# A Robust Bayesian Stopping Rule for monitoring Sequential Trials

P. Brutti , F. De Santis and S. Gubbiotti

December 2, 2011

**Abstract**

A standard Bayesian stopping rule for sequential trials is based on the posterior probability that a treatment effect exceeds a minimum relevant clinical threshold. In this paper we consider a robust version of this criterion by replacing the single prior distribution with a class of prior distributions. We 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. A surprising result is that, in some cases, the average sample sizes of the robust sequential approach are smaller than the non sequential sample sizes.

*Keywords*: Clinical trials, $\epsilon-$contamination priors, Robustness, Sample size determination, Sequential analysis.

## 1 Introduction

In clinical trials data are often collected gradually. Starting from the inclusion date, the follow-up period can last several months or years. Hence, results from patients recruited at the beginning of the trial become available for analysis and interpretation when enrolment of later subjects is still ongoing. It is common practice that the analysis is performed at the end of the trial when the preplanned total number of patients has been observed. However, the mechanism of data accumulation would make natural successive examinations (interim analysis) and the use of stopping rules with the purpose of early terminating the trial in case there is evidence of treatment efficacy (or, conversely, futility). In this sense, the conduct of the so-called *sequential clinical trials*, as defined by [1], "at any stage depends on the results so far obtained". The main ideas and methods

concerning the design and the analysis of the sequential clinical trials are reviewed in [2] which constitutes a fundamental reference in the frequentist context, together with [3] and [4]. As pointed out in [2], sequential methods are not as popular as one could figure out, due to some technical difficulties involved in their application, with respect to the fixed-sample techniques, much more popular in the standard analysis of clinical trials. These complications basically concern the adjustment for multiplicity when (one or more) interim looks at the data are scheduled (see for instance [5] and the references therein). In the present paper we adopt a Bayesian approach that allows us to avoid the typical drawbacks of frequentist methods related to multiplicity, as discussed by [6]. The application of Bayes theorem makes the information updating mechanism straightforward while data are sequentially accumulated. Several authors have dealt with the use of interim analyses in clinical trials and, more specifically, with sequential studies from this point of view (see for a comprehensive outlook [7], [8], [9] and [10]).

In this work we consider Bayesian methods for the monitoring of sequential trials. In particular, we restrict ourselves to phase II single-arm trials, such as efficacy trials, and we are only interested in the total study dimension. Although the general framework could be adopted for two arms randomized trials, the issue of patients allocation goes beyond the scope of this paper. As mentioned before, the main feature of sequential trials is that the total sample size (i.e. number of patients) is not fixed in advance. Hence, the number of observations progressively increases until the requirement of a predefined stopping rule is fulfilled. The main advantage of sequential procedures is that, in general, they require on average a smaller number of patients with respect to non sequential criteria. Specifically we consider a Bayesian method for the monitoring of sequential trials based on the posterior probability that a treatment effect exceeds a minimum relevant clinical threshold (as in [10]). We compare the sequential expected sample sizes with the optimal sample size of non sequential methods in a simulation study. Moreover, our purpose is to extend the robust (non sequential) sample size determination criteria presented in [11], [12] and [13] in a *sequential direction*. These papers deal with the issue of sensitivity to the prior choice, that is addresses by replacing a single prior distribution with a class of priors. The aim is to assess the impact of prior information on pre-posterior analysis and, consequently, on the choice of the optimal number of observations. We here adopt the same approach and we introduce a *sequential robust criterion*. The performance of this 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 Section 2 we describe the general set up and

we introduce notation. In Section 3.1 we provide details of the sequential method and we derive its robust version in Section 3.2, after pointing out the differences between the conditional approach and the sequential approach (Section 3.1.1). Comparisons between the sample sizes obtained using sequential and non sequential, robust and non robust criteria are discussed in Section 3.3 and further illustrated by a simulation study in Section 4.2, that is based on a real application regarding an efficacy trial on hypercholesterolemia. Finally, Section 5 contains some concluding remarks.

# 2 Preliminaries

Let us consider a phase II trial with the objective of establishing the efficacy of a new experimental treatment effect over a standard intervention (*superiority trial*). Let us assume that the parameter of interest $\theta$ represents a real-valued measure of treatment efficacy, large values of $\theta$ denoting superiority 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 $n^*$ using a specified criterion, we assume to observe an increasing number of individuals denoted by $n_j$, where $j = 1, \dots, J$ is the group index. For practical reasons, we fix a maximum total sample size $N_{max}$. Let $n_{j+1} = n_j + k_j$, with $k_j \geq 1$ the size of the *j-th* group and $n_J = N_{max}$. Without loss of generality in the following we take $k_j = 1$ for all $j = 1, \dots, J$, which simply means we consider each single patient sequentially. Let us denote by $Y_{n_j}$ a measure of treatment response based on the first $n_j$ patients that is supposed to be normally distributed with mean $\theta$ and variance $\sigma^2/n_j$, with a prefixed value for $\sigma^2$. Furthermore, let $y_{n_j}$ and $f(y_{n_j}; \theta)$ denote the observed data and the corresponding likelihood respectively, $j = 1, \dots, J$.

In a Bayesian perspective, we can formalize pre-experimental knowledge on the phenomenon of interest by considering a prior distribution on $\theta$, $\pi_A(\theta)$. For computational convenience, the most natural choice is a conjugate prior distribution with respect to the normal model. Hence, we assume for $\theta$ a normal density of mean $\theta_A$ and known variance $\sigma^2/n_A$, where, following the notation of [10], $n_A$ (prior sample size) expresses the "weight" of prior information. From Bayes theorem the posterior distribution of $\theta$ given the *j-th* observed response is

$$\pi_A(\theta|y_{n_j}) = N\left(\theta \,\middle|\, E_{n_j}, V_{n_j}\right), \tag{1}$$

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

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

are the posterior expectation and the posterior variance of $\theta$. Using iteratively (1) we update the information on $\theta$ as each value of the response $y_{n_j}$ is observed, for $j = 1, \dots, J$, and we use the posterior distribution to establish a stopping rule as proposed in the next section.

# 3 Bayesian stopping rules for sequential trials

## 3.1 Sequential criterion

In this section we first recall the sequential criterion described in [10]. Given the observed data $y_{n_j}$, let

$$P_{\pi_A, n_j}(\theta > \delta | y_{n_j}) = 1 - \Phi\left(\frac{\delta - E_{n_j}}{\sqrt{V_{n_j}}}\right) \tag{2}$$

be the posterior probability that $\theta$ exceeds a minimally relevant clinical value $\delta$, where $\Phi(\cdot)$ denotes the c.d.f. of the standard Normal random variable. The treatment is declared successful if the experiment shows sufficiently strong evidence that probability (2) is larger than a given threshold $\gamma \in (0, 1)$. Hence, we proceed according to the following **stopping rule**: if

$$P_{\pi_A, n_j}(\theta > \delta | y_{n_j}) > \gamma \tag{3}$$

the trial *stops with success*, otherwise the procedure is repeated for the $(j+1)$-*th* patient. It may happen that condition (3) is not fulfilled before the maximum preplanned number of patients $N_{max}$ is reached; in this case, the trial is terminated *without success*.

By adopting this sequential procedure, let $\mathbf{N}$ denote the random number of observations collected up to fulfilment of condition (3), i.e.

$$\mathbf{N} = \min\left\{n_j \in \mathbb{N} : \ P_{\pi_A, n_j}(\theta > \delta | Y_{n_j}) > \gamma, \, j = 1, \cdots, J\right\}. \tag{4}$$

Since it is not possible to derive the distribution of $\mathbf{N}$ analytically, to provide numerical examples in Section 4.2 we resort to simulation. In particular, we are interested in comparing the expected value of $\mathbf{N}$ with the optimal sample size that is obtained by the corresponding non-sequential criterion introduced in [11, 13], i.e.

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

where $\mathbb{E}(\cdot)$ is the expected value computed with respect to the distribution of $Y_n$ (see Section 3.1.1 for details on the distribution of the data). According to [10], we expect that the sequential procedure allows one to save observations with respect to the corresponding non sequential criterion, that is $\mathbb{E}(\mathbf{N}) \leq n^*$. This aspect will be further commented in Section 3.3.

### 3.1.1 Conditional approach or Predictive approach?

Before introducing a robust version of the sequential criterion of Section 3.1, a clarification is in order about the data drawing mechanism for simulating the distribution of $\mathbf{N}$. Two alternative approaches are briefly described below.

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

- *Predictive approach.* Data can be drawn sequentially from the marginal distribution, i.e.
$$m_D(y_n) = \int_\Theta f(y_n;\theta_D)\pi_D(\theta)d\theta,$$
where the prior distribution $\pi_D$ on $\theta$ (design prior) accounts for additional uncertainty involved in the choice of the design value $\theta_D$. Notice that $\pi_D$ must be a proper distribution in order to have $m_D$ well defined. Moreover, in the special case in which $\pi_D$ is a point-mass distribution centred on $\theta_D$, we retrieve the sampling distribution $f(\cdot;\theta_D)$ and we actually go back to the conditional approach.

We refer to [11], [14], [15] for more detailed discussion on these approaches. Before ending this section we now illustrate the possible distinction between the analysis prior $\pi_A$ and the design prior $\pi_D$. 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 $\pi_D$ and $\pi_A$ can be differently specified, as argued by several authors (see for instance, [11], [14], [15], [16], [17]). Here, we just recall the main distinctions between the two distributions, justified by their different role in pre-posterior analysis.

- The analysis prior ($\pi_A$) models pre-experimental information on $\theta$ 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 [15]). Alternatively, one can resort to a robust approach, as we suggest in next section, following [11]. Specifically, we consider classes of priors instead of single prior distributions for $\theta$.

- The design prior distribution ($\pi_D$) models uncertainty on the design value for $\theta$ and is used to obtain the marginal predictive distribution for pre-posterior computations. $\pi_D$ represents the design scenario we assume when planning the trial and it is required to be a proper distribution, otherwise $m_D$ is not well specified. Indeed, it is convenient to specify the design prior so that it is concentrated on the values of $\theta$ representing the goal of the trial, as suggested in [15]. For instance, in superiority trials the design prior assigns large probability to values of $\theta$ larger than $\delta$.

For further discussion on the distinction between the two prior distributions we refer to [11] and [13] and the references therein. In the present paper, we will consider the predictive approach: specifically in Section 4.2 we adopt a normal design prior, namely $\pi_D(\theta) = N(\theta|\theta_D, \sigma^2/n_D)$, which yields as a marginal distribution of the data $m_D(\cdot) = N(\cdot|\theta_D, \sigma^2\left(n^{-1} + n_D^{-1}\right))$.

## 3.2 Robust Sequential Criterion

The use of a robust approach is motivated by one of the most criticized features of Bayesian methods, that is the necessity of eliciting a specific prior distribution for posterior analysis. Then 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 [18, 19, 20, 21]. Applications to clinical trials are in [22, 23, 24, 25], while [11, 12, 13] are specifically centred on robust sample size determination. We recall here the general idea of the robust approach: if the range of posterior quantities of interest is small with respect to the various priors 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.

As mentioned above, in order to take into account the uncertainty involved in the specification of the prior distribution, we consider a class of prior distributions $\Gamma_A$ instead of a single prior $\pi_A$. In this way, we can derive a *robust* version of the sequential criterion of Section 3.1. The stopping rule based on condition (3) is extended as follows: we stop the trial at step $j$ if

$$\inf_{\pi_A \in \Gamma_A} P_{\pi, n_j}(\theta > \delta | y_{n_j}) > \gamma, \qquad j = 1, \dots, J \tag{6}$$

otherwise the recruitment proceeds to the $(j+1)$-*th* patient and so on. If criterion (6) is never fulfilled the trial stops after $N_{max}$ observations and the treatment is declared ineffective. Now, let us denote by $\mathbf{N}_\Gamma$ the random number of patients associated to the stopping rule in (6), i.e.

$$\mathbf{N}_\Gamma = \min \left\{ n \in \mathbb{N} : \inf_{\pi_A \in \Gamma_A} P_{\pi, n_j}(\theta > \delta | Y_{n_j}) > \gamma, \, j = 1, \dots, J \right\}. \tag{7}$$

This robust sequential criterion yields sample sizes that are uniformly larger than those determined with the non robust sequential procedure. Moreover we recall that the robust version of the non sequential criterion (5) is given by

$$n^*_\Gamma = \min \left\{ n \in \mathbb{N} : \, \mathbb{E} \left( \inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta | Y_n) \right) > \gamma \right\}, \tag{8}$$

In Section 3.3 we discuss the relationships between sequential and non sequential, robust and non robust sample sizes, whereas in Section 4.2 we illustrate the comparison by simulation results. In the next paragraph we consider a specific choice for the class $\Gamma_A$, i.e. the class of $\epsilon$-contamination prior distributions. This class has been widely studied in the literature on Bayesian robustness. See among others [26, 27, 13].

The **class of $\epsilon$-contamination prior distributions** is defined as follows

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

where $\pi_A$ 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, $Q$ can be the *class of all distributions*. Of course, other choices could be reasonable. However, as discussed in [13] 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, which would make the comparison with the fixed prior approach less interesting. Moreover, in general $Q$ can be regarded as a worst case. In order to calculate the inferior bound of the posterior probability involved in criterion (6), the results of [26] can be exploited, as discussed in details in [13].

## 3.3 Comparisons

In this section we compare the sample sizes obtained using sequential and non sequential, robust and non robust criteria. The main relationships are summarizes in **Figure 1**. First of all, let us focus on the vertical direction. As anticipated in Section 3.1, if we adopt a sequential procedure the study dimension is on average smaller than the optimal non sequential sample size, i.e. $\mathbb{E}(\mathbf{N}) \leq n^*$. A similar relationship holds for robust criteria, that is $\mathbb{E}(\mathbf{N}_\Gamma) \leq n^*_\Gamma$. Let us look now at each row of the scheme: the robust approach yields larger values of the sample size, regardless of the criterion being sequential or not. Indeed, as discussed in [13], when planning a non sequential trial, using a robust approach we actually account for 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, by considering increasingly wide classes of prior distribution we obtain larger values for the corresponding optimal robust sample sizes. As we will show by simulation in Section 4 analogous considerations also apply to the sequential case, i.e. $\mathbb{E}(\mathbf{N}) \leq \mathbb{E}(\mathbf{N}_\Gamma)$.

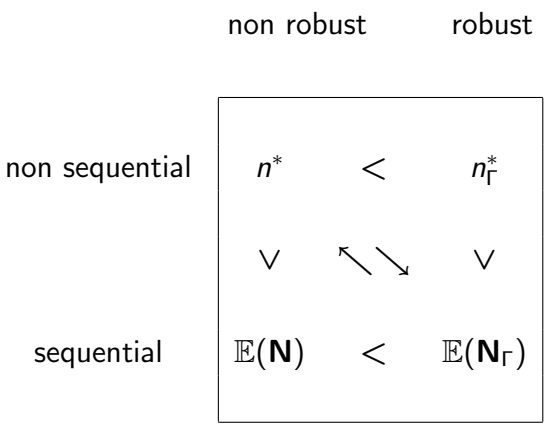


Figure 1: The chart summarizes the relationships between sequential and non sequential, robust and non robust sample sizes.

However, notice that previous remarks do not describe exhaustively all the possible

comparisons displayed in **Figure 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, $\mathbb{E}(\mathbf{N}_\Gamma)$. Depending on the choice of the class of prior distributions the latter can even entail an advantage in terms of observations saving with respect to the former. This will be illustrated by the example of Section 4.2. In particular, working with $\epsilon-$contamination classes offer an interesting key to analyse this comparison: we can assess the amount of contamination to be introduced when applying the sequential robust procedure, such that the number of observations is smaller than the non sequential non robust optimal sample size. In order to formalize this comparison, we define

$$K(\epsilon) = \frac{n^*}{\mathbb{E}(\mathbf{N}_{\Gamma_\epsilon})}$$

and study its behaviour as a function of $\epsilon$. Since, as argued before, $\mathbb{E}(\mathbf{N}_\Gamma)$ is larger for wider classes of prior distributions $\Gamma_\epsilon$, $K(\epsilon)$ decreases for increasing levels of contamination $\epsilon$ (see for instance **Figure** 5 in Section 4.2). In particular, we are interested in determining the **critical level** $\tilde{\epsilon}$, such that $K(\tilde{\epsilon}) = 1$ or, equivalently, $\mathbb{E}(\mathbf{N}_\Gamma) = n^*$, i.e. the level of contamination that makes the two criteria equivalent in terms of required number of patients. In summary, the interpretation of $K(\epsilon)$ is straightforward: 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 (that is quantified by $\epsilon$).

# 4 Example: efficacy trial on hypercholesterolemia

## 4.1 Monitoring a sequential trial

In this section, we show an example of an hypothetical efficacy trial based on [28] and [29] in which a new experimental treatment against hypercholesterolemia is considered. Let us suppose that treatment response is the reduction in total serum cholesterol in 4 weeks with respect to a baseline value (measured in $10^{-1}$ mmol/litre): the larger the reduction the more effective the treatment. Let us assume a normal distribution for the reduction. Let us also set the minimally clinical relevant reduction $\delta$ equal to 3, according to clinical experience.

In practice, in the following we consider a fictitious dataset of observed responses for 50 patients and we assume the data to be collected sequentially. Moreover, based on the results of previous studies, we elicit a normal prior distribution of parameters

$\theta_A = 2.5$, $\sigma^2 = 4$, $n_A = 10$ and we set a threshold on the posterior probability scale equal to $\gamma = 0.8$. Now we can proceed as described in Section 3.1: the trial stops as soon as we have evidence of efficacy, otherwise we continue up to the maximum number of patients, $N_{max} = 50$. Results are presented in **Figure 2**: the posterior probability that $\theta > \delta$ (black circles) is sequentially updated until it exceeds the threshold $\gamma$, that is after the 16-*th* patient is examined. Since condition (3) is fulfilled, the trial reaches success and is terminated. Adopting the robust version (6) of the sequential criterion, using a class $\Gamma_\epsilon$ with $\epsilon = (0.1, 0.3, 0.5)$, the required number of patients to satisfy the stopping rule increases to $(17, 25, 35)$ respectively, as shown in the three panels of **Figure 2** from top to bottom.

In **Figure 3** we display the predictive expectation of the posterior probability as a function of $n$. Given a threshold $\gamma = 0.8$, we obtain $n^* = 25$. Therefore, in this case the sequentially selected sample size (16) is smaller than $n^*$. Moreover, as expected using a robust approach increases the required number of observations: in this example $n^*_\Gamma = (39, 73)$ for $\epsilon = (0.1, 0.3)$ and the optimal robust sample size even exceeds 100 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.

## 4.2 Simulation study

In this section we illustrate by simulation the comparisons between the sample sizes obtained using sequential and non sequential, robust and non robust criteria. Let us consider a simulation study under the setting described in Section 4.1. As pointed out in Section 3.1, data are drawn from the marginal distribution $m_D$. First of all we need to specify a design prior: for illustrative purposes let us consider a normal density with $\theta_D = 4$, $\sigma^2 = 4$, $n_D = 8$. 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 $\mathbf{N}$ and $\mathbf{N}_\Gamma$ depending on the stopping rule (3) and on its robust version (6) respectively. The simulated distributions of the random variables $\mathbf{N}$ (light grey) and $\mathbf{N}_\Gamma$ (dark grey) are represented in **Figure 4**, for different choices of the level of contamination. As expected, we have $\mathbb{E}(\mathbf{N}) < \mathbb{E}(\mathbf{N}_\Gamma)$: for instance, when $\epsilon = 0.1$ (panel $(a)$) the simulated expected values is $\mathbb{E}(\mathbf{N}) = 15$ for the non robust criterion and $\mathbb{E}(\mathbf{N}_\Gamma) = 20$ for the robust criterion. Moreover, we notice that, by increasing the level of contamination $\epsilon$, the histogram of $\mathbf{N}_\Gamma$ tends to move towards larger values. In fact we obtain the values of $\mathbb{E}(\mathbf{N}_\Gamma)$ reported in **Table 1** for increasing levels of contamination $\epsilon$. From the histograms we also notice that, as

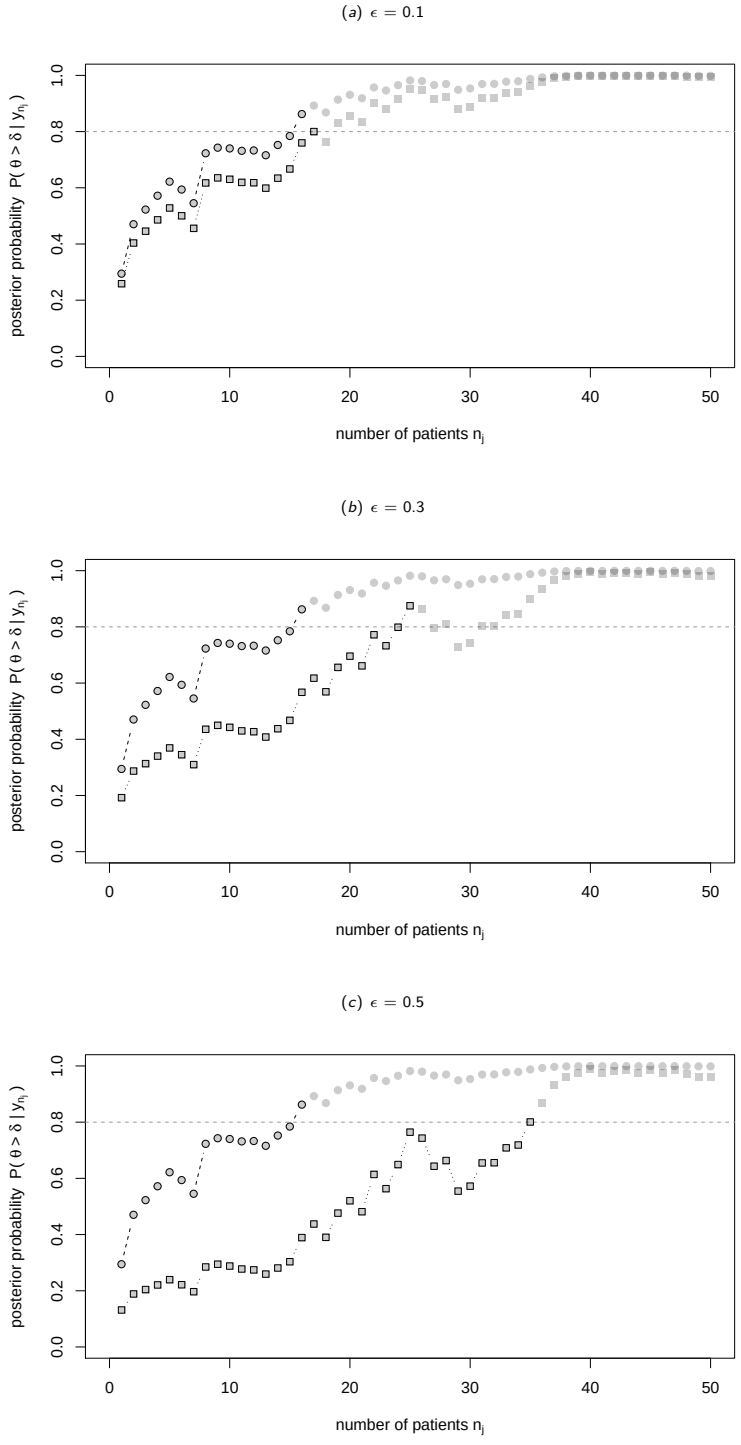


Figure 2: Posterior probability $P_{\pi_A}(\theta > \delta | y_{n_j})$ (circles) and inferior bound of the posterior probability $\inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta | y_{n_j})$ (squares) w.r.t. the sequentially increasing number of patients for $(a)$ $\epsilon = 0.1$, $(b)$ $\epsilon = 0.3$, $(c)$ $\epsilon = 0.5$.

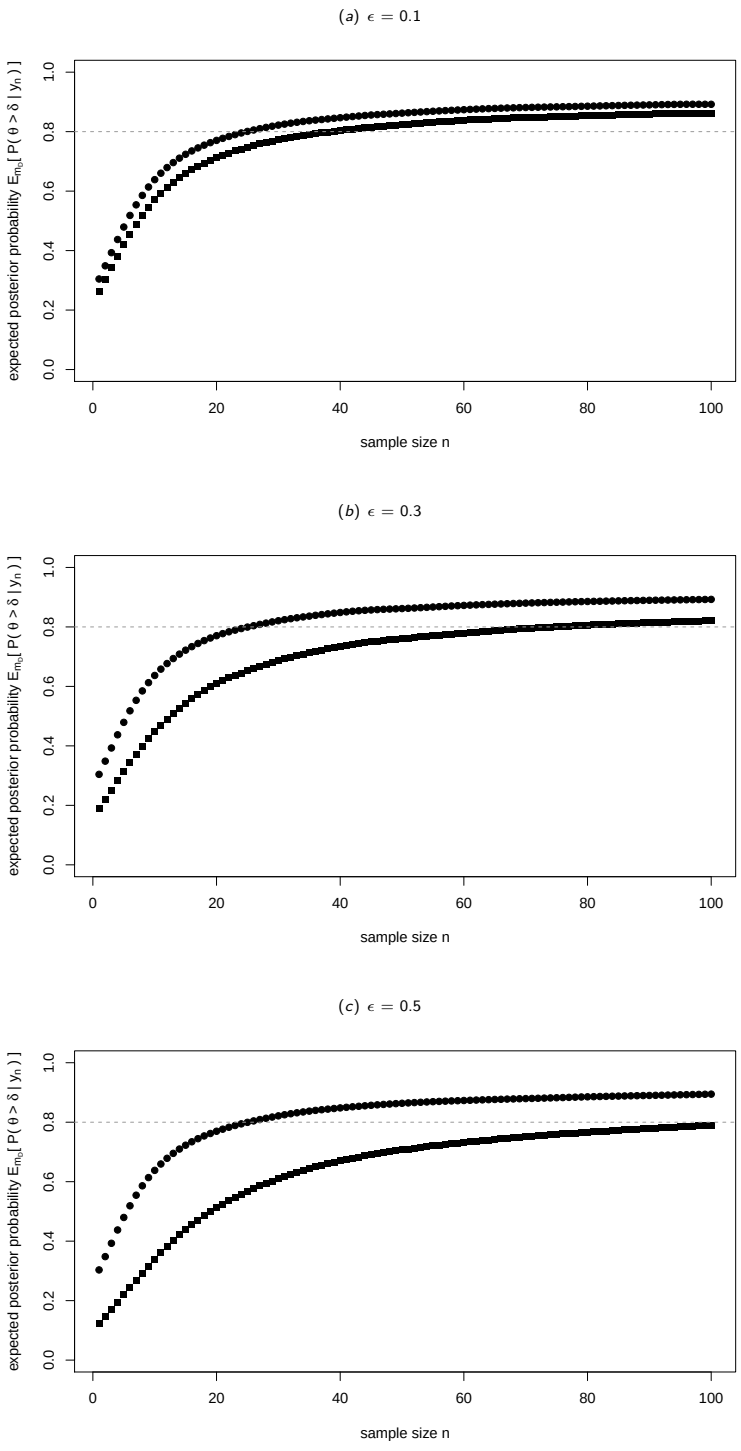


Figure 3: Predictive expected posterior probability as a function of the sample size using both the non robust criterion (circles) and the robust criterion (squares) respectively, with $\Gamma_\epsilon$, for $(a)$ $\epsilon = 0.1$, $(b)$ $\epsilon = 0.3$, $(c)$ $\epsilon = 0.5$.

$\epsilon$ increases, the variability of the distribution is inflated. Hence the wider $\Gamma_\epsilon$ (namely the larger its contamination level $\epsilon$), the larger the value of $\mathbb{E}(\mathbf{N}_\Gamma)$ is. As discussed in Section 3.3, this behaviour is consistent with the result highlighted in [13] for non sequential criteria. **Table 1** also compares the values of $\mathbb{E}(\mathbf{N}_\Gamma)$ with the corresponding optimal non sequential sample sizes $n^*_\Gamma$.

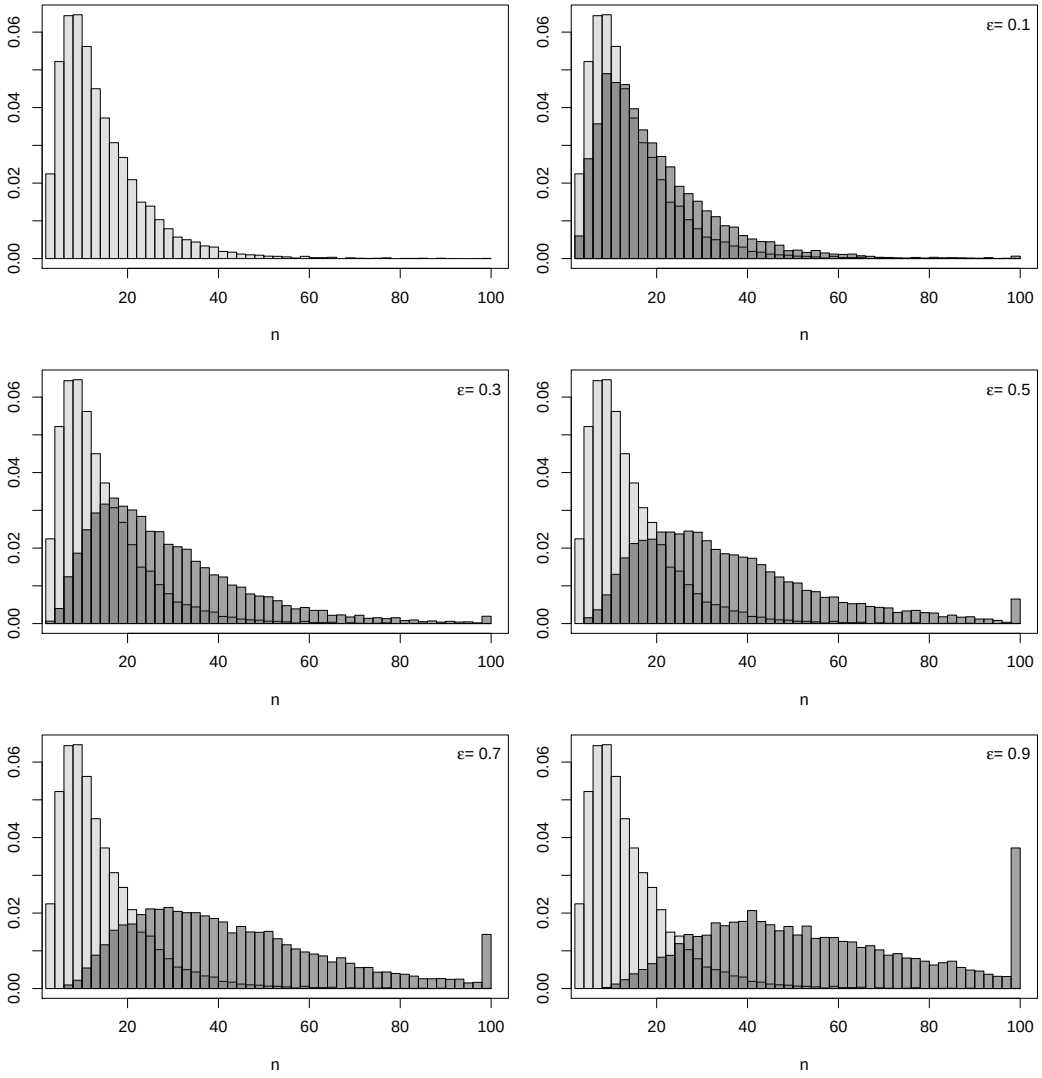


Figure 4: The simulated distribution of $\mathbf{N}$ (light grey) is compared with the simulated distribution of $\mathbf{N}_\Gamma$ (dark grey), with $\Gamma = \Gamma_\epsilon$ for several choices of $\epsilon$.

Up to this point we have retrieved in the example the main four relationships

| $\epsilon$ | **0** | **0.1** | **0.3** | **0.5** | **0.7** | **0.9** |
|---|---|---|---|---|---|---|
| $\mathbb{E}(\mathbf{N}_\Gamma)$ | 15 | 20 | 29 | 37 | 46 | 55 |
| $n^*_\Gamma$ | 25 | 39 | 73 | $> 100$ | $> 100$ | $> 100$ |

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

summarized in **Figure 1**. The last but most interesting comparison is the one between the non sequential non robust sample size, $n^*$, and the sequential robust sample size, $\mathbb{E}(\mathbf{N}_\Gamma)$. Analysing **Table 1** we see that for instance for $\epsilon = 0.3$ we have $\mathbb{E}(\mathbf{N}_\Gamma) = 29$ that is larger than $n^* = 25$ (corresponding to $\epsilon = 0$ in the table). But for a smaller level of contamination, for instance $\epsilon = 0.1$, the sequential robust expected sample size turns out to be smaller, $\mathbb{E}(\mathbf{N}_\Gamma) = 20$. This provides an example of the idea introduced in Section 3.3: when the class of priors $\Gamma_\epsilon$ 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.

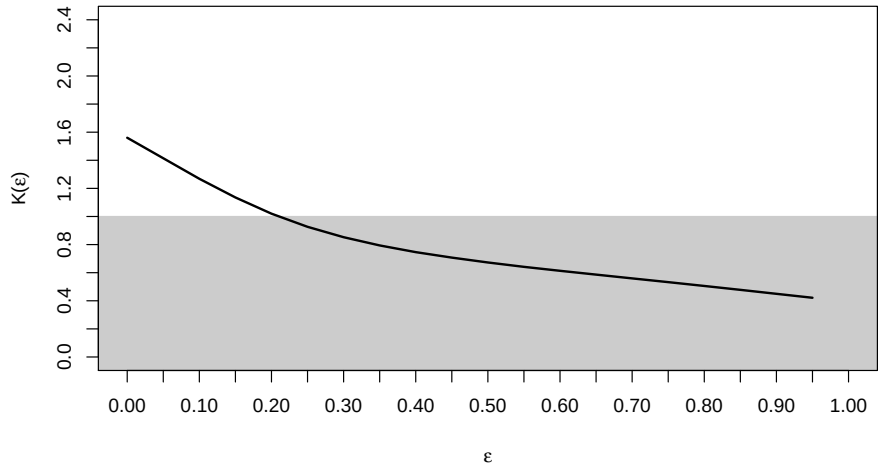


Figure 5: $K(\epsilon)$ is plotted with respect to $\epsilon$. The *critical level of contamination* is $\tilde{\epsilon} = 0.43$.

**Figure 5** shows the behaviour of $K(\epsilon)$ as a function of $\epsilon$: for those values of $\epsilon$ such that $K(\epsilon) > 1$ the robust sequential sample size is smaller than $n^*$, whereas for increasingly wide classes $\Gamma_\epsilon$, $K(\epsilon)$ 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 $\mathbb{E}(\mathbf{N}_\Gamma) = n^*$. Here we have $\tilde{\epsilon} = 0.24$: this value determines the largest class of $\epsilon-$contamination prior distributions yielding a robust 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. In other words, with the sequential approach introducing a degree of uncertainty on 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.

# 5 Conclusions

In this paper we have shown how a sequential procedure allows early termination when there is evidence of treatment efficacy and enables the experimenter to reach a much 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 already indicate a negative course that cannot be reverted even with extremely positive outcomes (see [10] for details).

In summary, the main focus of this 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 [13] for discussion). Here 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.

# References

[1] Armitage P. (1975) Sequential medical trials. *Oxford: Blackwell*.

[2] Whitehead J. (1997) The Design and Analysis of Sequential Clinical Trials *Wiley*.

[3] Pocock S.J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

[4] O'Brien and Fleming T.R. (1979) A Multiple Testing Procedure for Clinical Trials. *Biometrics* **35**(3), 549-556.

[5] Geller N.L., Pocock S.J. (1987) Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* **43**, 213-23.

[6] Berry D.A. (1985) Interim analyses in clinical trials: classical vs. Bayesian approaches. *Statistics in Medicine* **4**, 521-526.

[7] Freedman L.S., Spiegelhater D.J. (1989) Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* **10**, 357-367.

[8] Spiegelhater D.J., Freedman L.S., Parmar M.K.B. (1994) Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A* **157**, 357-416.

[9] Abrams K., Ashby D., Errington D. (1994) Simple Bayesian analysis in clinical trials: a tutorial. *Controlled Clinical Trials* **15**, 349-359.

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

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

[12] Brutti P., 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. DOI: 10.1016/j.jspi.2007.05.041.

[13] Brutti P., De Santis F., Gubbiotti S. Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine*, 2008; **27**, 2290-2306

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

[15] Wang F. and 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.

[16] Etzioni R. and Kadane J.B. (1993). Optimal experimental design for another's analysis. *JASA* **88**(424), 1404-11.

[17] Tsutakawa R.K. (1972) Design of experiment for bioassay. *JASA* **67**(339), 585-90.

[18] Berger, J.O. The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis* (J. Kadane, ed.), Amsterdam: North-Holland, 1984.

[19] Berger, J.O. Robust Bayesian analysis: sensitivity to the prior. *The Journal of Statistical Planning and Inference*, 1990; **25**, 303-328.

[20] Berger, J.O. , Rios Insua, D. and Ruggeri, F. Bayesian robustness. In *Robust Bayesian analysis* (D. Rios and F. Ruggeri, eds.). *Lecture Notes in Statistics*. New York: Springer-Verlag, 2000; **152**.

[21] Wasserman, L. Recent methodological advances in robust Bayesian inference. In *Bayesian Statistic 4* (J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith, eds). Oxford: Oxford University Press., 1992.

[22] Greenhouse, J.B. and Wasserman, L. Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine*, 1995; **14**, 1379-1391.

[23] Greenhouse, J.B. and Wasserman, L. A practical robust method for Bayesian model selection: a case study in the analysis of clinical trials (with discussion). In *Bayesian Robustness, IMS Lecture Notes - Monograph Series* (J.O. Berger et. al., eds.) Hayward: IMS, 1996; 331-342.

[24] Carlin, B.P., and Perez, M.E. Robust Bayesian analysis in medical and epidemiological settings. In *em Robust Bayesian analysis* (D. Rios and F. Ruggeri, eds.). *Lecture Notes in Statistics*. New York: Springer-Verlag, 2000; **152**.

[25] Carlin, B.P., and Sargent, D.J. Robust Bayesian approaches for clinical trails monitoring. *Statistics in Medicine*, 1996; **15**, 1093-1106.

[26] Sivaganesan, S. and Berger, J.O. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Annals of Statistics*, **17**, 868-889.

[27] Berger, J.O. and Berliner, L.M. Robust Bayes and empirical Bayes analysis with $\varepsilon$-contaminated priors. *Annals of Statistics*, 1986; **14**, 461-486.

[28] Facey (1992). A sequential procedure for a phase II efficacy trial in hypercholesterolemia. *Controlled Clinical Trials* **13**(2), 122-133.

[29] Mehta C.R. and Tsiatis A.A. (2001). Flexible sample size considerations using information based interim monitoring. *Drug Information Journal* **35**, 1095-1112.