Choosing the sample size for equivalence trials

Stefania Gubbiotti, Fulvio De Santis

Abstract

In clinical practice we are usually interested in showing that an innovative therapy is more effective than a standard one. However, in some cases we have to respond to different purposes, such as proving equivalence of two competing treatments. In this paper we deal with sample size determination (SSD) for *equivalence trials*. The first step is the definition of an *equivalence interval* $\mathcal{I}$, that is a set of values of the parameter of interest indicating a negligible difference between treatments effects. Hence, we declare success if an interval estimate of θ is entirely included in $\mathcal{I}$. By adapting the metodology presented in Brutti and De Santis (2008) to equivalence trials, we derive two alternative SSD criteria based on Bayesian credible intervals and we also introduce their robust version with respect to prior distribution specification. In this work we provide results for the normal model with conjugate priors and we illustrate an application, based on a real example by Spiegelhalter et al. (2004).

1 Introduction

An equivalence trial is designed to confirm the absence of a meaningful difference between treatments. For example, when a pharmaceutical company is aware that there is not pre-experimental evidence enough for demonstrating superiority of a new treatment, it can decide to prove that it is equivalent to the standard one. The idea is that the new drug has chances to be approved and put on the market if it guarantees additional advantages, for instance in terms of safety or costs.

In the case of two alternative treatments let us suppose that the parameter of interest, θ, is the unknown difference (on a specific scale) between two treatment effects. As suggested in a recent document by the European Agency for the Evaluation of Medicinal Products (CPMP/EWP/482/99 (2000)), in this setting it is more informative to conduct the analysis by means of an interval estimate for θ, although there are closely related methods using significance test procedures (as described, for example, in Julious (2004)). Specifically, first of all we define an interval of values of θ representing indifference (that is the *equivalence interval* $\mathcal{I} = [\theta_I, \theta_S]$). Then, we declare equivalence if an interval estimate of θ entirely lies within $\mathcal{I}$. This situation is schematically represented in Figure 1.

Note that equivalence bounds may be chosen either symmetrically or asymmetrically with respect to the value denoting no treatment difference (typically zero). There are in practice some difficulties associated with its specification, but a detailed discussion on this point goes beyond the scope of the present work.

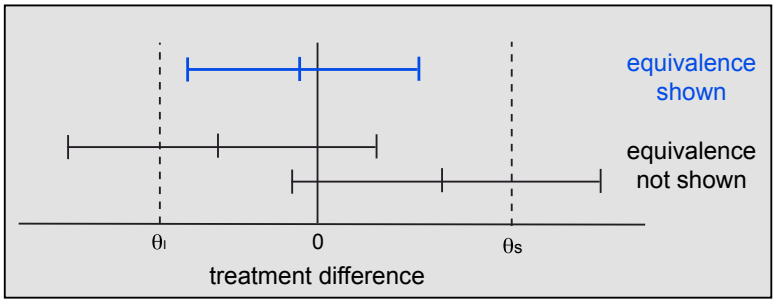

Figure 1: Equivalence trials

As discussed in Spiegelhalter et al. (2004), from a Bayesian perspective it is straightforward to define a region of equivalence and calculate the posterior probability that the treatment difference lies in this region. There is a large statistical literature on trials designed to establish equivalence between therapies using a Bayesian methodology: see for example Selwyn et al. (1981), Fluehler et al. (1983), Selwyn & Hall (1984), Breslow (1990), Grieve (1991) and Baudoin & O'Quigley (1994). A decision theoretic formulation is proposed in Lindley (1998): this approach is not considered in the present paper and in general it can give radically different conclusions, as noted in Spiegelhalter et al. (2004).

In this work we focus on the framework of an equivalence trial, with particular reference to the aspect of sample size determination (SSD). In Gould (1993) a Bayesian methodology for determining the sample sizes for event rate equivalence trials is proposed. Trials for demonstrating the equivalence of active standard and test treatments generally require large sample sizes that depend on the definition of equivalence and on the overall event rate, when the outcome is incidence of an event such as mortality. Planning of sample sizes for such trials requires the specification of a value for the overall event rate. This design value will often reflect the outcomes of previous trials of the standard treatment, and it is subject to uncertainty that needs some accommodation, to protect against an inadequate sample. For this reason the Author suggests to use Bayes and Empirical Bayes methods to incorporate information from one or more previous trials into the sample size calculation when equivalence means high confidence that the event rate ratio is less than some specified value.

In this paper we propose a predictive SSD approach, based on Bayesian credible intervals, by adapting the metodology presented in Brutti & De Santis (2008) to equivalence trials. The idea is to choose the minimum value of the sample size that guarantees to have the posterior credible interval of θ included in the equivalence interval. Since the interval limits are random quantities before the trial, we need to control the predictive distribution of the posterior credible interval bounds.

More specifically, we consider the *two priors approach*, first introduced in Tsutakawa

(1972) and Etzioni & Kadane (1993) within the Bayesian SSD methodologies and subsequently formalized by Wang & Gelfand (2002). According to this approach pre-experimental information is represented by the *analysis prior*, whereas uncertainty on the hypothesized parameter value at the design stage value is modeled by the so-called *design prior*. Note that, in general, design and analysis prior distributions do not coincide. This has been motivated and discussed in several papers, see for instance Wang & Gelfand (2002), Sahu & Smith (2006), De Santis (2006), O'Hagan & Stevens (2001), Brutti & De Santis (2008), Brutti et al. (2008). Here we do not go into further details of the discussion, but we adopt the two priors approach in defining the predictive SSD criteria, as described above: in particular the posterior distribution that we need to compute the credible intervals is based on the analysis prior, while the predictive distribution is induced by the design prior.

Moreover we deal with the issue of robustness with respect to the specification of a single analysis prior which is usually one of the major criticism towards Bayesian methods. The basic idea and motivations are discussed in De Santis (2006), Brutti & De Santis (2008) and Brutti et al. (2008). A robust version of the above criteria is obtained by replacing the analysis prior with a suitable class of prior distributions and referring to the bounds of the credible interval limits as the prior varies in the class that we call *robust bounds*. Due to the additional amount of uncertainty accounted for on the analysis prior, the resulting optimal sample sizes are larger than those obtained using a single prior, as shown in the application. Hence, it is interesting to quantiy the impact of misspecification of the analysis prior in terms of increase of the resulting sample sizes.

The outline of this paper is as follows. In Section 2.1, after describing the general setting of equivalence trials and introducing some notation, we present the Bayesian predictive SSD criteria. The robust criteria are then derived in Section 2.2. In Section 3 we provide results for the normal model with conjugate priors, using classes of restricted conjugate priors for the robust approach. Examples are illustrated and discussed in Section 4, with reference to the framework of a clinical trial analyzed in Spiegelhalter et al. (2004).

2 Predictive Bayesian approach to SSD

In the present section, we describe in details the setting of equivalence trials. Then we introduce the Bayesian SSD methodology. Let us suppose that the unknown parameter θ represents a measure of comparison between two alternative treatments, such as for instance the effects difference, the odds ratio or the hazard ratio. As anticipated in the Introduction, we consider an equivalence interval $\mathcal{I} = [\theta_I, \theta_S]$, that is a set of parameter values indicating a negligible difference between two competing treatments. We claim that the experiment is considered successful if it provides evidence that $\theta \in \mathcal{I}$, i.e. if an interval estimate of θ is entirely included into the equivalence interval.

Furthermore let us assume that pre-experimental information on θ is available. For instance we may want to take into account information provided by results of a previous study or opinions of some expert clinicians about the experimental treatment. Following the Bayesian approach, initial information can be formalized by specifying a prior probability distribution π_A for θ.

Let us consider the random sample $\mathbf{Y_n} = (Y_1, ..., Y_n)$, where Y_i is the random variable associated to treatments comparison, $f(\cdot;\theta)$ denotes its density or probability distribution function and n a prefixed number of patients to be recruited. Once the trial has been performed, the observed sample $\mathbf{y_n} = (y_1, ..., y_n)$, which is a realization of $\mathbf{Y_n}$, is available. Denoting the corresponding likelihood by $f(\mathbf{y_n};\theta)$, inference is then based on the posterior distribution that follows from Bayes theorem:

$$\pi_A(\theta|\mathbf{y_n}) = \frac{\pi_A(\theta) f(\mathbf{y_n};\theta)}{m_A(\mathbf{y_n})} \tag{1}$$

where the denominator is the marginal distribution $m_A(\mathbf{y_n}) = \int_\Theta \pi_A(\theta) f(\mathbf{y_n};\theta) d\theta$ and Θ denotes the parameter space. Let us assume for the sake of simplicity a unimodal continuous prior distribution. Given the posterior distribution of (1), we focus on the $(1-\alpha)$ credible interval for θ, that is

$$C_\alpha(\mathbf{y_n};\pi_A) = [l_n(\mathbf{y_n};\pi_A), u_n(\mathbf{y_n};\pi_A)]\,, \tag{2}$$

where $l_n(\mathbf{y_n};\pi_A)$ and $u_n(\mathbf{y_n};\pi_A)$ are the inferior and superior limit of the interval respectively. Note that $C_\alpha(\mathbf{y_n};\pi_A)$ can be for instance a HPD interval or an equal-tail interval. Finally the definition of success can be formalized as follows: we declare equivalence if

$$C_\alpha(\mathbf{y_n};\pi_A) \subseteq \mathcal{I},$$

that is if the following conditions

$$l_n(\mathbf{y_n};\pi_A) > \theta_I \quad \text{and} \quad u_n(\mathbf{y_n};\pi_A) < \theta_S \tag{3}$$

are simultaneously satisfied.

2.1 Criteria

Before the experiment is carried out, the limits of the posterior credible interval are random quantities, denoted by $l_n(\mathbf{Y_n};\pi_A)$ and $u_n(\mathbf{Y_n};\pi_A)$ to underline their dependence on the random sample $\mathbf{Y_n}$.

Thus, we need to take into account the randomness of the data using their marginal distribution. Adopting a conditional approach to SSD, it is possible to prefix a *design value* θ_D, that is a guess value for the parameter representing the objective of the experiment or,

in other words, the target effect to be detected. For an equivalence trial θ_D must be chosen inside the equivalence interval. For instance we could set $\theta_D = 0$ if $\mathcal{I}$ is symmetrical with respect to zero. In this case SSD criteria would be based on the sampling density $f_n(\cdot; \theta_D)$. However, according to a predictive approach, it is possible to model uncertainty on θ_D by specifying a design prior distribution π_D for θ. The design prior is used to average the sampling distribution, yielding the predictive distribution

$$m_D(\mathbf{y_n}) = \int_\theta f(\mathbf{y_n}; \theta)\pi_D(\theta)d\theta. \tag{4}$$

Notice that the sampling distribution of the data $f_n(\cdot; \theta_D)$ arises as a special case of $m_D(\cdot)$ when a point-mass design prior on the single value θ_D is chosen. In this sense $m_D(\cdot)$ generalizes $f_n(\cdot; \theta_D)$.

Now, from (3) we know that the success of the experiment relies on two simultaneous conditions: this reflects in the definition of the following criteria, based on predictive summaries of both $l_n(\mathbf{Y_n}; \pi_A)$ and $u_n(\mathbf{Y_n}; \pi_A)$. In particular we have:

1. **Predictive Expectation Criterion.**

   Let

   $$e_n^l = \mathbb{E}_{m_D}\left[l_n(\mathbf{Y_n}; \pi_A)\right] \text{ and } e_n^u = \mathbb{E}_{m_D}\left[u_n(\mathbf{Y_n}; \pi_A)\right] \tag{5}$$

   be the predictive expected value of $C_\alpha(\mathbf{y_n}; \pi_A)$ limits, computed with respect to the marginal m_D. The optimal sample size n_e^* is then selected as the minumum n such that the expected limits of the credible interval fall into the equivalence interval:

   $$n_e^* = \min\{n \in \mathbb{N} : e_n^l > \theta_I \text{ and } e_n^u < \theta_S\} \tag{6}$$

2. **Predictive Probability Criterion.**

   Based on the marginal m_D we define the predictive probability that the inferior limit is larger than θ_I and that the superior limit is smaller than θ_S, i.e.

   $$p_n^{l,u} = \mathbb{P}_{m_D}\left[l_n(\mathbf{Y_n}; \pi_A) > \theta_I, u_n(\mathbf{Y_n}; \pi_A) < \theta_S\right] \tag{7}$$

   Then, given a threshold $\gamma \in (0, 1)$, we select the optimal sample size n_p^* as the minumum n such that this probability is reasonably large, namely

   $$n_p^* = \min\{n \in \mathbb{N} : p_n^{l,u} > \gamma\}. \tag{8}$$

2.2 Robust criteria

Bayesian statistics is usually criticized because of the subjectivism due to the use of a specific prior distribution for posterior analysis. An attempt to address this objection

is represented by the *robust Bayesian approach* that is based on the following idea. In order to account for unavoidable uncertainty on pre-experimental information, the single prior is replaced by a whole class of distributions containing π_A. Then we compute the range of posterior quantities of interest over the selected class. If the posterior range is small, differences between distinct priors of the class are irrelevant and it is possible to use the starting prior with confidence. Otherwise, robustness is of concern and extra care is required. The robust approach can be adopted also for SSD. See for instance De Santis (2006), Brutti & De Santis (2008), Brutti et al. (2008).

In practice, in order to define a robust version of the SSD criteria introduced in Section 2.1, we replace π_A with a class of prior distributions Γ_A. Then we consider *robust bounds* of the posterior credible interval:

$$L_n(\mathbf{Y_n};\Gamma_A) = \inf_{\pi_A\in\Gamma_A} l_n(\mathbf{Y_n};\pi_A) \quad \text{and} \quad U_n(\mathbf{Y_n};\Gamma_A) = \sup_{\pi_A\in\Gamma_A} u_n(\mathbf{Y_n};\pi_A), \tag{9}$$

where $L_n(\mathbf{Y_n};\Gamma_A)$ is the lower bound of $l_n(\mathbf{Y_n};\pi_A)$ and $U_n(\mathbf{Y_n};\Gamma_A)$ is the upper bound $u_n(\mathbf{Y_n};\pi_A)$ as the prior π_A varies in Γ_A. Therefore we have robust evidence that θ belongs to $\mathcal{I}$ if $L_n(\mathbf{Y_n};\Gamma_A) > \theta_S$ and $U_n(\mathbf{Y_n};\Gamma_A) < \theta_I$, i.e. if we have $C_\alpha(\mathbf{y_n};\pi_A) \subseteq \mathcal{I}$ for any prior $\pi_A \in \Gamma_A$. Taking into account the double condition on both the robust bounds, we immediately derive the *robust* optimal sample sizes $n^*_{e,r}$ and $n^*_{p,r}$ for equivalence trials.

1. **Robust Predictive Expectation Criterion:**

$$n^*_{e,r} = \min\{n \in \mathbb{N} : e^L_n > \theta_I \ \text{ and } \ e^U_n < \theta_S\} \tag{10}$$

where

$$e^L_n = \mathbb{E}_{m_D}\left[L_n(\mathbf{Y_n};\Gamma_A)\right] \ \text{ and } \ e^U_n = \mathbb{E}_{m_D}\left[U_n(\mathbf{Y_n};\Gamma_A)\right] \tag{11}$$

2. **Robust Predictive Probability Criterion:** Given $\gamma \in (0,1)$,

$$n^*_{p,r} = \min\{n \in \mathbb{N} : p^{L,U}_n > \gamma\} \tag{12}$$

where

$$p^{L,U}_n = \mathbb{P}_{m_D}\left[L_n(\mathbf{Y_n};\Gamma_A) > \theta_I, U_n(\mathbf{Y_n};\Gamma_A) < \theta_S\right]. \tag{13}$$

The consequence of replacing π_A with Γ_A (which we assume to contain π_A), is that, in general, robust sample sizes are larger than single-prior sample sizes. Similarly, for any two classes of priors Γ_A and Γ'_A such that $\Gamma_A \subset \Gamma_{A'}$, optimal sample sizes determined with the latter class are larger than those obtained with the former. This will be illustrated in a practical application in Section 4, assuming the normal model with classes of restricted conjugate priors.

3 Results for the normal model

Let us assume that the data relevant to θ are summarized by a statistic Y_n with normal distribution of parameters $(\theta, \sigma^2/n)$. Let us denote its observed value with y_n. This notation is used for the sake of simplicity, being evident the distinction with the symbol denoting the $n-$th component of the random sample $\mathbf{Y_n}$ and of the observed sample $\mathbf{y_n}$ respectively. With no loss of generality, we consider a scale for the parameter θ, with $\theta = 0$ denoting equivalence. Note that this basic model provides an approximation that can be used, for instance, for binary data – with θ denoting the log odds ratio – and for survival data – with θ denoting the log hazard ratio – (see Spiegelhalter et al. (2004) and Section 4 for an application). Moreover for computational simplicity we adopt here conjugate prior distributions. Thus we assume that π_A is a normal density with expected value θ_A and variance σ^2/n_A, where n_A is the prior sample size (adopting the notation of Spiegelhalter et al. (2004)). From standard Bayesian conjugate analysis it follows that the resulting posterior distribution is

$$\pi_A(\theta|y_n) = N\left(\theta|\frac{n_A\theta_A + ny_n}{n_A + n}, \frac{\sigma^2}{n_A + n}\right)$$

For a given y_n, the posterior credible interval limits are

$$\begin{aligned} l_n(\mathbf{y_n}; \pi_A) &= \frac{ny_n + n_A\theta_A}{n + n_A} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n + n_A)}} \\ u_n(\mathbf{y_n}; \pi_A) &= \frac{ny_n + n_A\theta_A}{n + n_A} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n + n_A)}}, \end{aligned} \tag{14}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile.

Furthermore we assume that the design prior is $\pi_D(\theta) = N(\theta|\theta_D, \sigma^2/n_D)$; according to equation (4), the predictive distribution induced by π_D is a normal density of parameters $(\theta_D; \sigma^2(1/n + 1/n_D)$.

Since the objective of the trial is equivalence, the design prior is chosen to assign high probability to the equivalence interval values. For simplicity, in the following we set θ_D equal to the central value of the equivalence interval (for example $\theta_D = 0$, if $\mathcal{I}$ is centered on 0). On the other hand the analysis prior parameters are specified in order to model pre-experimental information on θ. Hence, π_A can be centered either on negative or positive values expressing respectively scepticism and enthusiasm towards one of the competing treatments. For example, let us suppose that a pharmaceutical company attempts to put a new drug on the market. The regulatory committee plans a clinical trial to show that the new drug is actually equivalent to the standard one. This yields an equivalence study with an optimistic analysis prior mean $\theta_A > 0$ and a design prior centered on 0. On the contrary, let us imagine that a pharmaceutical company wants to show that its new treatment is equivalent to a competing one, in terms of efficacy. This happens, for

instance, when the company, being aware that there is not evidence enough for proving superiority, aims at showing equivalence. Then the new drug has chances to be approved if it guarantees some other advantages, such as, for example in terms of safety or costs. In this case the design prior mean $\theta_D = 0$ represents the objective of the company, while the analysis prior expresses the opinion of an opponent, possibly fictitious. Note that in both situations the two-priors approach mentioned in Section 1 allows us to formalize two different points of view about the treatments difference (for further discussion on this interpretation see Etzioni & Kadane (1993)).

3.1 Criteria

It is now straighforward to compute the predictive quantities involved in the SSD criteria defined in Section 2.1. We have respectively:

$$
\begin{aligned}
1. \quad e_n^l &= \mathbb{E}_{m_D}\left[l_n(\mathbf{Y_n};\pi_A)\right] = \frac{n\theta_D + n_A\theta_A}{n+n_A} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} \\
e_n^u &= \mathbb{E}_{m_D}\left[u_n(\mathbf{Y_n};\pi_A)\right] = \frac{n\theta_D + n_A\theta_A}{n+n_A} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} \\
2. \quad p_n^{l,u} &= \mathbb{P}_{m_D}\left[l_n(\mathbf{Y_n};\pi_A) > \theta_I, u_n(Y_n) < \theta_S\right] = \\
&= \Phi\left(\frac{\theta_S(n+n_A) - z_{1-\alpha/2}\sigma\sqrt{n+n_A} - n_A\theta_A - n\theta_D}{n\sigma\sqrt{\frac{1}{n}+\frac{1}{n_D}}}\right) - \\
&- \Phi\left(\frac{\theta_I(n+n_A) + z_{1-\alpha/2}\sigma\sqrt{n+n_A} - n_A\theta_A - n\theta_D}{n\sigma\sqrt{\frac{1}{n}+\frac{1}{n_D}}}\right),
\end{aligned}
$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

3.2 Robust criteria

Let us suppose to replace the single analysis prior π_A with a class of distributions. For the sake of simplicity, we focus here on the class of restricted conjugate priors, defined as

$$\Gamma_{RC} = \{N(\theta|\theta_A, \sigma^2/n_A); n_A \in \left[n_A^L, n_A^U\right] \subset \mathbb{R}^+\}.$$

Under this assumption, we can exploit the results derived in Brutti & De Santis (2008) for computing the robust bounds of the credible interval in (9) and e_n^L and p_n^L, using the marginal distribution m_D. We recall these results in Appendix A, where we also derive analogous expressions for e_n^U and p_n^U. It is then immediate to apply the robust criteria defined in (10) and (12).

In next section we illustrate an application of the presented methodology, comparing the results obtained using robust and non robust criteria.

4 Example: CHART

The example considered in this paragraph is based on the CHART trial, first presented in Parmar et al. (1994) and further analysed in Parmar et al. (2001) and Spiegelhalter et al. (2004). Specifically we exploit the experimental setting described in Spiegelhalter et al. (2004) to elicit prior distributions and necessary clinical parameters required to plan an hypothetical equivalence trial, in order to draw a realistic design scenario. Then we revert the point of view of the original trial whose objective was superiority (see Spiegelhalter et al. (2004)).

First of all, let us review the general context of the CHART trial. In 1986 a new radiotherapy technique known as *continuous hyperfractionated accelerated radio therapy* (CHART) was introduced. The goal was to administrate radiotherapy continuously (no weekend breaks), in many small fractions (three a day) and accelerated (the course completed in 12 days), which clearly implies considerable logistical problems. Thus, the Medical Research Council wanted to compare CHART with conventional radiotherapy in lung cancer, to assess whether CHART provided a clinically important difference in survival that compensated for any additional toxicity and problems of delivering the treatment. The results were presented in terms of hazard ratio (HR), defined as the ratio of the hazard under CHART to the hazard under standard treatment. Hence, HRs less than one indicate superiority of CHART. In Spiegelhalter et al. (2004) a proportional hazards model is used, providing an approximate normal likelihood for the log HR: the estimated log HR has a normal density of expected value θ and variance σ^2/n, where n is the equivalent number of events in a trial balanced in recruitment and follow-up.

In order to specify the prior distribution and the equivalence interval the opinion of expert clinicians was considered. At the beginning, clinicians were enthusiastic about CHART. However oncologists who declined to participate in the trial expressed considerable scepticism. Opinions of eleven experts were collected and Spiegelhalter et al. (2004) suggest to average the corresponding distributions, obtaining as a summary a normal prior density centered on -0.28 with standard deviation of 0.23 (corresponding to an estimated HR of 0.76 with 95% interval from 0.48 to 1.19), which implies $n_A = 74.3$. Furthermore a sceptical prior was derived (see again Spiegelhalter et al. (2004)) with prior mean 0 and precision such that the prior probability that the true benefit exceeds the alternative hypothesis is 5%. This corresponds to a prior sample size $n_A = (1.65\sigma/\theta_A)^2 = 110$, noting that $\theta_A = \log(0.73) = -0.31$ and $\sigma = 2$. The eleven clinicians were also asked to specify the equivalence interval, namely "a range where they felt the two regimens were approxi-

mately equivalent". Upper and lower bounds for the equivalence intervals were averaged and the following results were obtained. The participants would be willing to use CHART routinely if it conferred at least 13.5% improvement in 2-year survival (from a baseline of 15%), and unwilling if less than 11% improvement. Thus the equivalence interval is from 11% to 13.5%, that is on the HR scale from 0.66 to 0.71, or on the log(HR) scale from -0.41 to -0.34.

However, in this case the above equivalence interval turns out to be too restrictive even if we choose a highly concentrated design prior on the central value of $\mathcal{I}$, for instance a normal density of mean -0.375 and standard deviation 0.067, with $n_D = 898$ (see Figure **??**). For instance, adopting the SSD criterion defined in (6), we obtain very large values for the optimal sample size – larger than 10000 units, which is absolutely unrealistic. Hence, with an illustrative intent we reset here the equivalence interval, in the light of the different purpose of the study. Let us imagine to assume the point of view of the CHART opponents: given the logistic problems connected with CHART, the supporters of the standard treatment could consider appropriate a wider interval, for instance from 5% to 15%, corresponding to $(-0.455; -0.164)$ on the log HR scale. In this case we manage to obtain smaller values for the optimal sample sizes, even if we specify a less demanding design prior, centered in $\theta_D = -0.3095$, and allowing for more uncertainty ($n_D = 51.9$, yielding a standard deviation of 0.278): the resulting optimal sample sizes are respectively $n_e^* = 682$ in correspondence of the clinical analysis prior and $n_e^* = 1037$ in correspondence of the sceptical analysis prior. Similar considerations apply when we consider the predictive probability criterion defined in (8): the original equivalence interval actually results unpractical (larger than 10000 units) and the optimal sample sizes are still too large when considering $\mathcal{I} = [\theta_I, \theta_S]$ (around 1000 units for a threshold $\gamma = 0.5$).

One may argue these conclusions are not satisfying yet from a practical point of view. Thus, we can change again our perspective and consider a less restrictive equivalence interval, for instance $I = [-0.41, 0.41]$. Note that we keep the lower bound of the previous interval and we make it symmetrical with respect to $\theta = 0$, ehich corresponds to no treatment difference: hence, we consider positive differences (on the log HR scale) as important as negative ones. Furthermore, let us assume a design prior of parameters $\theta_D = 0$ and $n_D = 100$. These design assumptions are represented in Figure 2, together with two choices for the analysis prior distribution. On the one hand the clinical prior – $\pi_A(\theta) = N(\theta|\theta_A = -0.28, n_A = 74.3)$ – is optimistic towards the superiority of CHART: thus, it can be interpreted as the point of view of a supporter of the innovative therapy, that we are willing to deny by proving equivalence of the two treatments. On the other hand the sceptical prior – $\pi_A(\theta) = N(\theta|\theta_A = 0, n_A = 50)$ – is centered on the equivalence interval, expressing a neutral opinion with respect to the treatment comparison.

Figure 3 represents the predictive expectation of the posterior credible intervals as n

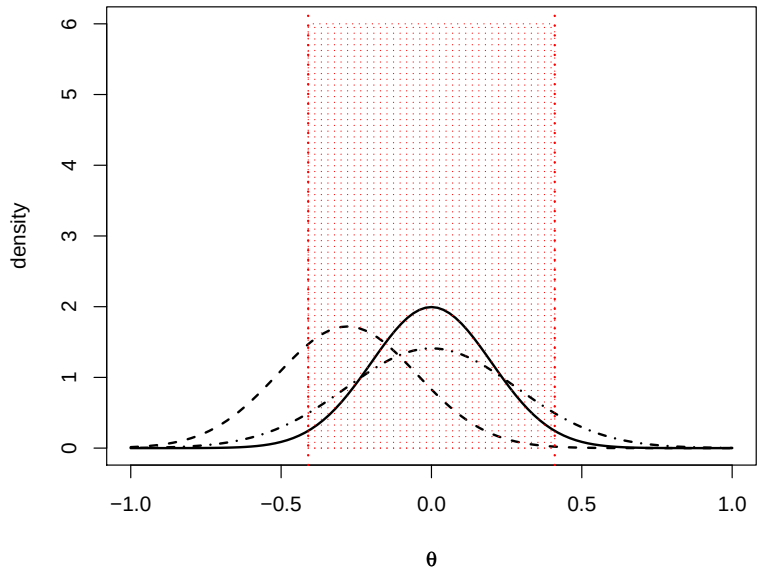

Figure 2: Clinical analysis prior (dashed line) with $\theta_A = -0.28$ and $n_A = 74.3$, sceptical analysis prior (dashed-dotted line) with $\theta_A = 0$ and $n_A = 50$, design prior (continuous line) with $\theta_D = 0$ and $n_D = 100$ and equivalence interval (dotted area) $\theta_I = -0.41$ and $\theta_S = -0.41$

increases: the optimal sample size is chosen in correspondence to the first interval limits (circled) falling into the equivalence interval (parallel dotted lines). In Figure 4 we show the behaviour of the predictive probability $p_n^{l,u}$ (continuous line) with respect to n and we pick the n_p^* as the smaller value of n corresponding to a larger probability than the threshold γ (for instance, $\gamma = 0.6$). Given the above design assumptions, adopting the predictive expectation and the predictive probability criterion, we have respectively $n_e^* = 105$ and $n_p^* = 182$ for the clinical prior and $n_p^* = 207$ and $n_e^* = 42$ for the sceptical prior.

Moreover in Table 1 we show the impact on the optimal sample sizes of different choices of the prior parameters (results corresponding to the above mentioned analysis priors are bolded). It is evident that, for a given analysis prior expectation, the larger n_A the smaller the optimal sample size. Conversely, given the prior variance, as θ_A approaches 0 the optimal sample sizes take smaller values. Note that a similar behaviour would result by considering positive values of θ_A decreasing towards 0, due to the symmetry of the equivalence interval. This can be interpreted as follows: both entusiastic and sceptical prior opinions on the innovative therapy require the same effort, in terms of experimental units, to be denied in demostrating equivalence.

Finally, note that in Figures 3 and 4 robust SSD criteria are also represented. As for the predictive expectation criterion (Figure 3), the gray vertical segments represent the expected robust credible intervals with respect to n with prefixed n_A^L and n_A^U. For example,

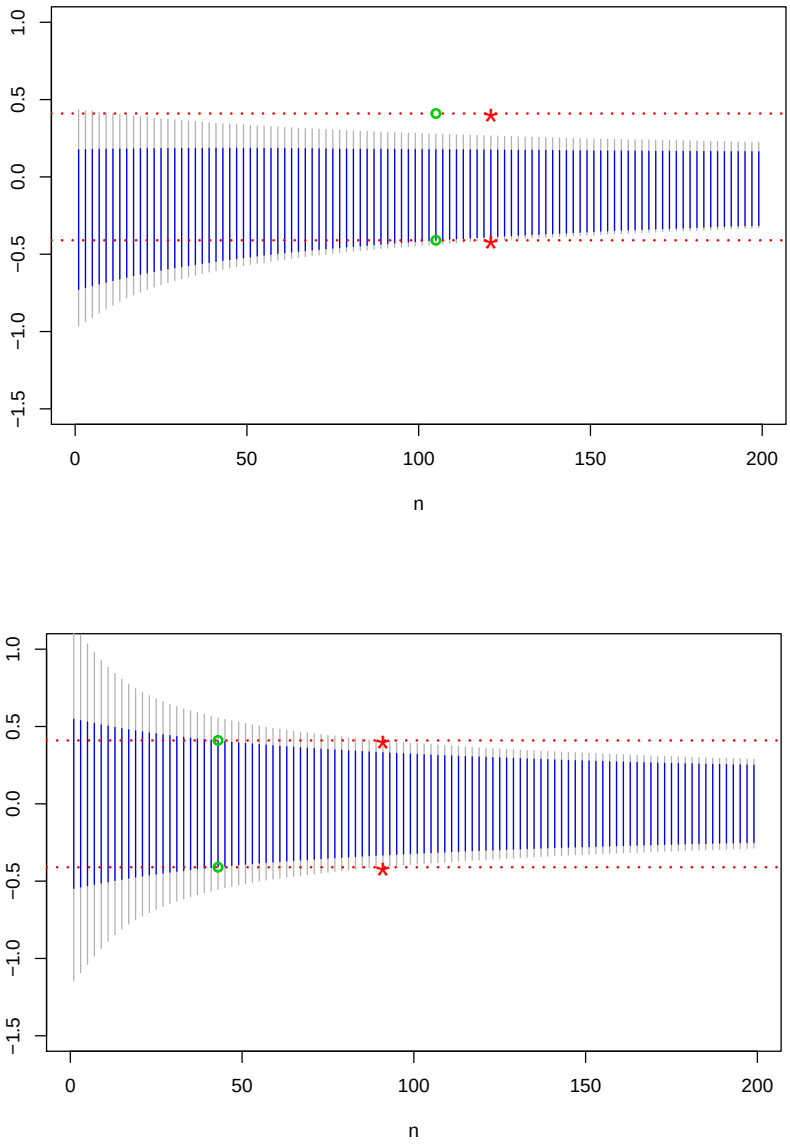

Figure 3: Robust and non robust SSD using the predictive expectation criterion using the clinical analysis prior (top panel) and the sceptical analysis prior (bottom panel), given the equivalence interval $[-0.41, -0.41]$ and design prior parameters $\theta_D = 0$ and $n_D = 100$.

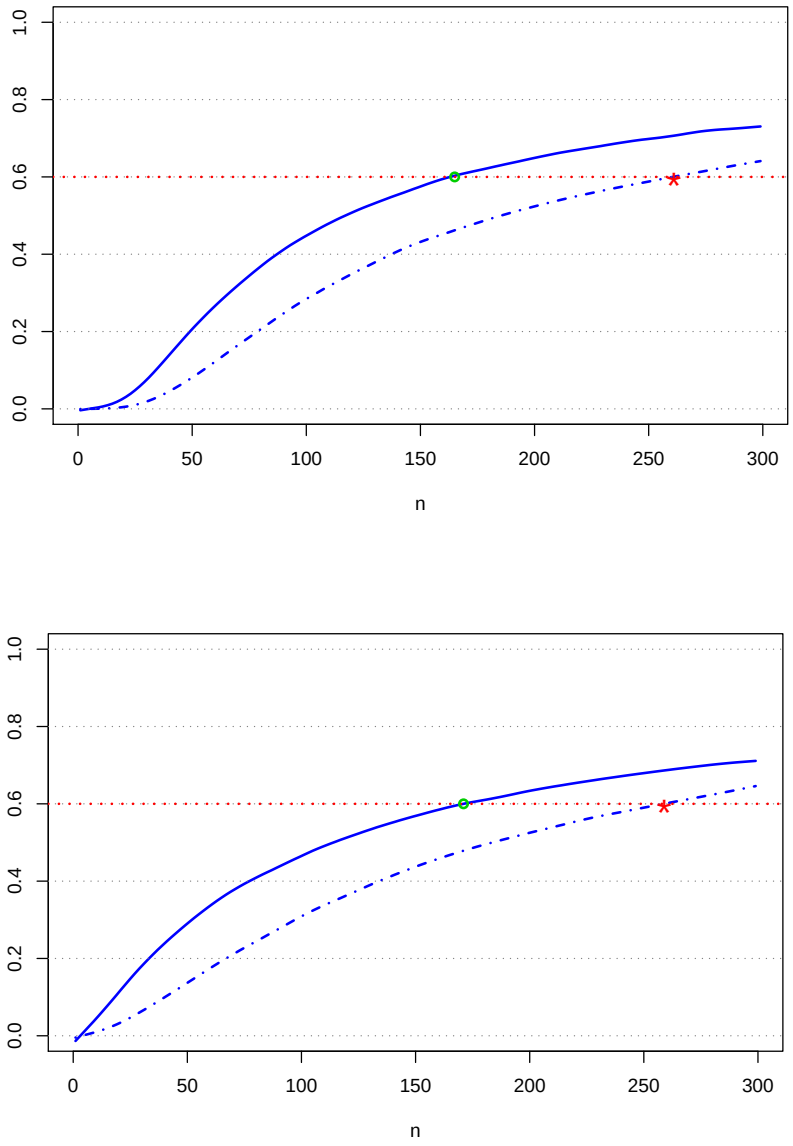

Figure 4: Robust and non robust SSD using the predictive predictive probability criterion using the clinical analysis prior (top panel) and the sceptical analysis prior (bottom panel), given the equivalence interval $[-0.41, -0.41]$, the sceptical analysis prior and design prior parameters $\theta_D = 0$ and $n_D = 100$.

	$\mathbf{n_e^*}$					$\mathbf{n_p^*}$				
	$\mathbf{n_A}$									
$\theta_\mathbf{A}$	10	30	**50**	**74.3**	90	10	30	**50**	**74.3**	90
-1.00	126	183	233	290	325	307	299	323	370	405
-0.50	105	126	145	164	176	303	267	245	236	239
-0.28	95	100	103	**105**	105	303	259	220	**182**	165
-0.20	91	89	86	82	78	302	257	214	165	139
-0.10	87	76	65	51	42	302	256	208	150	112
0.00	82	62	**42**	18	2	302	255	**207**	144	100

Table 1: Optimal sample sizes for equivalence interval $[-0.41, 0.41]$, with design parameters $\theta_D = 0$, $n_D = 100$ for several choices of the analysis prior parameters, using both the predictive expectation criterion and the predictive probability criterion (with $\gamma = 0.6$).

given $(n_A^L = 10, n_A^U = 200)$ the optimal robust sample sizes (indicated by a star) are respectively $n_{e,r}^* = 145$ for the clinical analysis prior and $n_{e,r}^* = 91$ for the sceptical one. Of course, if we consider a smaller class we obtain optimal sample size uniformly closer to the non robust ones: for instance, for $(n_A^L = 30, n_A^U = 100)$ we have $n_{e,r}^* = 122$ for the clinical analysis prior and $n_{e,r}^* = 67$ for the sceptical one. In Figure 4 the dashed line represents the probability that the robust bounds of the credible interval fall into the equivalence interval, as defined in (7)), using the restricted conjugate class with $(n_A^L = 10, n_A^U = 200)$. The optimal sample sizes (indicated by a star) are obviously larger than the non robust ones: in particular, we have $n_{p,r}^* = 261$ for the clinical analysis and $n_{p,r}^* = 263$ for the sceptical one. Again, if we consider a narrower class, with $(n_A^L = 30, n_A^U = 100)$ we get intermediate results: $n_{p,r}^* = 223$ for the clinical analysis prior and $n_{p,r}^* = 217$ for the sceptical one.

5 Discussion

In this paper we have considered a predictive methodology for sample size determination with specific reference to the setting of equivalence trials. Thanks to the predictive approach we are able to account for prior uncertainty and to model prior information, by specifying the design prior and the analysis prior. Specifically, due to the objective of the equivalence trials, we have pointed out that the design prior in this case should assign high probability to the values of the parameters indicating a negligible difference between the two treatments to be compared. As for the analysis prior, we have also addressed the issue of sensitivity to the prior specification by adopting a robust approach, as suggested before in De Santis (2006), Brutti & De Santis (2008) and Brutti et al. (2008). Some results

have been illustrated for the normal model with the class of restricted conjugate priors, although this methodology can be potentially extended to different models and classes of priors depending on the specific context of the application.

A slightly different problem is that of *bioequivalence*, which is very important in practice and very popular in the literature. Two different drugs or formulations of the same drug are defined bioequivalent if they are absorbed into the blood and become available at the drug action site at about the same rate and concentration (see for instance Berger & Hsu (1996)). In particular bioequivalence is of practical importance because the approval of most generic drugs in the USA and in the European Community requires the establishment of bioequivalence between the brand-name drug and the proposed generic version. This problem is theoretically interesting because it has been recognized as one for which the desired inference, instead of the usual significant difference, is practical equivalence. Hence, we hope to deal with this specific setting in future research.

A Appendix

A.1 Robust bounds

Let us denote by $l_n(y_n; w)$ and $u_n(y_n; w)$ the credible interval limits of (14) computed in correspondence of a given w, where $w = n_A^L, n_A^*, n_A^U$, with $n_A^* = \frac{4n^2(y_n-\theta_A)^2}{\sigma^2 z_{1-\alpha/2}^2} - n$.

Let us recall the results provided by Brutti & De Santis (2008) (Theorem 1). In this work it is shown that

$$L_n(\mathbf{y_n}; \Gamma_{RC}) = \begin{cases} l_n(y_n; n_A^L) & y_n < \theta_A + \xi_L \\ l_n(y_n; n_A^*) & \theta_A + \xi_L < y_n < \theta_A + \xi_U \\ l_n(y_n; n_A^U) & y_n > \theta_A + \xi_U \end{cases}$$

and

$$U_n(\mathbf{y_n}; \Gamma_{RC}) = \begin{cases} u_n(y_n; n_A^U) & y_n < \theta_A - \xi_U \\ u_n(y_n; n_A^*) & \theta_A - \xi_U < y_n < \theta_A - \xi_L \\ u_n(y_n; n_A^L) & y_n > \theta_A - \xi_L \end{cases}$$

where $\xi_k = \frac{z_{1-\alpha/2}}{2n}\sigma\left(n + n_A^k\right)^{1/2}$, for $k = L, U$.

Here, we derive the results used in Section 2.2 in order to obtain the robust SSD criteria. The following expression for e_n^L is also given by Brutti & De Santis (2008):

$$\begin{aligned} e_n^L &= \mathbb{E}_{m_D}(L_n(y_n)) = \\ &= l(\theta_D; n_A^L)\Phi(a_L) + l(\theta_D; n_A^U)(1 - \Phi(a_U)) + \theta_A\left[\Phi(a_U) - \Phi(a_L)\right] + \\ &+ \frac{1}{\sqrt{2\pi}\lambda_m}\left[\psi_U e^{-a_U^2} - \psi_L e^{-a_L^2}\right] - \frac{z_{1-\alpha/2}^2\sigma^2}{4n}\int_{\theta_A+\xi_L}^{\theta_A+\xi_U}\frac{1}{y_n - \theta_A}m_D(y_n)dy_n \end{aligned}$$

where $\lambda_m = \left(\sigma^2\left(n^{-1} + n_D^{-1}\right)\right)^{-1}$, $a_k = \sqrt{\lambda_m}(\theta_A - \theta_D + \xi_k)$, $\psi_k = \frac{n}{n+n_A^k}$ and $d_k = \theta_I + n_A^k/n(\theta_I - \theta_A) + z/n\sigma(n + n_A^k)^{1/2}$, for $k = L, U$.
It is then straightforward to compute e_n^U with an analogous procedure, as follows:

$$\begin{aligned} e_n^U &= \mathbb{E}_{m_D}(U_n(y_n)) \\ &= l(\theta_D; n_A^U)\Phi(c_U) + l(\theta_D; n_A^L)(1 - \Phi(c_L)) + \theta_A\left[\Phi(c_L) - \Phi(c_U)\right] + \\ &+ \frac{1}{\sqrt{2\pi}\lambda_m}\left[\psi_U e^{-c_L^2} - \psi_L e^{-c_U^2}\right] - \frac{3z_{1-\alpha/2}^2\sigma^2}{4n}\int_{\theta_A-\xi_U}^{\theta_A-\xi_L}\frac{1}{y_n - \theta_A}m_D(y_n)dy_n \end{aligned}$$

where $c_k = \sqrt{\lambda_m}(\theta_A - \theta_D - \xi_k)$ and $e_k = \theta_S + n_A^k/n(\theta_S - \theta_A) - z/n\sigma(n + n_A^k)^{1/2}$ for $k = L, U$.

Now, in order to derive $p_n^{L,U}$, for the sake of simplicity we define the following functions of w:

$$A(w) = \frac{\theta_S(n+w) - z\sigma(n+w)^{1/2} - w\theta_A - n\theta_D}{n\sigma\sqrt{\frac{1}{n} + \frac{1}{n_D}}}$$

and

$$B(w) = \frac{\theta_I(n+w) + z\sigma(n+w)^{1/2} - w\theta_A - n\theta_D}{n\sigma\sqrt{\frac{1}{n} + \frac{1}{n_D}}}.$$

Then, we have that

$$\begin{aligned}
p_n^{L,U} &= \Phi\left(\min\{A(n_A^U), c_U\}\right) - \Phi\left(B(n_A^L)\right) \cdot I_{\min\{A(n_A^U), c_U\} > B(n_A^L)} + \\
&+ \Phi\left(\min\{A(n_A^*), c_L\}\right) - \Phi\left(\max\{B(n_A^L), c_U\}\right) \cdot I_{\min\{A(n_A^*), c_L\} > \max\{B(n_A^L), c_U\}} + \\
&+ \Phi\left(\min\{A(n_A^L), c_L\}\right) - \Phi\left(\max\{B(n_A^L), a_L\}\right) \cdot I_{\min\{A(n_A^L), c_L\} > \max\{B(n_A^L), a_L\}} + \\
&+ \Phi\left(\min\{A(n_A^L), a_U\}\right) - \Phi\left(\max\{B(n_A^*), a_L\}\right) \cdot I_{\min\{A(n_A^L), a_U\} > \max\{B(n_A^*), a_L\}} + \\
&+ \Phi\left(A(n_A^L)\right) - \Phi\left(\max\{B(n_A^U), a_U\}\right) \cdot I_{A(n_A^L) > \max\{B(n_A^U), a_U\}}.
\end{aligned}$$

References

Baudoin, C. & O'Quigley, J. (1994), 'Symmetrical intervals and confidence intervals', *Biometrical Journal* **36**, 927–934.

Berger, R. L. & Hsu, J. C. (1996), 'Bioequivalence trials, intersection-union tests and equivalence confidence sets', *Statistical Science* **11**(4), 283–319.

Breslow, N. (1990), 'Biostatistics and Bayes', *Statistical Science* **5**, 269–284.

Brutti, P. & De Santis, F. (2008), 'Avoiding the range of equivalence in clinical trials: Robust Bayesian sample size determination for credible intervals', *The Journal of Statistical Planning and Inference* **138**, 1577–1591.

Brutti, P., De Santis, F. & Gubbiotti, S. (2008), 'Robust Bayesian sample size determination in clinical trials', *Statistics in Medicine* **27**, 2290–2306.

CPMP/EWP/482/99 (2000), 'Committee for proprietary medicinal products (london, 27 july 2000)', *www.emea.europa.eu/pdfs/human/ewp/048299en.pdf* .

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

Etzioni, R. & Kadane, J. B. (1993), 'Optimal experimental design for another's analysis', *Journal of the American Statistical Association* **88**(424), 1404–1411.

Fluehler, H., Grieve, A., Mandallaz, D., Mau, J. & Moser, H. (1983), 'Bayesian approach to bioequivalence assessment: an example', *Journal of Pharmaceutical Sciences* **72**, 1178–1181.

Gould, A. (1993), 'Sample sizes for event rate equivalence trials using prior information', *Statistics in Medicine* **12**, 2009–2023.

Grieve, A. (1991), 'Evaluation of bioequivalence studies', *European Journal of Clinical Pharmacology* **40**, 201–202.

Julious, S. A. (2004), 'Sample sizes for clinical trials with normal data', *Statistics in Medicine* **23**, 1921–1986.

Lindley, D. (1998), 'Decision analysis and bioequivalence trials', *Statistical Science* **13**, 136–141.

O'Hagan, A. & Stevens, J. W. (2001), 'Bayesian assessment of sample size for clinical trials for cost effectiveness', *Medical Decision Making* **21**, 219–230.

Parmar, M., Griffiths, G., Spiegelhalter, D., Souhami, R., Altman, D. & van der Scheuren, E. (2001), 'Monitoring large randomised clinical trials: a new approach using Bayesian methods.', *Lancet* **358**, 375–381.

Parmar, M., Spiegelhalter, D. & Freedman, L. (1994), 'The chart trials: Bayesian design and monitoring in practice', *Statistics in Medicine* **13**, 1297–1312.

Sahu, S. K. & Smith, T. M. F. (2006), 'A Bayesian method of sample size determination with practical applications', *Journal of the Royal Statistical Society. Ser. A* **17**(169), 235–253.

Selwyn, M. & Hall, N. (1984), 'On Bayesian methods for bioequivalence', *Biometrics* **40**, 1103–1108.

Selwyn, M. R., Dempster, A. & Hall, N. (1981), 'A Bayesian approach to bioequivalence for the 2×2 changeover design', *Biometrics* **37**, 11–21.

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

Tsutakawa, R. K. (1972), 'Design of experiment for bioassay', *Journal of the American Statistical Association* **67**(339), 585–590.

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.