To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

10

Information-based adaptive designs for the assessment of efficacy of antidepressant drugs:

A Bayesian framework for interim analysis accounting for enrolment rate

Gijs Santen, Erik van Zwet, Meindert Danhof, Oscar Della Pasqua

ABSTRACT

Clinical trials in depression often fail to detect drug effect, even for drugs that are known to be efficacious. Implementation of an interim analysis as a component of the assessment may help detecting failing studies early as well as prevent unnecessary continuation of enrolment once treatment effect is sufficiently large to demonstrate efficacy. Previously, we have shown how the posterior predictive power (PPP) within a Bayesian framework can be used to implement an interim analysis for clinical trials in depression. In the current investigation, we propose an adaptive approach to optimise the timing and decision criteria based on the use of simulated data and actual data of current study enrolment, taking into account the uncertainty in the estimation of the PPP.

The adaptive evaluation procedure for optimisation of the interim analysis was performed with a linear longitudinal mixed-effects model in which posterior predictive distributions are derived incorporating uncertainty into the estimation procedure by means of simulations and bootstrapping of the interim datasets.

First, the timing of the interim analysis and the optimal decision criteria are determined by simulating datasets conditional on the enrolment data of the actual clinical trial. Subsequently, choices regarding decision criteria and start times are made according to changes in the utility function using a range of decision criteria and start times.

The adaptive procedures are validated by evaluating the decisions reached upon application of the previously selected decision criteria on 're-enrolled' patients from two historical studies in depression from GlaxoSmithKline's clinical database.

The proposed approach allows the determination of the optimal timing and decision criteria for an interim analysis in clinical trials with antidepressant drugs. Validation procedures based on the re-enrolment of patients confirmed the relevance of the interim analysis as part of a clinical protocol to prevent continuation of inefficacious treatment arms whilst limiting the risk of premature stopping of efficacious treatment arms, and *vice versa*.

In contrast to the conventional methods currently available for interim analysis, the proposed adaptive approach allows the determination of the optimal timing and decision criteria to demonstrate efficacy or futility. The re-enrolment test provides evidence that this adaptive framework yields control of the risk of erroneous decisions. Advancement of clinical research with antidepressants requires solutions beyond established beliefs. We recommend the use of an adaptive Bayesian framework for interim analysis of antidepressant drugs. Optimisation of the timing and decision criteria at the interim stage is critical for the accuracy of the conclusions about drug efficacy or futility.

INTRODUCTION

The lack of novel antidepressants reflects one of the key problems of drug development: i.e., the consistently high failure and attrition rate. Despite the identification of novel targets and insight into the causes of depression, the ability to demonstrate the efficacy of new compounds in the patient population stumbles on major hurdles, some of which go beyond study implementation issues, such as power calculations. In addition, whilst selection of the appropriate dose range remains a key issue in the design of proof-of-concept studies in many therapeutic areas, this is certainly not the main problem in depression. This is clearly verified by the frequency of studies which do not demonstrate statistically significant separation from placebo, even when appropriate doses have been selected (Khan *et al.*, 2002b). The issues that arise from such evidence compel us to consider the nature and magnitude of factors associated with uncertainty in the estimation of true drug effect and its consequences for trial design and decision making.

Many factors contribute to the high failure rate of clinical trial in depression. An important factor which complicates the assessment of treatment effect of antidepressant drugs in clinical trials is the poor sensitivity of the clinical endpoint, the Hamilton depression rating scale (HAMD) (Hamilton, 1960). Furthermore, placebo response rate has been shown to compound the problem (Walsh *et al.*, 2002). Even though increased awareness of the implications of placebo have suggested it be abandoned as reference treatment in clinical trials in general (Hrobjartsson and Gotzsche, 2001, 2004), a different opinion has evolved with respect to placebo treatment in clinical trials in depression (Yang *et al.*, 2005; Khan *et al.*, 2002a).

From the heterogeneity and range of factors contributing to uncertainty in outcome, it is clear that fixed designs may not be advantageous. Flexible or adaptive designs can offer an opportunity to the implementation of a learning-confirming paradigm (Sheiner, 1997) in the early stage of development of antidepressant drugs. Adaptive designs rely on the accumulation of data to decide on how to modify aspects of the study without undermining the validity and integrity of the trial. Interim analysis forms an intrinsic part of such designs.

An overview of the adaptive design methodology is not in the scope of this manuscript. We refer to and adopt the terminology suggested by Dragalin (2006). In this framework, adaptive designs are divided according to the rules involved: allocation rules (how subjects will be allocated to available treatment arms), sampling rules (how many subjects will be sampled at the next stage), stopping rules (when to stop the trial due to evidence of efficacy, harm, futility) and decision rules (any other adaptation, such as change of endpoint). Adaptations based on decision rules are less common in the published literature.

In the current manuscript we demonstrate the implementation of an interim analysis with a stopping rule based on predictive power (Spiegelhalter *et al.*, 1986, 2004). The decision criteria of the stopping rule and the timing of the interim analysis are adapted on the enrolment rate, using simulated patient data. This approach is inspired by information-

based designs (Tsiatis, 2006), in which a decision is based on the amount of statistical information obtained. In our approach, the decision to perform an analysis is based on a combination of the enrolment rate and prior knowledge of parameter distributions (through simulations). Furthermore, we account for the uncertainty in the distribution of the posterior predictive power (predictive power). All data gathered up to the moment of the interim analysis is used, including data from patients that have not completed their treatment period. To scrutinise the proposed procedures on actual datasets, we propose a re-enrolment test, in which the value of historical data is further incorporated into the validation.

METHODS

Study data

Two placebo-controlled clinical trials in patients with major depression have been obtained from GlaxoSmithKline's clinical trial database. The studies were selected to include one positive and one negative trial, as concluded by standard statistical methods. Study 1 (Trivedi *et al.*, 2004) was a randomised placebo-controlled trial in which two doses (12.5 and 25 mg) of a controlled release (CR) formulation of paroxetine were tested for efficacy and safety. The HAM-D_{17} (Hamilton, 1967) was measured at weeks 1, 2, 3, 4, 6 and 8 after start of treatment. A total of 459 patients with major depression were evenly enrolled across the treatment arms.

Study 2 (DeVeaugh-Geiss *et al.*, 2000) was a randomised placebo-controlled trial in which lamotrigine was tested for efficacy and safety. The HAM-D_{17} (Hamilton, 1967) was measured at weeks 1, 2, 3, 4, 5, 6 and 7 after start of treatment. A total of 300 patients with major depression were evenly enrolled across the treatment arms.

Based on the findings from a previous investigation (chapter 9) and considering the burden due to computational time, we have decided to include only those observations made at weeks 2 and 6 in the analysis of both studies. To interpret the findings in the proposed interim analysis, it is important to consider the results of the actual studies. Therefore, estimates of the treatment effect were obtained by fitting a dual random effects model (DREM) to the data. An alternative approach to assess the significance of treatment effect is to determine the distribution of the posterior predictive power (see below) of the full dataset, as a function of the number of patients enrolled.

Linear mixed effects model

The decision on the timing, sample size and decision criteria were established using the dual random effects model (DREM) (equation 1), implemented in WinBUGS 1.4.1 (Lunn *et al.*, 2000). WinBUGS is a statistical program which allows straightforward application of Bayesian statistics by estimating the posterior distribution using a Markov chain Monte Carlo (MCMC) approach.

$$Y_{ij} = \mathsf{BAS}_i \cdot \beta_j + \theta_{z,j} + \eta 1_i + \eta 2_i \cdot j + \epsilon_{ij} \quad (1)$$

where BAS_i is the baseline for individual i, β_j is the baseline-time interaction at time j, $\theta_{z,j}$ represents the effect of treatment z at time j. $\eta 1_i$ and $\eta 2_i$ are the random effects of individual i (from a multivariate distribution with means 0 and unknown variance-covariance matrix) and ϵ is the measurement error (normally distributed with mean 0 and unknown variance).

Factors influencing the appropriate timing of an interim analysis

In the current investigation we use simulations which take into account how different factors influence the conclusions derived from an interim analysis as well as their consequences for the final outcome: **(1)** the enrolment rate from the study up to date, which determines the information-acquisition rate (and hence the timing of the interim analysis), **(2)** the duration of the study, **(3)** dropout and **(4)** the expected magnitude of the drug-placebo difference.

The aforementioned factors are used in conjunction with the model to simulate individual HAMD scores. Then, different timings and decision criteria are evaluated on these simulated datasets, which test the utility of the interim analysis and the statistical properties of the proposed criteria.

(1) Enrolment data is obtained from the study for which an interim analysis is performed. Simulated patient data using the DREM and parameter estimates from historical data are included into the analysis according to the observed enrolment rate of the study in progress, taking into account the duration of the study **(2)**.

(3) A dropout scenario is simulated according to findings in historical data, as proposed in previous work (chapter 9). In brief, the dropout scenario consists of dropout patterns according to missingness completely at random (MCAR) for the first 3 weeks, and a combination of MCAR and dropout according to missingness at random (MAR) in a 3:1 ratio (only dropout in 25% most severely depressed patients) thereafter. The dropout percentage was 4% per week.

(4) Throughout the manuscript, we have used a 2-point HAMD effect as a clinically relevant effect. In order to estimate the false positive rate (type I error) a treatment arm with no effect was also simulated.

Distribution of the posterior predictive power

The posterior predictive power (PPP) is obtained as follows. The last 1,000 samples of the posterior distributions are transferred to the language and environment for statistical computing R (R Development Core Team, 2007). R simulates new trials based on the posterior distribution of all parameters and calculates in how many of these trials the treatment is significantly different from placebo (p<0.05) using t-tests. The percentage of positive (statistically significant) trials is summarised as the posterior predictive

power (PPP). In order to obtain information about the uncertainty of the PPP and to describe its distribution, 100 bootstrap replicates of the dataset are created and the PPP is summarised for all 100 datasets. These results correspond to the distribution of the PPP.

Adapting timing and decision criteria

Utility function

We have defined the utility U of the interim analysis as its ability to separate between a treatment arm with no effect and a treatment arm with a clinically relevant effect (i.e., 2 points HAMD) (equations 2 and 3).

$$U_{\text{futility}} = P\{\text{accept } H_0 | \Delta\text{HAMD} = 0\} - P\{\text{accept } H_0 | \Delta\text{HAMD} = 2\} \quad (2)$$

$$U_{\text{efficacy}} = P\{\text{reject } H_0 | \Delta\text{HAMD} = 2\} - P\{\text{reject } H_0 | \Delta\text{HAMD} = 0\} \quad (3)$$

where the ΔHAMD is the difference between active and placebo in points HAMD, and H_0 is the null hypothesis of no difference between active and placebo.

Two additional restrictions were imposed: the false negative rate (type I error) should remain below 5%, and the false positive rate (type II error) should remain below 20%. Since these error rates are determined using repeated interim analyses, multiplicity is taken into account.

Optimisation of the decision criteria and timing of the interim analysis

Besides the time at which to start the interim analysis, three elements are required to optimise the decision criteria: the futility boundary, the efficacy boundary and the degree of evidence (percentage of the distribution of the PPP) that triggers a decision (figure 1). Using 100 of the simulated datasets described above, a range of values for each of these elements is evaluated as well as a range of starting times. The combination of starting time and decision criteria which provides the highest value for the utility function is subsequently used in the interim analysis of the study in progress. The interim analysis starting times to be evaluated were derived from the enrolment rate observed in the actual trial. Based on preliminary investigations, the range of decision criteria for futility has been set at 5, 10, ..., 50% PPP and for efficacy at 50, 55, ..., 95% PPP. The degree of evidence required to trigger a decision was evaluated in the range of 50, 55, ..., 95%. The final decision criteria and timing to be used in the interim analysis are those yielding the highest utility values for all the aforementioned combinations. However, since the utility function may continue to increase with progressively incoming data, the expected increase in utility has to be balanced against the benefits of stopping the clinical trial earlier and preventing patients from being accrued unnecessarily. We have chosen not to formalise these considerations by adding them as a negative factor to the utility functions (i.e., penalty or cost function), but it is certainly possible and future work will attempt to do this in a meaningful manner.

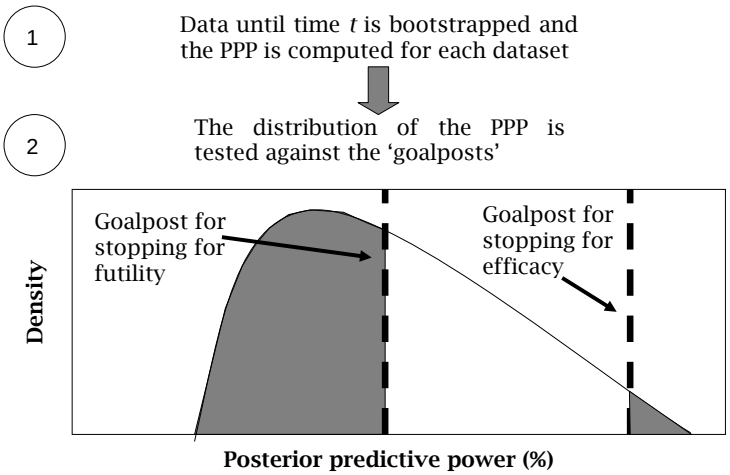


Figure 1. Illustration of the decision process. (1) 100 bootstrap replicates of the obtained dataset are created. The PPP is computed for each dataset and the distribution is plotted. (2) If a pre-specified percentage of the distribution of the PPP is beyond one of the boundaries (grey areas), the corresponding decision is triggered

Statistical evaluation of the approach using the re-enrolment test

As indicated above, the timing and decision criteria of the interim analysis were based on the observed enrolment rate in actual studies. Initially, we have considered the validation of the methodology by performing an analysis of the performance of the selected decision criteria and hence the proposed methodology. However, this approach provides only limited information. Instead, we propose an additional step to test the methodology in a more rigorous manner. Assuming exchangeability between patients and random enrolment, it is possible to 're-enrol' all patients in a different order. Based on this principle, the validation step will consist of running the interim analysis on 100 trials containing re-enrolled patients, each of which will be different since different patients will have entered the trial up to a specific time since the beginning of the trial. A truly robust interim analysis should perform consistently in the majority of the cases.

The percentage of treatment arms stopped for efficacy and futility is reported and may, if linked to the outcome of the trial, provide evidence that the proposed methodology is unbiased. These procedures investigate whether patients are in fact exchangeable, as is commonly assumed in statistical methods. The exchangeability will be visualised by plotting the time course of the median PPP for 100 trials in which patients were re-enrolled.

Impact of the application of the interim analysis

The savings in terms of patients and time it takes to reach a final decision will be calculated to assess the impact of an interim analysis. For these calculations, the time at which 50% of the trials with re-enrolled patients is stopped will be considered the interim analysis time.

RESULTS

Estimation of treatment effect in the clinical trials

To evaluate the decisions made by the interim analysis, it is important to first establish which conclusions should be drawn by analysing the full datasets. First, the DREM has been fitted to the datasets (table 1). Second, the distribution of the PPP has been determined for both studies (figure 2). These two analyses lead to the same conclusion: study 1 was a positive study, with a highly significant treatment effect for the 25 mg paroxetine dose and a less significant effect for the 12.5 mg paroxetine dose. Study 2 did not demonstrate any efficacy for lamotrigine. As an alternative to using the full distribution of the PPP, it is possible to focus on the point estimates from the original dataset, assuming 150 patients per treatment arm. In this case, the PPP is 73% and 98% for the 12.5 and 25 mg arms (study 1) respectively and 16% for lamotrigine (study 2).

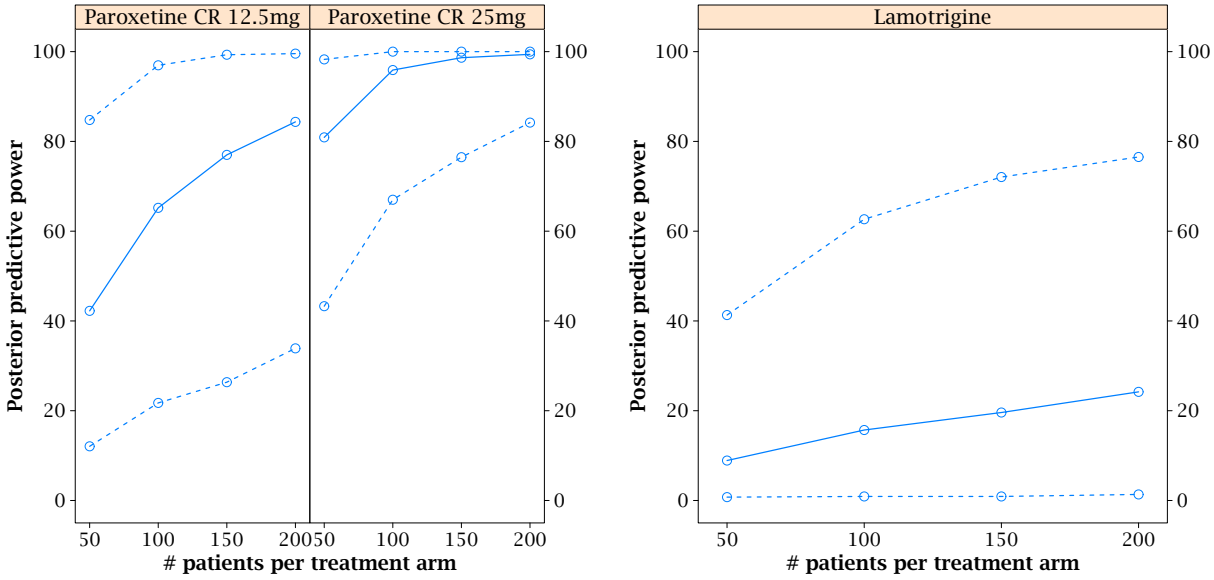
Application of the interim analysis

Study 1

Figure 3 shows the relation between the utility of the interim analysis, the day for starting the analysis and the decision criteria. The utility associated with the decision to stop for efficacy increases when the interim analysis procedure is delayed. In contrast, the utility associated with the decision to stop for futility is less sensitive to the timing of the interim analysis, although the decision criteria leading to the highest utility do depend on timing. Based on figure 3, a decision was made to start the interim analysis at day 70, with efficacy and futility boundaries of 45% and 60%. The degree of evidence required in order to trigger a decision was set at 85%.

Table 1. Results of the analysis of the data using the double random effects model (DREM). The posterior probability of inferiority (comparable to a one-sided *p*-value) is shown together with the mean and 95% credible interval for the treatment effect at the last scheduled visit

Study	Treatment	PPS	Treatment effect (95% Credible interval)
1	paroxetine (12.5 mg)	0.004	-2.0 (-3.5,-0.6)
	paroxetine (25 mg)	<0.001	-3.6 (-5.2,-2.1)
2	lamotrigine (max 200 mg)	0.36	-0.3 (-2.2,1.5)



(a) Study 1 (b) Study 2

Figure 2. Median (uninterrupted line) and 95% credible interval (interrupted lines) of the posterior predictive power as a function of the size of a future trial

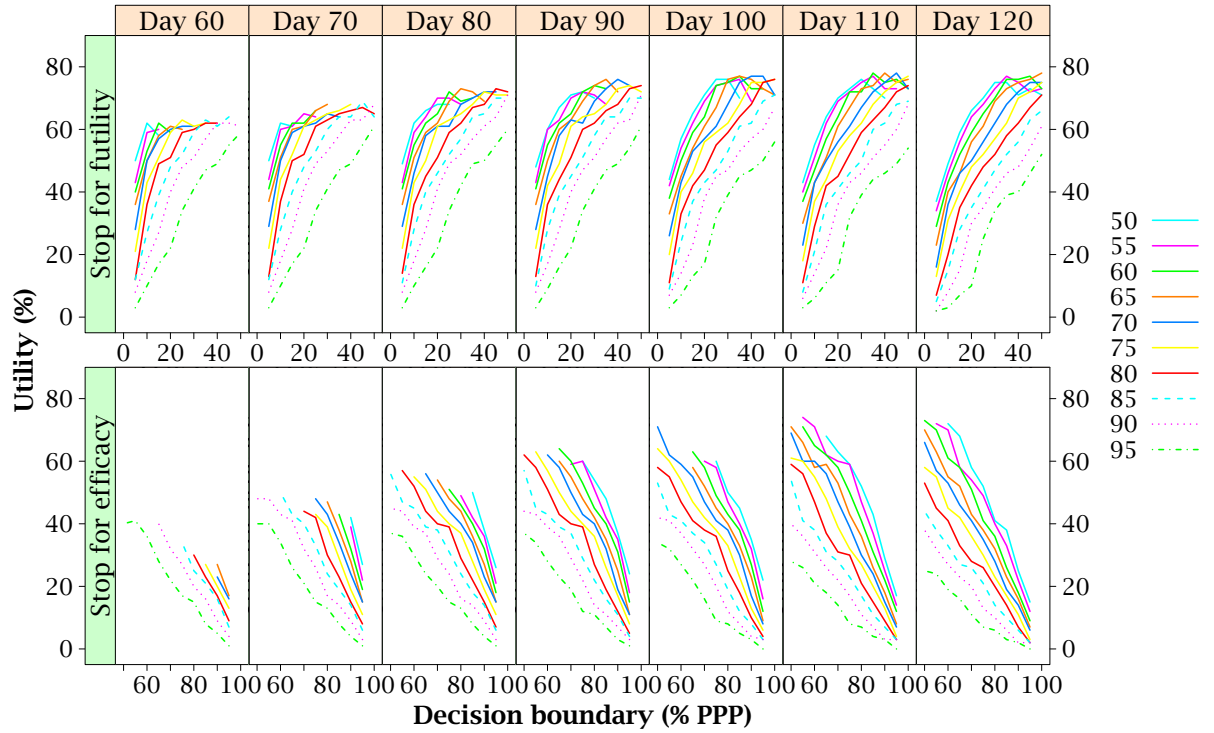


Figure 3. The influence of the stopping boundary on the utility of the interim analysis (100 simulations) for study 1. Different panels are shown for the start of the interim analysis in days after start of enrolment and for the different reasons (futility/efficacy). The different lines represent the percentage of the PPP that needs to be beyond a decision boundary to trigger a decision. Decision criteria leading to false positive rates > 5% or false negative rates > 20% are not depicted

Subsequently, validation of the procedures was performed using the re-enrolment test on the actual trial data. Figure 4 shows histograms of the frequency of futility/efficacy decisions and the timing at which these decisions were reached, as determined by the re-enrolment test. The 12.5 mg paroxetine arm has a treatment effect of approximately 2 points HAMD. Since the maximum futility stopping rate (type II error) was 20% for such a treatment effect, it is expected that the futility stopping rate will be in the same order of magnitude. In fact, an evaluation using the re-enrolment test shows that the interim analysis stops this treatment arm for futility in 7% of the trials (figure 4). Stopping for efficacy of the 12.5 mg arm occurs in 40% of the trials and for the 25 mg paroxetine arm in almost all of the replicates.

Study 2

The same methodology was applied to the data from study 2. This study had a clearly negative outcome and was therefore suitable to test whether the interim analysis would detect inefficacious treatment arms. The consequences of timing and decision criteria on utility are investigated in figure 5.

Based on the same considerations as for study 1, the interim analysis started at day 200 with boundaries at 40% and 65%, with the required degree of evidence set at 80%. The selected decision criteria were subsequently tested using the re-enrolment method. The study was found to be terminated in 83% of the trials evaluated using the re-enrolment test. A stopping decision for efficacy (false positive) was made in 2% of the trials, well under the maximum of 5%, which was deemed acceptable.

Consequences of the interim analysis

Table 2 illustrates the number of patients that were enrolled at the moment of an interim analysis decision to put its relevance into perspective. As shown, the impact of the interim analysis on the number of patients enrolled increases with a decreasing enrolment rate. Overall, the use of the proposed adaptive design with reassessment of interim analysis criteria leads to substantial savings in terms of the number of patients required for the interim analysis, as well as in terms of time it takes to reach an accurate decision about treatment effect.

Table 2. Impact of the interim analysis on the number of enrolled patients and the time saved (defined as the moment at which 50% of the re-enrolled datasets were terminated)

Study	Treatment	Day	Time completed	Patients enrolled
1	paroxetine (12.5 mg)	-	-	-
	paroxetine (25 mg)	70	37%	60%
2	lamotrigine (max 200 mg)	220	45%	44%

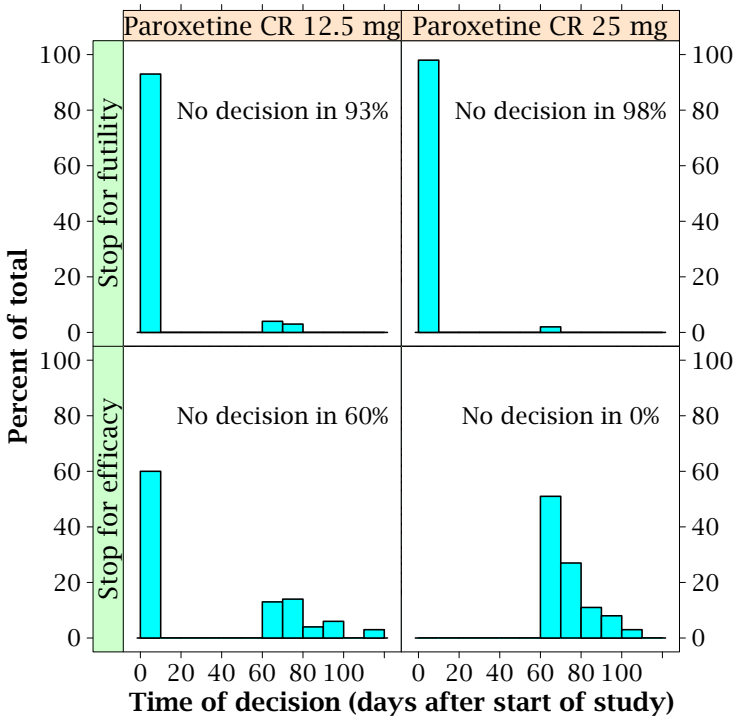


Figure 4. The probability of stopping a treatment for futility (top panels) and efficacy (bottom panels) *versus* time for paroxetine 12.5mg (left panels) and 25mg (right panels)

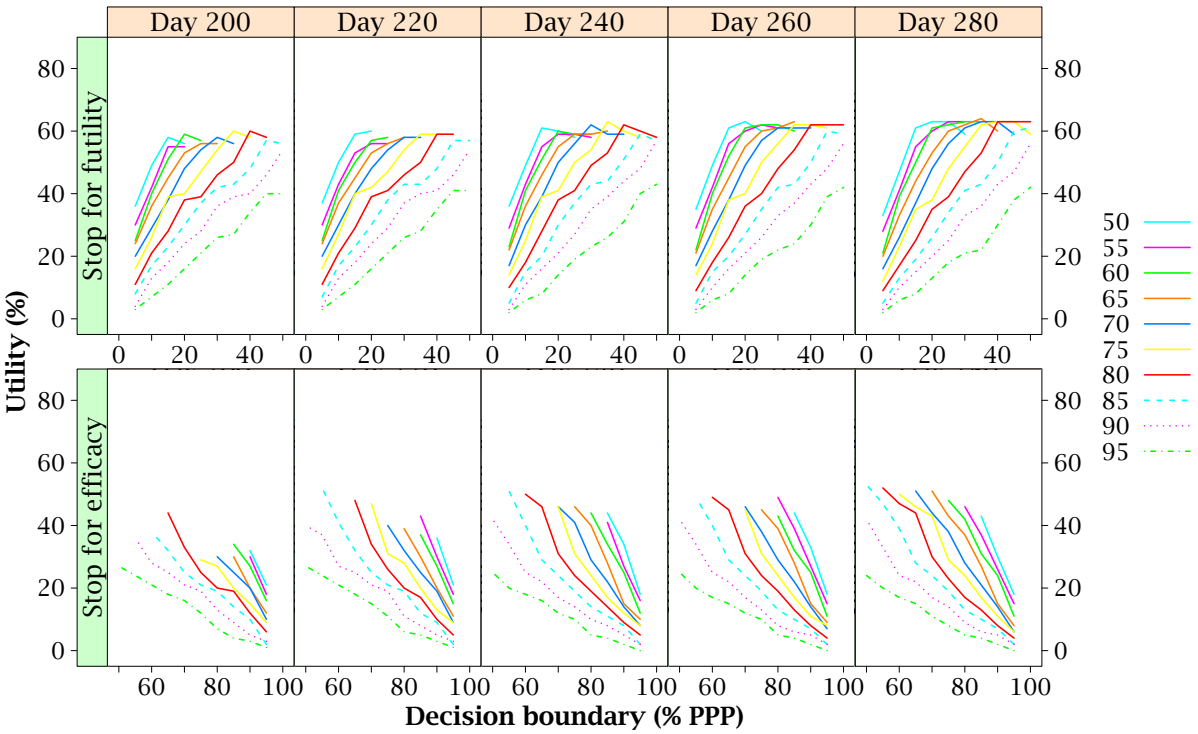
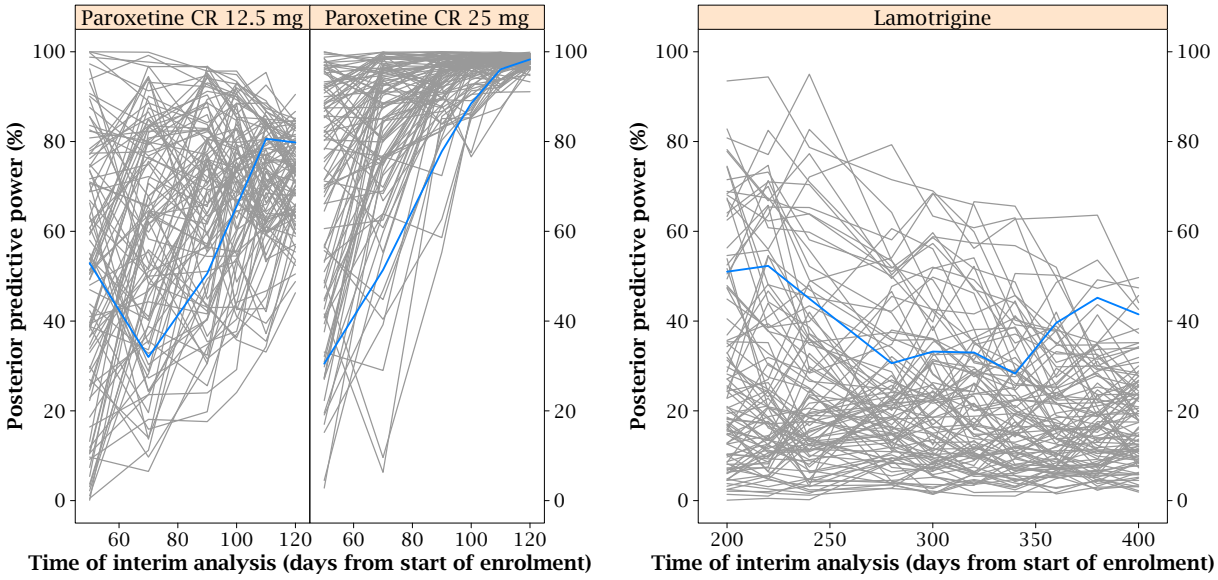


Figure 5. The influence of the stopping boundary on the utility of the interim analysis (100 simulations) for study 2. Different panels are shown for the start of the interim analysis in days after start of enrolment and for the different stopping reasons (futility/efficacy). The different lines represent the outcomes using different values for the required evidence to trigger a decision. Decision criteria leading to false positive rates > 5% or false negative rates > 20% are not depicted

Patients' exchangeability

The re-enrolment test allows investigation of the exchangeability of patients in clinical trials in depression. The time course of the point estimate of the PPP of 100 trials containing randomly re-enrolled patients was plotted for either study (figure 6). These data illustrate why it is difficult to perform an interim analysis in depression: the patient population is extremely heterogeneous. As expected, the PPP of the different trial replicates converges as more patients are enrolled in a trial, and more heterogeneity seems to be present when no treatment effect is visible (study 2 *versus* study 1). Clearly, undertaking an interim analysis at an earlier time than those estimated using the proposed methodology yields a high risk of inaccurate decision about the true treatment effect.



(a) Study 1

(b) Study 2

Figure 6. Posterior predictive power (point estimate) for 100 re-enrolled datasets *versus* time after start of enrolment

DISCUSSION

In contrast to the typical problems associated with dose finding studies, the high failure rate of clinical trials in depression seems to be independent of the selection of an appropriate dose range. In previous investigations we have shown that many factors explain the potential causes of such failure. First, the use of a HAM-D_7 subscale instead of the full HAM-D_{17} as clinical endpoint will improve the probability of a successful trial (chapter 3 and 4). Second, the power to detect treatment effect relies upon accurate assumptions about dropout patterns and can increase by using statistical models which take into account missing at random (MAR) (chapter 8). Third, statistical power can be considerably altered by elements of clinical trial designs, such as the primary statistical analysis and skewed enrolment (chapter 9). Indeed, from these results it is apparent that it may be more efficient to include a larger group of patients with fewer observations per patient than *vice versa*.

Another dimension of the problem is current practice and beliefs about the optimal requirements for the overall study design (e.g., run in phase, unbalanced randomisation ratios) and the confidence on the value of established decision criteria used for power calculations. Meta-analysis has provided evidence about some of these factors, such as the placebo run-in phase (Trivedi and Rush, 1994). Clinical trial simulations allow exploration of more possibilities, whilst controlling for each other factor. We have also verified the prospect of an interim-analysis as a study design factor. Indeed, for an indication where many clinical trials fail to demonstrate a statistically significant treatment effect, it is important to reach conclusive decisions before all patients have been accrued.

Information-based adaptive study designs offer advantages in those cases in which the magnitude of treatment effect size and variance cannot be warranted prospectively (Tsiatis, 2006). In the current investigation we have shown the value of an interim analysis as another opportunity to improve the efficiency of clinical trial design in the evaluation of antidepressant drugs. Of crucial relevance in the adaptation process is the optimisation of the information content available at the time of decision making, using enrolment rate information as well as simulated data.

The aim of the current investigation was to demonstrate that it is the ability to assess informative value of the data acquired until the moment an interim analysis takes place that determines the confidence in the outcome of the analysis, rather than just uncertainty about the appropriate dose selection. The proposed Bayesian framework relies on the availability of a method that incorporates the uncertainty of the posterior predictive power and accounts for enrolment rate. In addition, we show how simulated data can be used to optimise the adaptation elements in the trial, namely the timing of the interim analysis and the decision criteria for futility and efficacy. A novel approach is introduced for the evaluation and validation of the latter. We show that once the start of the interim analysis and decision criteria have been defined, the enrolment of patients in historical studies can be re-randomised to establish whether the proposed criteria effectively lead

to an appropriate decision, avoiding false positive and false negative decisions.

Another benefit of adaptive designs may be exploited by allocation and/or sampling rules. We have chosen not to investigate this in the current paper. In a preliminary investigation, we found that the number of patients has a relatively small impact. This is caused by the large variability in the estimates of the relevant model parameters (inter-individual variability, residual variability, treatment effect and placebo effect), especially when little data is available. This variability is such that it largely determines the outcome of the trial and minor variations in sample size do not play a key role. However, we do not preclude the value of sample size adjustments for trials with larger population size. Although other methods advocate sample size adjustment based on a point estimate of the observed variability in the data (Friede and Kieser, 2001), our approach uses the whole posterior distribution as a measure of uncertainty and variability. This may initially lead to less obvious dilemmas, but avoids a false sense of certainty at the moment of an interim analysis, which may later turn out to be incorrect. The introduction of allocation rules becomes interesting when a trial is designed to compare multiple arms to placebo, and a seamless transition to the next phase of drug development is contemplated. Before stopping a treatment arm for futility, a decision could be made to reduce patient allocation to that particular arm.

Implementation aspects

Implementation of this approach requires continuous monitoring or reassessment using the proposed simulation procedures on a weekly basis after a pre-specified (e.g., 20) number of patients per treatment arm has completed the assigned treatment. The simulation-optimisation procedure does not need to take more than a few hours. The results of each evaluation step can be submitted to a data monitoring committee (DMC), which can subsequently decide whether the information acquired at that stage is sufficient to formally perform the interim analysis. This decision depends on the ΔHAMD that is expected for the drug under evaluation, the threshold for clinical relevance and the purpose of the trial. Furthermore, we have accounted for type I and II error levels defined for statistical tests by excluding decision criteria leading to a false negative rate $> 20\%$ or a false positive rate $> 5\%$.

Another important aspect of the implementation of information based design is the need to adapt the number of recruiting centres. This requirement is critical to warrant the principle of exchangeability underlying the use of adaptive designs. A recent paper suggests a method to predict the consequences of additional recruiting centres on enrolment rate (Anisimov and Fedorov, 2007). Furthermore, these methods could be used to predict the enrolment for the coming weeks which could subsequently allow prediction of whether it is worthwhile to postpone the start of an interim analysis.

Assumptions and Limitations

The most important assumption in this work is that the model-parameters for the variability (both inter-individual and residual) used in the simulations are realistic, that is, reflect the variance of the actual data. Future work may investigate if it is appropriate to re-estimate these model parameters using the data acquired in the trial of interest. Nevertheless, the current approach would not be less valuable even if the variability in the actual dataset turned out to be much higher the previously observed. The consequences of such a discrepancy would be a flatter distribution of the PPP, which reduces the probability of an erroneous decision.

Another assumption is that the simulated dropout mechanism and dropout rates are similar to those observed in the study in progress. If the dropout rate is significantly higher, less information becomes available over time, which may increase the probability of false positive or false negative decisions. Our previous work (chapter 9) indicates, however, that the dropout mechanism in most depression trials is not expected to be extreme. The approach also assumes that the DREM is the best model to simulate the time course of HAMD scores in individual patients and that it resembles the actual study data. Previous work (chapter 8) has shown that data simulated by the DREM closely reflects clinical trial data, and indeed was better suitable to simulate new data than other models, such as the mixed model for repeated measures (MMRM) (Mallinckrodt *et al.*, 2004). The DREM was developed based on a functional data analysis into the variability between patients (chapter 7).

Advantages

Enrolment rate has been identified as a critical factor in determining the information content of interim datasets. The proposed approach takes the enrolment rate into account. The use of simulated data ensures therefore that an interim analysis is performed only when sufficient information is available. In addition, the interim analysis is based on the distribution of the PPP, which captures the uncertainty of the information of both the data and the estimates of model parameters. The difference between this approach and typical predictive power estimation procedures is illustrated in figure 7, which shows an interim analysis for study 1 without estimating the uncertainty in the PPP. A comparison between figures 7 and 3 reveals that a higher maximum utility is obtained when the distribution of the PPP is taken into account.

The impact of savings in time before a decision is made (table 2) is substantial. The consequences on the number of patients enrolled depend heavily on enrolment rate. If this is slow, as in study 2, enrolment may be stopped when only 44% of patients are enrolled, a significant gain compared to current practice. For study 1, the savings are less impressive. This underlines the importance of balancing the enrolment rate between the wish to complete the study quickly and the possibility to reduce the number of patients exposed to an experimental treatment when a negative or positive interim decision is

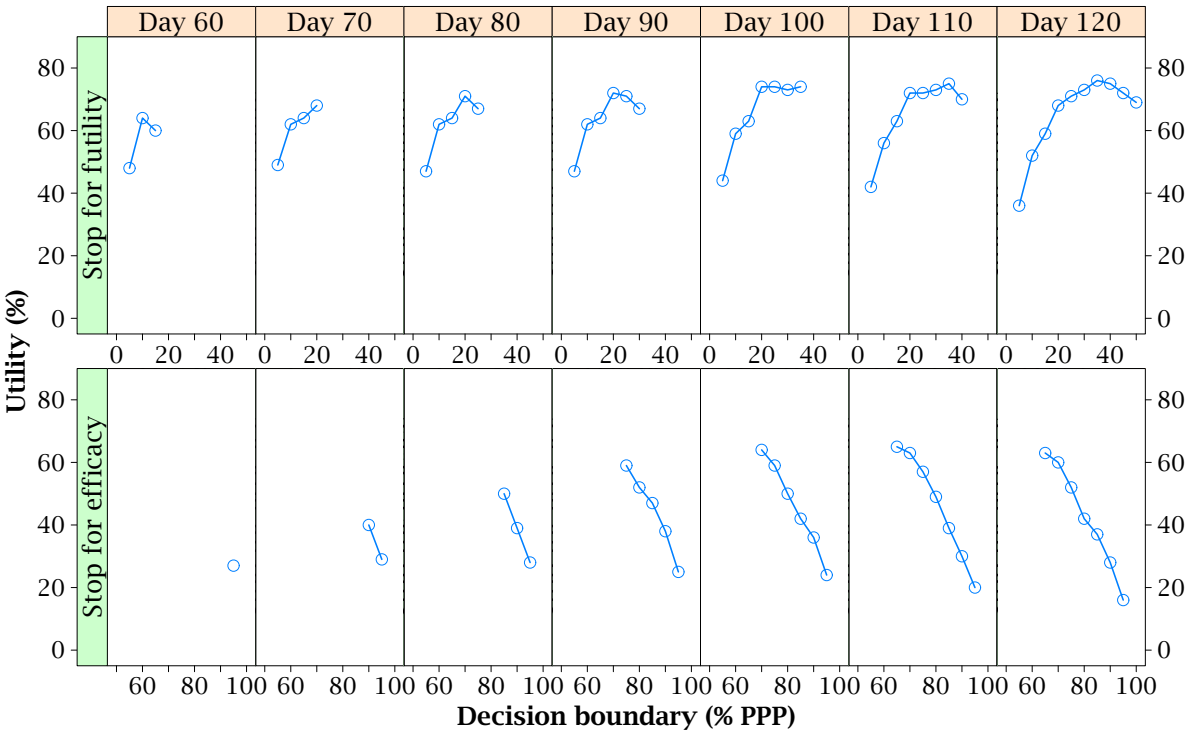


Figure 7. The influence of the stopping boundary on the utility of the interim analysis (100 simulations) for an interim analysis for study 1 without estimating the distribution of the PPP. Different panels are shown for the start of the interim analysis in days after start of enrolment and for the different reasons (futility/efficacy). Decision criteria leading to false positive rates > 5% or false negative rates > 20% are not depicted

reached. However, even if most patients have already been enrolled, the implementation of an interim analysis can potentially save considerable time and resources as well as reduce the burden for patient and investigator if the trial is stopped. Unfortunately, due to the high heterogeneity between patients and studies (as evidenced from the re-enrolment test in figure 6) earlier interim analyses with data from fewer patients are precluded.

The use of an integrated procedure for determining the timing of the interim analysis and its decision parameters has been validated using the re-enrolment test. This technique was developed because the investigators were not convinced by the prospect of using original datasets only to evaluate the validity of the methodology. Since the outcome of the interim analysis is expected to depend on the timing of the analysis, the enrolment rate and treatment effect size (including placebo), a randomisation of the order of enrolment of the patients maintains all these factors constant, whilst providing new datasets to test the methodology on. In fact, we recommend applying the re-enrolment test for any future interim analysis that is validated using historical datasets.

Confidence intervals and *p*-values are accepted methods to summarise treatment benefit. We have provided their Bayesian equivalents (table 1) and have also suggested another approach to summarise the evidence from clinical trials, by assessing the distribution of the PPP of the full dataset as a function of the number of enrolled patients (figure 2). This

is an appealing concept, because it relates directly to the questions many drug developers have: Based on the evidence available now, what is the probability of success for a new study? Although the incorporation of all uncertainty arising from the data allows for better judgement, it does complicate the interpretation. Most researchers prefer to have single numbers as summaries of an outcome, rather than a distribution of values. A Bayesian framework imposes further considerations on decision making, which refine current views about how to handle 'known unknowns'.

Many other methods have been proposed for interim analyses, all of which are based on a group sequential design and do not take enrolment rate into account (Friede and Kieser, 2001; Lachin, 2005). In practice, enrolment will continue after the first group has been enrolled. The impact of the interim analyses may thus be limited. Unfortunately, the consequences of these interim analyses are only reported as the adapted sample size, but since it is not known how enrolment progressed the actual benefits cannot be determined.

To our knowledge our approach is the only methodology that incorporates data from non-completers *and* enrolment rate. Because of the latter, the start of the interim analysis cannot be specified in the trial protocol, since the enrolment rate is not known beforehand. The method we propose to determine the appropriate timing and decision parameters for the interim analysis circumvents this issue. Future work will focus on the addition of sample-size re-estimation to the proposed design, the use of variability parameter estimates from acquired data in the assessment of the decision criteria and corresponding timing of the analysis. We also intend to incorporate other relevant decision factors into the utility function, such as sampling costs.

In conclusion, we show how to implement an interim analysis based on previous work (chapter 9). The proposed methodology has been validated on two historical studies and has been shown to be sensibly conservative towards stopping treatment arms. The extent of heterogeneity between patients as evident from the re-enrolment test precludes earlier interim analysis. We recommend implementation of this adaptive design approach on ethical and commercial grounds, since reductions in both time and enrolled patients are expected to be substantial and outweigh the associated controlled risks.

REFERENCES

Anisimov VV and Fedorov VV (2007) Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Stat Med* **26**:4958–4975.

DeVeaugh-Geiss J, Ascher J, and Brook S (2000) Safety and tolerability of lamotrigine in controlled monotherapy trials in mood disorders, in *39th ACNP Annual meeting*, San Juan, Puerto Rico.

Dragalin V (2006) Adaptive designs: Terminology and classification. *Drug Inf J* **40**:425–435.

Friede T and Kieser M (2001) A comparison of methods for adaptive sample size adjustment. *Stat Med* **20**:3861–3873.

Hamilton M (1960) A rating scale for depression. *J Neurol Neurosurg Psychiatry* **23**:56–62.

Hamilton M (1967) Development of a rating scale for primary depressive illness. *Brit J Soc Clin Psychol* **6**:278–296.

Hrobjartsson A and Gotzsche P (2004) Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment. *J Intern Med* **256**:91–100.

Hrobjartsson A and Gotzsche PC (2001) Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med* **344**:1594–1602.

Khan A, Khan S, and Brown WA (2002a) Are placebo controls necessary to test new antidepressants and anxiolytics? *Int J Neuropsychopharmacol* **5**:193–197.

Khan A, Leventhal RM, Khan SR, and Brown WA (2002b) Severity of depression and response to antidepressants and placebo: An analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* **22**:40–45.

Lachin JM (2005) A review of methods for futility stopping based on conditional power. *Stat Med* **24**:2747–2764.

Lunn DJ, Thomas A, Best N, and Spiegelhalter D (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat Comput* **10**:325–337.

Mallinckrodt C, Kaiser C, Watkin J, Molenberghs G, and Carroll R (2004) The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clin Trials* **1**:477–489.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Sheiner LB (1997) Learning versus confirming in clinical drug development. *Clin Pharmacol Ther* **61**:275–291.

Spiegelhalter D, Abrams K, and Myles J (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, New York.

Spiegelhalter DJ, Freedman LS, and Blackburn PR (1986) Monitoring clinical-trials - conditional or predictive power? *Control Clin Trials* **7**:8–17.

Trivedi MH, Pigott TA, Perera P, Dillingham KE, Carfagno ML, and Pitts CD (2004) Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder. *J Clin Psychiatry* **65**:1356–1364.

Trivedi MH and Rush J (1994) Does a placebo run-in or a placebo-treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology* **11**:33–43.

Tsiatis AA (2006) Information-based monitoring of clinical trials. *Stat Med* **25**:3236–3244.

Walsh BT, Seidman SN, Sysko R, and Gould M (2002) Placebo response in studies of major depression - Variable, substantial, and growing. *JAMA* **287**:1840–1847.

Yang HY, Cusin C, and Fava M (2005) Is there a placebo problem in antidepressant trials? *Curr Top Med Chem* **5**:1077–1086.