Bayesian design and conduct of phase II single-arm clinical trials with binary outcomes: A tutorial

Sarah Zohar, Satoshi Teramukai, Yinghui Zhou

BAYESIAN DESIGN AND CONDUCT OF PHASE II SINGLE-ARM CLINICAL TRIALS WITH BINARY OUTCOMES: A TUTORIAL

Sarah Zohar [1,2] , Satoshi Teramukai [3] , Yinghui Zhou[4]

(1) Inserm, CIC 9504, Centre d'Investigations Cliniques, Hôpital Saint Louis, Paris, France.

(2) Inserm, U717, Département de Biostatistique et Infomatique Médicale, Hôpital Saint Louis, AP-HP, Paris, France.

(3) Translational Research Center, Kyoto University Hospital, Kyoto, Japan.

(4) Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Earley Gate, Reading, RG6 6FN, U.K.

Abstract

The aim of phase II single-arm clinical trials of a new drug is to determine whether it has sufficient promising activity to warrant its further development. For the last several years Bayesian statistical methods have been proposed and used. Bayesian approaches are ideal for earlier phase trials as they take into account information that accrues during a trial. Predictive probabilities are then updated and so become more accurate as the trial progresses. Suitable priors can act as pseudo samples, which make small sample clinical trials more informative. Thus patients have better chances to receive better treatments. The goal of this paper is to provide a tutorial for statisticians who use Bayesian methods for the first time or investigators who have some statistical background. In addition, real data from three clinical trials are presented as examples to illustrate how to conduct a Bayesian approach for phase II single-arm clinical trials with binary outcomes.

Key Words: Phase II, single-arm, Bayesian design, response probability, binary outcomes

1. INTRODUCTION

Phase II clinical trials regroup a large panel of investigative studies, that is, form very small single-arm studies to randomized studies and compare several dose groups with a control group. The primary goal of phase II clinical trials is not to provide definitive evidence of treatment efficacy, but to propose a promising treatment for further investigations. In this context, most of the development of phase II designs has been in the area of oncology, where the severity of the disease makes early stopping particularly desirable [1].

In cancer trials, after determining the maximum tolerated dose, one continues to determine whether the new treatment has an anti-tumor effect. The objective of phase II cancer trials is to estimate the magnitude of this effect (in terms of tumor response or survival), and to gather additional information on safety. The most common phase II design is single-armed, that is, patients are treated with the new treatment, and the controls are the pooled patients previously treated with the current standard therapy. In this context, the main objective is to determine whether the new therapy is promising and requires further experimentation.

Several Bayesian designs for phase II single-arm cancer trials have been proposed over the past years [2-8]. Most of these reports only focus on one of the design aspects such as sample size determination, stopping rules, response probability determination, etc. However, Bayesian methodology has rarely been applied in practice. Guidelines and worked examples are surely needed to help the implementation of the methodology. The goal of this paper is to provide a tutorial of conducting a Bayesian approach for phase II single-arm trials with binary outcomes.

This paper describes some approaches that are simple to implement with basic statistical software. In section 2, we describe the Bayesian framework, the choice of the prior distribution, sample size determination, stopping criteria, and some practical procedures. We simplified the statistical parts in the main text, but details are given in the appendix. In section 3, we provide three examples to illustrate how to implement Bayesian methods. Practical guidelines for conducting Bayesian clinical trials are shown in section 4. Finally, a conclusion is given in section 5.

2. METHOD

2.1. The Bayesian framework

Statistical analysis can be performed sequentially, after each observed response (success or failure) from an individual patient or a group of patients. The probability of success can be modeled based on a beta model on the interval $[0,1]$ [2,9]. The uniform distribution is just the special case, which can be used to estimate probabilities of success when clinicians do not have a strong idea about the mean probability of response. Moreover, the beta distribution simplifies mathematical calculations since it provides conjugate priors *i.e.* the prior distribution has the same type of distribution as the posterior distribution.

The observed response is a binary variable, either a success or failure outcome from the administered treatment. With n being the number of observed patients, the number of observed responses s, is a binomial variable (n,π), where π is the probability of response.

<u>Bayesian inference</u>: According to the Bayesian framework, π is a random variable with the prior distribution $Beta(a,b)$ where a and $b>0$. The values of a and b are fixed at the beginning of the trial (details are in the next section). The mean and the variance of the $Beta(a,b)$ density are given by $E(\pi)=a/(a+b)$ and $Var(\pi)=ab/((a+b)^2(a+b+1))$.

After n patients are included into the trial, the posterior distribution of the probability of response, is given by $Beta(a+s,b+n-s)$, with its mean defined by $E_n(\pi)=(a+s)/(a+b+n)$.

Let W_{95} be the width of the 95% credible interval running from the 2.5^{th} to the 97.5^{th} percentiles. Credible intervals represent a measurement of precision used in Bayesian inference (mathematical details are given in appendix).

2.2. The choice of the prior distribution

Specification of the parameters in the prior probability model $Beta(a,b)$ is required. The prior distribution is a probability distribution that quantifies knowledge regarding unknown quantities in the absence of some evidence (eg, model parameters) prior to observing the data. The uncertainty about the real probability of response can be expressed as either a skeptical prior (skepticism about a large treatment effect) or an enthusiastic prior (the confidence of investigators about the new treatment [7,10,11]).

There are several different approaches for choosing values for a and b. Some approaches are quite simple and practical, such as to interpret $a+b$ as the total number of subjects from a previous trial with a being the number of successes and b being the number of failures [13]; set both a and b to be equal to 1 to form a vague uniform prior distribution with the prior mean being equal to 0.5. Others are complicated, such as to define a and b to form computer based elicitations using the 95% credible intervals (W_{95}) of the $Beta(a,b)$ density[11]. The following two approaches use the knowledge of the mean probability of response: construct a prior distribution with a being the mean probability of response and $b = 2 - a$. Secondly derive a and b from the mean and the variance of the mean probability of response. For instance, a and b are solutions of an equation where the mean of a prior is the mean probability of response with a fixed variance. The mean probability of response can easily be obtained from (i) preclinical trials, (ii) previous clinical trials and information from the medical literature and (iii) expert opinion interrogation [12].

2.3. Sample size determination

In a Bayesian approach, a particular sample size does not need to be chosen [14]. However, a maximum sample size, N, must be determined before the trial begins for practical reasons. In recent draft guidelines from the FDA on the use of Bayesian statistics in medical research, it is recommended that a minimum sample size should be decided in advance according to safety and effectiveness endpoints [16].

Several approaches have been proposed where sample size determination is based on the posterior estimation of the mean response rate [7,10,15]. We consider one of the following rules.
Rule 1: Specify N so that the trial is ended only if the posterior response probability distribution achieves a specified degree of precision.
Rule 2: Specify N so that the trial is ended only if the probability of obtaining a promising result is not smaller than a specific value.

2.4. Stopping criteria

Several stopping criteria have been developed and applied so that a trial can be stopped earlier because either the treatment works or the treatment does not work. They are based on either the posterior density of the response probability, or the predictive distribution of the number

of hypothetical responses from m additional patients [6,10,17-19]. We list some stopping criteria as follows:

Rule 1: Stop for efficacy; if there is a high probability that the estimated response probability is higher than a maximal response probability threshold.

Rule 2: Stop for inefficacy; if there is a high probability that the estimated response probability is lower than a minimal response probability threshold.

Rule 3: Stop for efficacy; if the value of cumulative predictive distribution (see appendix) of the number of responses from m additional patients is high, that is, stop if the part of the upper tail of the predictive distribution is high.

Rule 4: Stop for inefficacy; if the value of cumulative predictive distribution of the number of responses from m additional patients is high, that is, stop if the part of the lower tail of the predictive distribution is high.

Rule 5: Stop for satisfactory precision estimation of the response probability; if the maximal predictive gain of m future patients on the precision of the response probability (credible interval width) is low.

All parameters especially stopping rule thresholds are fixed by statisticians and are usually obtained from a simulation study prior to the trial using all of the study features.

In practice, these stopping criteria can be used simultaneously. For example,

(1) if at least one of the stopping criteria for inefficacy is fulfilled, then we recommend stopping the trial for inefficacy. Or

(2) if at least one of the stopping criteria for efficacy is fulfilled, then we recommend stopping the trial for efficacy. Or

(3) if (1) and (2) are not fulfilled and stopping rule 5 is, then we recommend stopping on the grounds of futility.

2.5. In practice: the conduct of the trial

Before beginning the trial:

1) Clinicians and statisticians should work together to define prior features of the beta distribution (see section 2.2).

2) The maximum sample size N can be determined from a calculation (see section 2.3) or be adapted to the inclusion capacity of collaborating centers.

3) If investigators decided to include stopping rules in the study, statisticians should inform the investigators in details which stopping rules are to be used for their study and fix stopping thresholds (see section 2.4).

4) If necessary, statisticians should assess operating characteristics of the trial by simulations.

Once the trial begins:

1) Patients are included sequentially in the study.

2) Once the patient's or group of patients' outcomes are known, statisticians can sequentially apply stopping rules if stopping rules are planned.

3) After the last patient's outcomes are known, statisticians can estimate the final updated response probability with credible intervals.

3. EXAMPLES

3.1. The DST trial

A single-arm clinical trial for evaluating the donor-specific transfusion (DST) on living-related liver transplantation is ongoing in Kyoto University Hospital. The primary endpoint is acute rejection on recipients within 6 months after transplantation and the "response" is defined as no acute rejection. We use a beta distribution with parameters (a=1 and b=1) as a vague uniform prior distribution because there was no information for the new treatment. The prior mean of the probability of success is 0.5. The planned maximum sample size is 35, which was determined by the estimated number of patients. The minimum sample size is 10, which was determined based on a minimum level of information to evaluate safety endpoints. A maximal response probability threshold is 0.75, based on the investigators' opinions. A minimal response probability threshold is 0.55, which was estimated based on a historical dataset deriving from 155 recipients of living-related liver transplantations. If 35 patients are recruited and the maximal response number threshold 26 (nearly equal to 35 times 0.75) is observed, then the 95% credible interval width ($W_{95,35}$) of the posterior probability will be 0.28.

Two stopping rules are planned, stopping rules 3 and 4 of section 2.4. For Stopping rule 3, the cumulative response probability is based on the maximal response number threshold being 26 and the maximum sample size being 35; If this cumulative probability is higher than 0.95 (τ_3=0.95 appendix), we will recommend stopping the trial for efficacy. For Stopping rule 4, the cumulative response probability is based on the minimal response number threshold being 19 (nearly equal to 35 times 0.55); If this cumulative probability is higher than 0.95 (τ_4=0.95 appendix), we will recommend stopping the trial for inefficacy. Six interim analyses are

planned. The first interim will start after 10 patients are recruited. At each interval between two interims, 5 additional patients will be recruited. The stopping boundary in this setting is shown in Figure 1. The upper boundary is to stop the trial earlier on the basis of efficacy. The lower boundary is to stop the trial earlier on the basis of inefficacy.

To assess operating characteristics of the design, we did some simulations. Table 1 shows operating characteristics from 1000 simulations, including the results from τ_3 and τ_4 being 0.80, 0.85, 0.90, and 0.95. The estimated type I error rates in the case where the true response probability is 0.55 are between 0.026 and 0.045, the powers in the case where the true response parameter is 0.75 are between 0.644 and 0.693, and the expected sample sizes range from 23.7 to 29.6. The simulation results show that the Bayesian design would have given acceptable type I error rate and moderate power from the frequentist point of view.

Simulations were also conducted for a skeptic prior, *Beta* (8,7) and an enthusiastic prior, *Beta* (12,4). Designs under the skeptic prior had least powers while designs under the enthusiastic prior achieved highest powers but with inflated type I error rates.

3.2. The arsenic trioxide trials

Two phase II single-arm trials aiming at estimating the success probability of arsenic trioxide in advanced Multiple Myeloma (MM) and in Acute Promyelocytic Leukaemia (APL), respectively, were previously conducted [20]. In these trials a Bayesian approach was conducted simultaneously as the standard approach. A success observation was defined as either complete or partial remission. At the beginning of the trials investigators were asked for the mean success probabilities in each disease, they guessed the success rate to be 10% in MM and 30% in APL. Two prior distributions for each trial were constructed, but in this example only one for each trial is presented. For the MM trial, the mean of the beta prior was equal to 0.1, and the variance was fixed to be equal to 0.0225. Therefore, the values of the beta distribution parameters were $a_{mm} = 0.3$ and $b_{mm} = 2.7$ (Figure 2 A). For the APL trial, the mean of the beta prior was equal to 0.3, and the variance was fixed to be equal to 0.0191. Therefore, the values were $a_{apl} = 3$ and $b_{apl} = 7$ (Figure 2 B). Thus, in the two trials the mean prior response probability was fixed equal to the mean response probability given by the investigators.

From October 1998 to April 1999, 12 patients were included in the MM trial. Table 2 summarises the sequential computation of the estimated posterior mean response probability calculated after each patient inclusion and Figure 2 C represents the posterior estimation of the distribution of the response probability. Among all of the stopping rules, stopping rule 2 recommended ending the trial for inefficacy after the inclusion of 7 patients (Figure 3 A), since there is a high probability that the estimated response probability is lower than a minimal response probability threshold, i.e. $P(\pi < 0.1 | data) > 0.9$. The estimated mean posterior response probability was 0.03. Stopping rule 4 recommended ending the trial for inefficacy after the inclusion of 9 patients, assuming that 5 more hypothetical patients are enrolled and treated after the interim analysis (Figure 3 C). That is, the probability of observing no response for the 5 future hypothetical patients was higher than the stopping critical value, 85%. Moreover, the stopping rule based on the maximal predictive gain on the width of the 95% credible intervals (rule 5) of 5 future patients did not recommended stopping for a satisfactory estimation of the response probability after the inclusion of 12 patients (Figure 3 E). The estimated mean posterior response probability was 0.025. If no stopping rule was applied, that is, the maximum sample size of 12 patients was reached, the trial would conclude inefficacy of the arsenic trioxide with the associated estimated mean posterior response probability being 0.020 (95% credible intervals: 0-0.12) which was far below the 10% response probability set initially by the investigators. In conclusion, the trial was ended for inefficacy of the arsenic trioxide in the treatment of Multiple Myeloma.

In the APL trial 20 patients were included from October 1998 to December 2001. The sequential computation of the estimated posterior mean response probability calculated after each patient's inclusion is given in Table 2 and Figure 2 D represents the posterior estimation of the response probability distribution. Among all of the stopping rules, Stopping rule 1 recommended ending the trial for efficacy after the inclusion of 10 patients (Figure 3 B), since there is a high probability that the estimated response probability is higher than a maximal response probability threshold, i.e. $P(\pi > 0.3 | data) > 0.9$. The estimated mean posterior response probability was 0.450. Stopping rule 3 recommended ending the trial for efficacy after the inclusion of 20 patients, using 5 more hypothetical patients (Figure 3 D). That is, the probability of observing two or more responses for the 5 future hypothetical patients was higher than the stopping critical value, 85%. The estimated mean posterior response probability was 0.600. Finally, the stopping rule based on the maximal predictive gain on the

width of the 95% credible intervals (rule 5) of 5 future patients recommended stopping for a satisfactory estimation of the response probability after the inclusion of 14 patients (Figure 3 E). The estimated mean posterior response probability was 0.500. If no stopping rule was applied, that is, the maximum sample size of 20 patients was reached, the trial would conclude efficacy of the arsenic trioxide with the associated estimated mean posterior response probability being 0.6 (95% credible intervals: 0.42-0.76), which was higher than the 30% response probability expected initially by the investigators. In conclusion, the trial was ended for efficacy of the arsenic trioxide in the treatment of Acute Promyelocytic Leukaemia.

4. PRACTICAL GUIDELINES

In the precedent sections several methods have been presented, in practice, the investigators who design the clinical trial has to choose which method is more adapted to their need.

During the design of the clinical trial a sensitivity analysis is recommended. The analysis should take look at the influence of some prior choice on the trial final recommendation, that is, values of *a* and *b* of the beta distribution. During this phase the choice of the method fixing the beta distribution parameters is to be advised following the available information on the tested drug or procedure. For instance, if the drug or procedure is already used for another disease or there are available published data an informative prior is recommended, by contrast, if the drug or procedure is new a vague prior will be then more appropriate.

Sample size determination using either rule 1 or rule 2 are depending on the wished trial result in terms of point estimate or precision. The use of stopping rules is optional, the authors, through, recommends to always use at least the stopping rules associated to the drug or the procedure inefficacy (rule 2 or/and rule 4 in the section 2.4). It would be unethical to continue and to include new patients using an inefficient drug or procedure.

Finally, most available statistical software provide a beta distribution procedure easily applicable. There is no macro or software that allows calculating and making the sensitivity analysis and the trial conduct, but this paper can be used as guideline in programming and conducting the clinical trial.

5. CONCLUSIONS

The aim of this paper is to give practical guidelines with realistic examples for conducting phase II single-arm Bayesian clinical trials with binary outcomes. The Bayesian process allows updating knowledge gradually rather than restricting revisions in a study design with fixed sample sizes. Moreover, it includes the ability to use predictive probabilities and make

inferences on response estimations. The trials conducted with Bayesians approaches can be updated at any time and without penalties [3].

When using a Bayesian design in trials, it is important to have a real collaboration between clinicians and statisticians. The clinicians should have some ideas about the mean response rate and recommend upper and lower response rate thresholds for which further clinical investigations will be recommended or not. This enables to construct a prior distribution that will be updated with patient accrual. Feasible sample size and patient accrual rate should be discussed. When reporting results, it is necessary to add figures or tables that show how the response rate progressed from the beginning of the trial to the end. Other important information such as stopping rules and quantification of the precision (probability or credible intervals) should be carefully considered. Simulations should be conducted prior to a trial. This enables investigators to see different design properties under several scenarios. Prior specifications and stopping rules could then be modified until a satisfactory design is achieved. Finally, a R script is available on request from the authors.

7. REFERENCES

1. Stallard N, Thall PF, Whitehead J: Decision theoretic designs for phase II clinical trials with multiple outcomes. Biometrics 55:971-7, 1999

2. Berry DA: Monitoring accumulating data in a clinical trial. Biometrics 45:1197-211, 1989

3. Berry DA: Bayesian clinical trials. Nat Rev Drug Discov 5:27-36, 2006

4. Brunier HC, Whitehead J: Sample sizes for phase II clinical trials derived from Bayesian decision theory. Stat Med 13:2493-502, 1994

5. Stallard N: Decision-theoretic designs for phase II clinical trials allowing for competing studies. Biometrics 59:402-9, 2003

6. Stallard N, Whitehead J, Todd S, et al: Stopping rules for phase II studies. Br J Clin Pharmacol 51:523-9, 2001

7. Thall PF, Simon R: A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. Control Clin Trials 15:463-81, 1994

8. Thall PF, Simon RM, Estey EH: New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J Clin Oncol 14:296-303, 1996

9. Gupta AK, Nadarajah S: Handbook Of Beta Distribution And Its Applications. New-York, Marcel Dekker, 2004

10. Tan SB, Machin D: Bayesian two-stage designs for phase II clinical trials. Stat Med 21:1991-2012, 2002

11. Thall PF, Estey EH: A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation. Stat Med 12:1197-211, 1993

12. Spiegelhalter DJ, Abrams KR, Myles JP: Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, John Wiley & Sons Ltd. , 2004

13. Gelman A, Carlin JB, Stern HH, et al: Bayesian Data Analysis. London, Chapman and Hall, 1995

14. Berry DA: A case for Bayesianism in clinical trials. Stat Med 12:1377-93; discussion 1395-404, 1993

15. Mayo MS, Gajewski BJ: Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. Control Clin Trials 25:157-67, 2004

16. FDA FaDA: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials - Draft Guidance for Industry and FDA Staff Rockville, MD. http://www.fda.gov/cdrh/osb/guidance/1601.html, 2006

17. Farge D, Marolleau JP, Zohar S, et al: Autologous bone marrow transplantation in the treatment of refractory systemic sclerosis: early results from a French multicentre phase I-II study. Br J Haematol 119:726-39, 2002

18. Zohar S, Chevret S: The continual reassessment method: comparison of Bayesian stopping rules for dose-ranging studies. Stat Med 20:2827-43, 2001

19. Chen C, Chaloner K: A Bayesian stopping rule for a single arm study: With a case study of stem cell transplantation. Stat Med 25:2956-66, 2006

20. Rousselot P, Larghero J, Arnulf B, et al: A clinical and pharmacological study of arsenic trioxide in advanced multiple myeloma patients. Leukemia 18:1518-21, 2004

21. Berger JO: Statistical decision theory and Bayesian analysis New York, Springer-Verlag, 1985

22. Berry DA, Stangl D: Bayesian Biostatistics New York, Marcel Dekker 1996

Appendix

<u>Credible intervals of π:</u> The credible interval of π is a probability that π lies in $1-\alpha$ interval [21]. A two-sided $1-\alpha\%$ credible interval with an equal probability in each tail area would comprise (π_L, π_U), where $P(\pi < \pi_L | data) = \alpha/2$ and $P(\pi > \pi_U | data) = 1 - \alpha/2$.

After the inclusion of j patients the width of the credible interval of the response probability is denoted as $W_{\alpha,j}$. In the case of the beta distribution this interval can be calculated from the

incomplete beta function $I_\alpha(a,b) = \dfrac{Beta_\alpha(a,b)}{Beta(a,b)}$ where $Beta_\alpha(a,b) = \int_0^\alpha t^{a-1}(1-t)^{b-1} dt$.

<u>Predictive distribution:</u> A Bayesian approach allows predicting some future observations on the basis of currently observed data. At an interim stage, suppose that we observed s responses from the first n observations and the posterior distribution for π is updated, which is $Beta(a+s, b+n-s)$. Using a beta-binomial model, the predictive probability of observed k responses from m future observations is defined by:

$$P(K=k|s,n,m) = \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}\binom{m}{k}\frac{\Gamma(a'+k)\Gamma(b'+m-k)}{\Gamma(a'+b'+m)}$$

where $a' = a+s$ and $b' = b+n-s$. The predictive distribution is used in trial decision making such as stopping rules (section 2.3), whether or not to conduct future trials [22].

<u>Sample size determination rules</u>

<u>Rule 1:</u> Select N so that the posterior probability $P(L < \pi < U | X_N)$ is bigger than p, where p is a desired coverage probability, $\Delta = U - L$ is a desired credible interval width. X_N is a hypothetical number of responses out of the N patients, where X_N is the nearest integer to $N \times R_U$ and R_U is a maximal response probability threshold.

<u>Rule 2:</u> Select N so that the probability $P(\pi > R_U | X_N)$ is bigger than λ, where λ is a specified power criterion, usually within the range of [0.8, 0.9]. X_N is a hypothetical number of responses out of the N patients, where X_N is the nearest integer to $N \times (R_U + \varepsilon)$ and R_U is a maximal response probability threshold and ε is a selected small value, usually within the range of [0, 0.1].

Stopping rules

Rule 1: $P(\pi > R_U | data) > \tau_1$ where R_U is a maximal response probability threshold, fixed before beginning the trial by investigators and τ_1 is arbitrarily chosen by statisticians, usually within the range of [0.8,1].

Rule 2: $P(\pi < R_L | data) > \tau_2$ where R_L is a minimal response probability threshold, fixed before beginning the trial by investigators and τ_2 is arbitrarily chosen by statisticians, usually within the range of [0.8,1].

Rule 3: $\sum_{i=lmax}^{m} P(K = i | s,n,m) > \tau_3$ where K is the number of responses from m future observations, $s+lmax$ would mean a maximal response number threshold which could be nearly equal to $N \times R_U$, and τ_3 is arbitrarily chosen by statisticians, usually within the range of [0.8,1]. The values of *lmax and m* are determined with clinicians in function of the number of remaining additional patients to be included in the trial at the time of the interim analysis.

Rule 4: $\sum_{i=0}^{lmin} P(K = i | s,n,m) > \tau_4$ where K is the number of responses from m future observations, $s+lmin$ would mean a minimal response number threshold, which could be nearly equal to $N \times R_L$ and τ_4 is arbitrarily chosen by statisticians, usually within the range of [0.8,1]. The value of *lmin and m* are also determined with clinicians in function of the number of remaining additional patients to be included in the trial at the time of the interim analysis.

Rule 5: $\max |W_{\alpha,j+m}(\pi) - W_{\alpha,j}(\pi)| < \xi$ where ξ is arbitrarily chosen by statisticians, usually within the range of [0.02,0.15].

Figure 1: Stopping boundary for the DST trial.
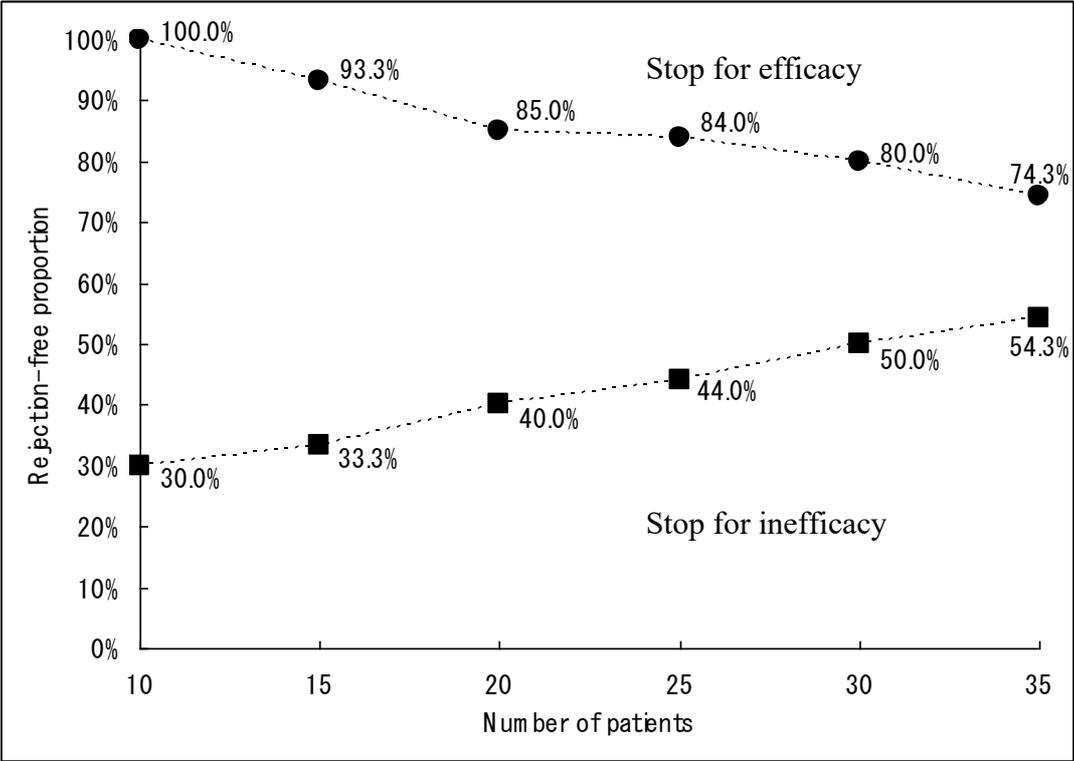
Table 1. Operating characteristics from simulations for the DST trial.

Stopping boundary: τ_3, τ_4	True response probability	P_-	P_+	$1-(P_- + P_+)$	Expected sample size
0.95, 0.95	0.55 (minimal)	0.553	0.026	0.421	29.4
	0.75 (maximal)	0.005	0.644	0.351	29.6
0.90, 0.90	0.55	0.574	0.037	0.389	27.4
	0.75	0.010	0.659	0.331	27.4
0.85, 0.85	0.55	0.590	0.043	0.367	25.1
	0.75	0.029	0.685	0.286	24.6
0.80, 0.80	0.55	0.611	0.045	0.344	23.9
	0.75	0.033	0.693	0.274	23.7

P_-: probability of early stopping due to inefficacy
P_+: probability of early stopping due to efficacy
$1-(P_- + P_+)$: probability of the trial being inconclusive (neither efficacious nor inefficacious)

Figure 2: Prior and posterior distributions of the response probability at the end in either the MM or the APL trial. The dashed vertical lines represent the estimation of the mean response probabilities.
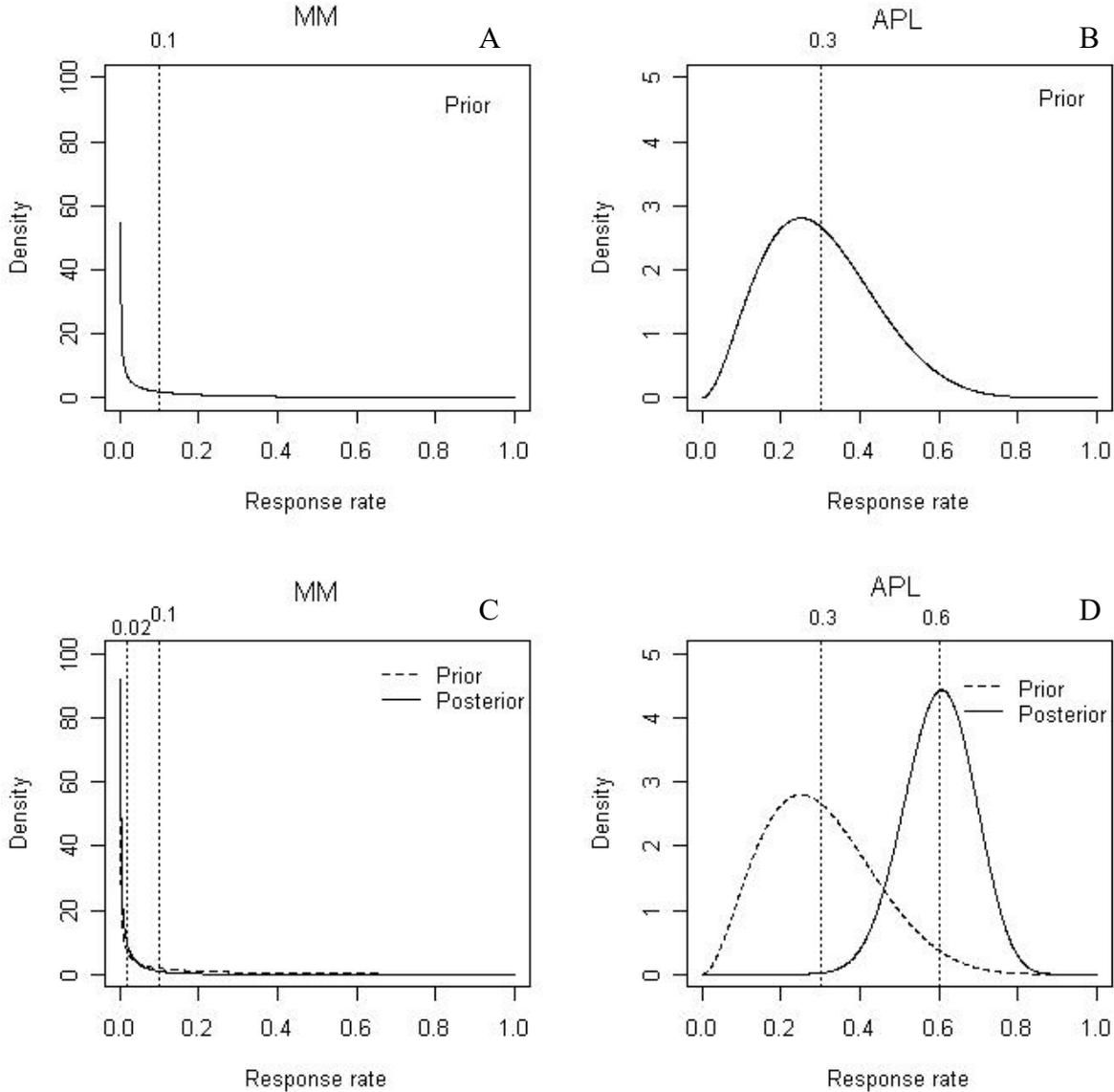
Table 2: Sequential estimations of the mean posterior response probabilities after the inclusion of each new patient in either the MM or the APL trial.

MM			APL		
	Mean prior response probability			Mean prior response probability	
	0.100			0.300	
Patient	Response*	Mean posterior response probability	Patient	Response	Mean posterior response probability
1	0	0.075	1	0	0.273
2	0	0.060	2	1	0.333
3	0	0.050	3	0	0.308
4	0	0.043	4	0	0.286
5	0	0.038	5	1	0.333
6	0	0.033	6	1	0.375
7	0	0.030	7	1	0.412
8	0	0.027	8	1	0.444
9	0	0.025	9	0	0.421
10	0	0.023	10	1	0.450
11	0	0.021	11	1	0.476
12	0	0.020	12	1	0.500
			13	0	0.478
			14	1	0.500
			15	1	0.520
			16	1	0.538
			17	1	0.556
			18	1	0.571
			19	1	0.586
			20	1	0.600

* 0=Failure and 1=Success

Figure 3: Sequential computation of the stopping rules of the MM and the APL trial.