To fail or not to fail : clinical trials in depression

Sante, G.W.E.

SECTION IV

CLINICAL TRIAL SIMULATION

Chapter

9

From trial and error to trial simulation:

An appraisal of current beliefs in clinical research practice in depression

Gijs Santen, Joe Horrigan, Meindert Danhof, Oscar Della Pasqua

ABSTRACT

Clinical trial simulation (CTS) is a model-based approach which allows the investigation of the influence of design characteristics on important aspects of clinical trials such as power and type I error.

The objective of this investigation is to evaluate in an integrated manner the impact of (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 on the outcome of clinical trials with antidepressant drugs. Based on current research practice, we have explored how varying scenarios affect the results and the conclusions about the relevance and statistical significance of treatment effect.

A dual random effects model (DREM) was used to simulate clinical trial data with various combinations of baseline conditions and study design characteristics. For comparative purposes, the simulated data was subsequently analysed using the DREM, the mixed model for repeated measures (MMRM), last observation carried forward (LOCF) and the Fisher exact test based on the percentage of responders/remitters. In addition to the analysis of a complete trial, the influence of interim evaluations was explored using posterior predictive distributions under the assumption of an enrolment rate of one patient per treatment arm per day.

The clinical trial simulations yielded evidence for the following facts: (1) Increasing the frequency of visits, and hence the number of assessments per patient does not increase the power to detect a treatment effect. (2) A skewed randomisation often used to reduce the number of patients on a placebo or comparator arm may lead to reduced statistical power. (3) Analysis of the percentage of responders leads to greatly reduced power compared to the linear mixed models. (4) An interim analysis is proposed to stop inefficacious treatment arms early, whilst preventing premature termination of effective treatments.

CTS is an important tool in optimising the design of clinical trials with antidepressants. Thus far, no other statistical approach has provided such comprehensive, integrated evaluation of the various factors contributing to failure in clinical research in depression. Our findings also strongly support the use of interim analyses as best practice in early clinical development of novel antidepressants.

INTRODUCTION

Drug discovery and drug development in major depressive disorder and other subtypes of depression are notoriously difficult. The nature of the disease and the lack of understanding of the key mechanistic pathways contribute to the problems encountered in drug *discovery*. Yet new hypotheses in the past years have led to drugs with innovative mechanisms of action which may be effective in the treatment of depression (Kramer *et al.*, 2004; Nielsen, 2006). However, even when an antidepressant with known efficacy is compared

to placebo in a clinical trial, historical evidence suggests that the likelihood of a successful trial is only 50% (Khan *et al.*, 2002). Clearly, drug *development* in depression is equally difficult. This is partly due to the nature of the disease itself, particularly through factors such as high heterogeneity between patients, the inherent large variability in the placebo effect, the high placebo response rate and the difficulties in measuring the severity of the disease in an objective and meaningful manner. Rather than inspiring innovative clinical trial designs, these difficulties seem to have led to a very conservative approach.

In fact, clinical beliefs have dominated trial practices, which are not corroborated by increasing scientific evidence about the compounding effect of different factors in the assessment of efficacy. Although some papers have suggested interesting innovative trial designs (Fava *et al.*, 2003), these have not yet been implemented. Indeed, the majority of clinical trials are conducted according to a fixed protocol, copied from one study to the next. Among the reasons for this conservatism, one should consider the high costs involved in these trials, which limits the willingness to experiment with trial designs and the fear of blame for the potential failure of a study, which could be assigned to the differences from the traditional study design.

In other fields, where similar concerns regarding the uncertainty about the relevance of confounding and design factors as well as costs play a role, the use of clinical trial simulation (CTS) has been advocated. This technique determines all possible outcomes under candidate trial designs, allowing such trial designs to be compared in a strictly quantitative and objective manner. In depression, numerous simulation studies have been performed to compare statistical models with respect to their type I and II errors (Lane, 2008; Mallinckrodt *et al.*, 2001a,b, 2004a,b), but none of these studies included an integrated evaluation of all relevant design factors.

It is important to stress that CTS allows investigation of factors that cannot be scrutinised by meta-analysis. First, designs which have not been implemented cannot be included in a meta-analysis. Second, it is difficult to separate the influence of multiple design factors, whereas CTS allows evaluation of a single factor at a time. Although meta-analyses can provide important information about differences in patient populations and treatment response, it is unfortunate that some investigators consider it sufficient to gather evidence on the impact of design factors from overall reviews, as often suggested in the discussion and interpretation of the findings of a meta-analysis.

In general, CTS utilises two types of models (Girard, 2005; de Ridder, 2005). The first type of model, the drug-action model, describes the effect that a drug has on an individual, taking into account the pharmacokinetics (PK) of the drug and the interaction between the drug and its target (pharmacodynamics, PD). Traditional PKPD models assume that the biological system does not change. However, in chronic diseases (which often involve long term studies) changes in the biological system due to disease progression cannot be discerned and must be taken into account. Such a time-related change or so-called time-dependency in response can be evaluated by dynamical or disease systems analysis (Danhof *et al.*, 2007; Post *et al.*, 2005).

The second type of model involved in CTS is the trial execution model. These models simulate other important aspects of the trial, such as dropout, compliance and protocol deviations. Probabilistic models are often used in these applications, ensuring that implementation factors are accounted for in the simulation of clinical studies.

Traditional power calculations only take a point estimate of the variability of the clinical endpoint into account, whilst neglecting the influence of disease progression and trial design factors and trial execution factors, such as dropout. Advanced CTS models have been applied to trials in oncology (Veyrat-Follet *et al.*, 2000), neuropathy (Lockwood *et al.*, 2003), Alzheimer's disease (Lockwood *et al.*, 2006), angina pectoris (Chabaud *et al.*, 2002), schizophrenia (Kimko *et al.*, 2000) and juvenile rheumatoid arthritis (Yim *et al.*, 2005). Unfortunately, the lack of knowledge about the mechanisms underlying treatment response as well as the difficulties in measuring depression severity have rendered the development of mechanistic PKPD models for this indication very difficult. Some attempts in this direction have been reported in the literature, but these models have either identifiability issues (Gruwez *et al.*, 2007) or their parameters lack direct clinical interpretation (Gomeni and Merlo-Pich, 2007), which complicates the development of meaningful simulation scenarios.

On the other hand, a variety of flexible statistical models is available for the analysis of longitudinal data from clinical trials, among which the mixed model for repeated measures (MMRM) is probably the best known (Mallinckrodt *et al.*, 2004a). In a previous investigation, we have shown that a dual random effects model (DREM) can describe and simulate depression data more adequately than the MMRM especially for the high and low HAMD scores (Hamilton, 1960, chapter 8). These models do not account for the concentration-effect relationship of drugs or for pharmacokinetic differences between subjects, which limits their use for predicting optimal doses and dosing regimens. Nevertheless, they offer an important advantage in that any conclusions made from simulations using these models will hold for any trial or experiment with a longitudinal endpoint which is normally distributed.

The objective of the current work is to demonstrate the value of CTS in the evaluation of clinical trial designs for antidepressant drugs. We also show that the results of CTS allow appropriate inferences for the implementation of clinical trials, irrespective of the assumptions about the dose-response relationship of the investigational drugs. Using historical clinical trial data, we evaluate in an integrated manner the impact of (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 as well as the relevance of (g) interim analysis in the evaluation of treatment effect. The results of these simulations are summarised as recommendations for the design of new clinical trials in depression.

METHODS

Study data

The simulations were based on two studies in major depression, which were extracted from GlaxoSmithKline's (GSK) clinical trial database. These studies are representative of trials in depression and correspond to a typical trial outcome, including treatment arms which show a clear separation from placebo (positive control) and treatment arms which do not yield significant separation from placebo (negative control). Our investigation was restricted to two studies due to limitations in computational power.

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. The HAM-D_{17} (Hamilton, 1967) was used as clinical endpoint and measured at weeks 1, 2, 3, 4, 6 and 8 after start of treatment. A total of 459 patients was evenly enrolled across the treatment arms. Further details can be found in the original publication (Trivedi *et al.*, 2004).

Study 2 (unpublished) was a randomised placebo-controlled trial in which paroxetine and fluoxetine were compared according to a dose-escalation design. The HAM-D_{17} (Hamilton, 1967) was measured at weeks 1, 2, 3, 4, 6, 9 and 12 after start of treatment. 140 Patients were enrolled in the placebo arm, and a total of 350 patients was enrolled in both active treatment arms. Further details can be found in the GlaxoSmithKline clinical trial register (http://ctr.gsk.co.uk, protocol number 128)

Whilst the performance of a given trial design *versus* another trial design is assumed to be independent of treatment effect size, the statistical power achieved under a particular design is dependent on treatment effect and cannot be determined beforehand when investigational drugs are being evaluated prospectively. Therefore, a hypothetical drug ('wonder drug' in table 1) showing significantly larger effect size has also been included in the simulation scenarios.

Model

The data from these studies were fit using the dual random effects model (DREM) described in detail elsewhere (chapter 8). Equation 1 describes the model parameterisation, as implemented in WinBUGS 1.4.1 (Lunn *et al.*, 2000). WinBUGS is a Bayesian statistical program that uses Markov chain Monte Carlo (MCMC) methods to determine the posterior distribution.

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

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).

Mean parameter estimates for all model parameters are shown in table 1.

Simulation of patient data

To mimic the typical demographics of patients enrolled into clinical studies, baseline values for the primary endpoint were simulated using a normal distribution with mean 20 and standard deviation 4, truncated between the inclusion criterion (19 or 25) as a lower boundary and 40 as a higher boundary. The resulting distribution reflected the observed baselines for patients in actual studies. No other patient covariates, such as age, gender or disease history were considered for the purposes of this first evaluation.

Starting from the selected baseline values, the time course of individual HAMD scores at each visit was simulated by including subject-specific random effects and residual error, as estimated from the posterior distributions (table 1) obtained from the model fitting described in the previous section. HAMD values were rounded to the nearest integer to mimic the real-life assessment.

Dropout model

The impact of dropout on the analysis of treatment effect is highly dependent on the nature and cause of dropout. Different dropout patterns exist that may or may not introduce bias in the analysis of treatment effect; missingness completely at random (MCAR) is completely random dropout, independent of any measured data. Missingness at random (MAR) is dropout related to observed data, whereas missingness not at random (MNAR) is dropout related to unobserved data. Rather than using extreme hypothetical or unrealis-

Table 1. Parameter values used for the simulations. The 2x2 variance-covariance matrix is given under the I.I.V. and σ/I.I.V. headings. The 'wonder drug' is defined as a drug with a 50% higher treatment effect than the best performing drug in that study

Study	Treatment	Baseline parameters (placebo) or treatment effect per week								I.I.V.	σ/I.I.V.
		1	2	3	4	6	8	9	12		
1	placebo	0.81	0.73	0.66	0.61	0.59	0.53	-	-	-	3.2
	paroxetine (12.5 mg)	0.1	0.7	1.4	1.7	2.4	1.5	-	-	Var/covar	
	paroxetine (25 mg)	0.0	1.4	1.8	2.3	3.9	2.9	-	-	23.1	-1.73
	'wonderdrug'	0.1	2.2	2.8	3.6	5.9	4.6	-	-	-1.73	1.22
2	placebo	0.85	0.75	0.70	0.63	0.61	-	0.59	0.54	-	3.64
	paroxetine (max 50 mg)	0.1	0.1	0.5	0.7	1.4	-	2.2	2.2	Var/covar	
	fluoxetine (max 80 mg)	0.0	0.8	1.1	0.7	1.8	-	2.8	2.0	19.8	-2.0
	'wonderdrug'	0.1	0.2	0.8	1.1	1.4	-	3.3	3.3	-2.0	1.1

tic scenarios to investigate the role of dropout in depression trials, a historical database of 11 clinical trials in major depressive disorder was used to explore dropout patterns. This analysis revealed that the dropout rate was not consistently higher or lower in the placebo-treated group. Also, dropout rates seemed to be constant over time. Figure 1 shows the probability of dropout against the depression severity, as indicated by HAMD quartiles. We found that for the first three weeks dropout seemed to adhere to a missing completely at random (MCAR) pattern. After week 3, however, the probability for dropout was clearly higher in the severely depressed patients, suggesting missingness at random (MAR). There did not seem to be any difference due to treatment type.

We have also considered another approach for the evaluation of dropout, which is based on the common notion among psychiatrists that dropout is mostly likely to occur because of the lack of effect in the placebo group (MAR) and because of side effects in the active treatment arms (MCAR).

Based on the aforementioned considerations, we have explored the following scenarios for evaluating the role of dropout on treatment effect:

(1) Drop out according to psychiatrists' current belief: For placebo dropout MCAR and MAR (with increasing probability of dropout as severity of depression increases) in the ratio 1:3. For the active treatment, MCAR en MAR were used in the ratio 3:1, as dropout was assumed to be dependent on the severity of side effects (MCAR in the absence of drug

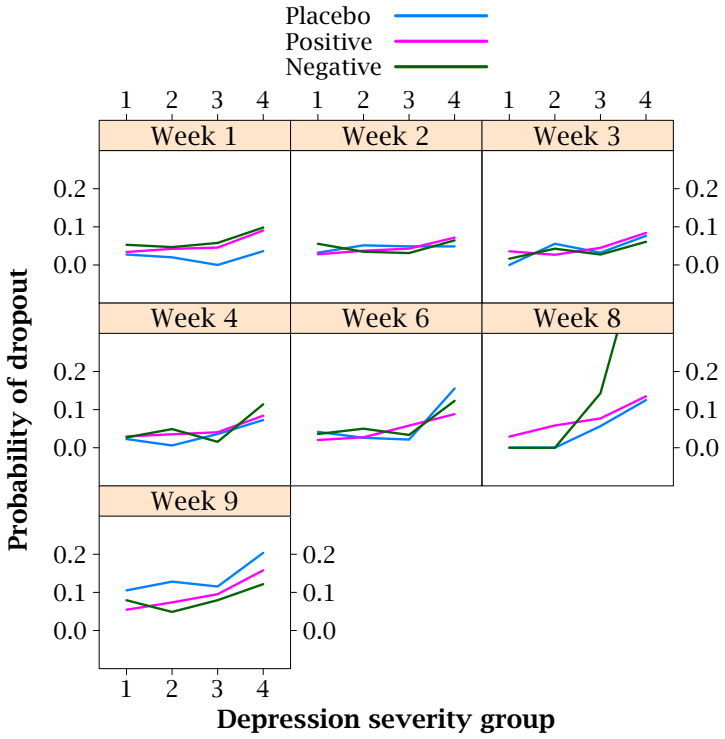


Figure 1. Probability of dropout *versus* depression severity. Panels depict the dropout rate at each weekly visit. Patients were split into 4 equally sized quartiles per treatment type (placebo, positive active and negative active) based on their HAMD score. Patients in group 1 had the lowest HAMD scores, patients in group 4 the highest (i.e., were most depressed)

concentration data).

(2) Drop out according to the findings from the aforementioned graphical evaluation: MCAR dropout in the first 3 weeks for placebo and active treatment arms. From week 4 onwards, dropout in the ratio 3:1 were used assuming MCAR and MAR (only patients in the highest quartile show significant probability of dropout)

(3) Drop out according to the findings from previous graphical evaluation, but replacing the MAR pattern by missingness not at random (MNAR).

For all schedules the dropout rate was set at 4% per week.

Investigated design characteristics

1. Treatment effect size

The treatment effects that were used in the simulations were based on those observed in study 1 and 2. Additionally, a treatment arm with 50% more effect than the best performing treatment arm in each study (the so-called 'wonder drug') and a treatment arm with no effect were included to investigate type I errors.

2. Number of patients

Assessment of the population size was performed using an equal randomisation ratio across treatment arms with 100, 125, 150 and 175 patients per arm. A commonly used 2:1 active:placebo unequal randomisation ratio was also evaluated. This scenario corresponded in total number of enrolled patients to the scenario with 125 patients assuming 3 treatment arms (75:150:150).

3. Duration of the study/frequency of visits

Trials in depression usually have a duration of 6, 8 or 12 weeks. Most recent studies, however, have an 8-week duration. The frequency of visits is mostly weekly at the start of the trial to enable detection of early drug effect, and may take place once in every 2 or 3 weeks at the end of the trial. In the current simulation we will investigate the effect of study duration, which depends on the time course of treatment effect, and the impact of the frequency of visits on trial outcome. The investigated designs are shown in table 2.

4. Endpoints

As extensively discussed in a previous investigation (chapter 3), the poor sensitivity of endpoints may prevent finding evidence for treatment effect in trials with limited population size. Therefore, simulations were performed using the full HAM-D_{17} and the response-based subscale HAM-D_7 (chapter 3). Data simulation for the subscale was performed in the same way as described above for the full HAM-D_{17}. In contrast to the full HAM-D_{17}, baseline values for this subscale were simulated under the assumption of a normal distribution with a mean of 13 and standard deviation 3, truncated between the inclusion criterion (11 or 15) and a maximum of 35. These values were selected as they mirrored the observed baseline distributions in the clinical studies used for model fitting.

Table 2. Visits, total number of visits and trial duration for the tested designs)

Design	Visits (week)	Number of visits	Total duration (weeks)
A	2,6	2	6
B	2,4,6	3	6
C	2,4,8	3	8
D	2,4,9	3	9
E	2,4,12	3	12
F	2,4,6,8	4	6
G	2,4,6,9	4	9
H	2,4,9,12	4	12
I	1,2,4,6,9	5	9
J	1,2,3,4,5,6,8	6	8
K	1,2,3,4,6,9,12	7	12

5. Inclusion criteria

As described previously, the inclusion criteria used for entry into the trial had boundaries defined for the HAM-D_{17} of 19 or 25. For the HAM-D_7 subscale the entry boundaries were 11 or 15. These inclusion criteria had an effect on the baseline-time interaction term that was present in the model. Since a baseline-treatment effect interaction could not be found in the original data, it was not included in the model.

6. (Primary) statistical analysis

The analysis of treatment effect in the simulated studies was based on the following methods:

- The dual random effects model (DREM). Fitting and estimation was performed in WinBUGS 1.4.1 (Lunn *et al.*, 2000)
- The mixed model for repeated measures (MMRM). Fitting and estimation were performed in PC SAS (v9.1 for Windows, SAS Institute, Cary, NC, USA)
- T-tests using LOCF imputation (LOCF)
- Fisher exact-tests on the number of responders/remitters

The latter two methods were implemented in the language and environment for statistical computing R (version 2.5.1 for Windows) (R Development Core Team, 2007). For the Fisher exact test, response was defined as a $\geq$50% reduction in HAMD on the last visit, relative to baseline. For remission a HAMD of $\leq$7 had to be measured at the last visit. For the HAM-D_7, the assessment of response was handled using the same definition, with remission being constrained to HAM-D_7 values of $\leq$3.

Practical implementation of the CTS

Combinations of all possible scenarios were simulated. This amounted to 780 scenarios: 5 (different numbers of patients) * 13 (different visit schemes) * 2 (different baseline inclusion criteria) * 2 (primary endpoints) * 3 (different dropout patterns).

All scenarios were subsequently analysed by four different statistical models. Each scenario was simulated and fitted 100 times. The scenarios were compared to each other with respect to power (% of simulations resulting in a statistically significant effect) and type I error (% of false positive results). Unfortunately, a super computer was not available for the purposes of our investigation. To cope with computing processing time in an effective manner, simulations were distributed over 20 PCs and integrated through the language and environment for statistical computing R (R Development Core Team, 2007).

Reduction in the frequency of scheduled visits

Preliminary investigations showed that a reduction of the frequency of visits was possible without loss in statistical power. Clearly, the consequences of a reduction in the frequency of visits depend on the dropout mechanism. Thus, the dropout mechanisms based on actual data that we have used in the simulations may be partly responsible for this outcome. We decided to investigate the consequences of a reduction of visits from 7 to 3 under more extreme dropout scenarios, as described in detail elsewhere (Lane, 2008, chapter 8). In brief, 3 scenarios for MAR and MNAR mechanisms were conceived. The first scenario assumes a gradual increase in dropout probability for 9 equally sized depression severity groups, the second scenario assumes gradual increase only for the most severe 4 categories and the third scenario assumes dropout only for the most severely depressed group of patients.

Interim analysis

In addition to accounting for confounders, CTS enables the evaluation of the requirements for optimally estimating treatment effect during the course of the trial using interim analyses. The implementation of the DREM in a Bayesian framework allows for the use of the posterior predictive distribution. This distribution takes into account all uncertainty in the parameters and is therefore suitable for an interim analysis. The following methodology was used to implement this interim analysis (figure 2).

First, an enrolment rate of 1 patient/treatment arm/day was assumed. Then, for each scenario, arbitrary time points of 75 and 100 days after start of enrolment were selected for an interim analysis. At each of the timings, all data acquired up to that point was included and analysed using the DREM. The last 1000 iterations of all parameters were imported into R, where 1000 trials were simulated with the expected number of patients. The results of the simulations were subsequently subjected to a *t*-test to investigate how many of these trials would result in a statistically significant outcome ($p<0.05$). The fraction of statistically significant results corresponds to the posterior predictive probability

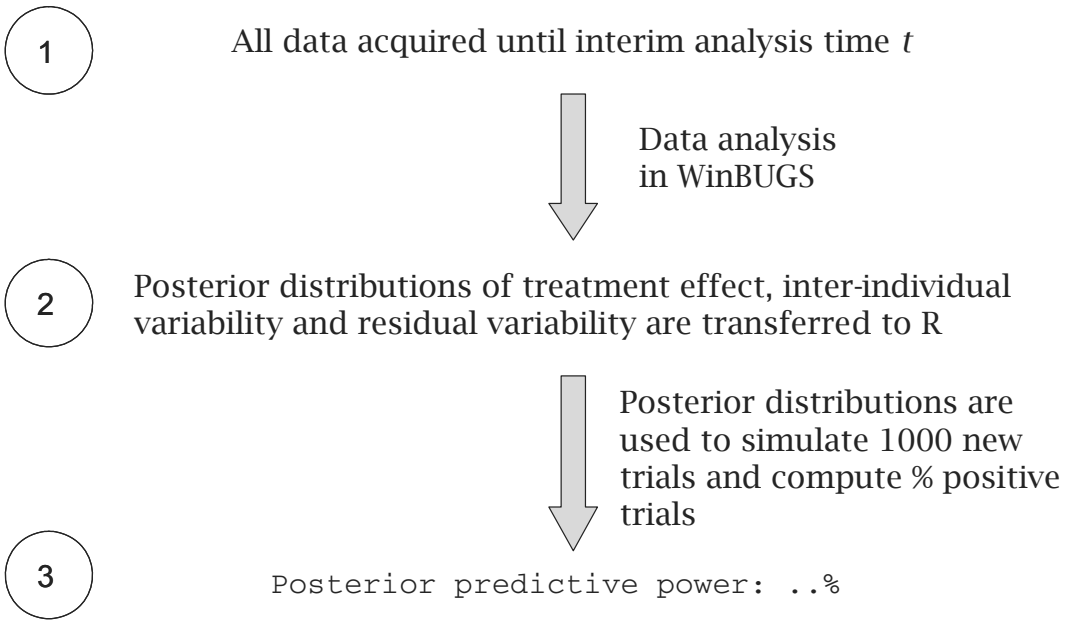


Figure 2. Computation of the posterior predictive power. (1) Data acquired until the interim analysis time *t* is analysed using WinBUGS, (2) the posterior distributions are used to simulate 1,000 new trials based on the posterior distribution, (3) the percentage of successful trials is summarised as the posterior predictive power

of success in a typical Bayesian framework, or the 'posterior predictive power (PPP)'.

In the current analysis a decision criterion of 40% was used for the PPP. Thus, all interim analyses resulting in a PPP<40% resulted in a stopping decision.

The objective of an interim analysis is to stop treatment arms which are not likely to separate from placebo. A side effect is that the power to detect a significant difference for effective arms will decrease, since there is a possibility that these may be dropped prematurely. The impact of an interim analysis will therefore be reported as the statistical power to stop a futile (inefficacious) treatment arm and as the reduction in statistical power for efficacious treatment arms.

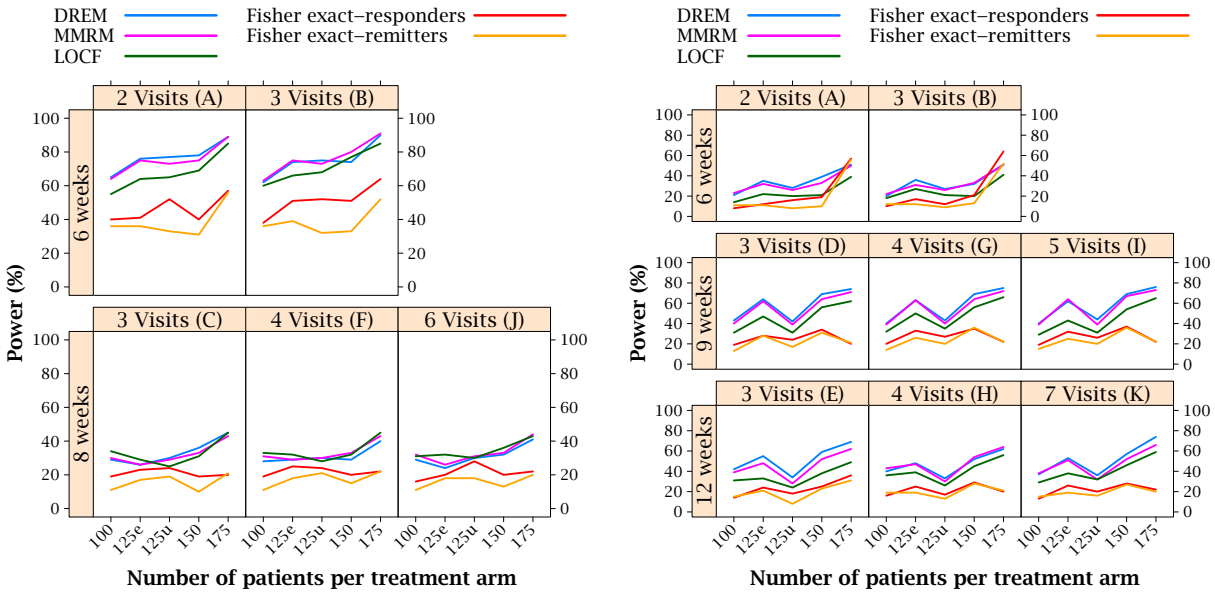
RESULTS

Differences between statistical endpoints

The sensitivity of the statistical techniques and the consequences of different population sizes on power to detect treatment effect are shown in figure 3. Clearly, from the different statistical approaches, the DREM and MMRM perform best, followed closely by the LOCF method. The Fisher exact tests based on either the percentage of remitters or the percentage of responders results in a much lower statistical power to detect treatment effect. On the other hand, figure 3 shows, as expected, that an increase in population size results in an increase in power.

Naturally, the statistical power is influenced by the dropout-mechanisms. The effects of sample size, frequency of visits and statistical endpoint however were qualitatively the same under the other two dropout scenarios (*data not shown*). Therefore only results from simulations under the second dropout scenario are reported.

Another important element is the evaluation of the false positive rate (type I error). As shown in figure 4 the type I error rates are well controlled under most scenarios. Although the statistical power of the tests based on the percentage of responders/remitters is considerably lower, the type I error is still comparable. In case of an unequal enrolment ratio (125u) the type I error of the Fisher exact test based on the percentage of responders



(a) Study 1

(b) Study 2

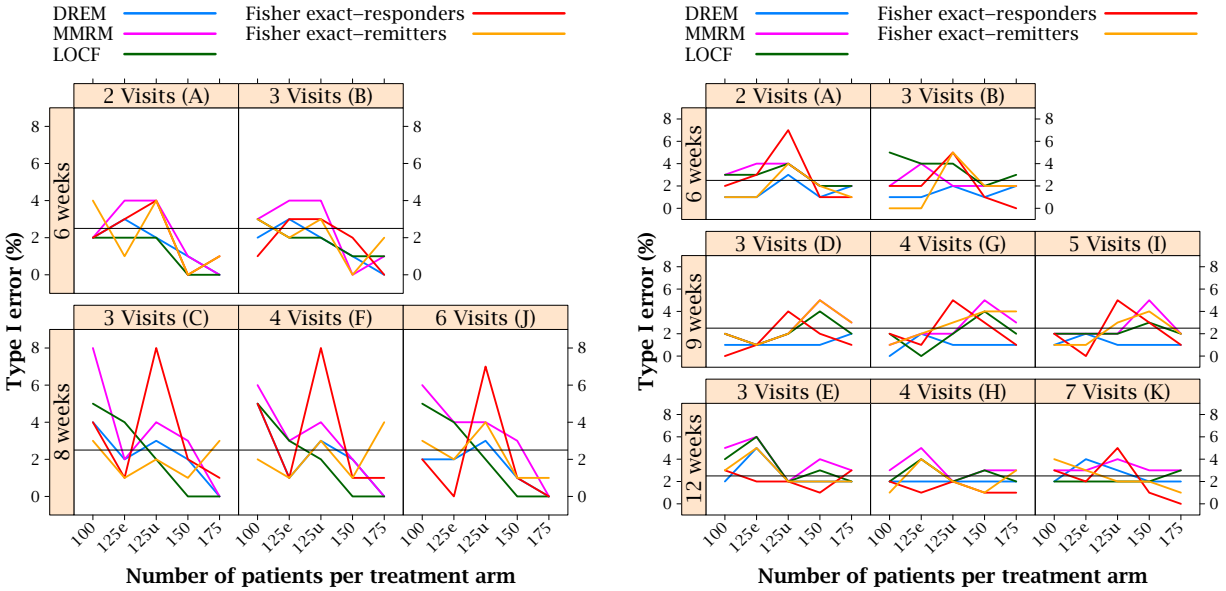
Figure 3. Power *versus* number of patients for the dual random effects model (DREM), mixed model for repeated measures (MMRM), LOCF and the Fisher exact test based on responders and remitters for treatment arm 1 in study 1 and study 2. Panels depict the different visit schemes. Panels on each row have the same duration and have increasing numbers of visits. Each point represents the summary of 100 simulated clinical trials. HAMD was used as clinical endpoint with a minimal inclusion criterion of 19. '125e' denotes the scenario with 125 patients evenly enrolled over all treatment arms, '125u' denotes the scenario with twice the number of patients in the active treatment groups as in the placebo group

is significantly inflated. Furthermore, when fewer than 150 patients are included, LOCF has a slightly inflated type I error.

Enrolment ratio

Figure 3 also allows investigation of the consequences of differences in randomisation ratio on study power by comparing the scenario equal and unequal enrolment (125e *versus* 125u), since the total number of enrolled patients is the same (n=375). The simulations show that the difference in power may be as much as 10%, in favour of equal randomisation. Although statistical power is an important characteristic of a clinical trial design, a comparison of bias may reveal more subtle effects of differences in study design features.

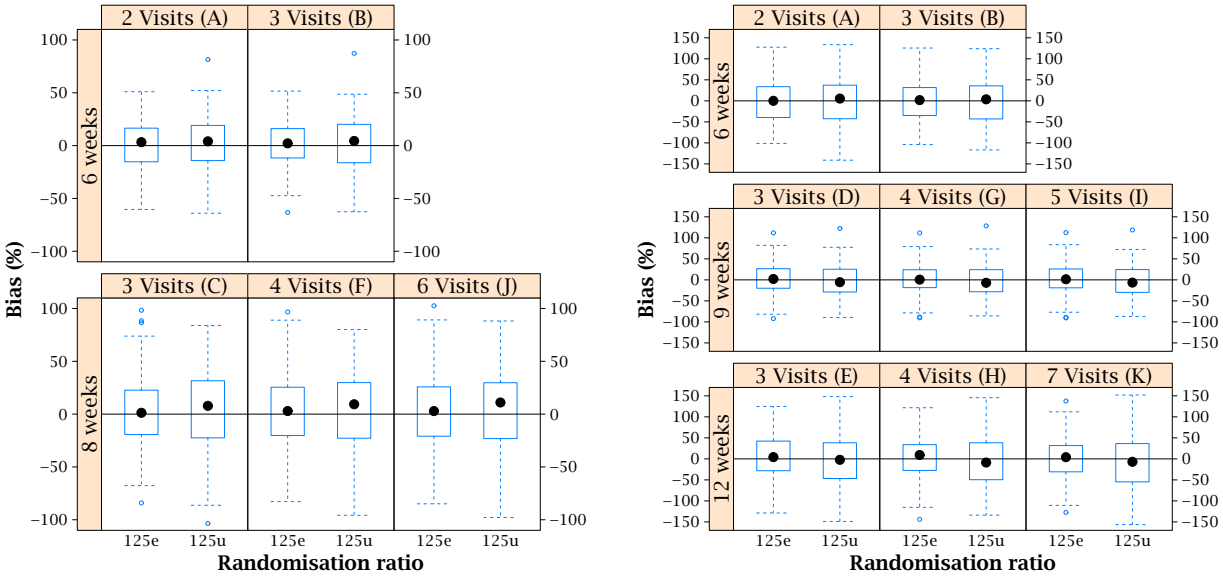
Box-plots of the bias of the estimate of treatment effect after equal and unequal enrolment resulting from 100 simulated datasets under all different observation schemes for treatment arm 2 in study 1 and 2 are shown in figure 5. The variability in the bias resulting from trials with unequal enrolment is consistently larger than those in trials with an equal randomisation ratio for each treatment arm. As indicated above, even if such differences are not very large, they may have implications for the statistical power of a trial.



(a) Study 1

(b) Study 2

Figure 4. Type I error *versus* number of patients for the dual random effects model (DREM), mixed model for repeated measures (MMRM), LOCF and the Fisher exact test based on responders and remitters in study 1 and study 2. Panels depict the different visit schemes. Panels on each row have the same duration and have increasing numbers of visits. Each point represents the summary of 100 simulated clinical trials. HAMD was used as clinical endpoint with a minimal inclusion criterion of 19. '125e' denotes the scenario with 125 patients evenly enrolled over all treatment arms, '125u' denotes the scenario with twice the number of patients in the active treatment groups



(a) Study 1

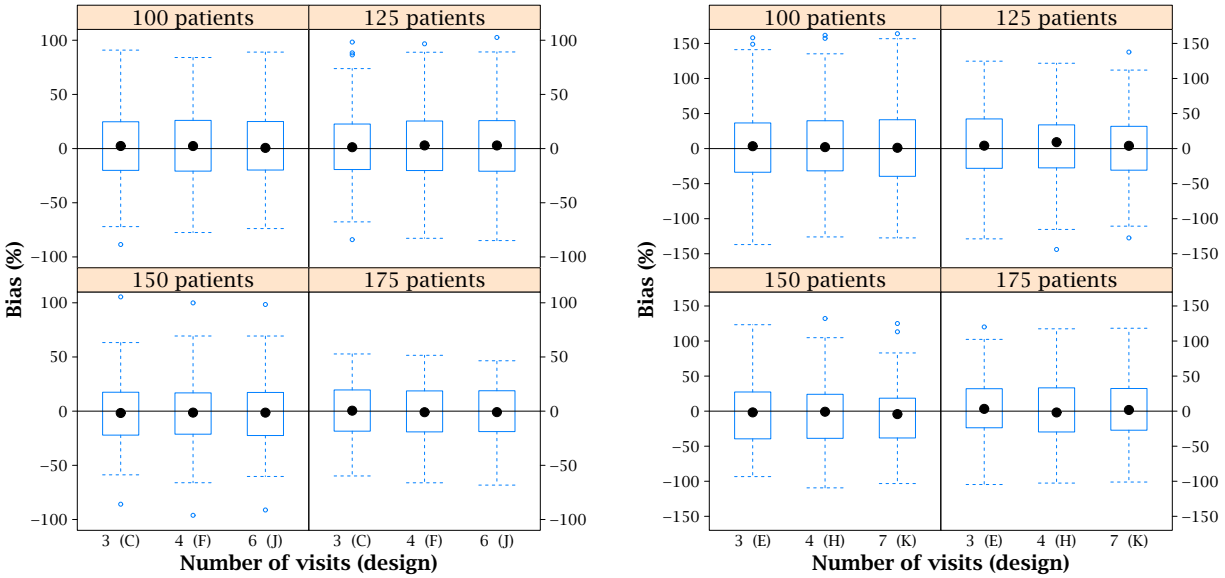
(b) Study 2

Figure 5. Box-plots of the bias for the estimates of the effect of treatment arm 2 for the even ('125e') and uneven ('125u') randomisation ratios for study 1 and study 2. Different panels are shown for the study designs. HAMD was used as endpoint with a minimal inclusion criterion of 19

Frequency of scheduled visits

Another important aspect that can be scrutinised in figure 3 is the effect of a reduction of the number of scheduled visits on the statistical power. The finding that treatment schedules terminating at week 6 (study 1) and week 9 (study 2) have a higher power to detect treatment effect than those terminating at the last visit in the original schedule (weeks 8 and 12 respectively) is explained by the larger treatment effect estimated at weeks 6 and 9 (table 1). More importantly, the effect of fewer measurements on statistical power appears to be negligible. Indeed, obtaining as few as two or three HAMD observations per patient yield a similar statistical power as in the case of a traditional sampling schedule (e.g., six or seven measurements).

To demonstrate the consequences of fewer measurements, the bias resulting from different schedules or visit schemes is shown in figure 6. Only scenarios including the last visit in the original observation scheme (weeks 9 and 12 respectively) are shown, ensuring the same absolute treatment effect so that these scenarios are comparable. These box-plots confirm that the number of scheduled visits does not influence the variability or the mean of the bias in the estimation of treatment effect. This was also the case for other dropout scenarios and simulated treatment effects (*data not shown*).



(a) Study 1

(b) Study 2

Figure 6. Consequences of the reduction of the number of measurements. The bias of the estimates of the effect of treatment arm 2 is shown for those study designs that include the last observation. Different panels are for the different number of enrolled patients. HAMD was used as endpoint with a minimal inclusion criterion of 19

Clinical endpoints and inclusion criteria

The consequences of using the HAM-D_7 subscale instead of the full HAM-D_{17} as clinical endpoint and of using a more stringent inclusion criterion are shown in figure 7 for study 2. The power to detect a significant treatment effect increases substantially with the use of the subscale as clinical endpoint. This difference between endpoints was not observed in study 1 (*data not shown*). The impact of using stringent inclusion criteria appears to be limited, most likely because the model did not take into account a differential treatment effect dependent on baseline severity.

Reduction in the frequency of scheduled visits

We have compared the results from simulated data in chapter 8 under the assumption that measurements were only made at weeks 2, 4 and 12 and under the full visit schedule (weeks 1, 2, 3, 4, 6, 9 and 12) (figure 8). As observed in figure 8, the bias in the estimate of the treatment effect with only 3 measurements is moderately larger only under more extreme dropout scenarios. Type I error rates increased only under unequal dropout scenarios (*data not shown*), confirming that a reduction of samples is possible without risking severely inflated type I and II errors.

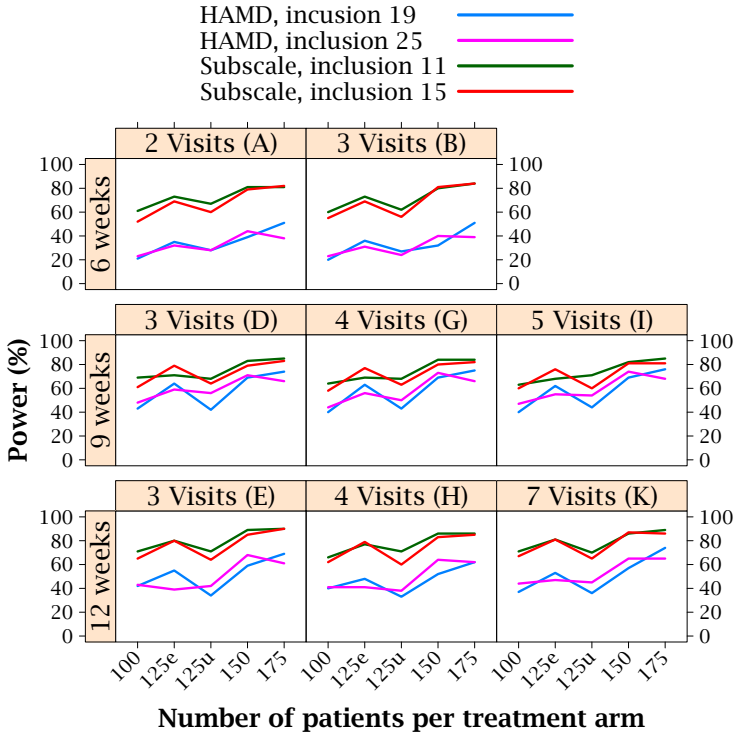


Figure 7. Influence of clinical endpoint and inclusion criterion on the estimated treatment effect of treatment arm 1 in study 2, analysed with the DREM. The power is plotted *versus* the number of patients, with different panels for visit schemes

Interim Analysis

The results of the interim analysis are presented as the power to stop a futile treatment arm and as the reduction in power to detect an effect caused by premature termination of an efficacious treatment arm. For the purposes of our evaluation, an analysis was assumed to take place after 75 and 100 days from the start of enrolment (figure 9). Consequently, enrolment was complete at the interim analysis planned on day 100 for the study enrolling 100 patients per treatment arm. The first column (futility) shows the probability of terminating the treatment arm with no drug effect. The interim analysis at 100 days has a power of 70% to terminate the inefficacious treatment arm. This percentage increases as the number of patients and study duration decrease. Indeed, when the 6-week and 8-week measurement schemes are compared it is striking to see that the 8-week measurement schemes clearly have gathered less information after 75 days (~20 completed patients per treatment arm) than the 6-week measurement schemes (~30 completed patients per treatment arm), resulting in fewer decisions to stop for futility (differences of up to 20%).

The remaining columns show the power conditional on the implementation of an interim analysis, which may lead to termination of efficacious arms for futility. To put this into perspective, the power of detecting a treatment effect without performing an interim analysis is also shown.

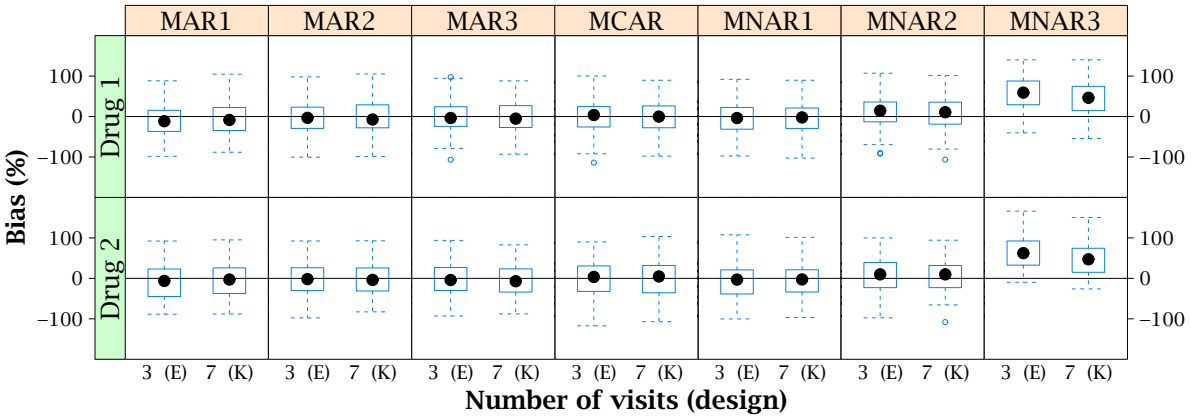


Figure 8. Percentage bias of the estimate of the treatment effect for simulations based the drug effects observed in study 2 *versus* the number of visits. Each panel shows the results for a different dropout scenario

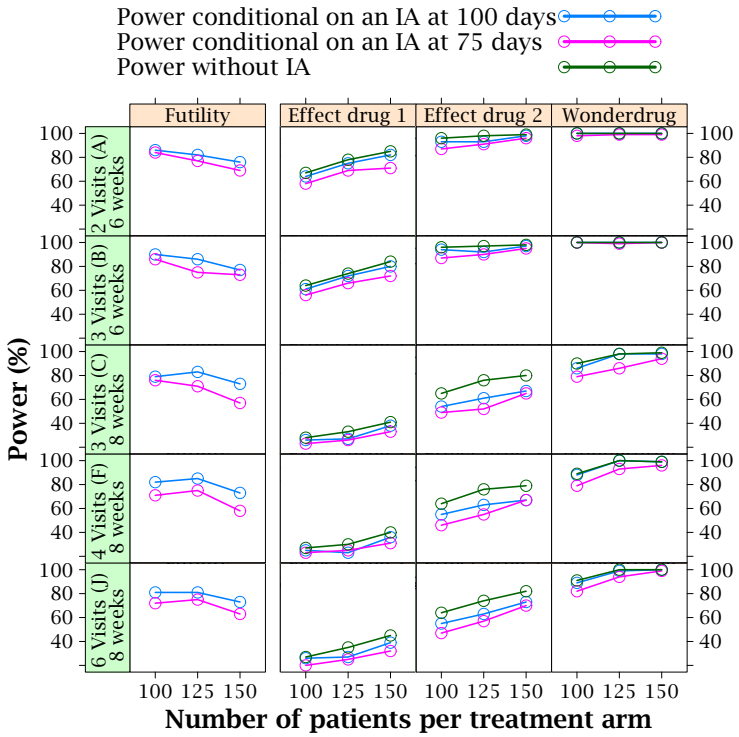


Figure 9. Results from the interim analysis based on study 1. The probability of early termination of an inefficacious treatment arm (in %) is shown in the first column *versus* the number of patients. Each row represents a different study design, the number of visits and the duration of the study protocol is specified. The remaining columns show the power to detect a treatment effect without interim analysis and the power conditional on an interim analysis at 75 days or 100 days. HAMD was used as endpoint as well as an inclusion criterion of 19

The findings observed for study 1 were repeated using simulations based on study 2 *(data not shown)*. Here it was also found that the use of shorter studies and reduced sampling frequency (6 and 9 weeks) performed significantly better than studies with longer duration (12 weeks).

DISCUSSION

There are numerous factors contributing to drug response and variability. The success of clinical trials in demonstrating significant treatment effect of novel compounds relies on our ability in understanding how each of these factors influence outcome. Clinical evidence from failure and attrition shows how intricate the interaction between these factors can be, which renders the planning, design and implementation of investigational studies difficult. Optimisation of clinical trials depends therefore on unravelling relevant factors, quantifying their impact and excluding those which result in noise or nuisance to signal detection. In this manuscript we illustrate how clinical trial simulations, a model-based approach, can be used to make inferences, test hypotheses and evaluate the performance of varying designs and conditions.

Simulation results

Most importantly perhaps is the realisation that limiting the number of visits per patient does not necessarily reduce statistical power. Even the reduction from 7 to 3 measurements did not lead to an important reduction in the ability to detect treatment effect. Not only does a reduction in the frequency of visits decrease costs, it also eases the burden on patients considerably. Another possible advantage is that the placebo effect may diminish due to the limited contact between investigators and patient, although these simulations do not allow us to investigate this. The limited impact of the frequency of scheduled visits on study power can be explained by the ratio between interindividual variability and residual variability. Since the variability between patients is larger than the error in measuring the HAMD, enrolling more patients is a more efficient manner of increasing power than increasing the frequency of assessments per patient.

A second valuable result from the simulations was that analysis of the percentage of responders or remitters using a Fisher exact tests results in reduced power. Furthermore, false positive (type I error) rates were higher, leading to even more concern, given that the percentage of responders is often reported as an outcome in clinical trials. Other authors have suggested than the percentage of patients in different stages of the disease is perhaps a better outcome than the analysis of the continuous HAMD (Bech *et al.*, 1984). We cannot dispute this recommendation using the results from a trial simulation, but a simple analysis of the number of responders or remitters based on the HAMD is clearly a far less sensitive measures than the use of a linear mixed model. Indeed, recently it has been shown that it may inflate drug-placebo differences dramatically (Kirsch and Moncrieff, 2007). This may seem contradictory, but the results presented here strengthen their conclusion that presenting a dichotomised endpoint based on an originally continuous endpoint may lead to false conclusions either way and should be discouraged. In contrast, we have previously shown that a survival analysis of depression data is feasible with similar sensitivity to detect treatment effect as compared to longitudinal models

(chapter 6). Survival analysis does not only dichotomise data, but takes the time dimension into account as well.

The third finding was that skewed enrolment in favour of an active treatment leads to a less precise estimate of the treatment effect (bias is more variable) and this may result in a reduction in power of up to 10%. This may be explained by a less precise estimation of the placebo effect, due to a reduced number of patients in this treatment arm. Since the outcome measure is the difference between treatment groups, this also reduces the precision of the estimates of drug effect. From an ethical perspective skewed enrolment may be desirable, since fewer patients are exposed to placebo treatment. However, the increase in the probability of failure may outweigh this. In any case, we do recommend simulations (with perhaps larger numbers than n=100) if a new clinical trial is to be designed, so that the consequences of the different options can be quantified and informed decisions can be reached, balancing ethical and statistical considerations.

The influence of the use of the HAM-D_7 rather than the HAM-D_{17} was also investigated. This subscale was found to be more sensitivity to treatment effect in a previous investigation (chapter 3). For study 1, the difference between the full HAMD and the response-based subscale was negligible. The similarity between the scales can be explained in this case by the effect size. Signal from a sufficiently large treatment effect may be equally detected irrespective of differences in the sensitivity of the endpoint. For study 2, however, an important difference was found. In this study the effect size is considerably smaller than study 1.

Demanding stricter inclusion criteria or enrolling only patients with higher HAMD scores at baseline did not seem to affect outcome. This is explained by the fact that no differential drug effect was included in the model for patients with high and low baseline, as no such effect was found in the data analysis. In a separate evaluation, in which the study population was divided into groups according to HAMD baseline, i.e., ≥ 23 and <23 (the median baseline), data were fitted with a model that included baseline as covariate. However, no improvement was observed in model fitting. Baseline HAMD was therefore not included as covariate in this analysis.

The interim analysis scenarios we propose in this paper prevented the progression of a futile treatment arm in 80% of the cases, whilst limiting the consequences for effective treatment arms. Future work will demonstrate the practical implementation of an adaptive trial design, in which the DREM is applied during an interim analysis using historical data. More than just adapting sample size, it is our interest to explore how adaptations can be made to optimise decision criteria as well as the actual timing of an interim analysis.

Limitations

Conclusions from clinical trial simulations are heavily dependent on the drug-action model and the trial execution model. The drug-action model we use does not take into account the mechanism of the interaction between the drug and its receptor or pharma-

cological target, nor does it consider pharmacokinetic factors. Clearly, incorporation of these elements would lead to an increased inferential value of our simulations, allowing for the investigation of a variety of additional trial design elements, such as dose, dosing interval and compliance. Because drug concentrations were not measured in the datasets which were at our disposal, these factors could not be included in the model. It is well known that steady state concentrations of paroxetine are highly variable (Kaye *et al.*, 1989). In fact, C_{max} ranged from 8.6 to 105 μg/ml, and the area under the curve (AUC) from 86.7 to 1911 μg/l*h. Such large differences in pharmacokinetics may partially explain the variability in treatment response that is commonly observed between patients. Moreover, investigations into the expression of the serotonin-receptor on blood-platelets have led to speculations that genetic influences at the receptor-level may also play an important role in the variability in responses. To investigate the consequences on study power if pharmacokinetic factors and exposure were taken into account, additional simulations were performed under the assumption of an exposure-response relationship according to an indirect response model. Results of these simulations indicated that the power of an analysis accounting for this relationship may be as high is 92%, whereas a standard *t*-test in the same scenario resulted in a power of 58% only (*data not shown*).

Naturally, the feasibility of reducing the frequency of scheduled visits per patient depends on the dropout scenarios that are present in a given clinical trial. We have investigated dropout scenarios that are sufficiently representative of those observed in clinical trials. In fact, we believe that the methodology presented in these simulations is more relevant to clinical practice than the assumption of a range of unlikely scenarios. Visualisation of the dropout of the different cohorts and each visit (figure 1) is a powerful manner of exploring the presence of dropout mechanisms. On the other hand, it is possible that dropout is due to some non-random non-observed mechanisms, which is why we have tested the robustness of a design with fewer measurements against more extreme dropout scenarios, such as those used in in Lane (2008) and chapter 8. The results of these investigations (figure 8) show that the bias in the estimate with a lower frequency of visits even under extreme assumptions of MNAR is only marginally higher.

Conclusions regarding study duration depend on the time course of drug effect, which was specified in the simulations. However, a reliable treatment effect is observed after 6 weeks in most studies (Montgomery, 2006). Furthermore, we argue that a drug which takes longer to demonstrate efficacy is not of interest to the treatment of depression. Advantages of short trial duration are numerous, including lower dropout, reduced costs, and increased usefulness of the interim analysis. Lower dropout is especially important since this means that dropout mechanisms will have less influence on study results.

The results of the CTS performed in this investigation depend on the parameters selected for the simulations. It is therefore interesting to consider whether the current results can be extrapolated to other populations, or even other disease areas. Especially the ratio between the inter-individual and residual variability plays a major part in this respect. Except for the baseline HAMD simulations, no assumptions were made considering

the population type.

The suggested changes may not be as inventive as those suggested by Fava *et al.* (2003). It was not possible to include this design in this exercise, because it requires the individual behaviour of patients to be captured, which is not feasible with the current state of knowledge. However, we hope that these simple adjustments may prove beneficial and can easily be implemented. As discussed elsewhere (Holbrook and Goldsmith, 2003), large changes in clinical trial design are unlikely to occur instantaneously due to the high financial stakes and patient time. Indeed, we hope that CTS will spur changes and perhaps create awareness in the field that change is necessary and not necessarily dangerous.

Recommendations

The following guidelines for the design of new clinical trials in depression can be derived from the wide range of simulated scenarios:

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 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. Consider the relevance of an interim analysis for early termination of futile treatment arms without compromising positive treatment arms

Most importantly, assess drug exposure and possibly the phenotype of patients for drug metabolism and pharmacodynamics. The availability of such data enables characterisation of exposure-response relationships and subsequent optimisation of dosing regimens (Danhof *et al.*, 2007). Although the exposure-response relationship of SSRIs in general and paroxetine in particular have remained obscure, it is hard to believe that psychiatrists still do not conceive such a relationship. Elucidation of the PKPD relationships of novel antidepressants will only be possible if sampling of pharmacokinetic and pharmacodynamic data is considered in the design of a clinical trial.

In conclusion, we have shown that CTS is a valuable tool to integrate and quantify the impact of multiple trial design factors. Such clear-cut results could not have been obtained by traditional meta-analysis, as confounding factors cannot be dissected independently. In an analogy to the impact of differential diagnostic tools following the advancements in laboratory technologies during the last century, research physicians must become aware that CTS is the instrument for differential diagnosis in clinical drug development. Differential diagnosis of clinical trials is the only alternative to trial and error.

REFERENCES

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.

Chabaud S, Girard P, Nony P, and Boissel JP (2002) Clinical trial simulation using therapeutic effect modeling: Application to ivabradine efficacy in patients with angina pectoris. *J Pharmacokinet Pharmacodyn* **29**:339–363.

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.

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.

Girard P (2005) Clinical trial simulation: A tool for understanding study failures and preventing them. *Basic Clin Pharmacol Toxicol* **96**:228–234.

Gomeni R and Merlo-Pich E (2007) Bayesian modelling and ROC analysis to predict placebo responders using clinical score measured in the initial weeks of treatment in depression trials. *Br J Clin Pharmacol* **63**:595–613.

Gruwez B, Poirier MF, Dauphin A, Olie JP, and Tod M (2007) A kinetic-pharmacodynamic model for clinical trial simulation of antidepressant action: Application to clomipramine-lithium interaction. *Contemp Clin Trials* **28**:276–287.

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.

Holbrook A and Goldsmith C (2003) Innovation and placebos in research: a new design of clinical trial. *Lancet* **362**:2036–2037.

Kaye CM, Haddock RE, Langley PF, Mellows G, Tasker TCG, Zussman BD, and Greb WH (1989) A review of the metabolism and pharmacokinetics of paroxetine in man. *Acta Psychiatr Scand* **80**:60–75.

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

Kimko HC, Reele SSB, Holford NHG, and Peck CC (2000) Prediction of the outcome of a phase 3 clinical trial of an antischizophrenic agent (quetiapine fumarate) by simulation with a population pharmacokinetic and pharmacodynamic model. *Clin Pharmacol Ther* **68**:568–577.

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

Kramer MS, Winokur A, Kelsey J, Preskorn SH, Rothschild AJ, Snavely D, Ghosh K, Ball WA, Reines SA, Munjack D, Apter JT, Cunningham L, Kling M, Bari M, Getson A, and Lee Y (2004) Demonstration of the efficacy and safety of a novel substance P (NK1) receptor antagonist in major depression. *Neuropsychopharmacology* **29**:385–392.

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

Lockwood P, Ewy W, Hermann D, and Holford N (2006) Application of clinical trial simulation to compare proof-of-concept study designs for drugs with a slow onset of effect; an example in Alzheimer's disease. *Pharm Res* **23**:2050–2059.

Lockwood PA, Cook JA, Ewy WE, and Mandema JW (2003) The use of clinical trial simulation to

support dose selection: Application to development of a new treatment for chronic neuropathic pain. *Pharm Res* **20**:1752–1759.

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, Clark W, and David S (2001a) Accounting for dropout bias using mixed-effects models. *J Biopharm Stat* **11**:9–21.

Mallinckrodt C, Kaiser C, Watkin J, Molenberghs G, and Carroll R (2004a) 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.

Mallinckrodt CH, Clark WS, and David SR (2001b) Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Inf J* **35**:1215–1225.

Mallinckrodt CH, Kaiser CJ, Watkin JG, Detke MJ, Molenberghs G, and Carroll RJ (2004b) Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. *Pharm Stat* **3**:171–186.

Montgomery SA (2006) Why do we need new and better antidepressants? *Int Clin Psychopharmacol* **21**:S1–S10.

Nielsen DM (2006) Corticotropin-releasing factor type-1 receptor antagonists: The next class of antidepressants? *Life Sci* **78**:909–919.

Post TM, Freijer JI, DeJongh J, and Danhof M (2005) Disease system analysis: Basic disease progression models in degenerative disease. *Pharm Res* **22**:1038–1049.

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.

de Ridder F (2005) Predicting the outcome of phase iii trials using phase ii data: A case study of clinical trial simulation in late stage drug development. *Basic Clin Pharmacol Toxicol* **96**:235–241.

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.

Veyrat-Follet C, Bruno R, Olivares R, Rhodes GR, and Chaikin P (2000) Clinical trial simulation of docetaxel in patients with cancer as a tool for dosage optimization. *Clin Pharmacol Ther* **68**:677–687.

Yim DS, Zhou HH, Buckwalter M, Nestorov I, Peck CC, and Lee H (2005) Population pharmacokinetic analysis and simulation of the time-concentration profile of etanercept in pediatric patients with juvenile rheumatoid arthritis. *J Clin Pharmacol* **45**:246–256.