P^B(\pi_D)$, we model both the prior information and the uncertainty on the design value, which can be formalized using -- eventually different — prior distributions. Now, if n_D tends to be infinitely large ($n_D \to \infty$), the design prior tends to a point-mass *i.*, a distribution that assigns probability 1 to the single point θ_D, and we get $\beta_P^B(\theta_D)$ conditional to the design value θ_D. On the other hand, if we keep n_D finite and we allow n_A to go to 0 ($n_A \to 0$), the analysis prior degenerates into a flat, non-in-

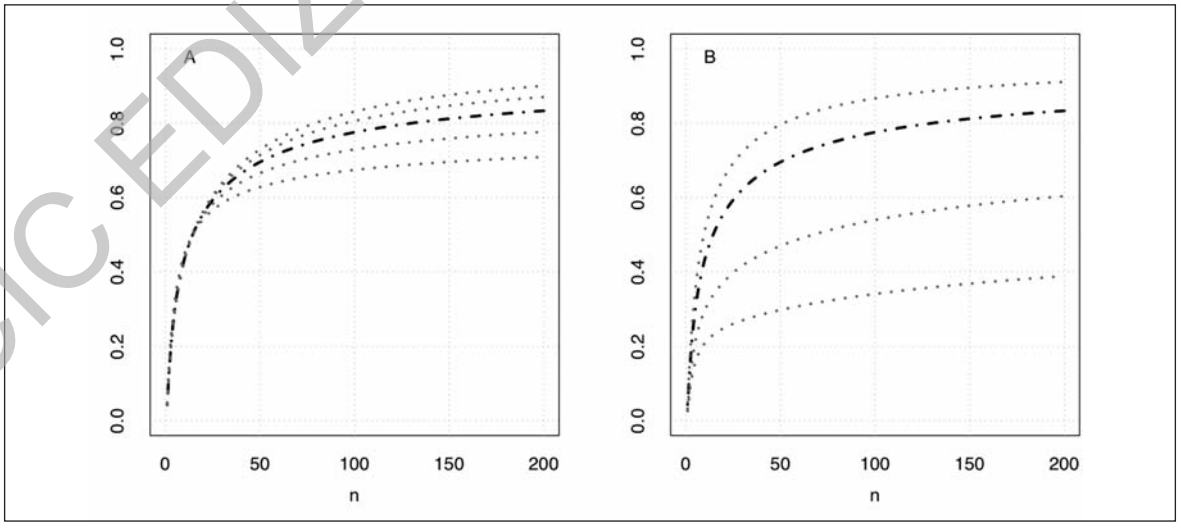


Figure 6. The Bayesian (dotted lines) predictive power curve $\beta_P^F(\pi_D)$, is plotted for different choices of the design prior π_D : A. $\theta_D = 0.56$ and from the bottom to the top $n_D = 10$, $n_D = 20$, $n_D = 34.5$, $n_D = 50$, and $n_D = 70$; B. $n_D = 34.5$ and from the bottom to the top $\theta_D = 0.1$, $\theta_D = 0.3$, $\theta_D = 0.56$ and $\theta_D = 0.7$.

Table 1. Classification of the power functions according to the use of prior information and to their ability to account for uncertainty on the design value: the predictive Bayesian power function can be thought as a general power function including the other three as special cases.

Modelling information:	Modelling uncertainty: design prior	
Analysis prior π_A	$n_D \to \infty : f(\cdot;\theta_D)$	$n_D < \infty : m_{\pi_D}(\cdot)$
Non-informative prior: $n_A \to 0$	$\beta_F^C(\theta_D)$	$\beta_F^P(\pi_D)$
Proper prior: $n_A > 0$	$\beta_B^C(\theta_D)$	$\beta_B^P(\pi_D)$

formative prior, and thus we obtain $\beta_C^F(\pi_D)$, the predictive frequentist power. The conditional frequentist power appears when we simultaneously allow n_D to approach infinity and n_A to approach 0. This means that both design uncertainty and prior information are actually ignored. Figure 5 allows us to compare the behaviour of the four power functions as the sample size n increases.

Discussion

In this work, we have reviewed frequentist and Bayesian power functions. First of all, we have presented the most common SSD criterion in clinical trials, based on the frequentist conditional power. We have argued that this criterion is not flexible enough. In particular, we have underlined that conditioning to a fixed design value, one takes no notice of uncertainty on this value (local optimality). This consideration has led us to introduce a predictive approach that allows us to incorporate uncertainty through a prior distribution, which guarantees a more careful choice of the optimal sample size. Moreover, we have noted that it is convenient to adopt a Bayesian approach to exploit available prior information directly in the SSD procedure. This allows one to take advantage of previous results or opinions of experts about the experiment and in a sense to "spare" sample units in the actual trial. Specifically, these ideas were formalized by introducing this two-prior approach. The basic principle of the two-priors approach is that initial uncertainty on the design value and pre-experimental information on the parameter of interest can be represented by two distinct prior distributions, a design prior and an analysis prior. By adopting this approach we have considered a general formulation of the power function: we have noted that the Predictive Bayesian power function previously introduced can be interpreted as a generalized power function that includes the other power functions as special cases. On one hand, if we take a point-mass design prior, we are actually conditioning the distribution to a single design value, that corresponds to work with a *conditional* power; whereas a proper design prior yields a predictive power. On the other hand, if we assume a non–informative analysis prior, we obtain equivalent results to the *frequentist* approach, whereas adopting a proper analysis prior we actually include prior information in the methodology and, specifically, in the power function that we consequently define *Bayesian*.

For the sake of simplicity, in this paper, we have focused on the normal model, since the main point is the interpretation and the comparison of the different forms of power functions. Any extension to alternative models would follow immediately.

Finally, an issue when dealing with the Bayesian method is its robustness with regard to the prior specification. As we have shown in the example, the choice of the prior parameters has an impact on the behaviour of the power function and therefore on the selected sample size. To deal with this problem, one can adopt a robust approach, *i.e.,* taking into account further uncertainty on the prior itself. See for example (13), (17), (18).

References

1. Spiegelhalter DJ, Abrams .R, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Wiley, 2004.
2. Tsiatis A A. The asymptotic joint distribution of the

efficient scores test for the proportional hazards model calculated over time. Biometrika 1981; 68: 311-315.
3. De Santis F, Perone Pacifico M. Accounting for historical information in designing experiments: the Bayesian approach. Annali Istituto Superiore di Sanità 2004; 40(2):173-179.
4. Guidance for the use of Bayesian statistics in medical device clinical trials. Draft guidance for industry and FDA staff, 2006.
5. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. JAMA 1995; 273(11):871-875.
6. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci 2000; 15(1):46-60.
7. De Santis F. Power priors and their use in clinical trials. Am Stat, 2006; 60(2): 122-129.
8. De Santis F. Using historical data for Bayesian sample size determination. J Roy Statist Soc. Ser. A 2007; 170(1):95-113.
9. Fayers PM, Ashby RD, Parmar M.B. Tutorial in Biostatistics: Bayesian data monitoring in clinical trials. Stat Med 1997;16:1413-1430.
10. Joseph L, du Berger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. Stat Med 1997;16:769-781.
11. De Santis F, Perone Pacifico M, Sambucini V. Optimal predictive sample size for case-control studies. J Roy Statist Soc. Ser. C 2004; 53:427-441.
12. M'Lan CE, Joseph L, Wolfson DB. Bayesian sample size determination for case-control studies JASA 2006; 101(474):760-772.
13. De Santis F. Sample size determination for robust Bayesian analysis. JASA 2006; 101(473):278-291.
14. Etzioni R, Kadane JB. Optimal experimental design for another's analysis. JASA 1993; 88(424):1404-1411.
15. Sahu SK, Smith TMF. A Bayesian method of sample size determination with practical applications. J Roy Statist Soc. Ser. A 2006; 169, Part2, 235-253.
16. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Statist Sci 2002; 17, n. 2, 193-208.
17. Brutti P, De Santis F, Gubbiotti S. Robust Bayesian sample size determination in clinical trials. Stat Med 2008; 27:2290-2306.
18. Greenhouse JB, Wasserman L. Robust Bayesian methods for monitoring clinical trials. Stat Med 1995; 14:1379-1391.
19. Lindley DV. The choice of sample size. The Statistician, 1997; 46, 129-138.
20. Raiffa H, Schlaifer R. Applied statistical decision theory. Wiley, 2000.
21. Tsutakawa RK. Design of experiment for bioassay. JASA 1972; 67, 339, 585-590.
22. Spiegelhalter DJ, Freedman JS. Bayesian approaches to clinical trials. Bayesian Statistics 3, Oxford University Press. Edited by: Bernardo J.M., De Groot M.H., Lindley D.V. and Smith A.F.M., 1988.
23. Brutti P, De Santis F. Avoiding the range of equivalence in clinical trials: robust Bayesian sample size determination for credible intervals. J Stat Plan Infer, 2008; 138, 1577-1591.