To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

2

Scope and outline of the investigations

SECTION I: GENERAL INTRODUCTION

Factors contributing to the failure of a clinical trial can be categorised into disease, drug and trial-related factors. Using historical data we aim to reduce the influence of trial-related factors, such as the insensitivity of clinical endpoints. In this thesis we show the relevance of existing knowledge in the evaluation of antidepressant drug effect. Be it to gain new insights into disease features and drug characteristics through retrospective review of existing data, or to enable the implementation of a full learning-confirming paradigm (Sheiner, 1997). Information derived from the vast pool of clinical trials can be formally integrated into the analysis of new trials and strengthen the conclusions about drug effect in a complex matrix of interacting factors, as is the case concerning treatment response in depression. The proposed approach is fundamentally different from mainstream clinical research, in which meta-analysis is used as a tool to retrieve information or to learn from existing data.

As mentioned in **chapter 1**, the failure rate of clinical trials in depression for registered antidepressants is 50%. Patients who have been enrolled into these studies, as well as the investigators that invested their time, have therefore not contributed to a further advancement of clinical pharmacology and medical practice. This is unacceptable, in particular for those patients assigned to placebo treatment. Also, drug development programs are delayed because of these failed studies, which cause efficacious compounds to be available later to those patients that need it. In the most extreme case, an efficacious drug may never reach the market. These issues have of course been taken up by drug development teams in industry, leading to reluctance in stopping the development programme of candidate compounds. Consequently, many extra studies are performed to make inefficacy absolutely certain, increasing the burden on patients and on R&D.

Clearly, a reduction of false negative and false positive results in clinical trials is critical to the advancement of research in depression. This thesis aims at contributing to this cause by investigating those factors that can be optimised or modified in prospective clinical trials. We assume that the existing evidence of attrition is sufficient to acknowledge the need to depart from current practices.

Firstly, the sensitivity of established disease endpoints are investigated to demonstrate how much the current measurement tools contribute to the failure rate in clinical trials. Then, the appropriateness and shortcomings of different statistical methods for data analysis are evaluated to explore their impact on bias, false positive (type I error) and false negative (type II error) rates. Finally, a pharmacometric approach is used to assess how these factors interact with other study design features and to define the requirements for optimisation of clinical trials in depression. These findings should form the basis for best practice in prospective clinical research with antidepressant drugs.

SECTION II: CLINICAL ENDPOINTS

The second section of this thesis focuses on the most important endpoints in depression studies: the Hamilton depression rating scale (HAMD) and the Montgomery Asberg depression rating scale (MADRS).

Especially in small studies, the sensitivity of clinical endpoints is of major importance. **Chapter 3** investigates the sensitivity of each of the 17 items of the HAMD to response. Using a graphical approach, seven items are identified which show a distinctly different time course between patients who respond to treatment, and those who do not. The sensitive items are grouped into a response-based subscale (HAM-D_7) and compared to the full HAMD with respect to their sensitivity to detect drug effect. This comparison is performed using a bootstrap approach, re-sampling the original dataset with a varying number of patients and observing the percentage of positive trials upon the use of both endpoints. It is found that the response based subscale (HAM-D_7) outperforms not only the HAMD, but also the Bech and Maier/Philip subscales.

In **chapter 4**, we show the properties of the MADRS in a similar manner as the HAMD in chapter 3. All items, with the exception of *reduced appetite* are found to be sensitive to clinical response. A comparison between the HAMD, its subscales and the MADRS reveals that the HAMD is the least treatment-sensitive endpoint, followed by the MADRS and the subscales of the HAMD. The relative sensitivity of the different subscales seems to depend upon the clinical data that is used. Nevertheless, the response-based subscale (HAM-D_7) proposed in chapter 3 is consistently shown to be more sensitive than the HAMD and MADRS.

Chapter 5 concludes the first section by exploring whether compounds with varying pharmacological properties can be differentiated by the HAMD. Differentiation of compounds has become an essential milestone in clinical development as novel targets are identified for depression. Using a database which consists of 11 clinical trials, the difference in time course of each item of the HAMD between responders and non-responders is shown to be insensitive to a variety of mechanisms of action. Subsequently, we examine the contribution of each individual item to the overall change in HAMD at completion of treatment. The magnitude of the between-subject variability of these contributions was much lower in responders than in non-responders. Furthermore, the variance associated with the individual items' contribution to the overall change in HAMD is larger within the classes of antidepressants than between them. These results may have major implications for the pursuit of novel targets with improved pharmacological properties, since neither the clinical scale nor any of the individual items seem to capture differences in pharmacological properties.

SECTION III: DATA ANALYSIS

The third section of this thesis discusses statistical methods that are available for the analysis of the clinical endpoints discussed in section II. We introduce the Bayesian cure rate model as an approach to distinguish responders from non-responders and parameterise drug effect by differences in response rate. Then, focus is shifted to the multiplicity of response patterns observed in the evaluation of the continuous scale. A functional data analysis is proposed to discern the dimensions of variability between patients over the course of treatment. Subsequently, we demonstrate how understanding of variability in longitudinal patterns can be incorporated in data analysis using a linear mixed model. In this evaluation a comparison is made with the mixed model for repeated measures under different mechanisms of dropout.

Chapter 6 reports the results of the first application of a fully parametric cure-rate model in a Bayesian framework. Given the magnitude of the variability and the lack of sensitivity of the HAMD items, a dichotomisation of the data is proposed that evaluates treatment effect by defining response in terms of its clinical relevance, i.e., a change of at least 50% from baseline. The cure-rate model allows estimation of the asymptotic fraction of non-responders to treatment, an endpoint related to the percentage of responders, which is frequently used in reports of clinical trials. The advantages of Bayesian statistics are fully exploited by using historical information of placebo data, reporting direct probabilities of superiority of treatment over placebo and by reporting confidence intervals that have meaningful interpretations in clinical practice.

Based on the diversity of response patterns observed in the evaluation of longitudinal data, an exploratory investigation applying functional data analysis is discussed in **chapter 7**. This chapter focuses on the trajectory or behaviour of response in individual patients as well as on how patients differ from each other. This approach shifts the current paradigm in the analysis of clinical response, in that it does not depart from the assumption of mean response as a 'natural' feature of populations, which is often defined in terms of arithmetic means or average response. It is shown that clear principal components exist in longitudinal response patterns, which are constant across the clinical studies analysed, despite the large variation in the overall time course. The first component corresponds largely to an additive random effect, as is often used in random effects models. Interestingly though, the second component points to a random slope effect, i.e., patients having a HAMD score above the average HAMD score at the start of the trial and lower at the end, and *vice versa*. Furthermore, the identification of these components demonstrates that response variability is a random phenomenon with a well-defined structure. The relevance of the HAM-D_7 subscale is further enhanced by confirming that it retains the same components identified in the time course of response for the full HAM-D_{17}.

In chapter 8, the understanding of the variance structure associated with the response patterns in longitudinal data is shown to provide the elements to characterise clinical

data better and describe the time course of response of individual patients. Simulated data is obtained from a dual random effects model (DREM), in which the random effects are defined according to the principal components identified in chapter 7. In the current approach, a scenario analysis is used to assess the consequences of seven different mechanisms of dropout, as dropout is one of the factors known to affect the results of a statistical analysis. The impact of varying degrees of sensitivity in the data analysis methods is illustrated by performing the analysis of the same datasets with the MMRM, the dual random effects model and a last observation carried forward approach. The linear mixed model performed with the least bias and yielded the lowest type I error. In contrast, the LOCF approach resulted in severe bias under likely scenarios of dropout, leading to inflated false positive (type I error) or false negative rates (type II error).

SECTION IV: CLINICAL TRIAL SIMULATION

Clinical trial simulation can be used as a tool to evaluate and optimise current clinical research practice. Its effectiveness in exploring multidimensional, complex problems has been demonstrated in many therapeutic areas. Taking into account the role of the factors discussed in previous sections and their potential impact on the outcome of clinical trials, simulation methods are applied to determine which characteristics of clinical studies ought to be modified to reduce bias, false positive and false negative rates (type I and type II errors, respectively). Novel implementation strategies and designs are explored that set aside current beliefs about the requirements for best practice in research. Furthermore, emphasis is given to the necessity to apply Bayesian adaptive methods including an interim analysis step, in which the likelihood of success is assessed prior to completion of a trial.

In **chapter 9** the impact and consequences of variations in different trial design elements are explored using clinical trial simulation based on the dual random effects model developed in chapter 8. Taking current clinical practice as a starting point, seven factors have been identified for evaluation: (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 analysis of treatment effect and (g) interim analysis. An important learning from this chapter is that a reduction in the frequency of visits during the trial has negligible influence on the estimation of drug effect. This finding may have further implications to study outcome, since it is conceivable that the placebo effect will diminish due to less frequent contact between patient and investigator. Furthermore, a design with equal enrolment across treatment arms may be more efficient than a design with skewed enrolment. Frequently, an ethical argument emerges to justify the importance of fewer patients enrolled into the placebo group. The simulations show that, in contrast to current views, this may lead to a lower power to detect drug effect, invalidating the scientific rationale for the study. A surprising finding was that under the relatively mild but realistic conditions of dropout, all models,

including those based on LOCF imputation, perform reasonably well. In addition, the use of statistical models which merely focus on the percentages of responders and/or remitters is shown to decrease the power to detect a statistically significant drug effect dramatically. Lastly, the introduction of an interim analysis into study design was shown to be possible, which enhances the efficacy of the trial design by giving the opportunity to stop inefficacious treatment arms early.

In **chapter 10**, we underline the value of using historical data in depression and the relevance of a framework for incorporating existing and arising information into the decision process in the context of Bayesian adaptive designs. In contrast to study design adaptations based on randomisation, population size, sampling rules or treatment allocation, an interim analysis method is proposed based on an adaptive stopping rule, which prevents inefficacious treatment arms from continuation at an early stage of the trial, without a relevant loss of statistical power.

The method consists of a simulation procedure followed by an optimisation step. Simulations based on the observed enrolment rate, planned visit frequency and a clinically relevant treatment effect are performed, enabling the subsequent assessment of the utility of the interim analysis associated with different timings and decision criteria. Historical data is used as a basis for these simulations. In this respect, the proposed methodology borrows concepts from information-based designs.

To investigate the accuracy and quality of the methodology, a validation step is performed using two historical datasets in which the exchangeability principle underlying the validity of adaptive designs is tested. New datasets are created to mimic different enrolment sequences (re-enrolment test). The interim analysis is run on these datasets and the fraction of stopping decisions for efficacy and futility is reported for each treatment arm. The proposed approach allows early decisions and warrants low type I and II errors. Moreover, these results reveal that accurate inferences about treatment effect in a depression trial require more than information on sample size or dose-response curve.

SECTION V: CONCLUSIONS AND PERSPECTIVES

The last section of this thesis provides an integral summary and discussion of all sections and chapters. Most importantly, recommendations are provided for optimising clinical trial designs in depression. In addition, an attempt is made to establish best practice requirements and to explore future perspectives for this challenging area.

REFERENCES

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