Bayesian sample size determination and re-estimation using mixtures of prior distributions

Pierpaolo Brutti*, Fulvio De Santis**, Stefania Gubbiotti** [1]

* *LUISS Guido Carli*

** *Sapienza Università di Roma*

Abstract

In this paper we propose a predictive Bayesian approach to sample size determination and re-estimation in clinical trials, in the presence of multiple sources of prior information. The method we suggest is based on the use of mixtures of prior distributions for the unknown quantity of interest, typically an unknown effect or an unknown effects-difference. Methodologies are developed using normal models with mixtures of conjugate priors. In particular we extend the sample size determination analysis of [1] and the sample size re-estimation technique of [2].

1 Introduction

In clinical trials one is often interested in assessing the efficacy of a treatment (phase II) or the superiority of a new treatment over a standard therapy (phase III). This problem is typically formalized as a (one-sided) testing problem on an unknown parameter, θ, that denotes the unknown treatment effect or the unknown effect-difference respectively. Here we are interested in sample size determination (SSD) for this testing problem. From a Bayesian perspective, efficacy or superiority of a treatment is typically assessed by considering the posterior probability that θ exceeds a minimal clinically relevant threshold, δ. If this posterior probability is sufficiently large, the experiment is considered successful. SSD for this problem is addressed by selecting the minimal number of units so that, before the experiment is performed, the chances of obtaining a successful experiment are sufficiently large.

In this context one of the crucial aspects is the choice of the probability distribution of the data in pre-posterior calculations. Two main approaches can be used: *conditional* and *predictive*. In the former, pre-posterior computations are made with the sampling distribution of the data, for some fixed design value of the unknown parameter; see for instance [3], [4], [5] and [1]. The *conditional* approach is often criticized for yielding locally optimal sample sizes, which do not

[1] **Address for correspondence**: Dipartimento di Statistica, Probabilità e Statistiche Applicate. Sapienza Università di Roma - Piazzale A. Moro, 5 - 00185, Roma - Italy. **E-mail**: stefania.gubbiotti@uniroma1.it

account for uncertainty on the design parameter value (see [6] for discussion). Hence, we prefer here a *predictive* approach. Uncertainty on the design values of the parameter is therefore modeled through a probability distribution. This distribution is used for averaging the sampling distribution of the data and for obtaining the prior predictive distribution, employed for sample size computation (see for instance [7]). A possible compromise between these two approaches is proposed in [8] and in [9]: according to a mixed Bayes-frequentist methodology, the prior uncertainty on the unknown parameters is taken into account only for the design, whereas a standard classical test statistics is used for final inference. However we focus here on a fully Bayesian approach to SSD that models prior uncertainty on the design value when planning the experiment and combines preexperimental informationwith data for final inference.

In general the present work is related to the literature on Bayesian experimental design. For general reviews and discussions see [6] and [10]. More specifically for Bayesian SSD see also [11], [12] and [13]. In [13], [14] and [15] a robust Bayesian approach is considered.

As far as the choice of the prior distribution of the unknown parameter for computation of the posterior probability is concerned, the method has been implemented and discussed using both noninformative and informative priors. For references and discussion see, for instance, [16], [5].

In this paper we suppose that multiple sources of prior information on θ are available, for instance, opinions of several clinical experts or results from historical studies. This framework has been recently considered by [1] for Phase II clinical trials with binary endpoints. As a prior for θ, the Authors proposed a mixtures of conjugate prior distributions, each representing information from a single source, with weights proportional to the degree of pre-experimental "reliability" of each source.

Here we propose an extension of the analysis in [1]. Specifically:

- we consider the predictive approach for pre-posterior sample size computations;
- we adopt the two-priors approach to SSD. In this regard see, for instance, [17], [18], [19], [20] and [7];
- we present results assuming normal endpoints (Section 3.1) and illustrate an application (Section 3.2).

The presence of multiple sources of prior information, motivates an adjustment of the sample sizes set at the start of the trial after that a portion of experimental outcome has become available. Hence, in addition to the above three points, we also consider the sample size re-estimation (SSRe) problem. We follow a predictive Bayesian approach close to the one proposed by [2] and based on the expected probability of ending up with a successful trial, given the information provided by the results of the interim analysis. One attractive feature of this approach is that results of the interim analysis allows for an update of the weights of the components of the mixture itself.

The outline of the paper is as follows. In Section 2 we introduce the setup and some notation. In Section 3 we present the predictive approach for SSD with correspondent results in the case of the normal model (Section 3.1). This methodology is illustrated in Section 3.2 for planning a trial focused on the effect of magnesium in acute myocardial infarction, given the information derived from several hystorical studies (see Spiegelhalter et al., 2004). In Section 4 we deal with the sample size re-estimation problem. Results for the normal model are derived in Section 4.1 and then applied in Section 4.2 to the B-14 trial on breast cancer (see [21]).

2 Preliminaries

Let Y_n be an estimator of θ, the unknown quantity of interest in a clinical trial. Let $f_n(\cdot;\theta)$ be the probability density or mass function of Y_n. We consider an *analysis prior* $\pi(\cdot)$ that formalizes pre-experimental knowledge on the unknown θ. Given the observed data, y_n, let

$$\pi(\theta|y_n) = \frac{\pi(\theta) \times f_n(y_n;\theta)}{m(y_n)},$$

be the posterior probability distribution of θ where, assuming with no loss in generality that θ is continuous, $m(y_n) = \int_\Theta f(y_n;\theta)\pi(\theta)d\theta$ is the marginal distribution of the data. Also, let $P_\pi(\cdot|y_n)$ be the posterior probability measure corresponding to $\pi(\cdot|y_n)$. We want to establish whether θ is greater than δ, a minimally clinical significant effect or effects-difference. We say the experiment is successful if, for a given $\gamma \in (0,1)$, we have that:

$$P_\pi\left(\theta > \delta|y_n\right) > \gamma.$$

Suppose now that K sources of prior knowledge are available for inference on θ, for instance, opinions of K clinicians or data from K historical studies on the efficacy/superiority of a new medical intervention. The information from each of these sources is formalized in terms of a prior distribution on θ: $\pi_i(\theta)$ for $i = 1,\ldots,K$. A standard way to summarize this knowledge is to consider a mixture of the K prior distributions:

$$\pi(\theta) = \sum_{i=1}^{K} \omega_{0,i}\pi_i(\theta), \qquad \omega_{0,i} > 0, \qquad \sum_{i=1}^{K} \omega_{0,i} = 1, \tag{1}$$

where $\omega_{0,i}$ is the prior weight assigned to the i-th component of the mixture, $i = 1,\ldots,K$.

It is straightforward to check that the posterior probability distribution of θ is:

$$\pi(\theta|y_n) = \sum_{i=1}^{K} \omega_{1,i}(y_n)\pi_i(\theta|y_n), \tag{2}$$

where

$$\pi_i(\theta|y_n) = \frac{\pi_i(\theta) \times f_n(y_n;\theta)}{m_i(y_n)} \qquad \text{and} \qquad m_i(y_n) = \int_\Theta f(y_n;\theta)\pi(\theta)d\theta$$

are the posterior probability distribution of θ and the marginal distribution of the data respectively. Moreover the weight of the i-th posterior distribution can be updated as

$$\omega_{1,i}(y_n) = \frac{\omega_{0,i} m_i(y_n)}{\sum_{r=1}^{K} \omega_{0,r} m_r(y_n)}, \qquad i = 1, \ldots, K.$$

Finally the mixture form of the analysis prior also reflects in the posterior quantity of interest

$$P_\pi\left(\theta > \delta | y_n\right) = \sum_{i=1}^{K} \omega_{1,i}(y_n) P_{\pi_i}\left(\theta > \delta | y_n\right),$$

where $P_{\pi_i}\left(\theta > \delta | y_n\right)$ is the posterior probability that θ exceeds δ under prior π_i, $i = 1, \ldots, K$.

3 Predictive approach to SSD

Before starting the experiment, Y_n and consequently, $\pi(\theta|Y_n)$ and $P_\pi(\theta > \delta|Y_n)$ are random. Hence we need to define predictive SSD criteria, by specifying the probability distribution of the data for pre-posterior computation. For this purpose we use the *design prior* π_D that models uncertainty on the unknown parameter when designing the experiment. The corresponding prior predictive or marginal density is:

$$m_D(y_n) = \int_\Theta f_n(y_n; \theta) \pi_D(\theta) d\theta.$$

At this point some remarks are in order.

(i) The prior predictive distribution m_D accounts for uncertainty on the design values of θ. This allows one to avoid local optimality that arises when one uses in pre-posterior computation $f_n(y_n; \tilde{\theta}_D)$, for a fixed design value $\tilde{\theta}_D$ (conditional approach).

(ii) If π_D is a point-mass probability on a fixed value $\tilde{\theta}_D$, then $m_D(y_n) \equiv f_n(y_n; \tilde{\theta}_D)$ and the predictive and the conditional approaches coincide.

(iii) If, in the place of f_n, we use a point-mass probability on a virtual experimental outcome, $\tilde{y}_n$, we end-up with the approach of [4], [5] and [1].

(iv) The design prior π_D and the analysis prior π do not necessarily coincide. The design prior models uncertainty on the design value before the experiment is performed and it is used to obtain a sampling distribution that accounts for this uncertainty. The analysis prior models prior information on θ, to be incorporated in the posterior distribution for final inference on θ. For motivations and discussion on the *two-priors* approach, see, among others, [19].

In this framework we choose a suitable predictive summary, for instance the expected value of $P_\pi\left(\theta > \delta | y_n\right)$, computed with respect to the marginal distribution,

$$e_n = \mathbb{E}_{m_D}\left[P_\pi(\theta > \delta | Y_n)\right]. \tag{3}$$

Using the prior in Equation (1), it is straightforward to check that Equation (3) is equal to:

$$e_n = \mathbb{E}_{m_D}\left[\sum_{i=1}^{K}\omega_{1,i}(Y_n)P_{\pi_i}\left(\theta > \delta|Y_n\right)\right] = \sum_{i=1}^{K}\mathbb{E}_{m_D}\left[\omega_{1,i}(Y_n)P_{\pi_i}\left(\theta > \delta|Y_n\right)\right]. \tag{4}$$

Hence, the predictive expectation e_n is the sum of the predictive expectations of the terms

$$\omega_{1,i}(Y_n)P_{\pi_i}\left(\theta > \delta|Y_n\right), \qquad i = 1, \ldots, K,$$

whose explicit expressions are given in Section 3.1 for normal models with mixtures of conjugate normal prior distributions.

Then according to the SSD *effect-size criterion* (introduced in [19]) the optimal sample size is selected as the minimum n such that the corresponding e_n exceeds a given threshold $\eta \in (0, 1)$

$$n^* = \min\left(n \in \mathbb{N} : e_n > \eta\right). \tag{5}$$

Note that a more stringent SSD criterion would be based on the predictive probability $P_\pi(\theta > \delta|Y_n)$, as done for instance in [5], [1] and [15]. However, for the sake of simplicity, we will focus on the expectation criterion just defined.

3.1 Results for Normal model

Assume now that

$$Y_n|\theta \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

and that each component of the prior is

$$\pi_i(\theta) = N\left(\theta|\mu_i, \frac{\sigma^2}{n_{0i}}\right), \quad i = 1, \ldots, K,$$

where $N(\cdot|a, b)$ denotes the density function of a normal random variable of parameters (a, b) and where, for simplicity, σ^2 is assumed to be known. From standard results on conjugate analysis for the normal model it follows that

$$\pi_i(\theta|y_n) = N(\theta|E_i(\theta|y_n), V_i(\theta|y_n)), \qquad m_i(y_n) = N\left(y_n|\mu_i, v_i\right)$$

and that

$$\omega_{1,i}(y_n) = \frac{\omega_{0,i}\phi(\frac{y_n-\mu_i}{\sqrt{v_i}})}{\sum_{r=1}^{K}\omega_{0,r}\phi(\frac{y_n-\mu_r}{\sqrt{v_r}})},$$

where

$$E_i(\theta|y_n) = \frac{n_{0i}\mu_i + ny_n}{n_{0i} + n} \qquad \text{and} \qquad V_i(\theta|y_n) = \frac{\sigma^2}{n_{0i} + n} \tag{6}$$

are the i-th posterior expectation and variance of θ, $v_i = \sigma^2(n_{0i}^{-1} + n^{-1})$ is the variance of the i-th predictive distribution m_i, $i = 1, \ldots, K$ and where $\phi(\cdot)$ is the standard normal density function. Furthermore,

$$P_i(\theta > \delta | y_n) = 1 - \Phi\left(\frac{\delta - E_i(\theta|y_n)}{\sqrt{V_i(\theta|y_n)}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Hence Equation (4) becomes

$$e_n = \sum_{i=1}^{K} \mathbb{E}_{m_D}\left\{\frac{\omega_{0,i}\phi(\frac{Y_n-\mu_i}{\sqrt{v_i}})}{\sum_{r=1}^{K}\omega_{0,r}\phi(\frac{Y_n-\mu_r}{\sqrt{v_r}})} \quad \left[1 - \Phi\left(\frac{\delta - E_i(\theta|Y_n)}{\sqrt{V_i(\theta|Y_n)}}\right)\right]\right\}. \tag{7}$$

In the following, normality will be also assumed for the design prior and, consequently, for the marginal distribution, namely,

$$\pi_D(\theta) = N\left(\theta|\mu_D, \frac{\sigma^2}{n_D}\right), \quad m_D(y_n) = N\left(y_n|\mu_D, \sigma^2\left(\frac{1}{n_D} + \frac{1}{n}\right)\right). \tag{8}$$

Before illustrating an application in Section 3.2, we briefly discuss the choice of the threshold η involved in the SSD criterion of Equation (5). Existence and actual values of the optimal sample size n^* depend crucially on the interplay between the threshold η, the choice of δ and of the design prior π_D.

More specifically, in order to tune η we start by evaluating the suprema of e_n, that is an increasing function of n, for given δ and design prior. In Appendix A we show that e_n converges to a quantity e_∞ that can be computed via a Monte Carlo approximation. Then, we propose to pick η as a pre-specified percentage of e_∞, say $\beta \cdot e_\infty$ with $\beta \in (0, 1)$, so as to ensure the existence of the optimal sample size n^*. In this way, on the one hand, the optimization problem defined in Equation (5) is actually well posed and, on the other, we only need to specify the value of easily interpretable quantities like δ, the clinically significant difference to be detected, and μ_D, the true treatment difference under the design prior.

3.2 Example (Magnesium): predictive SSD using a mixture of priors derived from previous studies

We revisit an example in [16] where the results of a mete-analysis are reinterpreted according to a Bayesian perspective, in order to show the degree of scepticism necessary to reach an opposite conclusion. A series of small randomized trials was conducted in order to prove a protective effect of intravenous magnesium sulphate after acute myocardial infarction. These studies culminated in a meta-analysis which showed a highly significant 55% reduction in odds of death. This was confirmed in 1992 by a larger study, the LIMIT-2 trial, which demonstrated a 24% reduction in

i	study	log(OR)	sd	n_{0i}
1	Morton	−0.65	1.06	3.6
2	Rasmussen	−1.02	0.41	24.3
3	Smith	−1.12	0.74	7.4
4	Abraham	−0.04	1.17	2.9
5	Feldstedt	0.21	0.48	17.6
6	Shechter	−2.05	0.9	4.9
7	Ceremuzynsky	1.03	1.02	3.8
8	LIMIT-2	−0.3	0.15	187

Table 1: Observed results (log odds ratio scale) in 8 studies on the protective effect of magnesium, standard deviation and effective number of events.

mortality in 2000 patients. All these results suggested an outstanding conclusion: a cheap, safe and simple treatment reduces mortality in a common condition. For this reason, further investigation was recommended. but the massive ISIS-4 trial did not actually show evidence of any benefit: the final result on 58000 patients showed a non significant protective effect of magnesium, also consistent across major subgroups. Here we simply draw on this framework in order to formalize the situation in which prior knowledge comes from different historical studies.

We focus on the log odds ratio (log OR) as parameter of interest θ. An estimate of θ is given by $\hat{\theta} = \log\left(\frac{(a+\frac{1}{2})(b+\frac{1}{2})}{(c+\frac{1}{2})(d+\frac{1}{2})}\right) = y_m$, where a and b denote respectively the number of observed events in the control arm and in the treatment arm, with $a+b=m$, and c and d are the respective numbers of patients in the two groups who did not experience any event. Then the corresponding statistic Y_m is asymptotically distributed as a normal density of mean $\hat{\theta}$ and variance σ^2/m, where σ^2 is set equal to 4. See [16] for further details.

So we use each historical study to elicit a conjugate normal prior distribution. We assume the estimated log odds ratios and the corresponding standard deviations summarized in Table 1 as the parameters of the normal prior components. The global analysis prior is then given by a mixture of these eight priors, with conveniently chosen weights. The prior components and the corresponding the mixture are represented in the left panel of Figure 1 and Figure 2 choosing respectively equal weigths or weights propotional to each prior sample size n_{0i}.

Note that, since the parameter of interest is the log OR of magnesium with respect to placebo, negative values on this scale support the idea of a benefit of magnesium administration. Nevertheless in this case we are actually interested in proving that θ is larger than a threshold δ, meaning that magnesium is not effective. This is not the standard situation of a superiority trial, but the methodology described in Section 2 and in Section 3 is essentially the same. Alternatevely the

problem can be reverted, defining the log odds ratio of placebo with respect to magnesium and focusing on $P_\pi(\theta < \delta)$ as a posterior quantity of interest.

At this point we specify a design prior expressing scepticism towards the treatment. A possible choice can be based on ISIS-4 trial: this yields a design prior which is a normal density with mean 0.058 and effective number of events 4319, resulting in a variance equals to 0.00092.

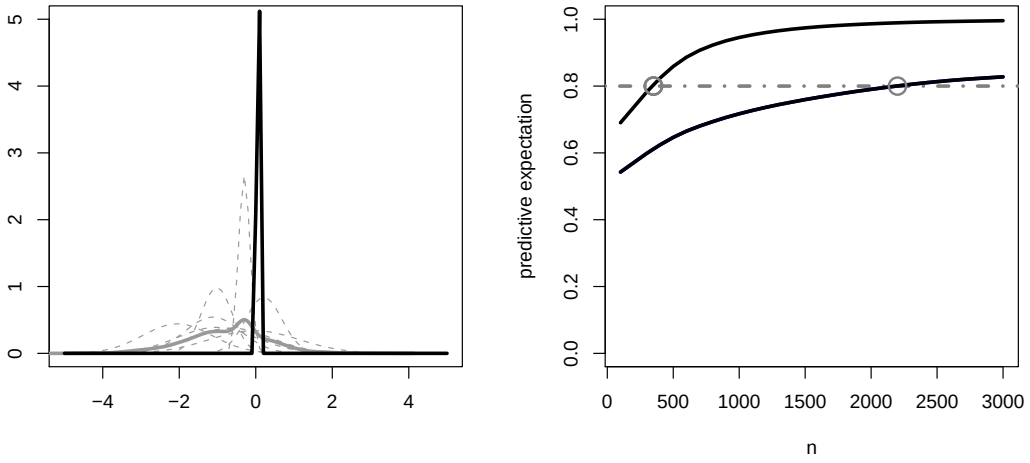

Figure 1: (*left panel*) Prior components (dashed gray lines), mixed prior with equal weights (continuous gray line) and design prior (black line). (*right panel*) Selection of the optimal sample size, using the mixed analysis prior with equal weights: $n^* = 350$ for $\delta = 0$ and $n^* = 2200$ for $\delta = -0.1$.

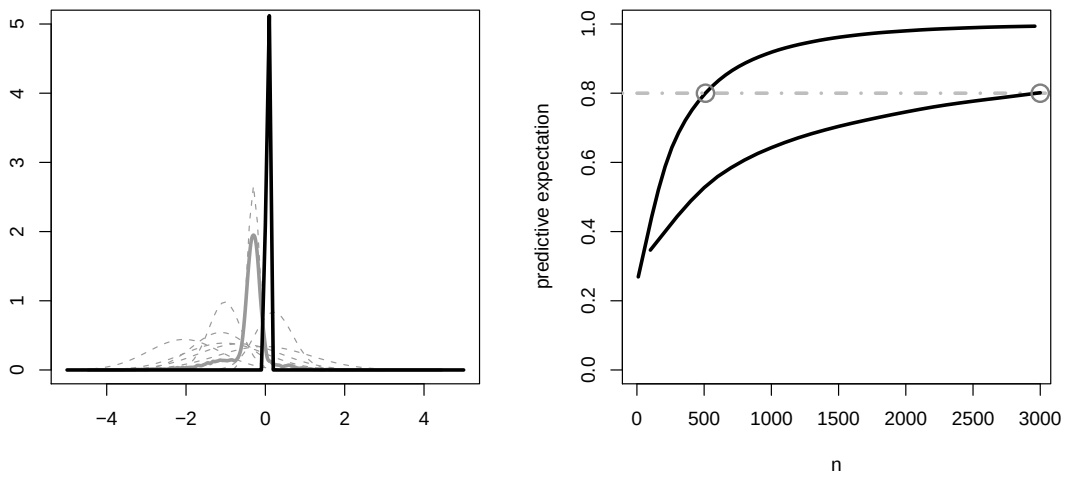

Figure 2: (*left panel*) Prior components (dashed gray lines), mixed prior (continuous gray line) with weights proportional to the number of events in each study and design prior (black line). (*right panel*) Selection of the optimal sample size, using the mixed analysis prior with weights proportional to the number of events in each study: $n^* = 510$ for $\delta = 0$ and $n^* = 3000$ for $\delta = -0.1$

In the right panel of Figure 1 the predicitive expectation e_n is plotted with respect to n, for two different choices of δ, and the optimal sample size is selected (circled) in correspondence of a prespecified threshold $\eta = 0.8$. Since the anaysis prior strongly supports the hypothesis of a protective effect of magnesium, we would need a sizeable number of events to be able to reach an opposite conclusion (about 1747 for $\delta = 0$). Moreover if we choose $\delta = -0.1$, the goal is less challenging and only 498 events are required.

Alternatevely we can choose prior weights proportional to the prior sample sizes n_{0i}. In this case we obtain the mixture represented in Figure 2; in the right panel the corresponding optimal sample size is selected. Notice that the prior component of LIMIT-2 trial is highly predominant in the mixed analysis prior ($n_{08} = 187$). This yields larger optimal sample sizes ($n^* = 2498$ for $\delta = 0$ and $n^* = 497$ for $\delta = -0.1$), since the analysis prior is more informative and closer to the design prior.

Moreover we considered two less informative design priors with smaller number of events, $n_D = 432$ and $n_D = 43$ respectively. For each different choice of the design parameters we computed the corresponding e_∞. We set consequently $\eta = \beta \cdot e_\infty$, as discussed in Section 3.1, for instance with $\beta = 0.80$. In this way we obtain the optimal sample sizes reported in Table 2. It is quite evident

δ	$\mathbf{n_D}$	$\mathbf{e_\infty}$	η	$\mathbf{n^*}$	
				equal weights	proportional weights
-0.1	4319	1	0.80	498	497
	432	0.95	0.76	509	510
	43	0.70	0.56	198	201
0	4319	0.97	0.78	1747	2489
	432	0.73	0.58	243	796
	43	0.58	0.46	42	228

Table 2: Optimal sample sizes for equal or proportional weights with respect to different design priors, choosing $\eta = \beta \cdot e_\infty$, with $\beta = 0.80$

that the more informative the design prior, the higher the maximum achievable value of e_n. Notice that, for example, for $n_D = 43$ and $\delta = 0$, e_∞ is equal to 0.58, so if we used a fixed $\eta = 0.80$, n^* would be undetermined. This shows how the criterion suggested in Section 3.1 for choosing η ensures a greater flexibility.

4 Predictive approach to SSRe

A predictive approach is now used for SSRe. Let us assume that, at a given time point, a fraction n_1 of the planned subjects have completed the trial. The objective is now to select the number n_2 of further sample units needed to complete successfully the experiment, by exploiting the information contributed by the first n_1 observed events; let us denote by y_{n_1} the corresponding observed statistic. The idea is to use as initial distribution at the interim analysis the posterior density of θ given y_{n_1}, $\pi(\theta|y_{n_1})$. Note that from Equation (2) it follows that $\pi(\theta|y_{n_1})$ can be written as a mixture of K different initial priors, whose weights are $\omega_{1,i}(y_{n_1})$, $i = 1, ..., K$.

In the second part of the trial n_2 events are to be observed, with $n_1 + n_2 = n$. The SSRe problem is to determine n_2. Given the observed value of y_{n_2} after n_2 events, the posterior distribution can be written as

$$\pi(\theta|y_{n_1}, y_{n_2}) = \sum_{i=1}^{K} \omega_{2,i}(y_{n_2}|y_{n_1}) \pi_i(\theta|y_{n_1}, y_{n_2})$$

where

$$\pi_i(\theta|y_{n_1}, y_{n_2}) = \frac{\pi_i(\theta|y_{n_1}) f_{n_2}(y_{n_2}; \theta)}{m_i(y_{n_2}|y_{n_1})} \tag{9}$$

and where the weights at the interim analysis are

$$\omega_{2,i}(y_{n_2}|y_{n_1}) = \frac{\omega_{1,i}(y_{n_1})m_i(y_{n_2}|y_{n_1})}{\sum_{r=1}^{K}\omega_{1,r}(y_{n_1})m_r(y_{n_2}|y_{n_1})}, \quad i = 1, ..., K.$$

The posterior predictive distribution of Y_{n_2} is

$$m_i(y_{n_2}|y_{n_1}) = \int_\Theta f_{n_2}(y_{n_2};\theta)\pi_i(\theta|y_{n_1})d\theta \tag{10}$$

and the posterior quantity of interest is

$$P_\pi(\theta > \delta|y_{n_1}, y_{n_2}) = \sum_{i=1}^{K}\omega_{2,i}(Y_{n_2}|y_{n_1})P_{\pi_i}(\theta > \delta|y_{n_1}, y_{n_2}).$$

Again, note that this quantity is random before y_{n_2} is observed. Hence, we introduce a predictive criterion to select the optimal additional sample size n_2^*:

$$n_2^* = \min\left(n \in \mathbb{N} : e_{n_1,n_2} > \eta\right) \quad \eta \in (0,1)$$

where

$$e_{n_1,n_2} = \mathbb{E}_{m_D}\left[P_\pi(\theta > \delta|y_{n_1}, Y_{n_2})\right] = \sum_{i=1}^{K}\mathbb{E}_{m_D}\left[\omega_{2,i}(Y_{n_2}|y_{n_1})P_{\pi_i}(\theta > \delta|y_{n_1}, Y_{n_2})\right]. \tag{11}$$

The expected value in Equation (11) is now computed with respect to the predictive distribution, m_D, induced by the design prior, π_D. Note that, at the interim stage, to obtain the predictive density m_D for SSRe we can use either $\pi_D(\theta)$ or $\pi_D(\theta|y_{n_1})$. In the former case we preserve the initial design goals, expressed by $\pi_D(\theta)$. In the latter we "adjust" design objectives according to the findings of the first part of the experiment. These two alternatives are discussed in the example of Section 4.2.

4.1 Results for Normal model

It is now straightforward to derive e_{n_1,n_2} for the normal model. From Equation (11) we have

$$e_{n_1,n_2} = \sum_{i=1}^{K}\mathbb{E}_{m_D}\left\{\frac{\omega_{1,i}(y_{n_1})\phi\left(\frac{Y_{n_2}-\mu_{i,2}}{\sqrt{v_{i,2}}}\right)}{\sum_{r=1}^{K}\omega_{1,r}(y_{n_1})\phi\left(\frac{Y_{n_2}-\mu_{r,2}}{\sqrt{v_{r,2}}}\right)} \cdot \left[1 - \Phi\left(\frac{\delta - E_{i,2}(\theta|y_{n_1}, Y_{n_2})}{\sqrt{V_{i,2}(\theta|y_{n_1}, Y_{n_2})}}\right)\right]\right\}. \tag{12}$$

This expression is essentially similar to the one in Equation (7), with updated posterior and predictive means and variances, given y_{n_1}. See Appendix B for further details. As for the choice of the threshold η, the criterion suggested at the end of Section 3.1 still holds.

In order to illustrate the proposed methodology for SSRe in Section 4.2 we consider an application in which the normal approximation for the log hazard ratio (log HR) is used and interim analysis data are available.

		$\mathbf{n_p}$	$\mathbf{n_t}$	$\mathbf{n_1}$	**log(HR)**	**sd**
I	after first interim	18	28	46	0.435	0.295
II	after second interim	24	43	67	0.567	0.244
III	after third interim	32	56	88	0.545	0.213
IV	after fourth interim	36	66	102	0.588	0.198
V	final results	50	85	135	0.519	0.172

Table 3: B14: Interim and final results arm on the log hazard ratio scale: n_p and n_t denote the number of events occurred in the placebo and in the treatment arm respectively and the total number of events is $n_1 = n_p + n_t$

4.2 Example (B-14): Predictive SSRe using a mixture of priors expressing opposite beliefs

Here we consider the B-14 study (see [21], [16]) in which data from four interim analysis and final results are available. The objective of the trial was to assess a long-term protective effect of tamoxifen in preventing the recurrence of breast cancer. A sequential randomized controlled study was performed, enrolling disease-free patients after 5 years of therapy. According to the sequential design, an interim analysis was scheduled approximatively every 1-1.5 years (using O'Brien-Fleming stopping boudaries). At the beginning of the trial the planned sample size was 115 events, to detect a 40% failure reduction (corresponding to a hazard ratio of 0.6) with 85% power. Assuming a 18% event rate, this yielded a total planned sample size approximately equal to 624 patients; finally the effective number of recruited patients was 1172, because of an accrual rate lower than expected.

Dignam et al. (1998) discussed a Bayesian interpretation of these results, under a range of prior assumptions. Using the normal approximation for the log hazard ratio estimator (see [16]) we choose here two normal priors expressing opposite beliefs, a sceptical prior $\pi_1(\theta) = N(\theta|0, 0.31)$ and an enthusiastic prior $\pi_2(\theta) = N(\theta| - 0.51, 0.31)$, where standard deviation is chosen to have 5% chance that the true difference exceeds a 40% reduction or, respectively, that a negative effect is observed ($\sigma^2 = 4$, $n_{01} = n_{02} = 41.4$). Furthermore, we center the design prior on the actual design value 0.51 (0.60 on the hazard ratio scale), with standard deviation equals to 0.19 ($\sigma^2 = 4$, $n_D = 115$).

The data at the four interim analyses and the final results of the trial are summarized in Table 3. After each interim analysis we re-estimate the optimal additional sample size n_2^*, needed to obtain that the predictive expectation of the probability $P(\theta < \delta|y_{n_1})$ is sufficiently large. For instance, we set $\delta = -0.22$, corresponding to a 20% reduction on the HR scale. For each interim analysis $n_1 = n_p + n_t$ denotes the total number of events observed so far, with n_p and n_t indicating the number of events in the placebo and in the treatment arm respectively.

First of all we assign equal weights to the two prior components of the mixture $\pi(\theta)$ defined

in Equation (1). The analysis prior and the design prior are represented in Figure 3. After the first interim analysis, in order to reach a conclusion favouring tamoxifen, it would be necessary to observe a large number of events (for example, $n_2^* = 59$, for a threshold $\eta = 0.75$ corresponding to the 80% of the supremum of e_{n_1,n_2}). Moreover after each interim analysis the additional number of units required to conclude in favour of a protective effect of tamoxifen becomes larger and larger (see Figure 4). This is coherent with the fact that the negative results actually observed at each step, make it more and more difficult to revert the evidence against tamoxifen.

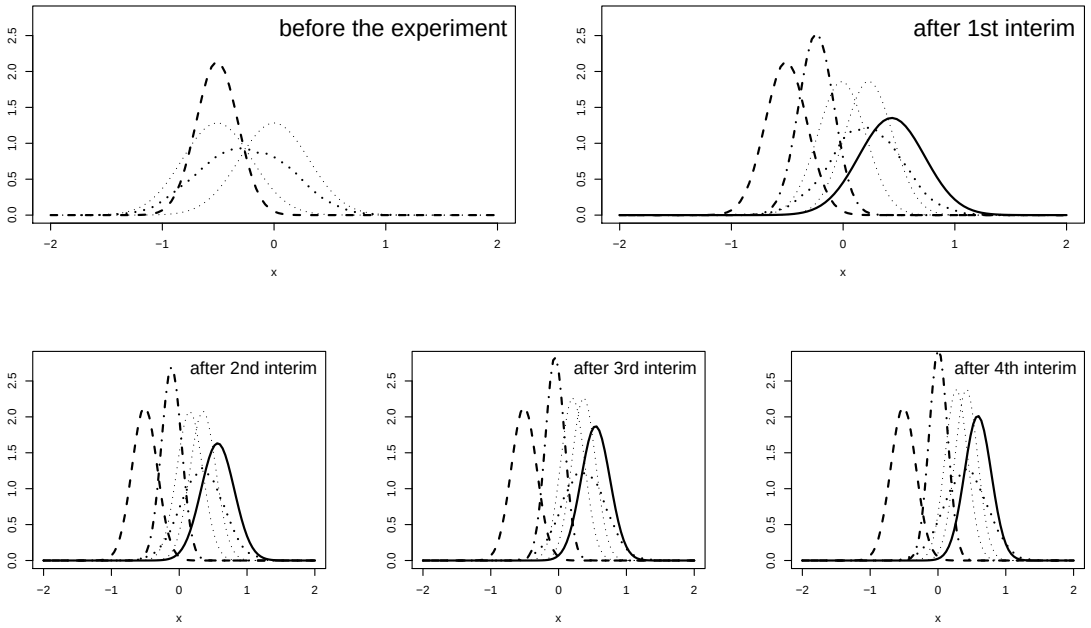

Figure 3: Information update at each interim point: the dotted lines represent the prior components of the mixed analysis prior, the dashed density is the fixed design prior, while the dashed-dotted curves indicate the progressive update of the design prior. The continuous line represents the likelihood at each step.

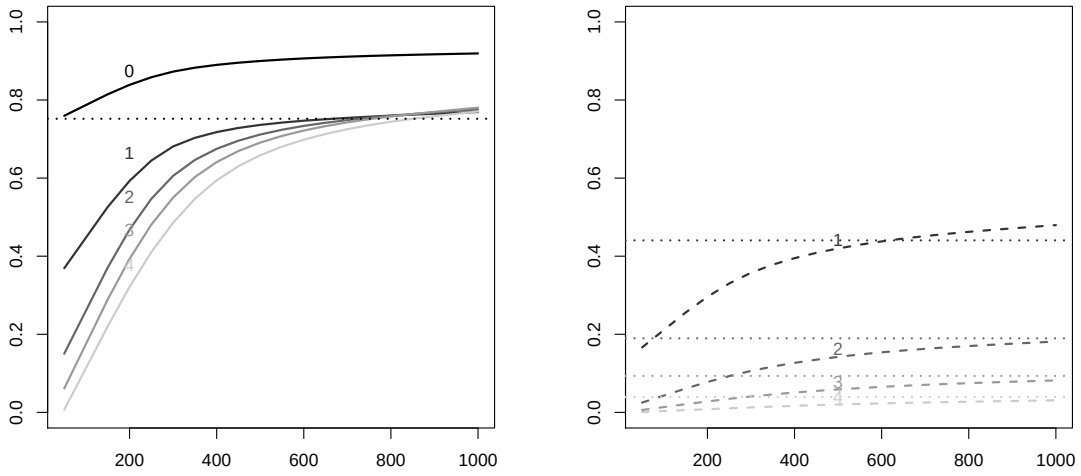

Figure 4: Sample size re-estimation at each interim time (denoted by the numbers from 0 to 4). Continuous lines are referred to the case of fixed design prior; dashed lines are referred to the case of updated design prior

In the right panel of Figure 4 the dashed lines represent the SSRe criteria when the design prior is also updated after each interim: the evidence of the data supports a conclusion opposite to the one we expected in designing the experiment and this affects e_{n_1,n_2}. In this case the previous threshold η is impractical already after the first interim, the optimal additional sample size is undetermined. If η is reduced to 0.44 (corresponding to $\beta = 0.8$), after the first interim, we have $n_2^* = 590$. In Table 4 we report the optimal re-estimated sample sizes for several choices of the initial weights, with fixed design prior. The weights of the sceptical component tend to be higher, due to the evidence of the data against a protective effect of tamoxifen. This corresponds to an increasing re-estimated number of required events after each interim analysis.

5 Discussion

In this paper we present a predictive methodology for sample size selection and adjustment in clinical trials. The two main features of the method are: (*i*) distinction between analysis and design prior; (*ii*) use of a mixture analysis prior.

The role of the design prior and its relationships with the analysis prior is discussed in several previous papers. Here we just want to remark the importance of a possible distinction between

		interim analysis			
	before	**I**	**II**	**III**	**IV**
$\mathbf{weight}_1$	1/2	0.87	0.94	0.95	0.96
$\mathbf{weight}_2$	1/2	0.13	0.06	0.05	0.04
$\mathbf{n_2^*}$	59	638	742	732	864
$\mathbf{weight}_1$	1/3	0.77	0.88	0.90	0.92
$\mathbf{weight}_2$	2/3	0.23	0.12	0.10	0.08
$\mathbf{n_2^*}$	36	543	671	714	796
$\mathbf{weight}_1$	2/3	0.93	0.97	0.97	0.98
$\mathbf{weight}_2$	1/3	0.07	0.03	0.03	0.02
$\mathbf{n_2^*}$	79	800	739	787	855
$\mathbf{weight}_1$	1/10	0.43	0.62	0.66	0.72
$\mathbf{weight}_2$	9/10	0.57	0.38	0.34	0.27
$\mathbf{n_2^*}$	10	503	592	669	759
$\mathbf{weight}_1$	9/10	0.98	0.99	0.99	0.996
$\mathbf{weight}_2$	1/10	0.02	0.01	0.01	0.004
$\mathbf{n_2^*}$	116	799	823	826	931

Table 4: Optimal re-estimated sample sizes for several choices of the initial weigths (weight_1 refers to the sceptical prior component, weight_2 to the enthusiastic one). Given that $e_{n_1,\infty} = 0.94$ and choosing $\beta = 0.80$, the threshold η is 0.75.

modelling pre-experimental knowledge on the unknown parameter (analysis prior) and modelling uncertainty on design goals of the trial (design prior). See [17], [18], [19], [20] and [7] for more details.

The use of priors mixtures allows one to take into account different sources of pre-experimental information and to combine them in a simple way. Sometimes these sources actually correspond to results of previous studies or to opinions of several experts. It is also possible to consider "conventional" priors that reflect opposite attitudes towards the trial such as enthusiasm and scepticism. In this way we are able to incorporate a large amount of information and uncertainty on the unknown treatment effect. One of the main advantages of this approach is that it typically avoids sample size underestimation and low predictive probability of trial success.

One critical aspect of the method is the choice of prior weights in the mixture. Of course, this is problem specific. However, we discuss in the examples some strategies. In Example of Section 3.2 we compare some alternative weights assignments, such as uniform weights and weights proportional to "prior sample sizes" of each historical study used to elicit the prior components. In Example of Section 4.2 we consider different combinations of weights for an enthusiastic and a sceptical prior and we examine their impact on resulting sample sizes.

The presence of several sources of prior knowledge makes it natural to plan an interim analysis and a sample size re-estimation step. This approach appears to us quite useful when available sources of prior knowledge (or experts opinions) are conflicting and when, initially, the weight of each prior in the mixture is not predominant over the others. In this case, the first portion of data (y_{n_1}) allows one to adjust both the starting prior distributions, π_i and their weights in the mixtures. Note also that, in principle, multiple sample size adjustments do not have drawbacks in a Bayesian perspective. In fact, from this point of view, repetition of the SSRe procedure implies just a sequential use of Bayes theorem. This is shown, for instance, in the Example of Section 4.2.

The predictive approach to SSD and SSRe based on mixtures of analysis priors presented in the paper can be potentially extended in several directions. First of all, this methodology can be applied to other models, such as Bernoulli and survival trials. (See also [1], where Beta mixtures are used for non-predictive SSD). A further possible extension is to consider mixtures of nonconjugate analysis priors. In these cases one typically needs to resort to numerical computational methods. See [19] for discussion. Comparison between sample size determined with exact results and normal approximations are also of interest.

A Asymptotic behaviour of e_n

As pointed out at the end of Section (3.1), in order to have a practical way to choose the threshold η, we need to study the asymptotic behaviour of (3). First of all notice that as $n \to \infty$ we have that:

- the posterior mean of the $i-$th component $E_i(\theta|Y_n)$ is asymptotically equivalent to Y_n;
- the posterior variance of the $i-$th component, $V_i(\theta|Y_n)$, tends to 0 (a.s.);
- the variance of the marginal distribution induced by the $i-$th prior component, v_i, converges to σ^2/n_{0i} (prior variance);
- the sequence of random variables Y_n, with marginal densities $m_D(\cdot)$, converges to $N\left(\mu_D, \frac{\sigma^2}{n_D}\right)$, whose density is here denoted as m_∞.

Hence, by the dominated convergence theorem, the limit of (3) is

$$\begin{aligned}\lim_{n\to\infty} e_n &= \sum_{i=1}^{K} \lim_{n\to\infty} \mathbb{E}_{m_D}\left\{\frac{\omega_{0,i}\phi\left(\frac{Y_n-\mu_i}{\sqrt{v_i}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{Y_n-\mu_r}{\sqrt{v_r}}\right)}\left[1-\Phi\left(\frac{\delta-E_i(\theta|Y_n)}{\sqrt{V_i(\theta|Y_n)}}\right)\right]\right\} \\ &= \sum_{i=1}^{K}\int_{\mathbb{R}} \lim_{n\to\infty}\left\{\frac{\omega_{0,i}\phi\left(\frac{y_n-\mu_i}{\sqrt{v_i}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{y_n-\mu_r}{\sqrt{v_r}}\right)}\left[1-\Phi\left(\frac{\delta-E_i(\theta|y_n)}{\sqrt{V_i(\theta|y_n)}}\right)\right]\right\} m_D(y_n)\mathrm{d}y_n. \quad (13)\end{aligned}$$

Note now that, as $n \to \infty$, the expression in square brackets in Equation (13) converges to 1 or 0 according to the sign of the argument of $\Phi(\cdot)$. Hence, taking into account the limiting distribution of Y_n, each term of the sum can be written as:

$$\int_{\mathbb{R}}\left\{\frac{\omega_{0,i}\phi(\frac{z-\mu_i}{\sqrt{v_i}})}{\sum_{r=1}^{K}\omega_{0,r}\phi(\frac{z-\mu_r}{\sqrt{v_r}})}\quad \mathbb{I}_{[\delta,\infty)}(z)\right\}\cdot m_\infty(z)dz.$$

Therefore Equation (13) can be written more synthetically as

$$\sum_{i=1}^{K}\mathbb{E}_{m_\infty}\left[\omega_{1,i}(Z)\cdot\mathbb{I}_{[\delta,\infty)}(Z)\right].$$

This quantity can be computed through a Monte Carlo approximation.

B SSRe: Results for the Normal model

We provide here the details for deriving e_{n_1,n_2} for the normal model (see Equation (12)). First of all, each posterior component of Equation (9) is

$$\pi_i(\theta|y_{n_1}, y_{n_2}) = N(\theta|E_{i,2}(\theta|y_{n_1}, y_{n_2}), V_{i,2}(\theta|y_{n_1}, y_{n_2}))$$

where the posterior mean and variance are respectively

$$E_{i,2}(\theta|y_{n_1}, y_{n_2}) = \frac{(n_{0i} + n_1)E_i(\theta|y_{n_1}) + n_2 y_{n_2}}{n_{0i} + n_1 + n_2}$$

and

$$V_{i,2}(\theta|y_{n_1}, y_{n_2}) = \frac{\sigma^2}{n_{0i} + n_1 + n_2}.$$

Moreover the marginal distribution in Equation (10) is a normal density of parameters $(\mu_{i,2}, v_{i,2})$, where the mean $\mu_{i,2}$ is equal to the posterior mean $E_i(\theta|y_{n_1})$ defined in Equation (6) and the variance is given by

$$v_{i,2} = \sigma^2 \left(\frac{1}{n_{0i} + n_1} + \frac{1}{n_2} \right)$$

for $i = 1, ..., K$. Note that the expected value in Equation (12) is computed with respect to the predictive distribution m_D, which is a normal distribution. As discussed in Section 4, m_D can be alternatevely derived using the design prior $\pi_D(\theta)$ or the posterior distribution $\pi_D(\theta|y_{n_1})$. In the first case we have again the predictive distribution of Equation (8), while in the second case we have

$$m_D(y_{n_2}|y_{n_1}) = N\left(y_{n_2} | \frac{\mu_D n_D + n_1 y_{n_1}}{n_D + n_1}, \sigma^2 \left(\frac{1}{n_D + n_1} + \frac{1}{n_2} \right) \right).$$

References

[1] Gajewski B.J., Mayo M.S.: Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. *Statistics in Medicine* 2006; **25**, 2554–2566.

[2] Wang M.D.: Sample size re-estimation by Bayesian prediction. *Biometrical Journal* 2006; **48**, 5, 1–13.

[3] Armitage P., Berry G. and Matthews J.N.S.: Statistical methods in medical research. IV Edition. Blackwell Science, 2002.

[4] Tan S.B. and Machin D.: Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine* 2002; **21**, 1991–2012.

[5] Mayo M.S., and Gajewski B.J.: Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Controlled Clinical* 2004 **25**, 2,157–167.

[6] Chaloner K. and Verdinelli I.: Bayesian experimental design: a review. *Statistical Science* 1995; **10**, 237–308.

[7] De Santis F.: Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association* 2006; **101**, 473, 278–291.

[8] Spiegelhalter D.J. and Freedman L.S.: A predictive approach to selecting the size of a clinical trial, base on subjective clinical opinion. *Statistics in Medicine* 1986; **5**, 1-13.

[9] Joseph L., du Berger R. and Belisle P.: Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16, 769–781.

[10] Das Gupta A.: Review of optimal Bayes designs. In *Design and Analysis of Experiments. Handbook of Statistics* 1996 **13**, 1099-1147.

[11] Joseph L. and Belisle P.: Bayesian sample size determination for normal means and difference between normal means. *The Statician*, 46, 209-226.

[12] Clarke B.S. and Yuan A.: A closed form expression for Bayesian sample sizes. *Annals of Statistics* 2006; **34**(3): 1293–1330

[13] De Santis F.: Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society, Ser. A* 2007; **170**, 1, 95–113.

[14] Brutti P. and De Santis F.: Avoiding the range of equivalence in Clinical trials: robust Bayesian sample size determination for credible intervals. *Journal of Statistical Planning and Inference* 2008; **138** 1577–1591.

[15] Brutti P., De Santis F., Gubbiotti S.: Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine* 2008; DOI: 10.1002/sim.3175.

[16] Spiegelhalter D.J., Abrams K.R., Myles J.P.: *Bayesian approaches to clinical trials and health-care evaluation* Wiley 2004.

[17] Etzioni R., Kadane J.B.: Optimal experimental design for anothers analysis. *Journal of the American Statistical Association* 1993; **88**, n.424, 1404–1411.

[18] O'Hagan A. and Stevens J.W.: Bayesian assessment of sample size for clinical trials for cost effectiveness. *Medical Decision Making* 2001; **21** 219–230.

[19] Wang M.D., Gelfand A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; **17**, 2, 193–208.

[20] Sahu, S.K. and Smith, T.M.F.: A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Ser. A (Statistics in Society)* 2006; **169**, (2), 235-253.

[21] Dignam J.J. , Bryant J., Wieand H.S., Fisher B., Wolmark N.: Early stopping of a clinical trial when there is evidence of no treatment benefit: Protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. Controlled Clinical Trials, 19, 575-88 (1998).