Open Access

Bayesian Monitoring and Bootstrap Trial Simulation: A New Paradigm to Implement Adaptive Trial Design for Testing Antidepressant Drugs

Emilio Merlo-Pich[1], Paolo Bettica[2] and Roberto Gomeni[*, 3]

[1]Discovery Performance Unit, Neurosciences CEDD, [2]Discovery Medicine Neurosciences CEDD, [3]Pharmacometrics & Clinical Pharmacology Modelling and Simulation, GlaxoSmithKline, Verona, Italy

Abstract: A novel methodology is proposed for continuous monitoring of efficacy data in ongoing antidepressant clinical trials and for decision making to support progression or discontinuation of the trial or one of the treatment arms.

The Posterior Probability of Superiority (PPS) resulting from the application of Monte Carlo Markov Chain approach to a longitudinal model describing the time course of placebo and antidepressant drugs was used to estimate criteria to discontinue a treatment arm or the trial for futility, and to predict the treatment effect at study-end while the trial was still ongoing.

The decision to stop the study was based on PPS, Predictive Power and on risk analysis based on a non-parametric bootstrap trial simulation. The performance of the Bayesian monitoring was evaluated by the retrospective analysis of 3 clinical trials. The Bootstrap-based methodology was compared to the Conditional Power and the Predictive Probability approaches.

The application of the proposed methodology showed the possibility to stop a trial for futility when about 50% of total information was available and to detect signal of a treatment effect when limited information (<40%) was available. The comparisons with the Condition Power and the Predictive Probability approaches indicated that the Bayesian Bootstrap method, based on data-driven assumptions for priors, provided a better control for the risk of inappropriate decisions.

The results suggest that the proposed methodology to monitor the accumulating information and to provide a scenario-based risk analysis could constitute a valuable approach to re-engineer the development process of novel drugs.

Keywords: Bayesian monitoring, posterior probability of superiority, bootstrap, trial simulation, futility stopping rule.

INTRODUCTION

Failed and negative trials are a recognized problem for the clinical development of novel antidepressant treatments. A trial is considered failed when the active treatment does not differentiate from placebo.

The reasons for the increasing number of negative or failed efficacy studies in Major Depressive Disorders (MDD) are poorly understood [1]. Contributing factors include, among other reasons, escalating placebo response rates, suboptimal dosing regimens, poor sensitivity of the clinical scores used to evaluate clinical efficacy (HAMD, MADRAS) and the inherently high statistical variance of studies, especially in the case of massive multi-centre trials [2, 3].

Another confounding factor specific to clinical trials in MDD is the lack of accurate and specific diagnosis criteria. This lead to a wide margin of subjectivity in patient selection and to an increased heterogeneity (inter-individual variability) in the MDD patient population enrolled in the trials; all factors that increase the risk of study failure by preventing the detection of a signal of a clinical response for novel antidepressant treatments.

Traditionally, the clinical trial design used to assess antidepressant efficacy is a randomized, double-blind, parallel-group, placebo-controlled study. This study design may not be the most efficient way to conduct a trial in MDD given the wide range of factors affecting the final study outcomes. For this purpose alternative strategies have been proposed based on the use of an adaptive approach [4]. Flexible or adaptive designs can offer an opportunity to the implementation for a learning/confirming paradigm [5] in the early stage of development of antidepressant drugs.

The data that accumulate during an ongoing clinical trial contain information about study outcomes already at a relatively early stage. The current approaches for data analysis rarely consider the use of this information to improve the overall efficiency of clinical trial. In contrast the Bayesian methodology could represent an ideal approach to capture early information that accrues during a trial, offering the opportunity to modify the study design, to stop or expand accrual, to unbalance randomization in favour to better-performing therapies, to drop treatment arms, or change the trial population to focus on patient subsets that are responding better to the experimental therapies [6, 7]. The availability of reliable electronic data capture technology is

*Address correspondence to this author at the Clinical Pharmacology Modelling & Simulation, GlaxoSmithKline, Via A. Fleming 4, 37135 Verona, Italy; Tel: + 39 348 531 7802; E-mail: roberto.a.gomeni@gsk.com

making monitoring approach feasible in the conduction of current clinical trials [8].

Interim monitoring is currently recognised as an important tool for early decision-making during the conduction of clinical trials [9, 10]. Most commonly, interim assessment is implemented using the group sequential [11] or stochastic curtailment approaches [12-14]. The Bayesian methodology has been used both in the group sequential [15] as well as stochastic curtailment frameworks [16]. Most Bayesian stochastic curtailment tests discussed in the literature are examples on mixed Bayesian-frequentist strategies known as predictive power tests. Choi [17] and Spiegelhalter [16] introduced early termination rules for binary endpoints based on a mixed Bayesian-frequentist approach. Berry [15] studied the use of futility rules in clinical trials with a binary outcome which are adjusted for important covariates. Johns [18] considered predictive power methods as well as methods relying on a comparison of sample proportions in clinical trials with binary outcomes. Under Bayesian and non-Bayesian assumption Spiegelhalter [19] proposed to use the Conditional Power to trigger decision at interim monitoring time. This approach is based on the conditional probability of rejecting the null hypothesis at the end of the trial. This probability can be computed under either the null or various alternative hypotheses. Recently, a generalisation of the predictive method [20, 21], called the Bayesian predictive or Predictive Probability approach has been proposed [22] to evaluate the interim monitoring outcomes.

One of the main reasons for the growing interest in Bayesian methods is the increase in computing power and the development of simulation based approaches such as Markov chain Monte Carlo (MCMC) methods [23-26]. Many of these models include hierarchical data structures where between-subject variation is modelled using random effects. Examples can be found in meta-analysis and generalised synthesis models [27], cluster randomised trials [28, 29], genetic epidemiology [30], institutional ranking [31] and subgroup analysis [32].

One important component of the Bayesian approach is the definition and selection of the appropriate ‘prior information’. According to Spiegelhalter [33], there are several possible sources of prior distributions, *i.e.* the Reference prior, the Sceptical prior and the Enthusiastic prior. The Reference prior or ‘non-informative’ represents minimal prior information. The Skeptical prior formalizes the belief that large treatment differences are unlikely. This can be set up, for example, as having a mean of no treatment effect, and only a small probability of the effect achieving a clinically relevant value. By contrast, the Enthusiastic prior can be specified, for example, with a mean effect equivalent to a clinically relevant effect, and only a small probability of no effect. The Reference prior, Skeptical prior, and Enthusiastic prior are essentially mathematical constructs, calibrated using an empirical approach. By contrast, a Clinical prior is intended to represent the current state of knowledge and it is generally based on good evidence, such as a meta-analysis of relevant randomised controlled trials [34].

Recently FDA has recommended criteria for the assessment of scientifically valid prior information in the preparation of novel trials for the use of medical device. In this guidance, priors were considered acceptable only if the novel trial shares the same protocol design, endpoints, target population, recruitment centres of previous trials [6, 35]. These requirements are rarely fulfilled in clinical trials for the development of novel antidepressants, mostly related to logistic problems [36]. When the transferability of prior information from other studies is questionable, an alternative approach consists in building priors based on partial data accrual within a given trial. In this case uninformative prior can be used to start the Bayesian monitoring process of data accrual.

The objective of this paper was to implement a bootstrap-based methodology for decision-making based on continuous monitoring of data accumulated during clinical trials for novel antidepressants and to compare the performance of the proposed methodology with the outcomes of other approaches, such as the Conditional Power and the Predictive Probability.

MATERIAL AND METHODS

In a depression trial the primary clinical endpoint is usually defined by the changes on HAMD-17 score over time. The partial longitudinal set of measurements collected for each individual up to the monitoring time (t) constitutes the information available for data analysis represented as the information score (IS):

$$[\mathrm{IS}]_t = \sum_{i,j} \mathrm{time}(i, j_i) \qquad (1)$$

where i is the subject index and ji is the sequence of measurement time-points for subject i. For example, a subject with partial data collected at week 0, 1, 2 and 4 will provide $[\mathrm{IS}]_4=7$. Similarly, the total expected information score $[\mathrm{IS}]_T$ can be estimated using the total number of subjects planned together with an estimate of the drop-out rate. Therefore, using the $[\mathrm{IS}]_t/[\mathrm{IS}]_T$ ratio, the fraction of the total information available at a given monitoring time-point can be computed.

Within the Bayesian framework, the trial’s result is driven by the posterior probability of a clinically relevant treatment effect given the available data. Based on monitoring assessment, the trial’s outcome is expected to be positive if the Posterior Probability of Superiority (PPS) of the active drug versus placebo is greater than a pre-specified threshold. PPS is defined as the probability that the difference between active and placebo (θ) will be greater than 0 given $X_{(n)}$ (the samples observed at the monitoring assessment):

$$\mathrm{PPS} = \mathrm{Prob}(\theta > 0 | X_{(n),} \text{prior on } \theta) > \eta \text{ (with } 0 < \eta < 1) \qquad (2)$$

The PPS can be calculated at the different monitoring time-points during the study progression, providing estimates of the treatment effect (separation from placebo) at study-end. Moreover, the amount of cumulated information at the time of monitoring assessments can be used to decide trial termination for futility. According to Gould, decision to stop the trial for futility requires at least 40% of the total planned information [37]. After this time the predictive power (PPW, estimated using the method proposed by Dmitrienko [22]), *i.e.*, the probability of rejecting the null hypothesis of no treatment effect at completion of the trial based on available information can be effectively used to support a decision. If this probability is sufficiently low, then the trial could be terminated and resources redirected more productively, otherwise the trial can be progressed [38, 39].

On this basis we can define the time for futility stopping criteria as the time when the 3 following conditions are simultaneously satisfied: 1) at least 40% of the planned information is accumulated; 2) PPS is lower than the agreed risk level of accepting superiority when this is not true (e.g. < 0.90) and 3) PPW is lower than the agreed risk level of rejecting a superiority when it is true (e.g. < 0.2).

Conditional Power and Predictive Probability

Conditional Power and Predictive Probability are estimated using a Bayesian framework by combining the data collected up to the monitoring time and a range of alternative prior assumptions. Table **1** shows the four alternatives normal priors used to evaluate the depression trials in the present analyses. These data were derived from the study design setting of the 810 trail (case 1). This trial was powered to demonstrate a clinically meaningful difference (3.5 points in the 17-item HAM-D) between paroxetine (two doses) and placebo. Sample size calculation was based on a standard deviation of 8 and normally distributed errors with a two-sided nominal significance level of 5% (actual significance level of 2.5%, adjusting for 2 treatment comparisons). The Reference prior assumes a uniform distribution and no difference between placebo and active treatments. The other priors assume a nominal difference of 3.5 in the HAMD-17 at week 8; while the standard deviation varies according to different level of expectation, *i.e.*, arbitrarily, large for Skeptical prior, driven by sample size calculation for Clinical prior and arbitrarily small for Enthusiastic prior. Conditional Power and Predictive Probability priors were estimated using the SAS macros [39].

Table 1. Parameters of the Normal Priors for the Mean Change in the HAMD-17 Score in the MDD Clinical Trial 810. Mean-1 and Mean-2 Represent the Expected HAMD-17 Change from Baseline at Week 8 in the Placebo and Active Treatment Arms and 'SD' Represent the Common Standard Deviation

Prior	Mean-1	Mean-2	SD	Effect Size
Reference	-10	-10	1000	0
Skeptical	-10	-13.5	12	0.29
Clinical	-10	-13.5	8	0.44
Enthusiastic	-10	-13.5	4	0.88

Bootstrap and Trial Simulation for Risk Assessment

Sometimes the decision to stop the trial for futility based on data not completely representative of the trial may lead to erroneous recommendations. This may occurs when a significant heterogeneity exists in the data remaining to be collected up to the study-end with respect to the data accumulated up to the monitoring time. For example, this may happen when the different centres have different starting time of recruitment, different recruitment rate and/or different quality of data.

To control the risk of a false negative decision to stop the trial, we propose a trial simulation method that uses the information available at various monitoring time points. This simulation estimates the expected PPS at the study-end for different typologies of subjects' response providing a framework to assess a scenario-based risk analysis.

The data remaining to be collected at the various monitoring time points were simulated using a non-parametric bootstrapping approach by resampling with replacement the original data collected up to the monitoring time [38, 40]. Two scenarios were considered: 1) the worst case scenario, where all the remaining subjects are assumed to show a placebo-like response, and 2) the best case scenario, where all the remaining subjects in the treatment arm are assumed to be antidepressant drug responders. Alternative intermediate scenario can be easily considered.

The Partial Database at Monitoring Time (PDMT) was initially split by treatment and only patients with a complete set of information (completers) were retained. Then, only patients responding to the treatment in the active treatment groups were selected (a responder is a patients with a HAMD-17 score reduction from baseline >=50%). These data constitute the Resampling Database (RD).

The number of unobserved patients in each treatment group was computed as the difference between the planned sample size and the number of patient included at the monitoring time (ex. Np for the placebo and Nt for the active treatment group). At this stage, bootstrap and trial simulation was implemented in a 4-step process:

1. Resample Np and Nt subjects with replacement from the Resampling Database.
2. Combine the Partial Database at Monitoring time with the simulated patients resampled in step 1 in a Working Database (WB)
3. Estimate the Posterior Probability of Superiority analysing the data in the Working Database *Repeat step 1 to 3 a number of time (100 for the present analysis)*
4. Compute the proportion of positive trials (the trials for which the PPS value is > 0.90). A schematic of the bootstrap process is shown in Fig. (**1**).

Once the futility stopping criteria has been satisfied (at least 40% of the total information, PPS < 0.90 and PPW < 0.20), a typical decision making based on the bootstrap could be:

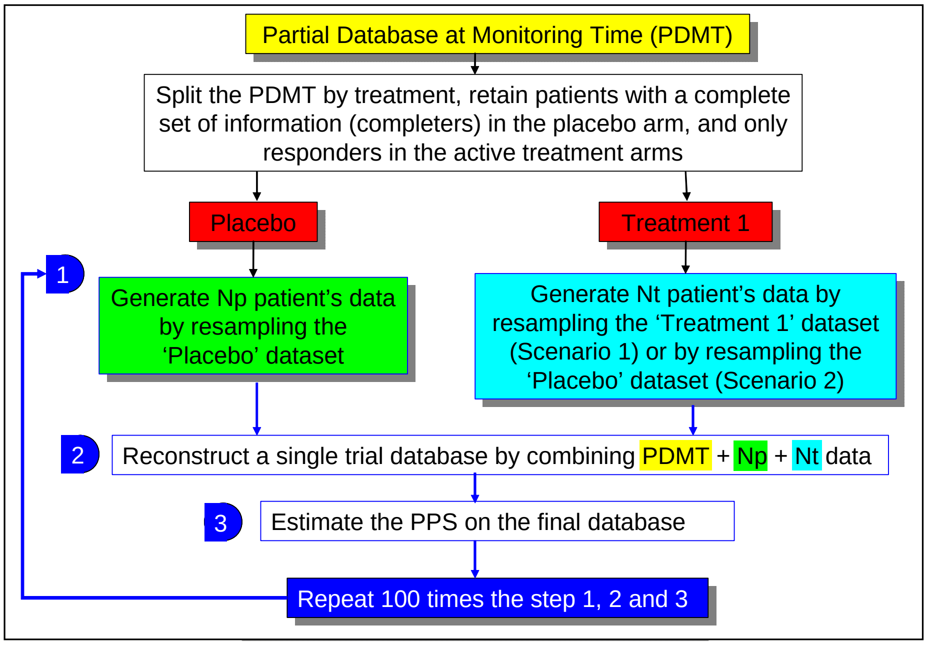
Partial Database at Monitoring Time (PDMT)
Split the PDMT by treatment, retain patients with a complete set of information (completers) in the placebo arm, and only responders in the active treatment arms
Placebo
Treatment 1
1
Generate Np patient's data by resampling the 'Placebo' dataset
Generate Nt patient's data by resampling the 'Treatment 1' dataset (Scenario 1) or by resampling the 'Placebo' dataset (Scenario 2)
2
Reconstruct a single trial database by combining PDMT + Np + Nt data
3
Estimate the PPS on the final database
Repeat 100 times the step 1, 2 and 3

Fig. (1). A schematic representation of the process of bootstrapping for a two arm trial (placebo and active treatment).

1. Stop the study (or terminate one arm) if the proportion of the expected positive trials in the best case scenario is lower than 80%. The threshold was arbitrarily set based on our experience regarding study team expectation for a clinically relevant signal.

2. Progress the study if the proportion of the expected positive trials in the best case scenario is greater than 80%. In this case, there is a reasonable chance to reverse the findings observed at monitoring time if the data remaining to be collected are particularly good.

Bayesian Longitudinal Model

The longitudinal structural model was defined as the combination of a Weibull and linear equations [41]. This model was used to fit the population HAMD-17 time-course using a Markov Chain Monte Carlo (MCMC) technique as implemented in the WinBUGS software package [25, 42]. The individual model parameters vector θ_i was assumed to follow a normal distributions with population parameters Φ. Specifically, Φ was assumed diagonal and arising from Gamma distribution (Γ) characterised by a relatively uninformative prior distribution $\Gamma(0.001, 0.001)$. This choice was motivated by the lack of rationale to transfer prior information from historical studies. The Bayesian analysis involved the estimation of the joint distribution of all parameters conditional on the observed data: p(θ, Φ|HAMD_data). Generating random samples from the joint posterior distribution allows the marginal distribution of each parameter to be completely characterized.

Retrospective Analysis of Clinical Trial Cases

The performance of Bayesian monitoring and trial simulation was evaluated on the retrospective analysis of 3 randomized, double-blind, parallel-group, placebo controlled, multi-center clinical trials (810, GSKX, and 002) on subjects suffering from MDD, including a total of 967 subjects. The primary efficacy measure was the change from baseline to study endpoint (week 8 for study 810 and GSKX and week 6 for study 002) as measured by the HAMD-17 total score.

Case 1 (Study 810): Efficacy and safety of paroxetine controlled release (CR) (12.5 and 25mg/day) versus placebo (N= 156, 154 and 149 respectively) were evaluated in this trial [43]. The primary efficacy analysis revealed that both the 12.5mg and the 25mg paroxetine CR treatment groups were associated with significant therapeutic effects (change in HAMD score) from baseline to study endpoint (LOCF: p = .038, 95% CI = -3.38 to -0.09 and p = .005, 95% CI = -4.06 to -0.74, respectively).

Case 2 (Study GSKX): Efficacy and safety of fix doses of Compound X or paroxetine (20mg) to placebo (N= 125, 117 and 118 respectively) were evaluated in this trial. This study did not meet the primary objective of demonstrating statistically significant efficacy of the Compound X and paroxetine compared to placebo for the primary efficacy endpoint with an adjusted mean difference 0.25, 95% CI (-1.74, 2.23), p=0.808 for Compound X and an adjusted mean difference –0.16, 95% CI (-2.18, 1.87), p=0.879 for paroxetine. While not significant at the Week 8 observed case (OC) endpoint, numerical superiority of paroxetine over placebo was observed (adjusted mean difference -1.45, 95% CI (-3.88, 0.99), p=0.243).

Case 3 (Study 002): Efficacy and safety of flex dose paroxetine (10-50mg/day) versus placebo (N= 138 and 135) were evaluated in this trial [44]. The primary efficacy analysis showed a significant difference between paroxetine and placebo from week 2 to week 6 on HAMD-21 change from baseline. The mean changes were -7.5, -9.4, -11.6 and

-12.7 estimated on the LOCF dataset. This trial was conducted in 4 centres providing a different response. The HAMD-21 scores at week 6 failed to separate from placebo in centre 3 in contrast to what was observed in the other 3 centres. In addition a strong unbalance in subject recruitment was observed: centres 4 start the recruitment when the recruitment in the first 3 centres was almost completed (Fig. **6**, right panel).

RESULTS

The longitudinal model used for a three arms study (Placebo, Active-1 and Active-2) was:

$$f_i(t) = A_i e^{-(t/td_i)^{b_i}} + h_{rec_i} t \qquad (3)$$

where A is the baseline HAMD-17 score, td is the time corresponding to 63.2% of the maximal change from baseline, b is the shape or sigmoidicity factor, h_{rec} is the recovery rate, i is the treatment arm index assuming the value of 0 (for placebo), 1 and 2 for the active arms. A total of 12 fixed-effect ($\mu_{j,}$ j=1 to 12) parameters were estimated together with their random components τ_j). The treatment effect (Pred) and the treatment differences (Dpred) with their posterior distribution were then estimated using the μ_j and τ_j values:

$$\text{Pred}_i(t) = \mu_{A_i} e^{-(t/\mu_{td_i})^{\mu b_i}} + \mu_{hrec_i} t \qquad (4)$$

$$\text{Dpred}_1(\text{t}) = \text{Pred}_2(\text{t}) - \text{Pred}_1(\text{t}) \qquad (5)$$

$$\text{Dpred}_2(\text{t}) = \text{Pred}_3(\text{t}) - \text{Pred}_1(\text{t}) \qquad (6)$$

where Dpred_1 and Dpred_2 are the predicted separation from the active arm 1 and 2 from placebo.

This model was applied to the analysis of selected clinical trials (Case 1, 2 and 3). For brevity, detailed results of individual and population longitudinal model fitting will be presented only for the Case 1 (study 810). For Case 1, 2 and 3 summary tables and graphical representation of the Bayesian monitoring and bootstrap analysis will be shown. The analysis of the Case 3 (study 002) was conducted on the HAMD-17 scores to ensure comparability of the results across the other trials presented.

Case 1 (Study 810)

The final population parameter distribution (fixed and random effect) estimated by fitting the longitudinal model to the HAMD-17 scores are shown in Table **2**.

All parameters were estimated with acceptable average precision. The treatment groups have comparable baseline values either in term of mean (fixed effect) or inter-individual variability (random effect). The treatment effect was well captured by the reduction of td (the time at which a drop of 63.2% from baseline is observed) as well as by the increase in the sigmoidicity factor (b) in the paroxetine arms.

The time-course of the observed (blue) and predicted HAMD-17 scores of representative subjects are shown in Fig. (**2**). Red lines indicate the median model predicted curve and the 95% credible intervals (precision of the estimate). These results illustrate the flexibility of the longitudinal model to describe individual HAMD-17 time-course profiles showing heterogeneous patterns, such as linear increase/decrease, sigmoid decline, bell-shape time-course and exponential decrease.

Table 2. Case 1: Mean Fixed and Random Effect Model Parameters with their Standard Deviation (Std)

Parameter	Fixed Effect		Random Effect	
	Mean	Std	Mean	Std
A [Placebo]	23.59	0.32	0.11	0.01
A [12.5mg]	23.18	0.29	0.09	0.01
A [25mg]	23.24	0.32	0.11	0.01
td [Placebo]	6.21	1.14	0.86	0.11
td [12.5mg]	4.46	0.39	0.71	0.07
td [25mg]	4.60	0.40	0.67	0.07
hrec [Placebo]	0.46	0.16	0.90	0.31
hrec [12.5mg]	0.59	0.09	0.73	0.11
hrec [25mg]	0.43	0.08	0.66	0.16
b [Placebo]	0.96	0.08	0.57	0.10
b [12.5mg]	1.07	0.07	0.46	0.07
b [25mg]	1.10	0.07	0.37	0.06

The results of the Bayesian monitoring (median separation from placebo in the 12.5 and 25mg paroxetine arms) conducted at different time-points with different levels of information are shown in Table **3**. The number of day from the enrolment of the first subject and the percentage of subjects remaining to be enrolled up to study termination are also shown at each monitoring time.

In this study the threshold defined by the futility stopping rule (>40% of the total information, PPS<0.90 and PPW<0.20) was never reached at any time-point. The two treatments displayed different behaviour, showing lower signal of response (separation from placebo) in the 12.5mg paroxetine arm.

In presence of a strong drug effect, as delivered by the 25mg paroxetine arm, the median separation from placebo at the study-end was already anticipated when only 18% of the total information was available. The robustness of the predicted treatment effect for each arm made at each monitoring time using the longitudinal model was illustrated by an example, *i.e.*, the comparison of the predictions obtained with the complete and the partial dataset (18% and 100% of total information, referring to monitoring day 80 and 172, respectively) (Table **2**). The good agreement between the median curves estimated using the complete and partial dataset indicates the robustness of the proposed longitudinal model to describe the HAMD-17 time-course.

Bayesian inference on treatment effect at various time-points was obtained by analysing the posterior distribution of the time-course of the difference between placebo and

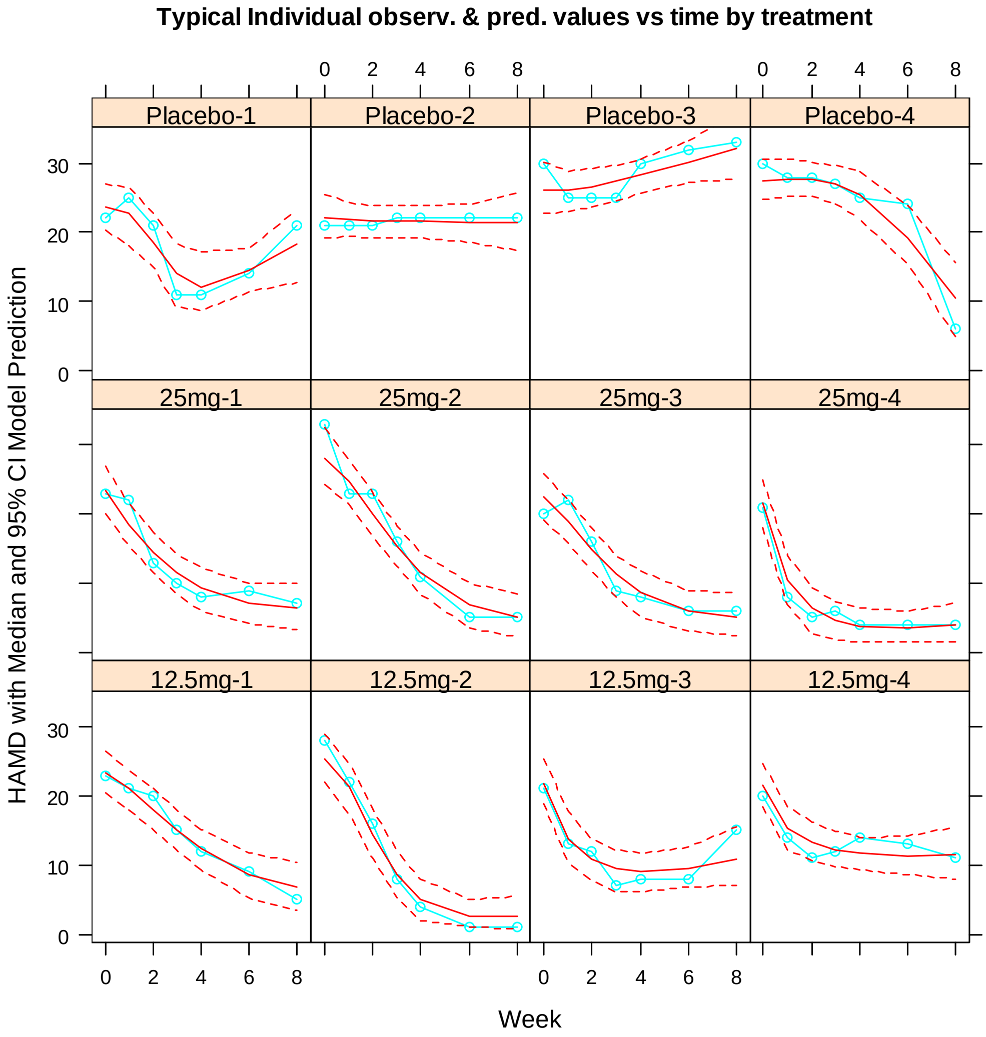

Fig. (2). Case 1: Typical individual observed HAMD-17 response profiles (blue curve) with the median model predictions (solid red line) and the 95% credible intervals (red dotted lines)

paroxetine in the 12.5mg and 25mg arms. As an example, the estimated time-course of these differences with the associated 95% credible intervals are shown at monitoring day 110 (51% of total information) in Fig. (**3**). The good predictive performance of the model is illustrated by the location of the median response and by the size and location of the credible intervals. In fact, the median curves defined by the incomplete dataset (monitoring day 110, blue lines) consistently overlaps with the median curves estimated with the full dataset (monitoring day 172, red line) and the 95% credible intervals of the full dataset (cyan shaded area) are included in the 95% credible intervals of the incomplete dataset (grey shaded area). As expected, the size of the 95% credible interval is function of the amount of the information, showing larger spread in the estimation done at monitoring day 110. The distance of the upper bound of the 95% credible interval from the zero line is a graphical indicator of the clinical relevance of the separation from placebo. In Fig. (**3**) the upper bound of the 95% credible interval of the 25mg arm estimated at monitoring day 110 is just above or overlapping the zero line suggesting a strong signal of clinical efficacy at the study-end.

The probability of observing HAMD-17 separation from placebo of the paroxetine treatment at study-end was estimated using the bootstrap-based trial simulation approach at each monitoring time. This probability was estimated as the percent of positive out of 100 simulated trials (probability of success). Scenario analysis was carried out assuming that the 'Worst Case' and the 'Best Case' were close to the 'Skeptical' and the 'Enthusiastic' assumptions, respectively. Fig. (**4**) shows the percent of positive trials by treatment for a given PPS in the best and the worst case scenario evaluated when 35% of the total information (monitoring day 96) were available. Assuming an agreed target PPS of 80% for detecting treatment superiority, the probability of success in the Worst Case scenario (left panel) was 85% for the paroxetine 25mg arm (red curve) and 42% for the paroxetine 12.5mg arm (blue curve). In the Best Case scenario (right panel) the probability of positive results was > 95% for both arms. This result indicates that even in the unlikely event that all the subjects remaining to be enrolled in the active treatment arms will respond as those receiving placebo (Worst Case scenario), there is still a high expectation of positive outcome for the 25mg paroxetine arm

Table 3. **Case 1: Results of Bayesian Monitoring**

Monitoring Day*	%Total Info	Parox.12.5 mg - Placebo Week 8				Parox.25mg - Placebo Week 8				% of patient Remaining to be Enrolled		
		Median Separation	2.5% - 97.5% Credib. Interval	PPS	PPW	Median Separation	2.5% - 97.5% Credib. Interval	PPS	PPW	Plac	Com pX	Parox
0	0%	-	-	-	-	-	-	-	-	100	100	100
66	9%	-	-	-	-	-	-	-	-	62	69	59
80	18%	-1.02	[-5.20, 3.35]	0.67	0.67	-4.35	[-8.54, 0.26]	0.97	-	41	50	50
96	35%	-1.60	[-5.30, 1.82]	0.82	0.43	-3.47	[-6.50, -0.16]	0.99	-	20	28	18
110	51%	-1.70	[-4.73, 1.13]	0.87	0.20	-2.80	[-5.43, -0.31]	0.98	-	8	8	7
127	67%	-2.10	[-4.37, 0.16]	0.97	-	-2.66	[-5.03, -0.32]	0.99	-	0	0	0
141	83%	-1.99	[-4.28, 0.25]	0.96	-	-2.82	[-5.14 -0.50]	0.99	-	0	0	0
158	94%	-1.69	[-3.80, 0.33]	0.95	-	-2.49	[-4.67, -0.28]	0.99	-	0	0	0
172	100%	-1.87	[-3.97, 0.25]	0.96	-	-3.02	[-5.11, -0.94]	1.00	-	0	0	0

* Number of days from the date of the enrolment of the first subject, - PPW not computed when PPS>0.90.

at study-end when this assessment is performed with very limited information (<40%).

The performance of the bootstrap approach was compared with the outcomes of the Conditional Power and Predictive Probability when 35% of the total information was available (Table **4**). All methods provided similar estimates of the outcome for the 25mg paroxetine, a well-known clinically effective dose [45]. Conversely, when the sub-clinical dose of 12.5 mg paroxetine was analysed, in presence of a weaker signal, only the bootstrap approach was sensitive to the different priors. Instead, Conditional Power systematically overestimated the rate of positive treatment results, while Predictive Probability systematically underestimated them.

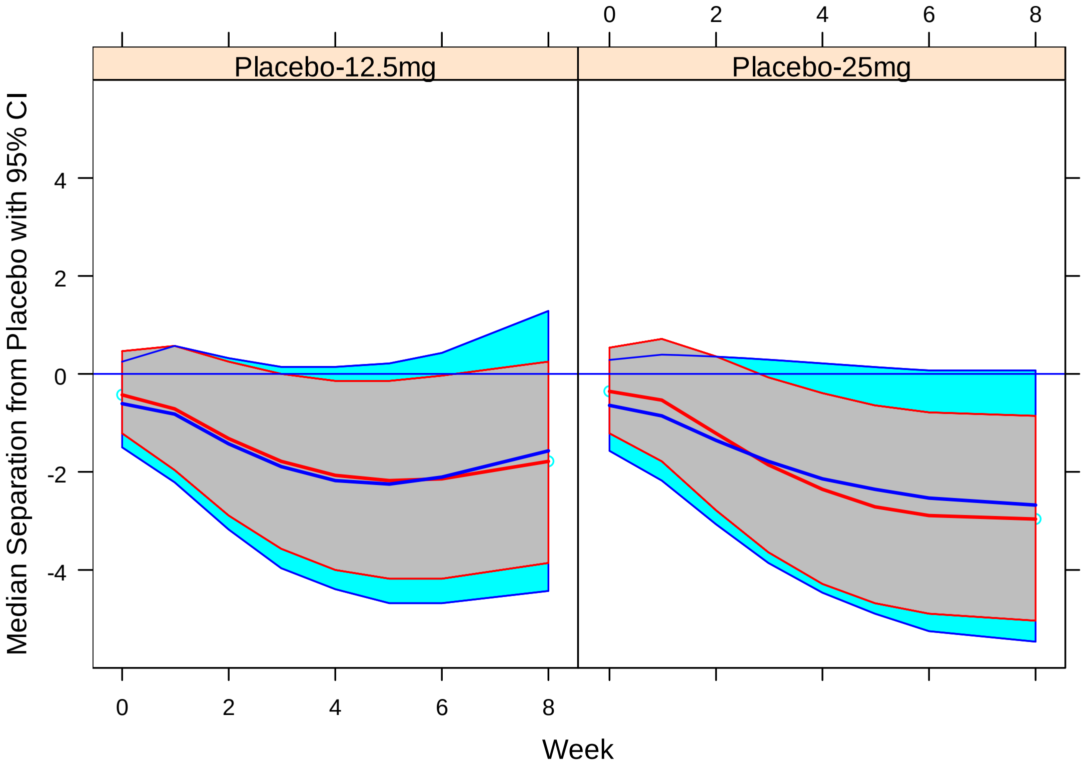

Fig. (3). Case 1: Median separation from placebo for the 12.5mg and 25mg arm with the 95% credible intervals. The solid blue and red lines represent the median separation from placebo at monitoring day 110 and 172. The grey area corresponds to the estimate done at the end of the study (monitoring day 172) while the cyan area corresponds to the analysis conducted when 51% of the data were collected (monitoring day 110).

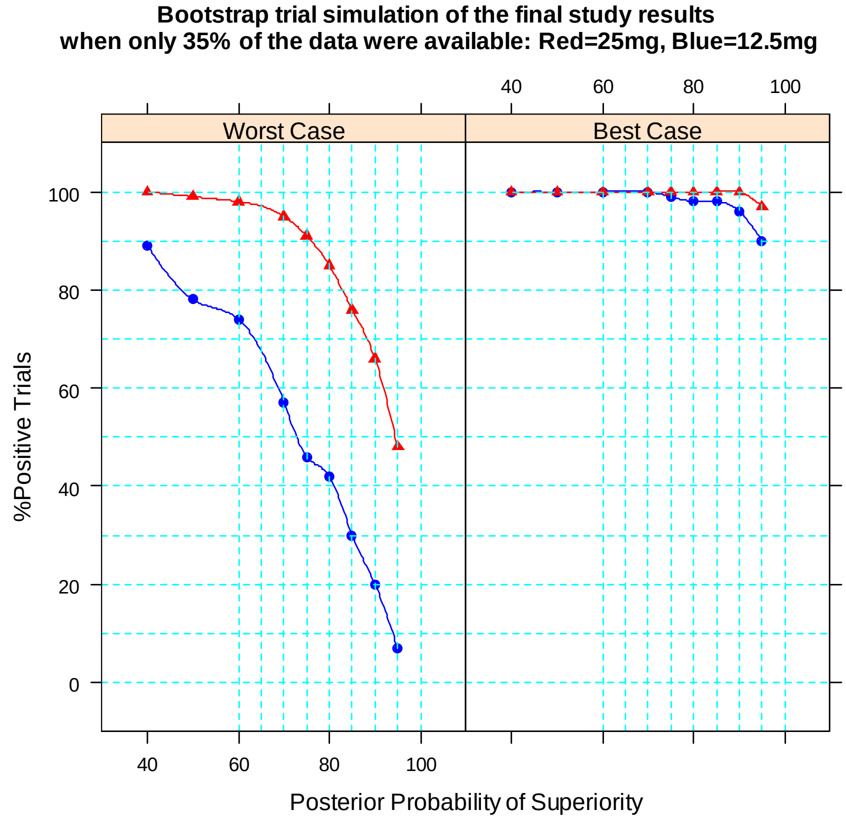

Fig. (4). Case 1: Bootstrap trial simulation of the final study results by treatment arm (red line: 25mg and blue line: 12.5mg) and worst/best simulation scenario, when 35% of the data were available.

As a conclusion, if the bootstrap-based monitoring approach were applied to Case 1 the recommendation at each monitoring time would have been to progress recruitment up to the planned study-end. This recommendation resulted from the combined evaluation of the PPS, the application of the futility stopping rule, and the bootstrap-based risk analysis.

Table 4. Bootstrap Analysis *vs* Conditional Power and Predictive Probability

	Conditional Power		**Predictive Probability**		**Bootstrap**	
	12.5mg	**25mg**	**12.5mg**	**25mg**	**12.5mg**	**25mg**
Reference	0.19	0.56	0.37	0.96		
Skeptical	0.87	0.99	0.38	0.97	0.42	0.85
Clinical	0.98	1	0.39	0.97		
Enthusiastic	1	1	0.43	0.97	0.8	1

Case 2 (Study GSKX)

The results of the Bayesian monitoring (median separation from placebo in paroxetine and Compound X arms) conducted at different time-points with different levels of information are shown in Table **5**. The number of day from the enrolment of the first subject and the percentage of subjects remaining to be enrolled up to study termination are also shown at each monitoring time.

The futility criteria (>40% of the total information, PPS<0.9 and PPW<0.2) for both arms were met at monitoring day 208, when 54% of the total information was available.

Data up to day 208 were used to estimate the time-course of the median separation from placebo for the Compound X and paroxetine arms with the associated 95% credible intervals (Fig. **5**). The location of the median separation from placebo, the size of the 95% credible intervals and the distance of the upper bound of the 95% credible interval from the zero line clearly indicate that there is no reasonable expectation for a clinical effect of Compound X. Similar considerations apply to paroxetine, but to a lesser degree, since the median change showed a trend towards clinical improvement.

Subsequently, the bootstrap trial simulation was performed to assess the risk of a false negative decision. The results of this analysis indicated that the percentage of expected positive trials in the worst and best case scenario for Compound X were 0% and 16% and for paroxetine were 6% and 100%, respectively. The performance of the bootstrap approach was compared with the outcomes of the Conditional Power and Predictive Probability when 54% of the total information was available (Table **6**).

These data support the decision to terminate the Compound X arm (and by the consequence the entire study) considering that, even in the very optimistic event to include only responders (best case scenario), the final study outcome

Table 5. Case 2: Results of the Bayesian monitoring. Bold Values Indicate when the Futility Stopping Criteria is Reached

Monitoring Day*	%Total info	CompX-Placebo Week 8				Paroxetine-Placebo Week 8				% of Patient Remaining to be Enrolled		
		Median Separation	2.5% - 97.5% Credib. Interval	PPS	PPW	Median Separation	2.5% - 97.5% Credib. Interval	PPS	PPW	Plac	CompX	Parox
0	0%	-	-	-	-	-	-	-	-	100	100	100
147	26%	-0.46	[-4.98, 4.64]	0.56	0.48	-4.70	[-9.12, 0.80]	0.97	-	60	61	59
177	39%	1.46	[-2.40, 5.88]	0.24	0.10	-1.46	[-5.50, 3.00]	0.75	0.35	42	51	41
208	**54%**	**1.65**	**[-1.84, 5.40]**	**0.18**	**0.01**	**-1.86**	**[-5.45, 1.97]**	**0.83**	**0.17**	**28**	**28**	**28**
238	69%	1.68	[-1.36, 4.89]	0.14	0.00	-1.58	[-4.87, 1.83]	0.82	0.02	12	16	20
269	85%	1.06	[-1.70, 3.87]	0.22	0.00	-2.02	[-5.07, 0.95]	0.90	0.00	2	3	3
300	96%	1.80	[-0.77, 4.48]	0.09	0.00	-1.21	[-4.01, 1.61]	0.80	0.00	0	0	2
324	100%	1.89	[-0.87, 4.54]	0.09	0.00	-1.36	[-4.32, 1.39]	0.83	0.00	0	0	0

* Number of days from the date of the enrolment of the first subject, - PPW not computed when PPS>0.90.

is expected to have a success rate of only 16%. The paroxetine arm showed a signal of superiority from placebo (as expected), but weak in intensity and high in variability. These data differ from the significant clinical effect reported in study 810 by paroxetine (see Case 1). The lack of expected clear clinical effects of paroxetine in Case 2 is indicative of failed trial.

Table 6. Bootstrap Analysis *vs* Conditional Power and Predictive Probability

	Conditional Power		Predictive Probability		Bootstrap	
	CompX	Parox	CompX	Parox	CompX	Parox
Reference	0.008	0.23	0.005	0.41		
Skeptical	0.14	0.74	0.006	0.42	0	0.06
Clinical	0.36	0.91	0.006	0.43		
Enthusiastic	0.96	1	0.009	0.49	0.16	1

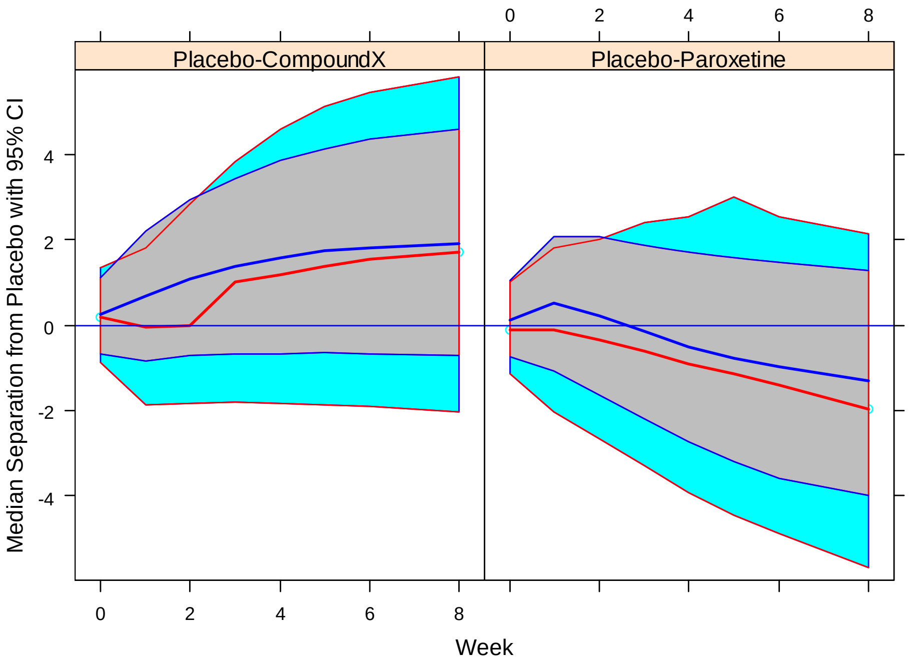

Fig. (5). Posterior probability of separation from placebo for the Compound X and the Paroxetine arm with the 95% credible intervals. The solid blue and red lines represent the median separation from placebo. The grey area corresponds to the estimate done at the end of the study on all the available data while the cyan area corresponds to the analysis conducted when 54% of the total data were available.

Table 7. Case 3: Results of the Bayesian monitoring. Bold Values Indicate when the Futility Sopping Criteria is Reached

Monitoring Day*	%Total Info	Paroxetine-Placebo Week 6				% of Patient Remaining to be Enrolled	
		Median Separation	2.5% - 97.5% Credib. Interval	PPS	PPW	Plac	Parox
0	0%	-	-	-	-	100	100
161	20%	1.28	[-2.26, 4.91]	0.24	0.491	73	75
253	45%	-2.19	[-6.65, 1.99]	0.84	0.33	53	52
312	**51%**	**-1.66**	**[-5.45, 2.11]**	**0.80**	**0.19**	**49**	**43**
373	**61%**	**-2.16**	**[-5.63, 1.36]**	**0.89**	**0.12**	**39**	**33**
434	73%	-2.81	[-5.71, 0.25]	0.97	-	25	24
526	90%	-2.98	[-5.20, -0.77]	0.99	-	12	9
810	100%	-2.96	[-5.12, -0.64]	0.99	-	0	0

* Number of days from the date of the enrolment of the first subject, - PPW not computed when PPS>0.90.

Case 3 (Study 002)

The results of the Bayesian monitoring (median separation of paroxetine from placebo) conducted at different time-points with different levels of information are shown in Table **7**. The number of day from the enrolment of the first subject and the percentage of subjects remaining to be enrolled up to study termination are also shown at each monitoring time.

The futility criteria (>40% of the total information, PPS<0.9 and PPW<0.2) was met at monitoring day 312, when 51% of the data were available. At this time, a consistent fraction of subjects remained to be recruited (49% in the placebo and 43% in the paroxetine arm). Therefore, it was decided to collect extra information (~10%) to increase the confidence in the decision process. At monitoring day 373, when 61% of the data was collected, the futility stopping criteria was again confirmed. At this time-point only 3 centres had been recruiting.

Data up to day 373 were used to estimate the time-course of the median separation of paroxetine from placebo with the 95% credible intervals (Fig. **6**, left panel). In this figure, the location of the median separation from placebo, the size of the 95% credible intervals and the distance of the upper bound of the 95% credible interval from the zero line indicate a trend for clinical improvement, supporting the decision indicated by application of the futility rule to stop the study. The results of the bootstrap trial simulation conducted at monitoring day 373 showed that the percentage of the expected positive trials was 100% in the best case scenario (all responders) and 21% in the worst case scenario (all subjects showing a placebo-like response). The performance of the bootstrap approach was compared with the outcomes of the Conditional Power and Predictive Probability when 61% of the total information was available (Table **8**).

These data indicated that, by study-end, there is a reasonable chance to obtain a clinically relevant separation from placebo. In fact, recruitment in the 4th centre started only after day 380 (Fig. **6**, right panel), significantly contributing to the final success of the study. Therefore, the results of the bootstrap trial simulation (as well as of the other methods) supporting the progression of the trial reversed the recommendation to stop the trial based on the application of the futility rule at monitoring day 373.

Table 8. Bootstrap Analysis *vs* Conditional Power and Predictive Probability

	Conditional Power	Predictive Probability	Bootstrap
Reference	0.99	1	
Skeptical	1	1	0.21
Clinical	1	1	
Enthusiastic	1	1	1

DISCUSSION

A model-based approach has been proposed as a tool for decisions making on termination, progression or adaptation of clinical trials in MDD by monitoring data accrual during the conduction of a trial.

The mixed Weibull/linear equation was identified as the most accurate model to describe the HAMD-17 longitudinal scores, consistently with recent analysis of placebo response in MDD trials [41]. This model adequately described not only the average treatment response but also the individual HAMD-17 time-course as shown by the comparison of the observed and model predicted scores, the goodness of fit plots, and the accuracy of the estimated parameters. The individual model predictions provided evidence supporting the flexibility of the model for the description of heterogeneous HAMD-17 time-course patterns (trajectories), such as linear increase/decrease, bell-shape and exponential decrease observed in the different treatment arms. In addition, the analysis conducted on the truncated databases at different calendar dates before study-end (simulating the accrual process) demonstrated the good predictability

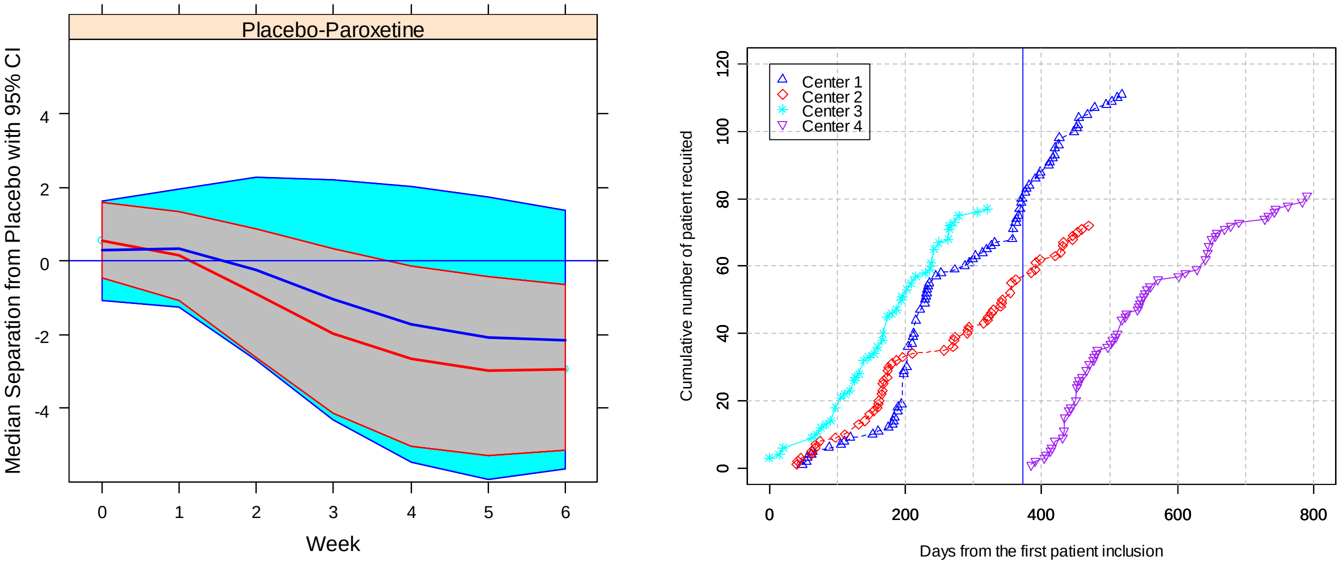

Fig. (6). Case 3: Left panel: Posterior probability of separation from placebo for paroxetine with the 95% credible intervals. The grey area corresponds to the estimate CI done at the end of the study while the cyan area corresponds to the analysis conducted after 373 days when 61% of the total data were available with no subject enrolled in centre 4. Right panel: Cumulative number of subjects recruited versus time by centre.

properties of the model when only partial data were available.

The good predictability properties of the model were used to implement a monitoring strategy aimed to support decision making during clinical trials in MDD. Data monitoring is today an established component of good practice in clinical trials [46-48]. Many statistical approaches have been proposed to data monitoring. Frequentist methods entail calculating critical values for the interim analyses that ensure maintenance of the desired Type I error over the repeated significance tests [49-51]. In contrast to these approaches, Bayesian methods provide a framework for incorporating and updating prior beliefs about drug/placebo effects as a function of the accumulated data [52]. The most commonly used Bayesian methods are the Conditional Power [34] and the Predictive Probability [22]. The decisions driven by the Conditional Power and Predictive Probability approaches remain highly correlated with the degree of subjective judgment employed to define the prior distribution settings. In alternative to these approaches, the bootstrap-based method tries to overcome these limitations by using data-driven priors. These priors were derived from the accrued data during the progression of the trial.

The decision making process is often based on a scenario analysis. This methodology provides an assessment of the risk of study success/failure by using boundaries for possible outcomes estimated with the best and the worst expectations and beliefs. In the Bayesian framework, the best case scenario is associated with optimistic expectation (*i.e.*, Enthusiastic) while the worst case scenario is associated with pessimistic expectation (*i.e.*, Skeptical).

In the Bootstrap-based method, the 'Skeptical' scenario assumes that the data remaining to be collected at the various monitoring time-points show a placebo-like response while the 'Enthusiastic' scenario assumes that all the remaining subjects in the treatment arm are drug responders. The method does not require parametric assumption on the shape of the distributions. Individual data are generated using non-parametric bootstrapping approach by resampling with replacement the original data collected up to the monitoring time. In the present work we compared the performances of three methods: Conditional Power, Predictive Probability, and Bayesian bootstrap.

In case of strong clinical signal (25mg paroxetine arm in study 810), all methods provided similar estimates of the risk to continue/discontinue the trial with no differences between worst and best scenarios. In case of week clinical signal (12.5mg paroxetine arm in study 810), only the bootstrap approach was sensitive to different priors, while Conditional Power systematically overestimated the rate of positive treatment results, and Predictive Probability systematically underestimated them. The results of the analysis are consistent with results recently published, showing that predictive probability decision rules based on 'Enthusiastic' priors are more likely to trigger an early stopping in futility monitoring [22]. Thus, over-Enthusiastic prior would be required for an appropriate decision making. In contrast, the Conditional Power overestimates the rate of positive trial outcomes (including when a 'Skeptical' scenario is considered), leading to unjustified recommendations to progress trial [22].

In the present work, the decision to discontinue a treatment arm or a trial for futility was based on the availability of about 40% of the total information and on the joint assessment of PPS and the predictive power (PPW), estimating the probability of rejecting the null hypothesis of equal treatment effects at the study-end given the interim observations and the assumptions about the prior distributions of the treatment effects.

This decision-making methodology was tested by conducting a retrospective analysis of 3 clinical trials. In the first trial (810) the conditions for discontinuing arm/trail for futility were never reached at any monitoring time-points. The strong signal of separation of the 25mg paroxetine arm from the placebo arm was estimated using bootstrap trial simulation with <35% of the total information collected in the trial. In the second trial (GSKX) a novel compound (Compound X) was tested in MDD. The criteria for discontinuing the Compound X arm was met at both 39% and 54% of total information. The bootstrap trial simulation estimated a success rate of 16% in the best case scenario, triggering the recommendation to stop the Compound X arm and, hence, to an early discontinuation of the trial. In the third trial (002) we provided an example of deviation from the exchangeability assumption due to the late inclusion of an extensively recruiting centre. The criteria for futility discontinuation were reached at both 51% and 61% of total information. However, the bootstrap trial simulation estimated a success rate of 100% and 21% in the best and the worst case scenario, respectively, recommending the continuation of the trial.

One of the critical issues in antidepressant drugs development is the complexity and relatively long duration of clinical trials (2 years on average). These factors represent serious drawback for the implementation of an effective portfolio management strategy. In early clinical development, ‘Proof of concept' and signal detection trials are carried out to determine if a treatment is clinically active or inactive before commitment of investment in late phase drug development. In this framework, the possibility of an early detection of lack of efficacy can trigger the decision to terminate the trial, to discontinue the development of the drug under investigation, and, more effectively, to redirect resources more productively. If the estimated probability of success is high, the clinical development plan can be accelerated by an early initiation of confirmatory trials without affecting the course of the PoC trial in any way. On this basis, the proposed methodology can constitute a novel alternative decision-making tool to improve the overall productivity of drug development by re-directing the (sometime limited) resources in the most promising projects.

In conclusion, the comparisons of Condition Power and Predictive Probability approaches indicated that the Bayesian bootstrap method, based on data-driven assumptions for priors, provided a better control for the risk of inappropriate decisions. These results suggest that the proposed approach to monitor the accumulating information could constitute a valuable alternative to the re-engineering of the development process of novel antidepressant drugs. The WinBUGS code for longitudinal models and the code used to implement the bootstrap trial simulation can be provided upon request from the authors.

REFERENCES

[1] Robinson DS, Rickels K. Concerns about clinical drug trials. J Clin Psychopharmacol 2000; 20: 593-6.

[2] Robinson DS, Khan A. Dosing strategies for antidepressant clinical trials: a commentary. J Clin Psychopharmacol 2004; 24: 1-3.

[3] Fava M, Evins AE, Dorer DJ, Schoenfeld DA. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. Psychother Psychosom 2003; 72: 115-27.

[4] Krishnan KRR. Efficient trial designs to reduce placebo requirements. Biol Psychiatry 2000; 47: 724-6.

[5] Sheiner LB. Learning versus confirming in clinical drug development. Clin Pharmacol Ther 1997; 61: 275-91.

[6] Berry DA. A guide to drug discovery: bayesian clinical trials. Nat Rev Drug Dis 2006; 5: 27-36.

[7] Krams M, Lees KR, Berry DA. The past is the future. Innovative designs in acute stroke therapy trials. Stroke 2005; 36: 1341-7.

[8] Marks RG. Validating electronic source data in clinical trials. Control Clin Trials 2004; 25: 437-46.

[9] Fayers PM, Ashby D, Parmar MK. Tutorial in biostatistics Bayesian data monitoring in clinical trials. Stat Med 1997; 16: 1413-30.

[10] Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. Stat Sci 1990; 5: 299-317.

[11] Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. London/Boca Raton, FL: Chapman & Hall/CRC 2000.

[12] Whitehead J. The Design and Analysis of Sequential Clinical Trials 2nd ed. Chichester, NY: Wiley 1997.

[13] Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical terms. Commun Stat Sequential Anal 1982; 1: 207-19.

[14] Betensky RA. Early stopping to accept H0 based on conditional power: approximations and comparisons. Biometrics 1997; 53: 794-806.

[15] Berry DA. Monitoring accumulating data in a clinical trial. Biometrics 1989; 45: 1197-211.

[16] Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? Control Clin Trials 1986; 7: 8-17.

[17] Choi SC, Smith PJ, Becker DP. Early decision in clinical trials when the treatment differences are small. Control Clin Trials 1985; 6: 280-8.

[18] Johns D, Anderson JS. Use of predictive probabilities in Phase II and Phase III clinical trials. J Biopharm Stat 1999; 9: 67-79.

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

[20] Geisser S. On the curtailment of sampling. Canadian J Stat 1992; 20: 297-309.

[21] Geisser S, Johnson W. Interim analysis for normally distributed observables. Multivariate Analysis and Its Applications. IMS Lecture Notes 1994; 263-79.

[22] Dmitrienko A, Wang M-D. Bayesian predictive approach to interim monitoring in clinical trials. Stat Med 2006; 25: 2178-95.

[23] Brooks SP. Markov chain Monte Carlo method and its application. Statistician 1998; 47: 69-100.

[24] Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50. MRC Biostatistics Unit: Cambridge 1996.

[25] Spiegelhalter DJ, Thomas A, Best NG, Lunn D. WinBUGS, Version 1.4, User Manual. MRC Biostatistics Unit: Cambridge 2001.

[26] Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian-analysis of realistically complex-models. J R Stat Soc Ser A Stat Soc 1996; 159: 323-42.

[27] Sutton AJ, Abrams KA, Jones DR, Sheldon TA, Song F. Methods for Meta-analysis in Medical Research. Chichester: Wiley 2000.

[28] Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Stat Med 2001; 20: 453-72.

[29] Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. Stat Med 2001; 20: 435-52.

[30] Burton PR, Tiller KJ, Gurrin LC, Cookson W, Musk AW, Palmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMS) and Gibbs sampling. Genet Epidemiol 1999; 17: 118-40.

[31] Goldstein H, Spiegelhalter DJ. League tables and their limitations - statistical issues in comparisons of institutional performance. J R Stat Soc Ser A Stat Soc 1996; 159: 385-409.

[32] Dixon DO, Simon R. Bayesian subset analysis in a colorectal cancer clinical trial. Stat Med 1992; 11: 13-22.

[33] Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000; 4(38): 1-136.

[34] Spiegelhalter DJ. Incorporating Bayesian Ideas into Health-Care Evaluation. Stat Sci 2004; 19(1): 156-74.

[35] Draft Guidance for Industry and FDA Staff Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials US Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health Division of Biostatistics Office of Surveillance and Biometrics, 2006. [cited 2009 May 13]. Available from: http://www.fda.gov/cdeh/osb/guidance/1604.pdf

[36] Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression: variable, substantial, and growing. JAMA 2002; 287: 1840-7.

[37] Gould AL. Timing of futility analyses for 'proof of concept' trials. Stat Med 2005; 24: 1815-35.

[38] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife and cross validation. Am Stat 1983; 37: 36-48.

[39] Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. Analysis of Clinical Trials Using SAS: A Practical Guide. Cary, NC: SAS Publishing 2005.

[40] Ette EI, Williams PJ, Lane JR. Population Pharmacokinetics III: design, analysis, and application of population pharmacokinetic studies. Ann Pharmacother 2004; 38: 2136-44.

[41] Gomeni R, Merlo-Pich E. Bayesian modelling and ROC analysis to predict placebo responders using clinical score measured in the initial weeks of treatment in Depression trials. Br J Clin Pharmacol 2007; 63: 595-613.

[42] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). J R Stat Soc B 2002; 64: 583-640.

[43] Trivedi MH, Pigotti TA, Perera P, Dillingham KE, Carfagno ML, Pitts CD. Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder. J Clin Psychiatry 2004; 65: 1356-64.

[44] Dunbar GC, Claghorn JL, Kiev A, Rickels K, Smith WT. A comparison of paroxetine and placebo in depressed outpatients. Acta Psychiatr Scand 1993; 87: 302-5.

[45] Rapaport MH, Schneider LS, Dunner DL, Davies JT, Pitts CD. Efficacy of controlled-release paroxetine in the treatment of late-life depression. J Clin Psychiatry 2003; 64: 1065-74.

[46] Ashby D, Tan S-B. Where's the utility in Bayesian data-monitoring of clinical trials? Clin Trials 2005; 2: 197-208.

[47] Ellenberg SS, Fleming T, DeMets DL. Data monitoring committees in clinical trials. New York: Wiley and Sons, 2002.

[48] Gary L, Rosner GL. Bayesian Monitoring of Clinical Trials with Failure-Time Endpoints. Biometrics 2005; 61: 239-45.

[49] Geiller NL, Pocock SJ. Interim analyses in randomised clinical trials: ramifications and guidelines for practitioner. Biometrics 1987; 43: 213-23.

[50] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549-56.

[51] Pocock SJ. Size of cancer clinical trials and stopping rules. Br J Cancer 1978; 38: 757-66.

[52] Berry DA. Interim analysis in clinical trials: the role of the likelihood principle. Am Stat 1987; 41: 117-22.

Received: December 15, 2008 Revised: April 20, 2009 Accepted: April 22, 2009