To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

11

To fail or not to fail - clinical trials in depression:

Summary, discussion and perspectives

CONTENTS

1 GENERAL INTRODUCTION

There are numerous factors contributing to the high failure rate in the development of novel antidepressant drugs. A detailed analysis of the problem reveals that these factors may be categorised into three main classes: disease, drug and trial design-related factors. The main objective of the research described in this thesis was to explore novel methodologies to overcome design-related factors and reduce the attrition rate in the evaluation of antidepressants in clinical drug development. The importance of trial design-related factors has been highlighted by Khan *et al.* (2002b), who have shown in a meta-analysis of the FDA database that even trials with registered and marketed antidepressants fail to demonstrate a statistically significant treatment effect in almost 50% of the cases. Irrespective of the advancements in the understanding of disease and the possibility to identify biomarkers for depression, our ability to demonstrate treatment effect and differentiate novel antidepressants in terms of their efficacy and safety profile requires methodological research aimed at minimising this unacceptable false negative rate.

Disease-related factors also have a significant contribution to the failure rate in depression trials. The criteria for enrolment of patients into clinical trials underestimate the heterogeneity of the patient population. Since patients are diagnosed by their symptoms, it is conceivable that differences in aetiology or variation in the underlying pathophysiological status lead to differential response to medication. Another explanatory factor of the high failure rate may be the subjective nature of the clinical symptomatology. The current clinical scales do not allow a precise and objective estimation of the severity of depression, which complicates the comparison of drug treatments between patients or even within patients over time. The limitations in the sensitivity and specificity of clinical endpoints compound and amplify the effect of trial design-factors, making the assessment of drug efficacy a major challenge.

In this thesis historical data has been used to investigate which factors can be modified in future clinical trial designs. The historical data was obtained from GlaxoSmithKline's clinical trial database and consisted of several placebo-controlled randomised trials in major depressive disorder. Mostly, these trials were performed with paroxetine as the model drug. An active comparator was included whenever required. The studies range from 6-12 weeks in duration, from 50 to 350 patients per treatment arm, and used the Hamilton depression rating scale (HAMD) (Hamilton, 1960, 1967) as clinical endpoint, in some cases accompanied by the Montgomery Asberg depression rating scale (MADRS) (Montgomery and Asberg, 1979).

In the second part of this thesis we have focused on the development of novel clinical endpoints used in depression, based on the HAMD and MADRS. The approach allowed identification of new subscales which were shown to outperform the HAMD, MADRS and to some extent other published subscales, such as the Bech and Rafaelsen (1980) and Maier and Philipp (1985) subscales. Moreover, evaluation of drug response for compounds with different mechanisms of action revealed that the HAMD is not sensitive to

the differences in the pharmacological properties of the compounds. This finding may have considerable implications for the differentiation of compounds in clinical trials.

The third part of this thesis focuses on the use of novel statistical methods for the analysis of clinical trial data. Particular attention is given to Bayesian statistical concepts, with the introduction of a Bayesian cure rate model and linear mixed effects models. It is shown that there is an important difference in the sensitivity of the various statistical approaches in terms of their ability to detect drug effect. Subsequently, functional data analysis is used to explore the heterogeneity in the time course of response of individual patients and a new model is developed which describes data from clinical trials in depression more adequately.

The value of clinical trial simulation (CTS) and adaptive designs in depression research is introduced in the fourth part of this thesis. CTS is performed to investigate the importance of numerous clinical trial design characteristics and to evaluate the impact of interim analysis in defining efficacy or futility of treatment arms before study completion. The method proposed for the implementation of the interim analysis includes the use of historical data, a longitudinal model of response and enrolment rate.

2 INVESTIGATION AND COMPARISON OF THE HAMD AND MADRS

The investigations described in section two of this thesis focuses on the clinical endpoints used in depression. The literature on the HAMD provides considerable evidence about the multidimensionality of the scale, in that it measures more than one aspect of the disease (Bech and Rafaelsen, 1980; Moller, 2001) and hence may not be a sensitive measure to detect a drug effect (Bech and Rafaelsen, 1980; Bech, 2006; Faries *et al.*, 2000; Khan *et al.*, 2002a, 2004). In fact, it has been suggested that the HAMD is becoming 'a lead weight' rather than a 'gold standard' (Bagby *et al.*, 2004). In chapter 3 we reveal the varying degree of sensitivity of the individual items of the HAMD to treatment response. This was achieved by comparing the time course of the score distribution of the individual items between responders and non-responders. In this analysis differences in sensitivity are based on a predefined threshold for response, irrespective of treatment allocation (i.e., active or placebo). The hypothesis underlying this approach was that the time course of response for sensitive items would be markedly different between responders and non-responders. Sensitivity was assessed by a graphical approach, in which the response rate is characterised, and by the Fisher exact test, based upon which the extent of response on the last observation is ranked. High heterogeneity was found with respect to the sensitivity to response of individual items. Therefore, two subscales were constructed including the seven response-sensitive items (HAM-D_7). The first subscale is a 'response-based' subscale, derived from the graphical approach, whilst the second subscale, which reflects the 'extent of response', was based on *post hoc* statistical criteria. A bootstrapping approach using the , mixed model for repeated measures (MMRM) (Mallinckrodt *et al.*, 2004) compared the two proposed subscales to those currently available in the literature

(i.e., the Bech and Rafaelsen (1980) and Maier and Philipp (1985) subscales) and to the full HAM-D_{17} with regard to their ability to detect drug effect. The response-based subscale outperformed the other subscales and each of the subscales outperformed the total HAM-D_{17}. The gains were such that the number of patients enrolled in a treatment arm could be reduced by 30%, whilst maintaining the same level of statistical power. The difference between the 6-item Bech subscale and the proposed HAM-D_7 is the inclusion of the *suicide* item. Given the apparent link between SSRIs and suicidal behaviour (Healy, 2006; Lenzer, 2006), it is noteworthy that inclusion of the *suicide* item in a subscale increases the difference between paroxetine and placebo. One may infer from these findings that SSRIs do reduce the scores associated with the severity of suicidal thoughts (*suicide* item) in the population responding to treatment, since it increases the SSRI effect size when it is included in the endpoint. The potential link with suicidal behaviour might therefore be a consequence of lack of response or limited to a specific sub-population. To our knowledge, little attention has been paid to whether patients are truly responding to therapy when suicide is attempted.

In chapter 4, the sensitivity of the items of the MADRS to treatment response was investigated using the same graphical method as applied in chapter 3. A comparison of the items of the HAMD and MADRS which aim to capture the same disease domain revealed that no differences exist between these scales with respect to the time course of the score distributions of the items in responders and non-responders, respectively. Furthermore, all MADRS items were found to be sensitive to response and therefore no subscale needs to be devised for this clinical endpoint. Using the same bootstrapping approach as in chapter 3 and the MMRM, the MADRS was compared to the HAMD and its subscales. The MADRS outperformed the HAMD, but the subscales of the HAMD were even more sensitive in detecting treatment effect. This may be due to the domain of physical symptoms (*somatic symptoms general*) contained in each of the subscales, which is not present in the MADRS. These differences in sensitivity indicate that by changing the clinical endpoint, the number of patients could be reduced by up to 30% without losing statistical power to detect a treatment effect. The savings due to a reduction in population size are particularly important for those clinical trials in which large population sizes are required, such as non-inferiority trials and head-to-head comparisons.

3 THE MISSING LINK BETWEEN THE HAMD AND RECEPTOR TARGETS

Having established that the use of subscales may reduce the number of false negative results in clinical trials in depression, another endpoint-related issue was investigated in chapter 5, that is, the specificity of the HAMD to detect different mechanisms of action. Although global disease markers such as the HAMD and MADRS are important, new antidepressants have markedly different pharmacological properties which may lead to selective, more specific treatment effect. Therefore, one potential approach to address compound differentiation is to measure the different aspects of depression separately.

Such an approach may provide clues with regard to the relationship between pharmacological mechanisms and clinical effect. In fact, this has been proposed recently and defined as componential analysis (Katz, 1998; Katz *et al.*, 2004).

On the other hand, efforts in early drug discovery are focused on devising more specific ligands with better efficacy-safety ratios. Very few clinical scientists have inquired about the relation between receptor pharmacology and the sensitivity and the specificity of clinical measure. Hence, the successful differentiation of compounds depends on whether the clinical endpoint currently used will be able to distinguish the differences between these ligands.

Therefore, we decided to investigate whether the HAMD can differentiate the effects of drugs with distinct mechanisms of action. To this end an aggregated database containing 5035 patients was created including treatment arms with placebo and drugs with three different mechanisms of action: tricyclic antidepressants (TCAs), serotonin-specific reuptake inhibitors (SSRIs) and the anticonvulsant lamotrigine. Two separate methods were developed to explore whether the differences in mechanisms of action were reflected by the HAMD and its individual items. Based on dichotomisation of the patient population into responders and non-responders, we first evaluated the time course of the score distribution of the individual items of the HAMD, separately for each mechanism of action. This graphical evaluation, similar to the approach used in chapters 3 and 4, revealed no important differences between the mechanisms of action, with the exception of lamotrigine, for which the frequency of low scores on response-sensitive items was lower than observed for the other compounds. This observation may be explained by the lack of efficacy of lamotrigine in this patient population, as demonstrated by the negative results obtained in all clinical trials included in the database.

Subsequently, we calculated the contribution of each item to the total change in HAMD at the last observation for each patient. Using descriptive summary statistics, box-plots were constructed to compare the contributions of all items for responders and non-responders, by mechanism of action. An analysis of variance (ANOVA) was performed to test whether differences between the mechanisms of action reached statistical significance. The variability in the contribution of each item relative to total change observed for each class of antidepressant was clearly greater than the variability observed between different mechanisms of action. Furthermore, the contribution of individual items to the total change in HAMD was found to be more variable in responders than in non-responders. This substantiates the fact that the time course of response in responders is more predictable than that of non-responders, supporting the use of model-based approaches to characterise treatment response in depression.

Until now, research was not conclusive as to the question whether the HAMD behaved differentially to specific mechanisms of action (Khan *et al.*, 2004; Moller, 2001; Nelson *et al.*, 2005). Our results confirm that HAMD does not reflect differences in pharmacological properties. Hence, no direct link can be established between clinical endpoint and mechanism of action, which raises an important question about the probability of dif-

ferentiating compounds in development. These findings cannot, however, exclude that specificity may exist for novel mechanisms of action which have not been tested thus far. Given the distal relationship between receptor systems currently targeted by antidepressants and the multidimensional nature of the clinical endpoint, it is not surprising that no mechanism-specific effects can be identified for the HAMD. In contrast, endpoints such as the componential approach (Katz, 1998; Katz *et al.*, 2004), which focuses on specific domains of depressive symptoms may provide more insight into the relationship between pharmacological properties and symptomatology.

It is clear from our investigation that new endpoints for clinical trials in depression must be developed which do not only describe symptoms, but accurately characterise disease severity on different levels. In rheumatology, the disease activity score (DAS) is such an endpoint. The DAS integrates biomarkers (sedimentation rate), symptoms (number of swollen joints, number of joints tender to touch) and clinical scales (patient assessment of disease activity using a visual analog scale) to form a single measure of disease severity (van der Heijde *et al.*, 1993). Similarly, scales of disease severity could be developed in depression by integrating imaging parameters, such as receptor occupancy and baseline fMRI activity, with biomarkers, such as the dexamethasone response test or other phenotypical measures, and clinical scales specific to the different dimensions of the disease. The availability of such a scale is essential to the development and differentiation of novel compounds.

In summary, the main findings in the first section of this thesis show that subscales of the HAMD provide a more sensitive measure of drug effect than the full HAMD and even the MADRS. A reduction of the number of patients to be enrolled in a trial that is required to separate treatment effect from placebo and a reduction of false negative results are not the only consequences of the use of the suggested HAM-D_7 as clinical endpoint in future clinical trials. Increased precision in the estimates of treatment effect may also reveal dose-response relationships which may have been obscured by the noise in the endpoints used so far. Often, no differences are observed between the different dose levels of antidepressants using the full HAMD. The differences in the sensitivity of the subscales may be sufficient to elucidate dose-response relationships in dose ranging studies, as required in Phase IIb trials. As will be discussed later, the use of more sensitive endpoints in conjunction with measurements of drug exposure in efficacy trials may enable characterisation of the exposure-response relationship for antidepressants.

4 STATISTICAL ANALYSIS: THE CURE RATE MODEL

The third part of this thesis focuses on another important aspect of the clinical trial design: the statistical model used for analysis of the primary endpoint. Chapter 6 reports the application of a Bayesian parametric cure rate model (CRM) (Chen *et al.*, 1999) on data from clinical trials in depression. This model is based on a survival approach, which captures the time it takes for a patient to respond to medication. Additionally, the proportion of non-responders (cure rate) is estimated, and treatments are assumed to exert

their effect by reducing this proportion. Given the poor sensitivity of the HAMD scale, dichotomisation of treatment response may offer an opportunity to make clinically meaningful inferences from a continuous endpoint with limited sensitivity. Fitting of data from two trials to the CRM showed that drug treatment (paroxetine and fluoxetine) reduces the proportion of non-responders. It is shown that the sensitivity of the model to detect drug effect is comparable to the Cox proportional hazard model (Cox, 1972), the mixed model for repeated measures (MMRM) (Mallinckrodt *et al.*, 2004) and *t*-tests with last observation carried forward imputation. The advantage of this model over longitudinal models such as the MMRM is that it does not rely on the HAMD as a continuous marker for the severity of depression. As discussed in the first section of this thesis, substantial noise in the HAMD is caused by the presence of response-insensitive items, which may lead to inaccurate estimation of true treatment effect.

A preliminary assessment of the treatment effect based on partial data, before completion of the trial, showed that accurate estimates can be obtained well before the complete dataset was available. In fact, in most cases treatment effect could be established when 50% of the trial was completed (calculated as the difference between the first and last observation) and 70% of patients had been enrolled. Various methods were tested to determine whether the estimate of treatment effect was sufficiently accurate to perform an interim analysis, but no approach could be found which produced consistent results as determined using the re-enrolment test (chapter 10). The application of this model as interim analysis tended to result in uncontrollable false positive and false negative results, leading to efficacious treatment arms being terminated for futility, and inefficacious arms being terminated for efficacy (*data not shown*). Another important contributing factor which stopped development of the CRM as interim analysis is that clinicians are used to express treatment effect as a difference in change in HAMD, rather than the asymptotic non-response rate, as is the case in the cure rate model. Nevertheless, the method can be applied for data analysis at completion of treatment. The proportion of non-responders offers a different measure of efficacy, particularly when comparing treatment effect on different sub-populations. The use of continuous endpoints and change relative to baseline captures differences in effect size, but may not necessarily allow for differentiation of compounds in terms of their effect on different subpopulations, as expressed by the proportion of non-responders.

5 STATISTICAL ANALYSIS: LINEAR MODELS

Various approaches were conceived in an attempt to develop a statistical model which could overcome the limitations described above, ensuring accurate estimation of treatment effect at an interim stage of the trial. Initially, we have considered the highly variable trajectory of response as a transition between disease states (e.g., improvement, relapse), in which the progression of response over time might behave according to Markov or semi-Markov processes. In a Markov or hidden-Markov chain, the probability of transition

between states depends only on the current state, implying that future behaviour is memoryless and hence independent of any previous trajectory. In contrast, in a semi-Markov chain the probability of transition between states depends on the permanence or time spent in a given state. Surprisingly, none of these properties seemed to describe the time course of response in individual patients, nor could drug effect be parameterised accurately (*data not shown*). These findings could be partly explained by uncertainty on the timing and state in which a patient enters the trial as well as by the duration of the study, as compared to the overall duration of the depressive episode.

Challenged by the complexity of the HAMD trajectory, we have undertaken a different approach, that of functional data analysis (chapter 7). Rather than focusing on the average behaviour of patients in a clinical trial, functional data analysis aims to describe the variability between patients and was used to explore longitudinal patterns or time course of response in individual patients. This approach revealed different components of the variability in time course of response as assessed by the HAMD, of which the two main components were an additive variability (patients exhibiting a higher or lower score than the average score) and a slope, indicating that some variability resulted from patients performing worse than the average patient at the start of the trial, but switching over to better performance at the end, and *vice versa*. Other findings were that the main components of variability were the same for responders and non-responders, which provides an indication that these two groups are not two entirely different groups, but are rather the result from a continuous spectrum of patients showing a differing degree of response to medication. An explanation for these findings may be found in the variability of drug exposure, which is not commonly measured in depression trials. Part of non-response may be due to insufficient exposure. Future attempts to elucidate an exposure-response relationship may provide more insight in these matters. The approach used in the first part of the thesis, where responders and non-responders were compared to each other, did however provide valuable information regarding the degree of sensitivity of the individual items to change of disease severity. In that sense, it is more appropriate to introduce the concept of *improvers* and *non-improvers*, to underline that variability in the time course of HAMD is not directly linked to lack of clinical response nor is it a trait belonging to a specific patient. Most likely, differences in drug exposure contribute to the distinct behaviour of the two population phenotypes.

Another important finding in chapter 7 was that the main components of variability were also present in the response-based subscale (HAM-D_7). These results indicate that the main characteristics of the time course of response for an individual patient are not determined by the remaining items in the HAM-D_{17} and hence only add noise to the measurement. This strengthens the value of the HAMD subscales as primary endpoint in the evaluation of efficacy of antidepressant drugs, instead of the full HAMD or MADRS.

Based on the two main principal components that were identified in chapter 7, a dual random effects model (DREM) was proposed in chapter 8 to model longitudinal data in depression. In contrast to current methodologies, which focus primarily on the estimation of

treatment effect size, we reveal that the DREM also provides a more accurate description of individual patient data, especially for low and high HAMD values. In addition, measures of goodness-of-fit showed an increased performance compared to the MMRM and hierarchical (single) random effects model. Calculation of normalised prediction discrepancy errors (NPDE) (Brendel *et al.*, 2006) shows that the DREM is also more appropriate to simulate new patient data than the MMRM. The ability to simulate realistic response profiles is essential for predicting treatment effect in individual patients as well as for the implementation of adaptive clinical trial designs for the development of antidepressant compounds. A simulation exercise was performed to evaluate the influence of seven dropout scenarios on the assessment of treatment effect, as proposed by Lane (2008). Data were analysed using the DREM, MMRM and last observation carried forward (LOCF) techniques. The simulations of individual patient data were carried out using parameter estimates obtained previously from fitting of two clinical trials. This procedure allowed estimation of the false negative rate (type II error), false positive rate (type I error) and bias in the results. The DREM and MMRM were found to produce similar results, controlling the type I & II errors and bias better than the LOCF. Under extreme dropout scenarios and unequal dropout ratios however, we observed either a significant increase in the type I error, or a severe reduction in study power for all three models. As indicated above, despite the minor differences found between the MMRM and DREM, the latter does yield better fits for individual patient curves especially for very low and high HAMD scores, rendering this model valuable for simulation purposes.

This finding concludes the second section of this thesis, having established a new model to analyse clinical data in depression trials based on a functional data analysis approach. Furthermore, it confirms the importance of the overall time course of response in the estimation of treatment effect, as opposed to methods which solely rely on the mean differences between treatment arms at last measurement, corrected for baseline.

6 CLINICAL TRIAL SIMULATIONS

In section four of this thesis we show how the impact of the different factors presented in the previous sections can be evaluated by clinical trial simulations. In this section we also illustrate how prior knowledge can be formally incorporated into statistical inferences and how Bayesian statistics can take explicit account of uncertainty, improving the decisions accounting for the observed treatment effect in a clinical trial.

Meta-analyses are often used to summarise prior knowledge and review estimates of the treatment effect size in mainstream clinical research, particularly when comparing treatment options. Meta-analyses can also be used to investigate the influence of clinical trial design features, such as the placebo run-in phase (Lee *et al.*, 2004; Trivedi and Rush, 1994). However, these techniques allow scrutiny of only those design factors that have been implemented, and even then the effect of confounding factors cannot be excluded. Moreover, they rely on mean parameter estimates, yielding results which ignore the underlying differences between patients. In contrast to meta-analysis, clinical trial simulation

(CTS) allows for the investigation of the impact of a range of design characteristics on the power to detect a treatment effect prior to exposing patients to an experimental drug. In a field where most clinical trials have a conservative design, this methodology offers a unique opportunity to evaluate innovative designs. In the clinical trial simulation reported in chapter 9, we have investigated the effect of the following design features on statistical power and false positive rate: (a) sample size (number of patients), (b) randomisation ratio across treatment arms, (c) frequency of assessments (number of visits), (d) dropout mechanisms, (e) clinical endpoint, (f) statistical method for the assessment of treatment effect and (g) interim analysis in an adaptive design. Based on data from two historical studies, simulations were performed that included combinations of all design factors mentioned above.

The main findings are the following:

1. Reduction of the number of visits from 6-7 to as few as two observations per patient does not cause bias or affect the precision of estimates of treatment effect. The reduction in the number of visits can have a considerable impact on the burden on patients and investigators. The decreased interaction between patients and investigators may also lead to a reduced placebo effect. These results are not affected even under more extreme dropout scenarios.
2. Analysis of the percentage of responders/remitters using a Fisher-exact test instead of applying continuous models such as the DREM and MMRM resulted in a severely reduced power. However, this endpoint is often reported since the percentage of responders/remitters is considered more relevant in clinical practice. Bech *et al.* (1984) have suggested to report the change in disease state rather than the change in HAMD, because of its clinical relevance. Although this may be true, a dichotomisation of the HAMD does not necessarily discriminate between disease states. Indeed, Kirsch and Moncrieff (2007) advise not to use dichotomised endpoints based on continuous scales for this reason. The reduction in statistical power observed in the CTS is a further reason not to report clinical trials in this manner.
3. Data analysis based on LOCF was found to cause only minor bias under the mild, but realistic dropout scenarios used in the simulations. Under more extreme dropout scenarios and unequal dropout rates between treatments LOCF may, however, result in highly biased estimates, as shown in chapter 8.
4. Skewed enrolment over treatment arms may cause an increase in bias and lower statistical power. Many studies have employed skewed randomisation in order to reduce the number of patients exposed to placebo. Even though this may seem appropriate from an ethical point of view, our results show a severe reduction in power, leading to a failed trial. Therefore, we strongly recommend the use of simulation scenarios to evaluate the consequences of skewed randomisation prior to implementation of a study, if an unequal randomisation ratio is considered. Undoubtedly, the ethical arguments underlying these choices need to be revisited.

Additionally, we have assessed the relevance of an interim analysis in demonstrating the efficacy or futility of a treatment arm prior to completion of treatment. An approach was proposed which relies on the posterior predictive power (PPP) estimated from the posterior distributions of the parameters of the DREM. Besides allowing for the incorporation of the uncertainty associated with the treatment effect, the method also takes into account the uncertainty associated with all model parameters in the PPP. A simple stopping rule resulted in 60-80% of inefficacious treatment arms being dropped without jeopardizing the treatment arms that did show significant separation from placebo.

Other clinical trial design adaptations have also been suggested in recent publications. Fava *et al.* (2003) have designed an attractive and innovative new study design which aims at resolving the high placebo response rate. Unfortunately, to our knowledge, this design has not yet been implemented, possibly because of a reluctance of the field to make significant alterations in clinical trial design. We anticipate that the results obtained from comprehensive simulation scenarios using historical data are sufficiently robust to persuade the psychiatric community about the fact that trial design adaptations are not only feasible and safe, but also necessary and worthwhile.

7 THE USE OF INTERIM ANALYSIS IN ADAPTIVE DESIGNS

Further elaboration of the requirements for implementing an interim analysis in the context of an adaptive trial design is provided in chapter 10. Whereas the timing of an interim analysis and the corresponding decision criteria are chosen *a priori* in most clinical trial protocols, in chapter 10 we show how Bayesian methodology can support the optimisation of the timing and criteria to be used for the purposes of futility and efficacy in an interim analysis. Additionally, we propose a method which takes into account the uncertainty in the posterior predictive power (PPP). Although the PPP itself factors the variability in parameter estimates, it remains a point estimate. Incorporation of uncertainty is achieved by bootstrapping interim datasets, the summary of which reflects the uncertainty in the estimation of the PPP.

The impact of the timing at which an interim analysis is started and the utility associated with selected decision criteria are determined using simulation procedures. These simulations take several factors into account, such as the observed enrolment rate, number of projected patients and sampling scheme of the trial. The risks and benefits of the interim analysis can be accurately quantified by evaluating the behaviour of the utility function at different starting times with a range of decision criteria. This procedure can guide accurate decisions about the timing of the interim analysis as well as the optimal decision criteria. In this evaluation, only information about the enrolment rate is derived from the actual trial. All other factors are explored using simulated data, which in turn is derived from model fits of the DREM to historical data.

A re-enrolment test is introduced to substantiate the principle of patient exchangeability and validate the proposed methodology. The concept of the re-enrolment test assumes

that the order in which patients are enrolled is random. Therefore, a re-ordering of these patients results in slightly different datasets at the proposed timing of the interim analyses, allowing for an evaluation of the selected decision criteria and hence an appraisal of power, type I and type II error. For instance, if an interim analysis is performed when half of the trial has been completed, datasets can be created containing different subsets of patients at the proposed time. In addition, the re-enrolment test accounts for the extent of the heterogeneity between patients in the trials. The more heterogeneous the population, the larger the number of re-enrolled patients required before a consistent signal is observed.

It was found that the impact of an interim analysis depends largely on the enrolment rate of patients in the trial. Slow enrolment allows for a treatment arm to be stopped for futility with less than 50% of the population to be enrolled in the study, which constitutes a substantial saving in terms of patients, time and resources. Clearly, the impact of an interim analysis may be reduced with faster enrolment rates, as proportionally more patients will have been enrolled at the moment in which an interim analysis becomes feasible. Also in these circumstances, the interim analysis was shown to be able to make a stopping decision whilst controlling type I and II errors.

8 IMPLEMENTATION OF INTERIM ANALYSIS

Implementation of adaptive designs deserves careful consideration. The decision about the right moment to start the interim analysis requires active involvement of an independent data monitoring committee (DMC). One possibility is to define a preliminary meeting of the DMC in the protocol when, e.g., 20 patients per treatment arm have completed the full observation schedule. Since the enrolment rate is not a blinded factor in the execution of clinical trials, simulations taking into account the actual enrolment rate can be carried out to explore the statistical power as well as the false negative and false positive rates for an analysis associated with the amount of data available thus far, conditional on a range of decision criteria. Using models that predict the enrolment rate (Anisimov and Fedorov, 2007), estimates of the statistical power, false negative and false positive rates due to future enrolment can also be determined.

Independent of the purpose and phase of the trial and of the expected or desirable clinical benefit, this approach offers the appropriate statistical tools for a data monitoring committee (DMC) to decide when an interim analysis should be initiated. In addition, these procedures also contribute to a closer control of the integrity and validity of the trial, as the method enables immediate quantification of the impact of variations in enrolment rate. Undoubtedly, implementation of the steps described above will reduce the unacceptable false negative rates currently observed in clinical trials.

We have also demonstrated that the implementation of an interim analysis in the Bayesian context automatically enables extension of the framework to accommodate decision modelling. In fact, a possible extension of the method present in this thesis would

include penalties or cost functions to the modelling procedure (Patel and Ankolekar, 2007; Fenwick *et al.*, 2008; Willan and Pinto, 2005). Since accurate estimates of false negative and false positive rates can be derived from the simulated data, the consequences of additional information *versus* the uncertainty associated with each subset of patients can be coupled to a penalty. In the case of decisions associated with false positive results, the cost is that future trials will turn out to be futile. In contrast, decisions associated with false negative results may lead to (moderately) efficacious drugs being terminated, whereas these were potentially effective antidepressants. The cost function in this event should not only encompass the loss of return of investment, but also the loss of a novel therapeutic option for the patient population. Such a loss must be balanced against the gains from making an early stopping decision, i.e., immediate savings in terms of patients, time, and resources. The burden for enrolled patients, the costs of enrolling new patients, the cost per sample and the costs of delaying a decision are factors that should also be taken into account. An additional level of complexity could be introduced by incorporating evidence of adverse events. Adverse events are generally divided into three levels of severity, each of which could be given a weighing factor in the cost function, multiplied by the expected probability for the incidence of these events.

9 PRACTICAL RECOMMENDATIONS FOR CLINICAL TRIAL DESIGN

The main recommendations proposed for the design of future studies are the following:

1. Use the HAM-D_7 as primary endpoint
2. Limit the study duration to a maximum of 6 to 8 weeks
3. Apply equal enrolment ratio across treatment arms. Alternatively, perform clinical trial simulations to investigate the consequences of unequal enrolment
4. Reduce the frequency of visits for the assessment of the primary endpoint to only 2 to 3 times per patient (excluding baseline)
5. Use the DREM as statistical model for the analysis of the primary endpoint
6. Implement adaptive designs in which the timing and decision criteria for the interim analysis take into account enrolment rate and incorporate uncertainty in the estimation of treatment effect size.

Given the approximate cost per patient (£ 10,000) in a clinical trial, the proposed recommendations translate in financial savings of approximately £ 400,000 per treatment arm (30% reduction in the number of patients in a standard trial, with 150 patients/arm), excluding potential savings due to a reduction of the number of visits. Most importantly, these recommendations warrant a significant decrease in the rate of false negatives, which have even larger financial and ethical repercussions. The reduction in false negatives represents savings equivalent to the cost of a whole trial or even of a whole development

programme if the decision associated with the results implies termination of the development of the asset. From an ethical perspective, our recommendations raise the scientific standards of research in depression, allowing translation of the oxymoron that is the benefit of placebo, in which patients allocated to the placebo arm in a randomised controlled trial not only contribute to ruling out the use of ineffective compounds, but also may ultimately benefit from new treatments once their efficacy has been accurately demonstrated.

10 GENERALISATION TO OTHER DISEASE AREAS

In physics, the influence that the observer and experimental setting may have on experimental results has since long been accepted. In medical research such acceptance has yet to come. The meticulous evaluation of design-related factors presented in this thesis illustrates the importance of such factors in explaining failure and attrition rate in the development of antidepressants. In fact, our findings underline the extent of the interaction between drug and experiment and anticipate the potential consequences for the future of drug discovery and development if these factors remain immutable.

The methods used in the analysis of the sensitivity and specificity of the HAMD and MADRS may be applied on clinical scales in different diseases, such as schizophrenia and other psychiatric disorders. Similar to what was demonstrated in depression, insight in the dimensionality and sensitivity of the items to response for these clinical scales may lead to the development of more sensitive measures of treatment effect. The bootstrap technique proposed to determine the differences between endpoints is also generalisable. This technique allows exploration of the consequences of the relation between the number of patients included in a trial and the power to find a treatment effect.

The other parts of this thesis, especially the linear models that have been applied throughout, may be applied to any disease with an endpoint which is (approximately) normally distributed. In the absence of well-established concentration-effect relationships, the use of linear mixed effects models is the most straightforward statistical option. Findings that were made in the last chapter can be extrapolated to other endpoints, but the specific statistical characteristics of the endpoint, such as the ratio between the residual variability (within-subject variability) and between-subject variability (the random effect), must be considered. Also, the interim analysis, based on Bayesian posterior predictive power, can be applied to any normally distributed endpoint and may be extended to non-normally distributed endpoints. Moreover, the concepts used for establishing the timing of the interim analysis as well as the use of the re-enrolment test can be generalised to any disease area and statistical model.

It is evident from our analysis that model-based data analysis offers a major advantage to drug development. We have shown that in addition to the typical issues related to model parameterisation, parameter estimation and model validation, Bayesian methods allow further consideration of how uncertainty affects decision making. As such, this decision theoretic approach can be expanded to other therapeutic areas. It is anticipated

that the concept of Bayesian optimal designs will also evolve. Effective drug development in the 21^{st} century will require maximisation of the difference between the costs of any given implementation step, including the execution of a trial, and the value of the information gained from the results.

11 OTHER CONTRIBUTORS TO THE HIGH FAILURE RATE OF CLINICAL TRIALS

Because of the limitations set by the use of historical data, some important issues contributing to the high failure rate of clinical trials in depression could not be investigated. One of the main difficulties in the clinical development of antidepressants is the heterogeneity of patients. Although inclusion criteria in typical clinical trials are aimed at limiting this heterogeneity by including patients without other psychiatric diseases and by excluding patients with mild depression (typically HAMD has to be ≥ 18), high variability exists between patients as shown by the re-enrolment test in chapter 10. Over the years, the growing popularity of antidepressants and acceptability of the diagnosis have created a patient population that is more knowledgeable about mood disorders. This in itself renders comparisons of historical trials to current clinical trials difficult, as is evidenced by the time dependent change in placebo response rate (Walsh *et al.*, 2002). The main cause for the heterogeneity lies in the fact that depression is diagnosed based on a set of subjective symptoms. The underlying mechanistic pathways remain obscure and are therefore not taken into account during diagnosis. Discrimination of psychotic depressed patients *versus* non-psychotic depressed patients has been shown to be feasible on the basis of a dexamethasone suppression test (Nelson and Davis, 1997). However, a recent review concludes that more research is needed to increase the specificity of biomarkers in order to contribute to the diagnostic process (Mossner *et al.*, 2007).

Imaging techniques such as PET and functional MRI may also increase the ability to distinguish between different types of patients. As discussed previously in section 4.4 of chapter 1, the research in this area is mostly exploratory and correlations to disease severity have not been attempted in a systematic manner. Improved understanding about the differences between patients and underlying mechanisms of disease requires the development of an index of disease severity which integrates imaging parameters, such as receptor occupancy and baseline fMRI activity, with biomarkers, such as the dexamethasone response test or other phenotypical measures, and clinical scales specific to the different dimensions of the disease. A higher homogeneity of patients in clinical trials may result in lower failure rates and better understanding of the specific treatment requirements for different patients groups. Ideally, such an index should enable individualisation of treatment and dosing regimen.

The area of pharmacokinetic-pharmacodynamic modelling can improve our understanding of the impact of drug and disease-related factors on treatment response in depression. In conjunction with concepts from control theory and pharmacostatistical

methods, one could explore exposure-response relationships, rather than relying solely on dose-response curves to describe the interaction between the underlying disease process and drug effect (Danhof *et al.*, 2007). In fact, in drug development any effective strategy must endeavour the characterisation of the exposure-response relationship. Dose is a poor descriptor of response for biological systems which are associated to time-dependent activation and transduction mechanisms. In that sense, PKPD modelling can contribute to differentiating and explaining why a given pharmacological intervention may not translate into clinical benefit. Moreover, it can function as a framework for the scaling of response across species and validation of novel endpoints.

It is unfortunate that in drug development past experience seems to limit future behaviour. Industry continues to focus on compliance to milestones rather than addressing key development questions. A strategy is missing that integrates knowledge from receptor pharmacology to recent development in neurosciences and conceptualises the interaction between target and biological system, stimulus and response. In addition to current attempts to achieve proof-of-concept milestones, it is essential to develop an approach that builds upon thorough understanding of drug-receptor interaction *in vivo* (proof-of-mechanism, drug-related parameters) and of the transduction process upon target blockade or target activation (proof-of-principle or proof-of-pharmacology, system-related parameters). The rationale for dose range and dosing regimen can hardly be justified without considering receptor occupancy profile and concentration-effect relationships for the pharmacological effects or biomarkers.

It is essential to realise that it is pointless to optimise trial design-related factors without ensuring that relevant exposure levels are maintained during the evaluation of treatment response. No matter how sensitive the statistical model, accurate data analysis depends upon the aforementioned assumption. In contrast to hypothesis testing, PKPD models enable parameterisation of drug effect taking into account pharmacological mechanisms and may provide more relevant estimates of efficacy and safety. However, the assessment of the exposure-response relationship in efficacy trials still faces an unfounded resistance by investigators and clinical scientists, who do not seem to be aware of the value of sparse sampling techniques. As highlighted previously, the assessment of efficacy will remain vulnerable to failure without information on drug exposure and without evidence of patient compliance to the prescribed regimen.

We conclude this thesis with the opening statement. To fail or not to fail in clinical trials in depression does not depend solely on the identification of novel targets and availability of better candidate compounds. Clinical trials in depression have hardly changed in the past 50 years. In that sense the acceptance of the randomised clinical trials methodology which evolved over the last 30 years may be a curse: it does not stimulate new ideas and methods, but instead seems to imply that current practice is the best way to demonstrate drug effect. A reduction in attrition rate and failure will require a shift in the whole clinical development paradigm of depression research. Future efforts should focus on elucidating the mechanisms underlying drug response and establishing the exact

relationship between drug dose, exposure and effect. The assumption that drug exposure plays a minor role in determining treatment response and that the available clinical scales are mandatory for the purpose of drug development must be abandoned in favour of a more pharmacologically oriented approach.

REFERENCES

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

Bagby RM, Ryder AG, Schuller DR, and Marshall MB (2004) The Hamilton depression rating scale: Has the gold standard become a lead weight? *Am J Psychiatry* **161**:2163–2177.

Bech P (2006) Rating scales in depression: limitations and pitfalls. *Dialogues Clin Neurosci* **8**:207–215.

Bech P, Allerup P, Reisby N, and Gram LF (1984) Assessment of symptom change from improvement curves on the Hamilton depression scale in trials with antidepressants. *Psychopharmacology (Berl)* **84**:276–281.

Bech P and Rafaelsen OJ (1980) The use of rating-scales exemplified by a comparison of the Hamilton and the Bech-Rafaelsen melancholia scale. *Acta Psychiatr Scand* **62**:128–132.

Brendel K, Comets E, Laffont C, Laveille C, and Mentré F (2006) Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharm Res* **23**:2036–2049.

Chen MH, Ibrahim JG, and Sinha D (1999) A new Bayesian model for survival data with a surviving fraction. *J Am Stat Assoc* **94**:909–919.

Cox DR (1972) Regression models and life-tables. *J R Stat Soc Ser B-Methodol* **34**:187–220.

Danhof M, de Jongh J, De Lange ECM, Della Pasqua O, Ploeger BA, and Voskuyl RA (2007) Mechanism-based pharmacokinetic-pharmacodynamic modeling: Biophase distribution, receptor theory, and dynamical systems analysis. *Annu Rev Pharmacol Toxicol* **47**:357–400.

Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, and Potter WZ (2000) The responsiveness of the Hamilton depression rating scale. *J Psychiatr Res* **34**:3–10.

Fava M, Evins AE, Dorer DJ, and Schoenfeld DA (2003) The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychother Psychosom* **72**:115–127.

Fenwick E, Claxton K, and Sculpher M (2008) The value of implementation and the value of information: combined and uneven development. *Med Decis Making* **28**:21–32.

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.

Healy D (2006) Drug regulation - Did regulators fail over selective serotonin reuptake inhibitors? *Br Med J* **333**:92–95.

van der Heijde DMFM, van 't Hof M, van Riel PLCM, and van de Putte LBA (1993) Development of a disease-activity score based on judgment in clinical practice by rheumatologists. *J Rheumatol* **20**:579–581.

Katz MM (1998) Need for a new paradigm for the clinical trials of antidepressants. *Neuropsychopharmacology* **19**:517–522.

Katz MM, Houston JP, Brannan S, Bowden CL, Berman N, Swann AC, and Frazer A (2004) A multivantaged behavioural method for measuring onset and sequence of the clinical actions of antidepressants. *Int J Neuropsychopharmacol* **7**:471–479.

Khan A, Brodhead AE, and Kolts RL (2004) Relative sensitivity of the Montgomery-Asberg depression rating scale, the Hamilton depression rating scale and the clinical global impressions rating scale in antidepressant clinical trials: a replication analysis. *Int Clin Psychopharmacol* **19**:157–160.

Khan A, Khan S, Shankles E, and Polissar N (2002a) Relative sensitivity of the Montgomery-Asberg depression rating scale, the Hamilton depression rating scale and the clinical global impressions rating scale in antidepressant clinical trials. *Int Clin Psychopharmacol* **17**:281–285.

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.

Kirsch I and Moncrieff J (2007) Clinical trials and the response rate illusion. *Contemp Clin Trials* **28**:348–351.

Lane P (2008) Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharm Stat* **7**:93–106.

Lee S, Walker JR, Jakul L, and Sexton K (2004) Does elimination of placebo responders in a placebo run-in increase the treatment effect in randomized clinical trials? A meta-analytic evaluation. *Depress Anxiety* **19**:10–19.

Lenzer J (2006) Manufacturer admits increase in suicidal behaviour in patients taking paroxetine. *BMJ* **332**:1175.

Maier W and Philipp M (1985) Comparative-analysis of observer depression scales. *Acta Psychiatr Scand* **72**:239–245.

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.

Moller H (2001) Methodological aspects in the assessment of severity of depression by the Hamilton depression scale. *Eur Arch Psychiatry Clin Neurosci* **251 Suppl 2**:II13–II20.

Montgomery SA and Asberg M (1979) A new depression scale designed to be sensitive to change. *Br J Psychiatry* **134**:382–389.

Mossner R, Mikova O, Koutsilieri E, Saoud M, Ehlis AC, Mueller N, Fallgatter AJ, and Riederer P (2007) Consensus paper of the WFSBP task force on biological markers: Biological markers in depression. *World J Biol Psychiatry* **8**:141–174.

Nelson JC and Davis JM (1997) DST studies in psychotic depression: A meta-analysis. *Am J Psychiatry* **154**:1497–1503.

Nelson JC, Portera L, and Leon AC (2005) Are there differences in the symptoms that respond to a selective serotonin or norepinephrine reuptake inhibitor? *Biol Psychiatry* **57**:1535–1542.

Patel NR and Ankolekar S (2007) A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. *Stat Med* **26**:4976–4988.

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.

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.

Willan AR and Pinto EM (2005) The value of information and optimal clinical trial design. *Stat Med* **24**:1791–1806.