METRON (2013) 71:81–95
DOI 10.1007/s40300-013-0008-x

Robust Bayesian monitoring of sequential trials

P. Brutti · F. De Santis · S. Gubbiotti

Received: 21 October 2012 / Accepted: 1 March 2013 / Published online: 28 May 2013

Abstract In sequential experiments the sample size is not planned in advance. Data are progressively collected and a stopping rule based on the observed results is defined in order to terminate the study. In a Bayesian framework, it is straightforward to monitor an ongoing experiment looking at the posterior probability that a parameter of interest θ, belongs to a given set. Specifically, in this paper we focus on the context of phase II clinical trials, where θ represents treatment efficacy. The Bayesian stopping rule we adopt involves the posterior probability that θ exceeds a clinically relevant threshold. Moreover, we introduce a robust version of this criterion by replacing the single prior distribution with a class of prior distributions. A simulation study is performed to compare the average sample sizes of the robust sequential approach both with the sample sizes of the non robust approach and of the non sequential approach. An interesting result is that, when the class of prior distributions is sufficiently narrow, the average sample sizes of the robust sequential approach can be smaller than the non sequential sample sizes.

Keywords Clinical trials · ϵ-contamination priors · Robustness · Sample size · Sequential analysis

1 Introduction

In many experimental contexts data are collected steadily over a period of time, but it is common practice that the analysis is performed at the end of the experiment, once the preplanned sample size is reached. However, from a practical point of view, it is quite natural to monitor results during the data accumulation process and, consequently, to take into consideration the possibility of early stopping or modifying the ongoing study design, due to ethical, administrative and economic reasons. Relevant examples of application can be found in almost any area in which an experiment or a survey is carried out over a period of time and intermediate

P. Brutti · F. De Santis · S. Gubbiotti (✉)
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
e-mail: stefania.gubbiotti@uniroma1.it

analyses are planned. In the medical field this includes clinical trials, together with animal experiments and epidemiological studies; in industrial applications there are established sequential methods for acceptance sampling for quality control and for life testing in reliability. A comprehensive review of statistical methods for group sequential trials is provided by Jennison and Turnbull [18], which constitutes a milestone mainly in the frequentist literature, together with the previous works by Armitage [1], Pocock [21], Whitehead [28] and O'Brien and Fleming [19]. As pointed out in Whitehead [28], sequential methods are not as popular as standard fixed-sample techniques. This is due to some technical complications that basically concern the adjustment for multiplicity when (one or more) interim looks at the data are scheduled, yielding potential bias in statistical procedures, especially related to the control of type I error (see for instance Geller and Pocock [16] and the references therein). However, as discussed in Berry [6], a possibility to avoid multiplicity issues is that of adopting a Bayesian perspective as, for instance, in Freedman and Spiegelhalter [15] and Spiegelhalter et al. [23]. In Berry's words [7], indeed, there is "no price to pay for looking at the data" without waiting for the end of recruitment, since the techniques for inference both at the interim analyses and upon termination are much more straightforward, provided that the sequential design does not affect the Bayesian inference. A complete review on Bayesian adaptive methods can be found in the recent book by Berry et al. [8] with specific reference to the context of clinical trials.

In this work we consider Bayesian methods for the monitoring of sequential trials. Specifically we deal with the case of continuous endpoints. Typical examples of normally distributed data (possibly after a suitable transformation) are, for instance, tumour shrinkage (see the application in Sect. 3), blood pressure, lung function or concentration of some chemical in the blood. In particular, in phase II trials, when there is no control group, the parameter of interest is the mean response to a medical intervention to be compared with a clinically relevant value. However, if the goal of the trial is to compare pair-matched patients in terms of a given measurement or two measurements on the same patients (as for instance in crossover studies), we essentially retrieve the one-sample problem. Conversely, in placebo-controlled trials, the parameter of interest is a measure of comparison between treatment effects under the two arms, such as the difference in the mean responses or, in case of binary or survival data, the log odds ratio and the log hazard ratio respectively (using normal approximations, as explained in Spiegelhalter et al. [23]). For the sake of simplicity, in the following we will treat the one-sample problem, although the proposed methodology could be extended to the case of controlled trials, that also involves the issue of patients allocation. Here, we are only interested in evaluating the total study dimension and we assume that the goal of a phase II single-arm trial is to prove efficacy of a new experimental treatment. In practice, data are collected gradually, starting from the inclusion date, during a follow-up period that can last several months or years and the number of observations progressively increases until the requirement of a predefined stopping rule is fulfilled. As in Spiegelhalter et al. [23], we consider a Bayesian stopping rule based on the posterior probability that a treatment effect exceeds a minimum relevant clinical threshold. Our first goal is to illustrate the main advantage of sequential procedures, i.e. the fact that the average number of patients required in the study is smaller than for non sequential criteria. Therefore, we compare the expected sample sizes of sequential methods with the optimal sample sizes of non sequential methods in a simulation study. Moreover, we consider the issue of sensitivity to the prior choice that was addressed in Brutti and De Santis [9], Brutti et al. [10] and De Santis [13] by replacing a single prior distribution with a class of priors. The goal is to assess the impact of prior information on pre-posterior analysis and, consequently, on the choice of the optimal number of observations. Here we extend the robust non sequential sample size determination criterion in a *sequential direction*. The performance of the proposed *sequential robust criterion* is evaluated in terms of expected number of observations, that is compared via simulation to the optimal sample sizes of (sequential and non sequential) non robust methods.

The outline of this paper is as follows. In Sect. 2.1 we describe the general set up and we introduce notation. After pointing out the differences between the conditional approach and the predictive approach, in Sect. 2.2 we provide details of the Bayesian stopping rule that constitutes the starting point for the derivation of its robust version (Sect. 2.3). Comparisons between the sample sizes obtained using sequential and non sequential, robust and non robust criteria discussed in Sect. 2.4, are further illustrated by a simulation study (Sect. 3.3), based on a real application set up regarding a phase II trial in which the continuous endpoint is tumour shrinkage. Finally, Sect. 4 contains some concluding remarks.

2 Bayesian stopping rules for sequential trials

2.1 Preliminaries

Let us consider a phase II trial with the objective of evaluating the efficacy of a new experimental treatment. Let us assume that the parameter of interest θ represents a real-valued measure of efficacy, large values of θ denoting benefit of the new treatment. We assume that groups of patients are sequentially accrued and evaluated for response. In order to terminate the trial, a stopping rule is defined according to the study objective. Hence, instead of prefixing an optimal sample size n^* based on a specified criterion, we assume to observe k_j individuals at each step of the sequential trial. We denote by $n_j = k_1 + \cdots + k_j$ the total number of observations up to step j. For practical reasons, we fix a maximum total sample size n_{max}. Without loss of generality in the following we take $k_j = 1$, for all $j = 1, \ldots, J$, i.e. we consider each patient sequentially. Let us denote by Y_{n_j} a continuous measure of treatment response, based on the first n_j patients. Furthermore, let y_{n_j} and $f(y_{n_j}; \theta)$ be the observed data and the corresponding likelihood function respectively, $j = 1, \ldots, J$.

In a Bayesian perspective, we can formalize pre-experimental knowledge on the phenomenon of interest by considering a prior distribution on θ, $\pi_A(\theta)$. Hence, from Bayes theorem, the posterior distribution of θ given the jth observed response is

$$\pi_A(\theta|y_{n_j}) \propto \pi_A(\theta) \cdot f(y_{n_j}; \theta). \tag{1}$$

Using iteratively (1), we update the information on θ as each value y_{n_j} is observed, for $j = 1, \ldots, J$, and we use the posterior distribution to define a stopping rule as described in the next section.

2.2 Sequential criterion

Let us first recall the sequential criterion illustrated in Spiegelhalter et al. [23]. Given the observed data y_{n_j}, let $P_{\pi_A,n_j}(\theta > \delta|y_{n_j})$ be the posterior probability that θ exceeds a minimally relevant clinical value δ. The treatment is declared successful if the experiment shows sufficiently strong evidence that $\theta > \delta$, i.e. if the probability of interest is larger than a given threshold $\gamma \in (0, 1)$. In summary, we proceed according to the following *stopping rule*: if the observed value y_{n_j} is such that

$$P_{\pi_A,n_j}(\theta > \delta|y_{n_j}) > \gamma \tag{2}$$

the trial *stops with success*, otherwise the $(j+1)th$ patient is enrolled. It may happen that condition (2) is not fulfilled before the maximum preplanned number of patients n_{max} is reached; in this case, the trial is terminated *without success*. More formally, let us introduce the random set

$$S_{\pi_A}(\delta, \gamma) = \left\{n_j \in \mathbb{N} : P_{\pi_A, n_j}(\theta > \delta | Y_{n_j}) > \gamma, \quad j = 1, \ldots, J\right\}.$$

For a given analysis prior, it contains all the integer numbers such that the random posterior probability of the event $(\theta > \delta)$ exceeds γ. Now the random number of observations N is defined as

$$N = \begin{cases} \min S_{\pi_A}(\delta, \gamma) & \text{if } S_{\pi_A}(\delta, \gamma) \neq \emptyset \\ n_{max} & \text{otherwise} \end{cases}.$$

Since it is not possible to derive the distribution of N analytically, we resort to simulation to provide numerical examples in Sect. 3.3. In particular, we are interested in comparing the expected value of N with the optimal sample size n^* that is obtained by the corresponding non-sequential criterion introduced in Brutti et al. [10], i.e.

$$n^* = \min\left\{n \in \mathbb{N} : \ \mathbb{E}\left(P_{\pi_A}(\theta > \delta | Y_n)\right) > \gamma\right\}, \tag{3}$$

where $\mathbb{E}(\cdot)$ is the expected value computed with respect to the distribution of Y_n (see Sect. 2.2.1 for details on the distribution of the data). According to Spiegelhalter et al. [23], we expect that the sequential procedure allows one to save observations with respect to the corresponding non sequential criterion, that is $\mathbb{E}(N) \leq n^*$ (see Sect. 2.4 for discussion).

2.2.1 Conditional approach or predictive approach?

Before introducing a robust version of the sequential criterion of Sect. 2.2, we discuss the data drawing mechanism for simulating the distribution of N. Two alternative approaches are briefly described below.

- *Conditional approach.* Data can be drawn sequentially from the sampling distribution $f(\cdot; \theta_D)$, where θ_D is a design target value for treatment effect. For instance, in superiority trials, θ_D is chosen among those values of the parameter denoting an effective treatment (i.e. values larger than δ).
- *Predictive approach.* Data can be drawn sequentially from the marginal distribution, i.e.

  $$m_D(y_n) = \int_{\Theta} f(y_n; \theta)\pi_D(\theta) d\theta,$$

  where the prior distribution π_D on θ (design prior) accounts for additional uncertainty involved in the choice of the design value θ_D. Notice that π_D must be a proper distribution in order to have m_D well defined. Moreover, in the special case in which π_D is a point-mass distribution centred on θ_D, we retrieve the sampling distribution $f(\cdot; \theta_D)$ and we actually go back to the conditional approach.

We refer to De Santis [13], O'Hagan and Stevens [20], Wang and Gelfand [25] for more detailed discussion on these approaches. Before ending this section we stress the importance of the distinction between the analysis prior π_A and the design prior π_D. Although most of Bayesian sample size determination methods make use of one prior distribution for computing both the posterior distribution and the marginal distribution, in general π_D and π_A can be differently specified, as argued by several authors (see for instance, De Santis [13], Etzioni and Kadane [14], O'Hagan and Stevens [20], Tsutakawa [24], Wang and Gelfand [25]). Here, we just recall the main difference between the two distributions, justified by their different role in pre-posterior analysis. For further discussion we refer to Brutti et al. [10] and De Santis [13] and the references therein.

- The analysis prior (π_A) models pre-experimental *information* on θ that one wants to account for in determining the posterior distribution. One of the most common choices is to base prior elicitation on previous studies results, but it is also possible to use the analysis prior to formalize the subjective opinion of experts on the phenomenon of interest. However, incorporation of "external" evidence on final inference has been often criticized. The most straightforward solution is that of using noninformative analysis priors (see, for instance Wang and Gelfand [25]). Alternatively, De Santis [13] suggests to resort to a robust approach. Specifically, in next section we consider classes of priors instead of single prior distributions for θ.
- The design prior distribution (π_D) models *uncertainty* on the design value for θ and is used to obtain the marginal predictive distribution for pre-posterior computations. Since π_D represents the design scenario we assume when planning the trial, it is convenient to specify a prior that it is well concentrated on the values of θ representing the goal of the trial, as suggested in Wang and Gelfand [25]. Consequently, in our setting, the design prior should assign large probability to values of θ larger than δ.

In the present paper, we consider the predictive approach, based on two distinct prior distributions.

2.3 Robust sequential criterion

The use of a robust Bayesian approach is motivated by one of the most criticized features of Bayesian methods: the necessity of eliciting a specific prior distribution for posterior analysis. In order to assess the impact of the choice of the prior distribution we proceed as follows: (i) we replace the single prior by a class of distributions that gives a more flexible and realistic representation of pre-experimental knowledge, (ii) we study changes in posterior inference as the prior varies over the class. General principles of the robust Bayesian approach are discussed in Berger [2,3], Berger et al. [5], Wasserman [27]. Applications to clinical trials are in Carlin and Perez [11], Carlin and Sargent [12], Greenhouse and Wassernan [17], while Brutti and De Santis [9], Brutti et al. [10], De Santis [13] are specifically centred on robust sample size determination. The general idea is that if the range of variations of posterior quantities of interest is small (as the prior varies in the class), then one can use the single prior, relying on the robustness of the final conclusions. Conversely, if differences between the various priors in the class are relevant, one should be aware of the sensitivity of the posterior results to the prior choice and consequently refine prior knowledge. In our problem, we are mainly concerned with robustness with respect to the analysis prior, π_A. Therefore, in order to take into account the uncertainty involved in its specification, we consider a class of prior distributions Γ_A instead of a single prior π_A. In this way, we can derive a *robust* version of the sequential criterion of Sect. 2.2 by extending the stopping rule based on condition (2) as follows: we stop the trial at step j if the observed y_{n_j} is such that

$$\inf_{\pi_A \in \Gamma_A} P_{\pi_A, n_j}(\theta > \delta | y_{n_j}) > \gamma, \quad j = 1, \dots, J, \tag{4}$$

otherwise the recruitment proceeds to the $(j+1)th$ patient and so on. The idea is that we stop the trial at step j only if, as π_A varies in Γ_A, the minimal evidence in favor of the new treatment (measured by $\inf_{\pi_A \in \Gamma_A} P_{\pi_A, n_j}(\theta > \delta | y_{n_j})$) is sufficiently large. If criterion (4) is never fulfilled the trial stopsafter n_{max} observations and the treatment is declared ineffective.

Denote by $S_{\Gamma_A}(\delta, \gamma) = \left\{n_j \in \mathbb{N} : \inf_{\pi_A \in \Gamma_A} P_{\pi_A, n_j}(\theta > \delta | Y_{n_j}) > \gamma,\ j = 1, \dots, J\right\}$. Then the random number of patients N_Γ associated to the robust stopping rule is defined as

$$N_\Gamma = \begin{cases} \min S_{\Gamma_A}(\delta, \gamma) & \text{if } S_{\Gamma_A}(\delta, \gamma) \neq \emptyset \\ n_{max} & \text{otherwise} \end{cases}$$

This robust sequential criterion yields sample sizes that, on average, are larger than those determined with the non robust sequential procedure. Finally, let us recall that the robust version of the non sequential criterion (3) is given by

$$n^*_\Gamma = \min\left\{ n \in \mathbb{N} :\ \mathbb{E}\left(\inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta | Y_n) \right) > \gamma \right\}, \tag{5}$$

A specific choice for Γ_A *is the class of* ϵ *-contamination prior distributions*, widely studied in the literature on Bayesian robustness (see among others Berger and Berliner [4], Sivaganesan and Berger [22]). It is defined as

$$\Gamma_\epsilon = \{\pi : \pi(\theta) = (1 - \epsilon)\pi_A + \epsilon q; q \in Q\}$$

where π_A is a base prior distribution, $\epsilon \in [0, 1]$ is the level of contamination and Q is a conveniently chosen class of distributions. In the most general case, Q is the *class of all distributions* and can be regarded as a worst case, although other choices could be reasonable. However, as discussed in Brutti et al. [10] in the specific context of sample size determination, small differences with respect to the non robust case have been encountered when considering other contaminant classes, such as unimodal distributions or unimodal symmetric distributions. In our specific set up this would make the comparison with the fixed prior approach less interesting. From a technical point of view, in order to calculate the inferior bound of the posterior probability involved in criterion (4), the results of Sivaganesan and Berger [22] can be exploited, as discussed in details in Brutti et al. [10] with reference to the normal case.

2.4 Comparisons

In this section we compare the sample sizes obtained using sequential and non sequential, robust and non robust criteria. The main relationships are summarized in Fig. 1. First of all, let us focus on the vertical direction. As anticipated in Sect. 2.2, if we adopt a sequential procedure the study dimension is on average smaller than the optimal non sequential sample size, i.e. $\mathbb{E}(N) \leq n^*$. A similar relationship holds for robust criteria, that is $\mathbb{E}(N_\Gamma) \leq n^*_\Gamma$. Let us look now at the rows of the table: the robust approach yields larger values of the sample size, regardless of the criterion being sequential or not. Indeed, as discussed in Brutti et al. [10], when planning a non sequential trial, using a robust approach we actually account for

Fig. 1 The chart summarizes the relationships between sequential and non sequential, robust and non robust sample sizes

	non robust		robust
non sequential	n^*	$\leq$	n^*_Γ
	$\geq$	$\nwarrow\searrow$	$\geq$
sequential	$\mathbb{E}(N)$	$\leq$	$\mathbb{E}(N_\Gamma)$

additional uncertainty in the analysis prior specification and this implies an increase in the number of required observations, that is $n^* \leq n^*_\Gamma$. Moreover, as the "amplitude" of the class of priors increases, the optimal robust sample sizes n^*_Γ become larger and larger. As we will show by simulation in Sect. 3 analogous considerations also apply to the sequential case, i.e. $\mathbb{E}(N) \leq \mathbb{E}(N_\Gamma)$.

The previous remarks do not describe exhaustively all the possible comparisons displayed in Fig. 1. It is interesting to investigate, in fact, the relationship between the non sequential non robust sample size, n^*, and the expected number of observations required by the sequential robust criterion, $\mathbb{E}(N_\Gamma)$. Depending on the choice of the class of prior distributions, the latter can even entail an advantage with respect to the former, in terms of observations saving. This will be illustrated by the example of Sect. 3.3. In particular, working with ϵ-contamination classes offers an interesting key to analyse this comparison: we can assess the maximal amount of contamination ϵ so that $\mathbb{E}(N_\Gamma) \leq n^*$, i.e. the maximal level of contamination that makes the sequential robust approach convenient with respect to the non sequential non robust approach. More formally, we define

$$K(\epsilon) = \frac{n^*}{\mathbb{E}(N_\Gamma)}$$

and study its behaviour as a function of ϵ. Since, as argued before, $\mathbb{E}(N_\Gamma)$ is larger for wider classes of prior distributions Γ_ϵ, $K(\epsilon)$ decreases for increasing levels of contamination ϵ (see for instance Fig. 6 in Sect. 3.3). In particular, we are interested in determining the *critical level* $\tilde{\epsilon}$, such that $K(\tilde{\epsilon}) = 1$ or, equivalently, $\mathbb{E}(N_\Gamma) = n^*$, i.e. the level of contamination that makes the two criteria equivalent in terms of required number of patients. In summary, if $\epsilon < \tilde{\epsilon}$, then $K(\epsilon) > 1$ and we conclude that using a sequential procedure allows us to keep the average required number of observations smaller than n^*, even if we are introducing in the analysis prior specification a certain amount of uncertainty, quantified by ϵ.

3 The case of normal endpoints

3.1 Results for normal endpoints

In this section we explicit the results of Sect. 2.1 referring to the case of normal likelihoods. We suppose that the measure of treatment efficacy Y_{n_j}, based on the first n_j patients, is normally distributed with mean θ and known variance σ^2/n_j. Moreover, for computational convenience, the most natural choice for the analysis prior π_A is a conjugate prior distribution with respect to the normal model. Hence, we assume for θ a normal density of mean θ_A and known variance σ^2/n_A. Following the notation of Spiegelhalter et al. [23], we refer to n_A as to the prior sample size , i.e. the weight of prior information. Then the posterior distribution of Eq. (1) is

$$\pi_A(\theta|y_{n_j}) = N\left(\theta \left|E_{n_j}, V_{n_j}\right.\right), \tag{6}$$

where $N(\cdot|a, b)$ denotes a normal density of mean a and variance b and

$$E_{n_j} = \frac{n_A\theta_A + n_j y_{n_j}}{n_A + n_j} \quad \text{and} \quad V_{n_j} = \frac{\sigma^2}{n_A + n_j},$$

are the posterior expectation and the posterior variance of θ. Consequently, the probability involved in condition (2) is simply given by

$$P_{\pi_A,n_j}(\theta > \delta | y_{n_j}) = 1 - \Phi\left(\frac{\delta - E_{n_j}}{\sqrt{V_{n_j}}}\right) \tag{7}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (c.d.f.) of the standard Normal random variable.

Finally, as discussed in Sect. 2.2.1, we need to specify a design prior to model uncertainty on the design value for θ. For the sake of simplicity, we adopt a normal design prior of mean θ_D and prior sample size n_D. From standard results on conjugate analysis, this yields as a marginal distribution of the data $m_D(\cdot) = N(\cdot|\theta_D, \sigma^2(n^{-1} + n_D^{-1}))$.

3.2 Example: monitoring of a phase II cancer trial

In this section, we consider an example based on Wason et al. [26], where the primary continuous endpoint of a two-stage phase II cancer trial is the measurement of tumour shrinkage. Our purpose is to illustrate the methodology described in Sect. 3, using the set up of Wason et al. [26] as a basis to fix sensible values for the design parameters and for the required prior assumptions. The primary endpoint is the percentage decrease in the sum of lesion diameters, that is assumed to be normally distributed with mean θ and known variance σ^2. A positive value of the endpoint represents shrinkage in tumour size: the larger the percentage reduction, the more effective the treatment. The minimally clinical relevant reduction δ is set equal to 10 (based on the alternative hypothesis of the original example).

We start considering a fictitious dataset of observed responses for 200 patients and we assume the data to be collected sequentially. In practice, the dataset was actually simulated under a relatively enthusiastic scenario, by setting $\theta_D = 12$, $\sigma^2 = 20$, $n_D = 10$. Note that this is the same scenario used in the simulation study of the next section. Moreover, we elicit a normal prior distribution centered on $\theta_A = 3$, corresponding to an almost negligible shrinkage. In order to have a quite flat prior density, we assume a prior sample size as small as $n_A = 1$. This analysis prior expresses scepticism about treatment benefit: specifically it assigns 25 % chance to negative values of the reduction, i.e. increase in tumour mass, as represented by the black area in Fig. 2. On the other hand, a priori the probability that θ exceeds δ, highlighted in grey in Fig. 2, is pretty small and equal to 0.06. Finally, the required threshold on the posterior probability scale is equal to $\gamma = 0.8$.

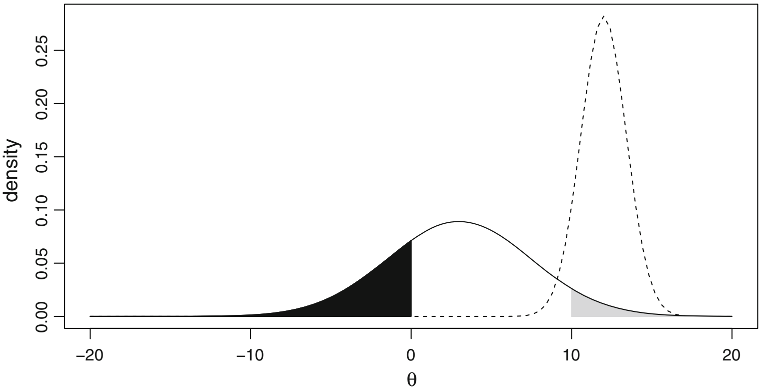

Fig. 2 Analysis (*continuous curve*) and design (*dashed curve*) prior distributions

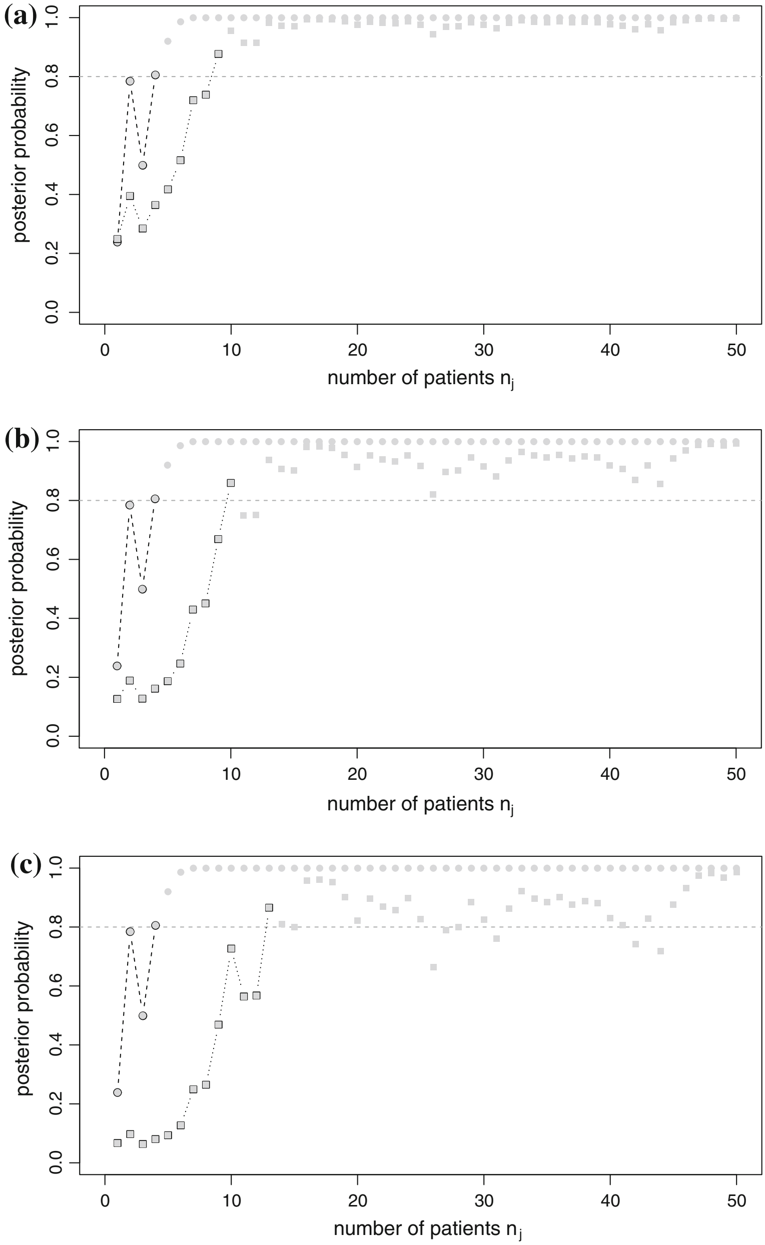

Fig. 3 Posterior probability $P_{\pi_A,n_j}(\theta > \delta|y_{n_j})$ (*circles*) and inferior bound of the posterior probability $\inf_{\pi_A \in \Gamma_A} P_{\pi_A,n_j}(\theta > \delta|y_{n_j})$ (*squares*) w.r.t. the sequentially increasing number of patients (up to the first 50 patients of the dataset) for **a** $\epsilon = 0.1$, **b** $\epsilon = 0.3$, **c** $\epsilon = 0.5$

Now we can proceed as described in Sect. 2.2: the trial stops as soon as we have evidence of efficacy, otherwise we continue up to the maximum number of patients, in this case the total size of our fictitious dataset, $n_{max} = 200$. Results are presented in Fig. 3: the posterior probability that $\theta > \delta$ (black circles) is sequentially updated until it exceeds the threshold

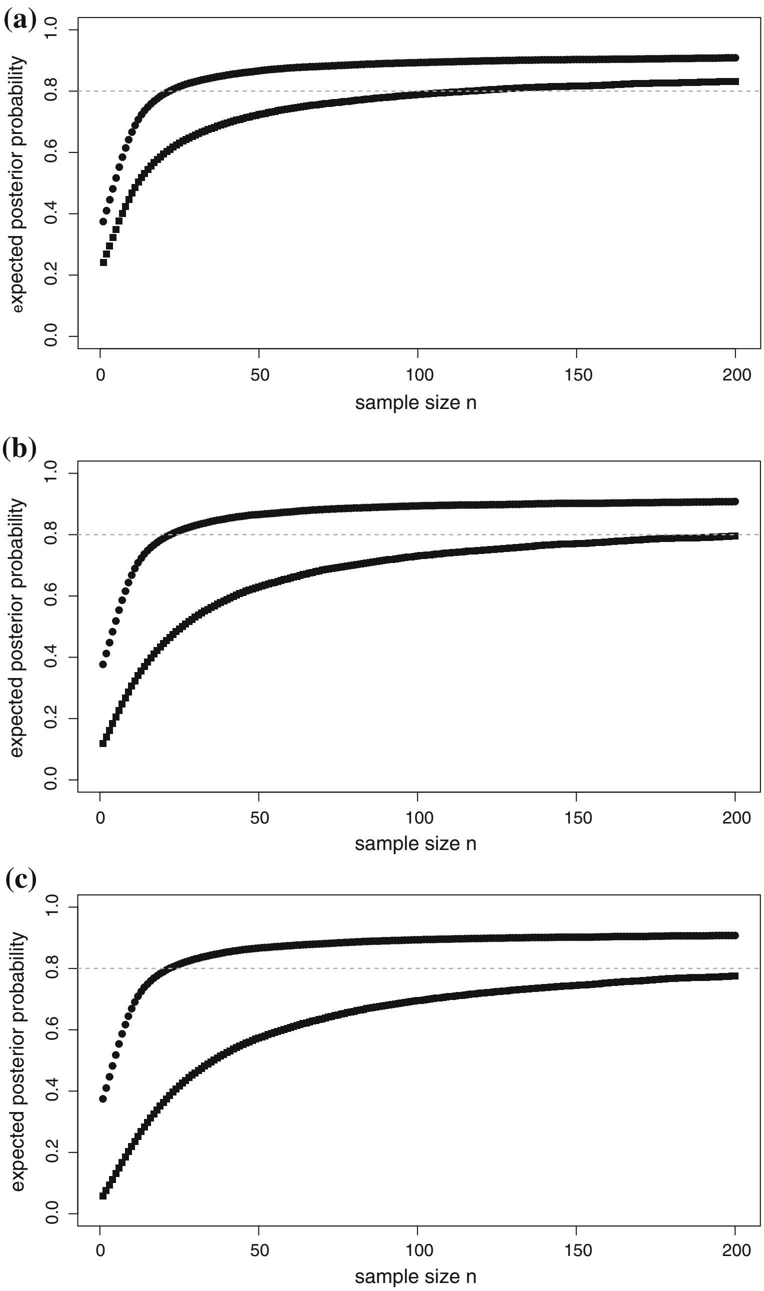

Fig. 4 Predictive expected posterior probability as a function of the sample size using both the non robust criterion (*circles*) and the robust criterion (*squares*) respectively, with Γ_ϵ, for **a** $\epsilon = 0.1$, **b** $\epsilon = 0.3$, **c** $\epsilon = 0.5$

γ, that is after the $4th$ patient is examined. Since condition (2) is fulfilled, the trial reaches success and is terminated. Adopting the robust version (4) of the sequential criterion, using a class Γ_ϵ with $\epsilon = (0.1, 0.3, 0.5)$, the required number of patients to satisfy the stopping rule increases to (9, 10, 13) respectively, as shown in the three panels of Fig. 3 from top to bottom.

In Fig. 4 we display the predictive expectation of the posterior probability involved in the non sequential criterion defined by (3) as a function of n. Given a threshold $\gamma = 0.8$, we obtain $n^* = 22$. Therefore, in this case the sequentially selected sample size (3) is smaller than n^*. Moreover, as expected, with the robust approach the required number of observations increases: in this example $n^*_\Gamma = (109, 197)$ for $\epsilon = (0.1, 0.3)$ and the optimal robust sample size even exceeds 200 units for $\epsilon = 0.5$. In the next section we show by simulation that these relationships hold in general, regardless of the criterion being sequential or not.

3.3 Simulation study

In this section we compare the sample sizes obtained using sequential and non sequential, robust and non robust criteria. We consider a simulation study under the setting described in Sect. 3.2. As pointed out in Sect. 2.2, we adopt a predictive approach and the data are drawn from the marginal distribution m_D. First of all we need to specify a design prior: for illustrative purposes we consider a normal density with $\theta_D = 12$, $\sigma^2 = 20$, $n_D = 10$, displayed in Fig. 2 together with the analysis prior. Hence, we simulate a large number of datasets, say $M = 10000$, and for each given dataset we apply the previously described sequential procedure. This yields M simulated values of N and N_Γ depending on the stopping rule (2) and on its robust version (4) respectively. The simulated distributions of the random variables N (light grey) and N_Γ (dark grey) are represented in Fig. 5, for different choices of the level of contamination. As expected, we have $\mathbb{E}(N) < \mathbb{E}(N_\Gamma)$: for instance, when $\epsilon = 0.1$ the expected value is $\mathbb{E}(N) = 7$ for the non robust criterion and $\mathbb{E}(N_\Gamma) = 28$ for the robust criterion. Moreover, we notice that, by increasing the level of contamination ϵ, the histogram of N_Γ is shifted towards larger values, and we consequently obtain larger and larger values of $\mathbb{E}(N_\Gamma)$, as reported in Table 1. We also notice that, as ϵ increases, the variability of the distribution is inflated. Hence the wider Γ_ϵ (namely the larger its contamination level ϵ), the larger the value of $\mathbb{E}(N_\Gamma)$ is. As discussed in Sect. 2.4, this behaviour is consistent with the result highlighted in Brutti et al. [10] for non sequential criteria. Table 1 also compares the values of $\mathbb{E}(N_\Gamma)$ with the corresponding optimal non sequential sample sizes n^*_Γ. Notice that, to be fair, we have considered the same maximum number of observations, in this case $n_{max} = 200$, for both criteria.

So far we have retrieved the main four relationships summarized in Fig. 1. The last but the most interesting comparison is the one between the non sequential non robust sample size, n^*, and the sequential robust sample size, $\mathbb{E}(N_\Gamma)$. Figure 6 shows the behaviour of $K(\epsilon)$ as a function of ϵ: for those values of ϵ such that $K(\epsilon) > 1$, the robust sequential sample size is smaller than n^*, whereas for increasingly wide classes Γ_ϵ, $K(\epsilon)$ decreases up to values smaller than 1. Now, we are interested in determining the critical level of contamination $\tilde{\epsilon}$, that is the amount of contamination such that $\mathbb{E}(N_\Gamma)$ and n^* coincide. Here we have $\tilde{\epsilon} = 0.06$: this value determines the largest class of ϵ-contaminated prior distributions yielding a robust (average) sequential sample size as small as the non robust non sequential one. In practice, this means that, for levels of contamination smaller than $\tilde{\epsilon}$, working sequentially we can afford a robust procedure, that is to say we pay the same price in terms of required observations.

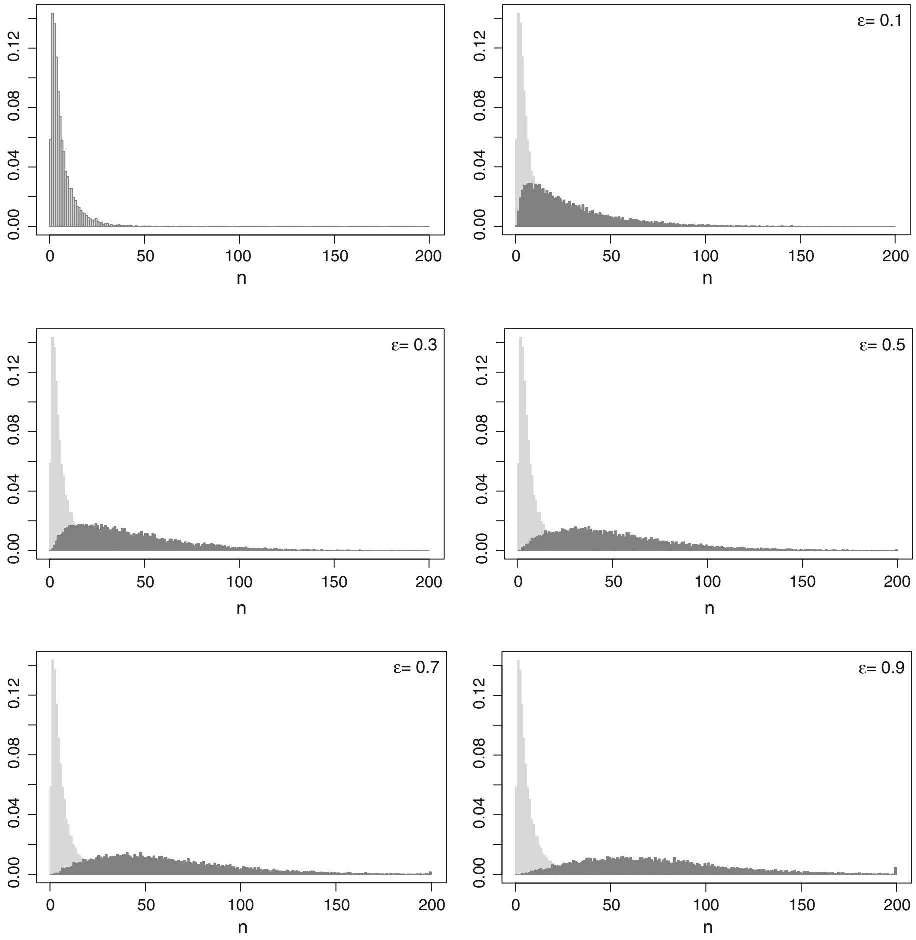

Fig. 5 The simulated distribution of N (*light grey*) is compared with the simulated distribution of N_Γ (*dark grey*), with $\Gamma = \Gamma_\epsilon$ for several choices of ϵ

Table 1 Optimal sample sizes for increasing levels of contamination using sequential and non sequential robust criteria, with $\theta_D = 4$, $n_D = 8$, $\sigma^2 = 4$, $\theta_A = 2.5$, $n_A = 10$, $\gamma = 0.8$

ϵ	0	0.1	0.3	0.5	0.7	0.9
$\mathbb{E}(N_\Gamma)$	7	28	42	52	60	74
n^*_Γ	22	109	197	>200	>200	>200

In other words, with the sequential approach the additional uncertainty introduced in the prior, until the level $\tilde{\epsilon}$, that does not imply a larger number of observations with respect to the non sequential single-prior approach. This provides an example of the idea anticipated in Sect. 2.4: when the class of priors Γ_ϵ is *sufficiently small*, the robust sequential criterion allows one to save observations (on average) with respect to the non robust and non sequential optimal sample size.

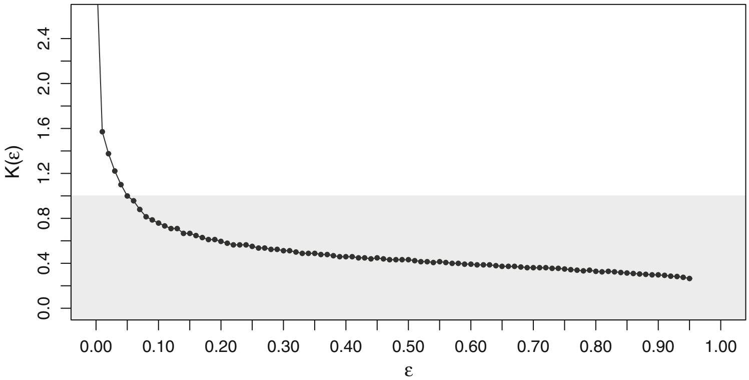

Fig. 6 $K(\epsilon)$ is plotted with respect to ϵ. The *critical level of contamination* is $\tilde{\epsilon} = 0.06$

4 Conclusions

One of the major advantages of Bayesian methods with respect to standard frequentist methods is its flexibility. As discussed in Berry et al. [8], "Bayesian inferences are flexible in that they can be updated continually as data accumulate. For example, the reason for stopping a trial affects frequentist measures but not Bayesian inferences". This technical feature of the Bayesian approach, a consequence of its respect of the likelihood principle, has a strong impact on the practical application to the actual way of conducting a trial. The Authors assertion sounds quite definitive: "In a Bayesian approach, a sample size need not be chosen in advance; before a trial, the only decision required is whether or not to start it." But they also warn that flexibility does not imply a total lack of constraints. Some kinds of deviations from the original plan are possible: the sample size can be adjusted, the drugs or devices involved can be modified, the definition of the patient population can change, etc. Such modifications can result in weaker conclusions, unless they are prespecified, as the Authors recommend, but "Bayesian analyses may still be possible in situations where frequentist analyses are not".

In this paper we have focused on the specific aspect of sample size, showing how a sequential procedure allows early termination only when there is evidence of treatment efficacy and how it enables the experimenter to reach an earlier conclusion than in a typical study with fixed sample size. This is very natural in a Bayesian context, since updating information on the parameter of interest as patients are enrolled, treated and evaluated for response, just translates in a sequential application of Bayes theorem and in a straightforward condition on a quantity of interest to be checked. An interesting extension of the proposed methodology could be a slight complication of the stopping rule, to include the possibility of early stopping for futility. This would allow to anticipate trial termination in case the ongoing results indicate a negative course that cannot be reverted even with extremely positive outcomes (see Spiegelhalter et al. [23] for details).

The main focus of our work is the introduction of a sequential procedure adopting a robust approach, in order to control the impact of the prior specification on the conclusions in terms of the required number of observations. However, the preplanned optimal sample size turns out to be inflated with respect to the non robust one and it sometimes becomes huge and therefore unreasonable (see Brutti et al. [10] for discussion). Here it comes the advantage of using a sequential procedure that, at the same time, allows one to deal with the issue of robustness, keeping the required number of observations feasible, indeed sparing experimental units with respect to the non sequential non robust method.

In future research, we plan to address the unknown variance case for normal endpoints. Moreover, it would be interesting, and particularly relevant for real applications, to extend this methodology to binary endpoints that typically characterize phase II clinical trials.

Acknowledgments We would like to thank Ludovico Piccinato for his helpful comments and suggestions on a previuos draft of the article.

References

1. Armitage, P.: Sequential Medical Trials. Blackwell, Oxford (1975)
2. Berger, J.O.: The robust Bayesian viewpoint (with discussion). In: Kadane, J. (ed.) Robustness of Bayesian Analysis. North-Holland, Amsterdam (1984)
3. Berger, J.O.: Robust Bayesian analysis: sensitivity to the prior. J. Stat. Plan. Inference **25**, 303–328 (1990)
4. Berger, J.O., Berliner, L.M.: Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Ann. Stat. **14**, 461–486 (1986)
5. Berger, J.O., Betro, B., Moreno, E., Pericchi, L.R., Ruggeri, F., Salinetti, G., Wasserman, L.: Bayesian Robustness, IMS Lecture Notes, Monograph Series, 29. IMS, Hayward (1996)
6. Berry, D.A.: Interim analyses in clinical trials: classical vs. Bayesian approaches. Stat. Med. **4**, 521–526 (1985)
7. Berry, D.A.: Interim analyses in clinical trials: the role of the likelihood principle. Am. Stat. **41**, 117–122 (1987)
8. Berry, S.M., Carlin, B.P., Lee, J.J., Muller, P.: Bayesian Adaptive Methods for Clinical Trials. Chapman & Hall/ CRC Biostatistics Series, London (2010)
9. Brutti, P., De Santis, F.: Avoiding the range of equivalence in Clinical trials: robust Bayesian sample size determination for credible intervals. J. Stat. Plan. Inference **138**, 1577–1591 (2008)
10. Brutti, P., De Santis, F., Gubbiotti, S.: Robust Bayesian sample size determination in clinical trials. Stat. Med. **27**, 2290–2306 (2008)
11. Carlin, B.P., Perez, M.E.: Robust Bayesian analysis in medical and epidemiological settings. In: Rios, D., Ruggeri, F. (eds.) Em Robust Bayesian analysis. Lecture Notes in Statistics, vol. 152. Springer, New York (2000)
12. Carlin, B.P., Sargent, D.J.: Robust Bayesian approaches for clinical trails monitoring. Stat. Med. **15**, 1093–1106 (1996)
13. De Santis, F.: Sample size determination for robust Bayesian analysis. J. Am. Stat. Assoc. **101**(473), 278–291 (2006)
14. Etzioni, R., Kadane, J.B.: Optimal experimental design for another's analysis. J. Am. Stat. Assoc. **88**(424), 1404–1411 (1993)
15. Freedman, L.S., Spiegelhater, D.J.: Comparison of Bayesian with group sequential methods for monitoring clinical trials. Controlled Clin. Trials **10**, 357–367 (1989)
16. Geller, N.L., Pocock, S.J.: Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. Biometrics **43**, 213–23 (1987)
17. Greenhouse, J.B., Wasserman, L.: Robust Bayesian methods for monitoring clinical trials. Stat. Med. **14**, 1379–1391 (1995)
18. Jennison, C., Turnbull, B.W.: Group sequential methods with applications to clinical trials. CRC press, Boca Raton (2001)
19. O'Brien, P.C., Fleming, T.R.: A Multiple testing procedure for clinical trials. Biometrics **35**(3), 549–556 (1979)
20. O'Hagan, A., Stevens, J.W.: Bayesian assessment of sample size for clinical trials for cost effectiveness. Med. Decis. Mak. **21**, 219–230 (2001)
21. Pocock, S.J.: Group sequential methods in the design and analysis of clinical trials. Biometrika **64**, 191–199 (1977)
22. Sivaganesan, S., Berger, J.O.: Ranges of posterior measures for priors with unimodal contaminations. Ann. Stat. **17**, 868–889 (1989)
23. Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian approaches to clinical trials and health-care evaluation. Wiley, New York (2004)
24. Tsutakawa, R.K.: Design of experiment for bioassay. J. Am. Stat. Assoc. **67**(339), 585–590 (1972)

25. Wang, F., Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Stati. Sci. **17**(2), 193–208 (2002)
26. Wason, J.M.S., Mander, A.P., Eisen, T.G.: Reducing the sample sizes in two-stage phase II cancer trials by using continuous tumour shrinkage end-points. Eur. J. Cancer **47**, 983–989 (2011)
27. Wasserman, L.: Recent methodological advances in robust Bayesian inference. In: Berger, J.O., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistic 4. Oxford University Press, Oxford (1992)
28. Whitehead, J.: The design and analysis of sequential clinical trials. Wiley, New York (1997)