To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

8

Relevance of a new hierarchical model for the analysis of longitudinal data with dropout in depression trials

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

Submitted for publication

ABSTRACT

A number of issues related to the design and analysis of clinical studies contribute to the high failure rate observed in the evaluation of antidepressant drugs. In contrast to statistical methods in which response is determined by differences relative to placebo at completion of treatment, increasing evidence exists that treatment effect may be better characterised by individual longitudinal data. Longitudinal models offer many advantages as they provide information about individual patients in the trial. These models can be especially useful for simulation purposes, but require attention with regard to dropout. Based on the results of a functional principal component analysis, we propose the use of a dual random effects model (DREM) that accounts for the presence of different dropout scenarios. The objective of this investigation was to compare the analysis of efficacy data and evaluate the impact of dropout on type I error and power using the DREM, the mixed model for repeated measures (MMRM) and last observation carried forward (LOCF) methods.

Historical data from clinical trials in depression was used for model fitting. The goodness-of-fit of all models was compared using graphical and statistical approaches. Individual HAMD scores over time were simulated using the DREM. The effect of dropout was investigated according to seven different scenarios under the assumption of missingness completely at random (MCAR), missingness at random (MAR) and missingness not at random (MNAR). Subsequent data fitting included the interactions treatment-time and baseline-time as fixed effects for the DREM and MMRM model.

Diagnostic plots reveal that the DREM describes individual patient data better than the MMRM or a single random effects model. In addition, simulations show that there is little difference between the DREM and MMRM with respect to the fixed effect estimates in all scenarios. The DREM was found to outperform the MMRM with regard to type I error and power. As expected, LOCF showed higher type I errors or reduced power under various scenarios.

A considerable improvement in the goodness-of-fit is observed for the DREM, as compared to the MMRM. Although this difference represents only a minor variation in type I error and power, we recommend the use of DREM, especially for the purpose of simulations in clinical trial design. The main advantages include its simplicity and parameterisation, which may facilitate the interpretation of model estimates by the non-statistical community. The use of LOCF is strongly discouraged since the estimates may be biased under likely dropout scenarios.

INTRODUCTION

The analysis of treatment efficacy in chronic diseases, such as depression, ought to consider the time course of response rather than be based merely upon the differences from baseline at a specific time point. A considerable number of publications have shown that

the latter method misrepresents the true treatment effect (Mazumdar *et al.*, 1999; Wood *et al.*, 2005; Huson *et al.*, 2007). This possible bias is further increased by the use of intent-to-treat (ITT) analysis, which requires that all subjects randomised to a treatment arm should be included in the analysis. A shift in the paradigm for statistical analysis of longitudinal data demands, therefore, better understanding of the natural history of disease and appropriate consideration about the phenomenon of dropout and data censoring.

In depression, it is an established fact that up to half of clinical trials fail in spite of adequate active treatment (Khan *et al.*, 2002). One of the possible reasons for this failure is the high dropout rate observed in these studies. The percentage of patients completing the trial after at least one efficacy measurement is as low as 50% in some cases, with percentages higher than 70% being the exception rather than the rule. The consequence of these high dropout rates is that methodologies which take missingness adequately into account will improve the quality of the analysis of depression trials, and may decrease this high failure rate.

Up to a few years ago, last observation carried forward (LOCF) imputation was the standard method to handle missing data. It has been demonstrated that LOCF imputation provides an unbiased estimate of the effect of a drug in the presence of dropout according to missingness completely at random (MCAR, not attributable to any specific cause) only when the dropout rate in all treatment arms is the same (Molenberghs *et al.*, 2004). Further bias is expected in the presence of dropout according to missingness at random (MAR, depending on observed data) and missingness not at random (MNAR, depending on censored data). Although this is well-accepted it is commonly believed that this bias is of a conservative nature and it is still often used in regulatory submissions, in spite of the efforts to change this (Mallinckrodt, 2006).

More advanced methodologies to deal with censored data are becoming available as a result of elaborate software packages and faster computers. The era in which statisticians were forced to resort to LOCF imputation because of constraints other than regulatory requirements, is now over. In recent years, the mixed model for repeated measures (MMRM) (Mallinckrodt *et al.*, 2001a), a marginal linear mixed model (Verbeke and Molenberghs, 2000; Laird and Ware, 1982) has gained considerable popularity because of its ability to use all obtained data during a trial to provide unbiased estimates of drug effect in the presence of both MCAR and MAR dropout mechanisms (Mallinckrodt *et al.*, 2004a; Kinon *et al.*, 2006; Thase *et al.*, 2006; Davis *et al.*, 2005). Several simulation studies have shown the robustness of the MMRM in these situations (Mallinckrodt *et al.*, 2001b, 2004b), and a recent simulation study (Lane, 2008) has shown that the MMRM performs much better than LOCF over a range of scenarios of missingness. However, none of these studies have questioned whether the MMRM fits individual data accurately.

Elsewhere, we have applied functional principal component analysis to individual curves from patients in anti-depressant trials (chapter 7). Rather than focusing on the average behaviour of patients, as in standard longitudinal data analysis, functional data analysis

investigates the differences between patients instead. Information about the nature and extent of inter-individual variability may lead to more appropriate models as well as model parameterisations. In that investigation, it was found that the first principal component from the functional data analysis corresponds to an additive random effect. In hierarchical linear mixed models such an additive random effect is commonly included. Given the presence of a second component, which was found to describe a random slope effect, the current manuscript proposes the use of a model with two random effects, a dual random effects model (DREM) to fit depression data. Taking into account the high dropout rate observed in clinical trials, it is anticipated that the combination of random effects will provide a more accurate description of individual patient data.

An extension to the hierarchical linear mixed-effects model can be used to implement such a model. As mentioned above, this approach generally assumes that a random (subject-specific) effect exists which is additive, i.e., a subject is expected to have measurements which are predominantly above or below the population average. Even though this parameterisation leads to issues with respect to methods based on maximum likelihood theory (Molenberghs and Verbeke, 2004), we believe that such issues may be overcome within the Bayesian context. Two important advantages of the use of Bayesian statistics include the possibility of the computation of the posterior predictive distribution, instead of a single point estimate as measure of treatment effect, as well as the explicit calculation and interpretation of probabilities in general. Bayesian statistics is gaining in popularity due to the fact that flexible software exists such as WinBUGS (Lunn *et al.*, 2000) which implements Markov chain Monte Carlo (MCMC) sampling from the posterior distribution. In the current investigation, both the model with one additive random effect (random effect model, REM) and its extension, the DREM, are implemented in WinBUGS. After a comparison of the MMRM, REM and DREM with respect to their ability to describe individual data, simulations based upon the DREM will be performed to investigate the power and bias of the MMRM, DREM and LOCF methods under seven different scenarios of dropout according to MCAR, MAR and MNAR (Lane, 2008). False positive rates (type I error) will be investigated through simulation of a hypothetical treatment arm which has no treatment effect and as many patients as the active treatment arms. Moreover, the current investigation focuses on exploring whether the DREM is the most appropriate model for the simulation of new data. The use of simulated patient data is a powerful tool for the evaluation of study characteristics before the implementation of a study protocol. It is also essential for the optimisation of adaptive designs and interim analyses.

METHODS

Study data

Data from two double-blind, placebo-controlled randomised clinical trials of patients with major depression were retrieved from GSK's 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 17-item Hamilton depression rating scale (HAMD) (Hamilton, 1967) was measured at weeks 1, 2, 3, 4, 6 and 8 after start of treatment. A total of 459 patients with major depression were evenly enrolled across the treatment arms.

Study 2 (unpublished, see http://ctr.gsk.co.uk, protocol number 128) was a randomised placebo-controlled trial in which paroxetine and fluoxetine were compared. The HAMD 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 350 patients were enrolled in both active treatment arms.

All data manipulation and graphs were performed in R, the language and environment for statistical computing (R Development Core Team, 2007).

Data fitting and parameter estimation

First, the mixed model for repeated measures (MMRM), a hierarchical random effects model (REM) and the dual random effects model (DREM) were fitted to the data. The fixed effects in these models were the interactions between time and treatment, and between time and baseline. The MMRM was implemented using *proc mixed* in PC SAS (v9.1 for Windows, SAS Institute, Cary, NC, USA). The REM and DREM were implemented in WinBUGS version 1.4.1 (Lunn *et al.*, 2000). The equations describing each model are given below. Throughout the Bayesian analyses flat normal priors with little precision were used for the fixed effects, and uniform priors on the scale of the standard deviation, since these are generally assumed not to influence the posterior distributions of the parameters of interest, and therefore the inference.

The mixed model for repeated measures is represented by equation 1:

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

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. A further assumption is that the fixed effects are drawn from a multivariate distribution with the same unstructured covariance matrix for all individuals.

The random effects model is represented by equation 2:

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

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. $\eta 1_i$ is the random effect of individual i (normally distributed with mean 0 and unknown variance) and ϵ is the measurement error (normally distributed with mean 0 and unknown variance).

The dual random effects model is represented by equation 3:

$$Y_{ij} = \mathsf{BAS}_i \cdot \beta_j + \theta_{z,j} + \eta 1_i + \eta 2_i \cdot j + \epsilon_{ij} \tag{3}$$

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). The second random effect $\eta 2$ is multiplied by the observation number, which corresponds to the random slope effect identified in the functional principal component analysis (chapter 7).

Comparison of the models

The performance of the MMRM, REM and DREM are compared using two graphical approaches. First, model predicted HAMD scores will be plotted against the observed HAMD scores. Second, the time course of individual response profiles and corresponding model fits will be compared between the three models. Since our main objective is to evaluate model performance for the purposes of simulation, it was deemed appropriate to apply a statistical diagnostic measure which focuses on the simulation abilities of a model. Recently, normalised prediction discrepancy errors (NPDE) have been proposed by Brendel *et al.* (2006). Briefly, this method determines whether simulated datasets are exchangeable with the original dataset using graphical diagnostics and statistical tests. Since the maximum likelihood estimates of the MMRM and the REM are the same for normally distributed data, model comparison will be limited to the REM and DREM.

Simulations

The next part of the manuscript investigates the operational characteristics (type I and II error) of all models under various scenarios of dropout. As the DREM is considered the most appropriate model to generate new data, it is used to simulate new patients. Subsequently, dropout is introduced according to seven scenarios, followed by a fit of all models to the simulated datasets. A more detailed description of these procedures is provided below.

New trials were simulated in R based on the means of the posterior distributions of the parameters estimated in the DREM. First, baseline values for all patients were simulated using a normal distribution (mean 20, standard deviation 4) truncated between 19 and 40. These values were based on observed patient data in the historical trials. The simulated baseline values were subsequently multiplied by the baseline-time fixed effect for each time point, yielding individual response time profiles. Fixed treatment effects were then added to the simulated individual response profiles. The random subject-specific effects were simulated from a multivariate normal distribution based on the parameters fitted from the data. Finally, measurement error was introduced by sampling from a normal distribution. The resulting HAM-D_{17} values were then rounded to the nearest integer to

represent the discrete nature of the endpoint.

In a second stage, the simulated data was exposed to seven different dropout mechanisms, as described in detail elsewhere (Lane, 2008). In brief, the dropout rate was fixed at a 3.5% per week, with the total dropout at the end of the trial being approximately the same as in the original studies. The choice for the dropout rate was based on an analysis of data from 8 clinical studies which was available to us, which showed no reproducible time-dependency of the dropout rate over time. For MCAR, a completely random dropout mechanism was used. For MAR and MNAR, 3 different scenarios each were used. For all scenarios the patients in each treatment arm were divided into 9 equally sized dropout cohorts. For MAR, the preceding observation was used to determine the probability of dropout, whereas for MNAR the value of the current (to be censored) observation was used to determine this probability. The probabilities of dropout were calculated according to the following steps: In scenario A (MAR1/MNAR1) the likelihood of dropout increased linearly with severity of depression. In scenario B (MAR2/MNAR2) only patients in the 4 most severely depressed categories were subject to dropout, again increasing linearly with the severity of depression. In scenario C (MAR3/MNAR3) dropout was present only in the most severely depressed patient population. The slope for the linear increase was calculated to result in a dropout percentage of 3.5% per week. Note that the dropout percentage of 3.5% was applied per week rather than per visit, as this corresponds to the total dropout rates observed in the two trials which have different trial durations (8 versus 12 weeks). Furthermore, like Lane (2008), we have investigated the consequences of unequal dropout between treatment arms. Therefore, dropout rates were varied in ratios of 1:1, 1:2 and 2:1 for placebo and active treatment respectively, with the resulting dropout rate of 3.5% per week.

For *unequal* dropout this resulted in problems for the simulation of scenario C if the interval between measurements was longer than two weeks. In this case, the dropout rate in a treatment arm with the higher dropout rate would have to be >11% in order to result in a total dropout rate of 3.5% per week. However, because only the 100%/9 cohorts = 11% most severely depressed patients are subject to dropout in scenario C, this dropout rate could not be achieved. In these circumstances all patients in the most severely depressed group were dropped from the trial.

The resulting datasets were fitted using the DREM, MMRM and LOCF. LOCF was implemented by carrying the last observation forward to the last occasion when a subject was removed from the study, followed by a *t*-test for the differences between the treatment arms and placebo as suggested by Molenberghs *et al.* (2004). This simulation-missingness-fitting procedure was repeated 100 times for the calculation of bias and type II error (positive treatments) and 1000 times for the type I error (false positive rates) since the latter were generally low.

To summarise the results, box-plots of the bias of the estimate of the treatment effect at the last observation were created. In addition, graphs were used to report the power to detect a statistically significant difference (equal to 100-type II error) and the type I error.

For the type I error, we only take into account cases in which $p<0.05$ and where active treatment outperforms placebo. This results in one-sided type I error rates which have an expected value of 2.5%.

RESULTS

Model fitting & diagnostics

The parameter estimates for the fixed effects were similar, irrespective whether the DREM, MMRM, or REM was used (*data not shown*). Figure 1A shows a plot of the individual predicted HAMD scores versus the observed scores and figure 1B shows the time course of the HAMD and the model fits for 49 subjects.

Since the maximum likelihood estimates of the MMRM and the REM are the same for normally distributed data and the MMRM does not produce individual predictions because of its parameterisation, the focus in figure 1A should be on the comparison of the REM and the DREM. Clearly, the second random effect in the DREM diminishes the bias observed in the REM for the prediction of low and high HAMD values. This is further illustrated by the fact that the observed total number of responders (more than 50% decrease from baseline HAMD) in the original dataset was 119 and this same number based on the individual predicted HAMD values for the REM amounted to only 87. If the number of responders was computed based on the individual HAMD predictions of the DREM however, it equalled 109. The MMRM does not allow this sort of calculation because it does not provide individual predicted values. From figure 1B it is clear that the DREM provides a slightly better description of the data than the REM in some cases.

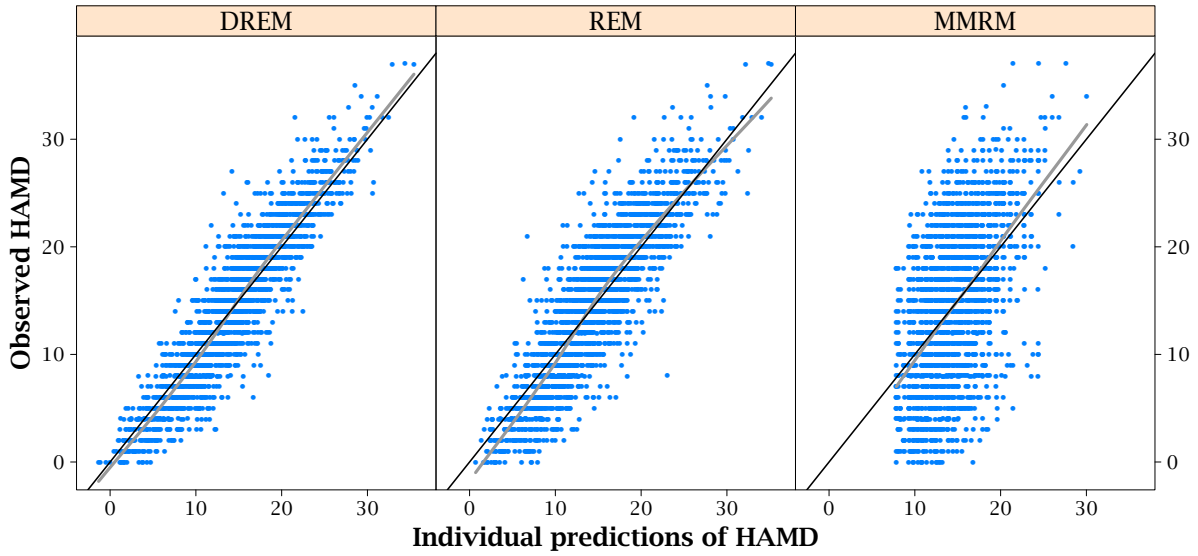
Figure 2 shows the normalised prediction discrepancy errors (NPDE) for the REM and DREM. The NPDE for the DREM follow a standard normal distribution, whereas the NPDE for the REM show that the variability in the data is lower in the simulated datasets than in the original data. The mean and variance of the distribution of the NPDE were tested for being different from their expected values using a Wilcoxon signed rank test for the mean and a Fisher test for the variance. The Wilcoxon test showed that the mean of the NPDE for both models did not differ significantly from 0. The Fisher test however showed that the variance of the NPDE for the REM differed significantly from 1 ($p<0.001$), but did not reveal such a discrepancy for the DREM.

Simulation outcome

Based on the aforementioned results, simulations of individual patient data were performed according to the DREM. The means of the posterior distributions of all parameters which were used for the subsequent simulations are shown in table 1.

Study 1: Estimated Bias

Box-plots of the bias of the estimates of the treatment effect at week 8 (n=100 simulations) under the seven dropout scenarios for the various dropout ratios are shown in figure 3.

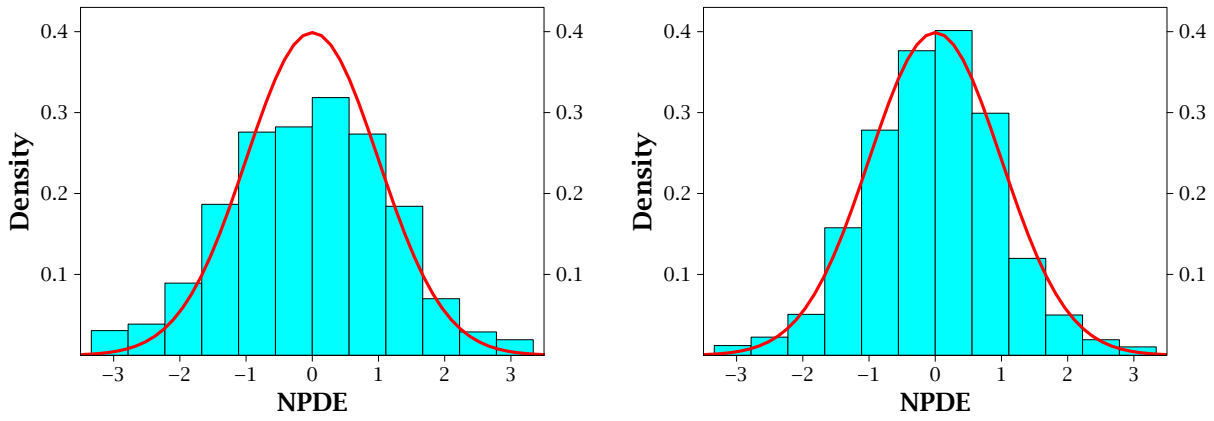


(a) observed HAMD *versus* predicted HAMD

(b) individual model fits

Figure 1. (a) Observed HAMD scores *versus* individual predicted HAMD scores for the three models. The black line is the line of unity and the grey line is a smoothing function. (b) Individual HAMD time profiles for 49 patients. Observations are represented by dots

These simulations were based on study 1, in which 150 patients were included in each treatment arm and a total dropout rate of 25% was achieved. The bias of the estimates of the DREM and MMRM are very similar under all dropout scenarios. The LOCF method is consistently biased when placebo and active dropout rates are not equal. The estimates obtained using the DREM and MMRM are unbiased when the dropout mechanism is MCAR or MAR, but the more extreme scenarios of MNAR do yield biased estimates when placebo and active dropout rates are not the same. The bias for treatment 1 is consistently more variable than that for treatment 2, which can be explained by the higher treatment effect size for this treatment arm.



(a) REM (b) DREM

Figure 2. Distribution of the normalised prediction discrepancy errors. As a reference, the density function of the standard normal distribution (mean 0, variance 1) is shown

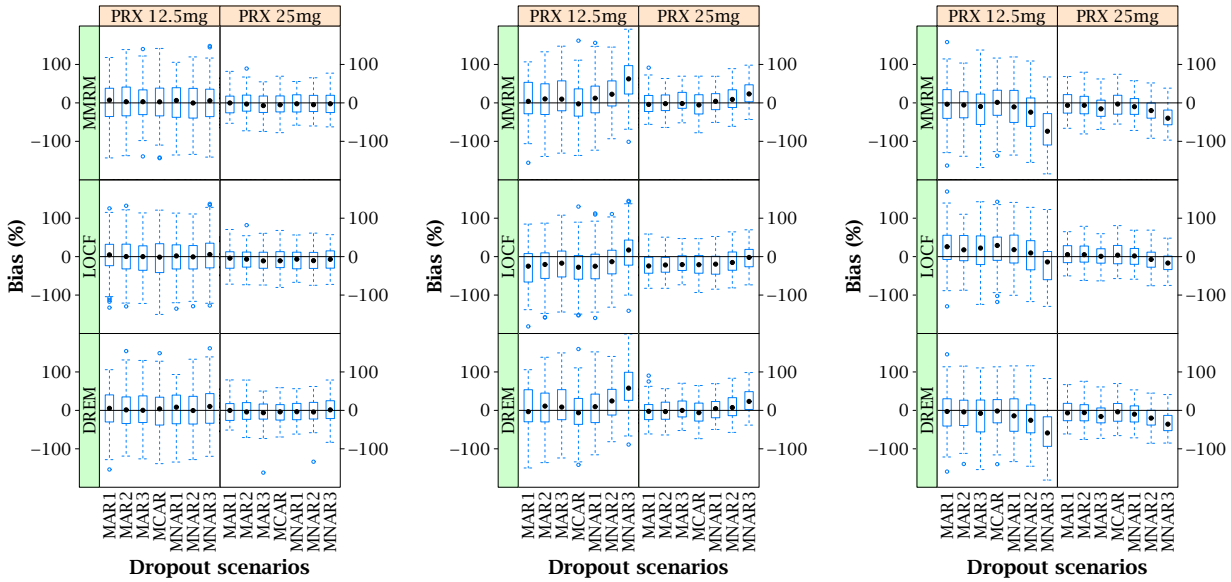
Table 1. Parameter values used for the simulations. The 2x2 variance-covariance matrix is given under the I.I.V. and σ/I.I.V. headings

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	-	-	23.1	-1.73
	paroxetine (25 mg)	0.0	1.4	1.8	2.3	3.9	2.9	-	-	-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	19.8	-2.0
	fluoxetine (max 80 mg)	0.0	0.8	1.1	0.7	1.8	-	2.8	2.0	-2.0	1.1

Study 1: Type I error and power

The false positive rates (n=1,000 simulations) and the power (1 - type II error, n=100 simulations) for paroxetine CR 12.5 mg are summarised in figure 4. Since the type I error is one-sided, its expected level is 2.5%. Based on MCAR scenario with equal dropout rates, the predicted power of the treatment arm was estimated to be 40%.

It is apparent that the type I error can inflate under various dropout scenarios. LOCF has inflated one-sided type I errors of up to 6.2% (> 2 times the nominal level) when dropout under placebo is higher than under active treatment. The MMRM and DREM control the type I error under MAR (except when active dropout is higher than placebo dropout), but the most extreme MNAR scenarios lead to an inflated type I error (18.9%) when dropout under active treatment is higher than under placebo. The power to detect



(a) 1:1 placebo:active dropout (b) 1:2 placebo:active dropout (c) 2:1 placebo:active dropout

Figure 3. Box-plots of the bias for each treatment arm of study 1 *versus* the dropout scenarios for the MMRM, DREM and LOCF methods (100 simulations). PRX = paroxetine CR

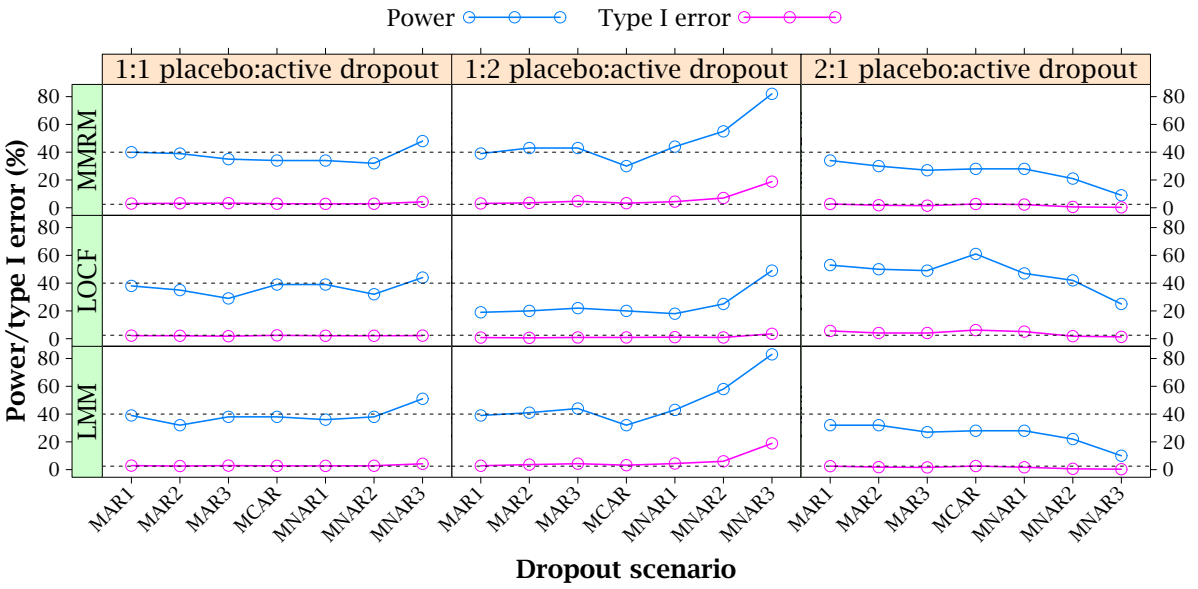
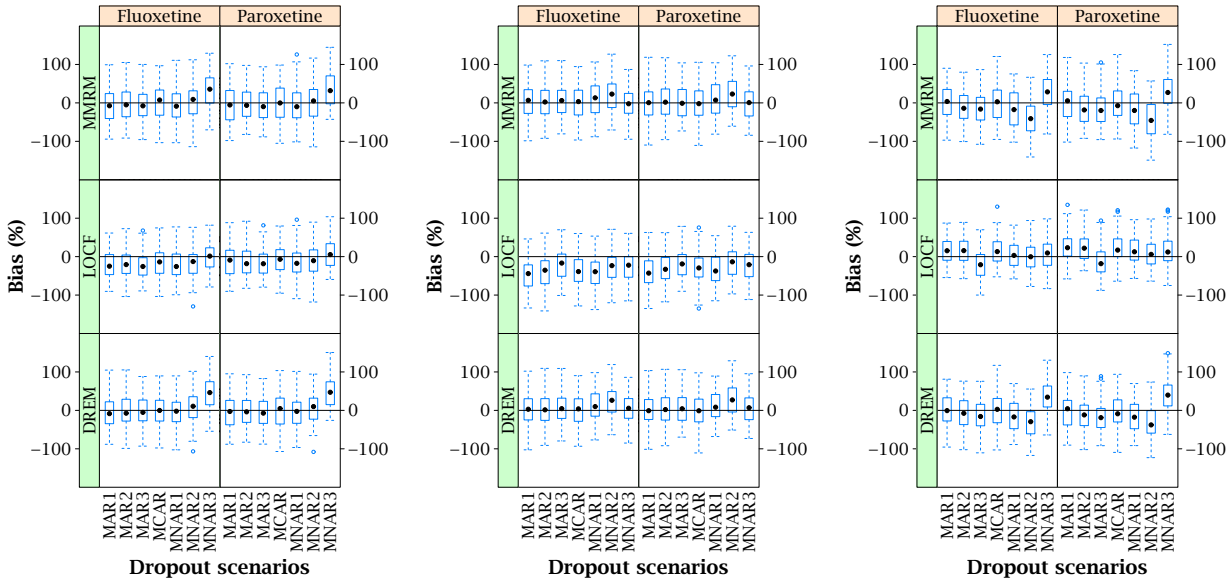


Figure 4. Power and type I error associated with the dropout scenarios, based on study 1. Different panels represent the models (horizontally) and the dropout ratios (vertically)

a significant treatment effect was generally the same for MMRM and DREM, and worse for LOCF (except when the anti-conservative bias led to a higher power).

Study 2: Estimated bias

Box-plots of the bias of the estimates of the treatment effect at week 12 (n=100 simulations) under the seven dropout scenarios and various dropout ratios are shown in figure 5. These simulations were based on study 2, in which 140 patients were included in



(a) 1:1 placebo:active dropout (b) 1:2 placebo:active dropout (c) 2:1 placebo:active dropout

Figure 5. Box-plots of the bias for each treatment arm of study 2 *versus* the dropout scenarios for the MMRM, DREM and LOCF methods (100 simulations)

the placebo arm and 350 in each active treatment arm. A total dropout rate of 35% was simulated in this trial.

The bias of the estimates based on study 2 is clearly larger than the bias observed in the simulations based on study 1 (figure 3). This may be due to the increased dropout rate, caused by the difference in duration between the two studies. Again, the DREM and MMRM methods are indistinguishable, and a bias is observed in the case of equal dropout only under the more extreme MNAR scenario. When the dropout rates are unequal, estimates under the less extreme MNAR scenarios are also biased. For LOCF the situation is considerably worse, especially in the unequal dropout scenarios. The bias caused by LOCF is mostly conservative but becomes anti-conservative when the placebo dropout rate is twice the active dropout rate.

Study 2: Type I error and power

The false positive rates (type I error, n=1,000 simulations) and the power (1 - type II error, n=100 simulations) for fluoxetine are summarised in figure 6. Since the type I error is one-sided, its expected level is 2.5%. Based on MCAR scenario with equal dropout rates, the predicted power of the treatment arm was estimated to be 70%.

The type I error in study 2 is considerably higher than in study 1 (figure 4). When the dropout rates are equal between treatments or higher for placebo, the type I error is controlled except for the most extreme MNAR scenario. For the case in which the dropout rate is twice as high in placebo treatment, LOCF analysis shows inflated one-sided type I error (up to 13.3 %) except for the most extreme MAR mechanism. The MMRM and DREM yield type I error which is acceptable for most scenarios, except for the two more extreme

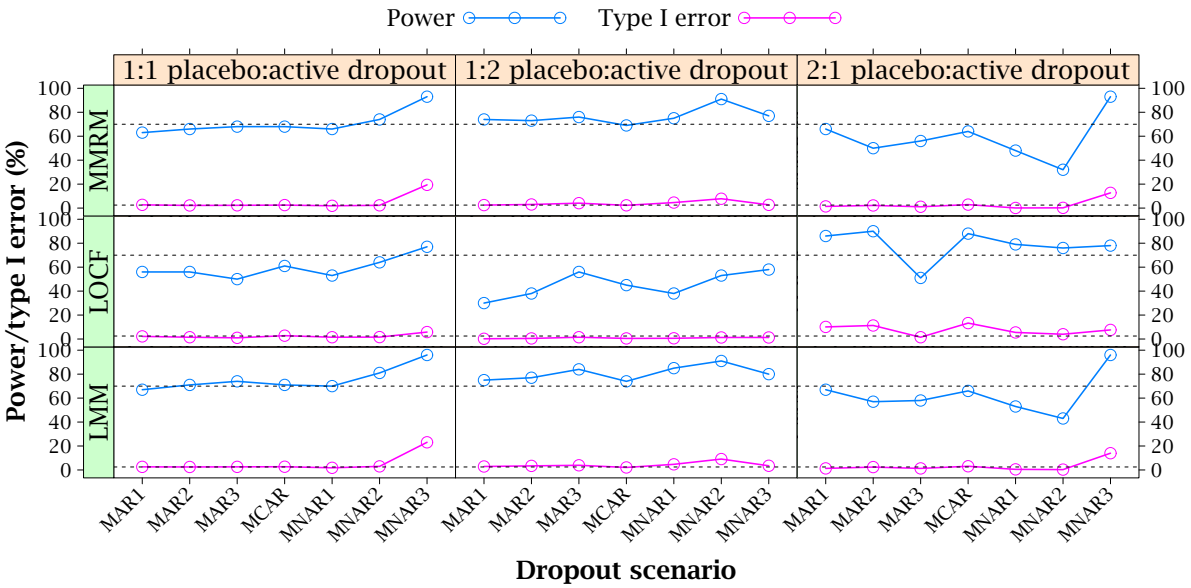


Figure 6. Power and type I error associated with the dropout scenarios, based on study 2. Different panels represent the models (horizontally) and the dropout ratios (vertically)

MNAR scenarios. The power to detect a significant treatment effect was similar for the MMRM and DREM, although differences of up to 11% were observed which were mostly in favour of the DREM. The LOCF had considerably less power to detect a treatment effect in those cases where its bias was conservative.

DISCUSSION

The availability of longitudinal models which describe individual response profiles is essential to ensure accurate estimation of treatment effect. Our results show the relevance of the DREM as statistical analysis method for the evaluation of longitudinal clinical trial data in depression. These findings are particularly important for study design optimisation (clinical trial simulations), as well as for the implementation of adaptive study designs and interim analyses.

Relevance of longitudinal models in depression trials

At present, the analysis of clinical data often relies on flawed statistical methods, which measure treatment as a change relative to placebo or control group at the completion of treatment. Of particular concern is the use of last observation carried forward (LOCF) imputation (Wood *et al.*, 2004). It has been shown that results from any analysis under this imputation method may be biased, depending on the underlying dropout mechanism and whether dropout rates are equal across treatment groups (Molenberghs *et al.*, 2004). The intent-to-treat (ITT) paradigm increases the bias under LOCF because patients with only a single efficacy measurement are also included in the analysis. The application of LOCF is an anachronism, caused by the reluctance of some statisticians and investigators

to use methodologies which are more appropriate from a statistical point of view. Despite numerous examples from the Food and Drug Administration in which alternative statistical methodologies other than LOCF are applied, false beliefs about the conservatism of LOCF in a regulatory context seem to enhance this reluctance to adopt novel statistical approaches. Recent statistical developments, such as the MMRM, have provided the means to deal with missing data. A reduction of the bias minimising dropout in clinical trials (Wisniewski *et al.*, 2006) seems less realistic and is likely to yield inconsistent results.

Based on a functional principal component analysis (chapter 7) a dual-random effects model (DREM) was suggested to analyse longitudinal depression data. As dropout is a major concern in this area, we have also evaluated the robustness of the DREM, MMRM and LOCF models against various dropout scenarios. Of the three models which have been compared, the mixed model for repeated measures is gaining increasing popularity, although LOCF methods are still frequently used. The relatively poor uptake of the MMRM may be partly due to the way such statistical models summarise results. Traditionally, clinical researchers seem to prefer to assess the quality of a statistical analysis by reviewing the fit of the model to the data, a process which is visualised graphically by goodness-of-fit and fits for individual patients as in figure 1. Although the MMRM provides unbiased treatment effect estimates in the presence of MAR, it is not possible to produce goodness-of-fit plots based on individual predictions. This is due to model parameterisation, which prevents prediction of individual observations.

Causes of dropout in depression

Patient dropout may be caused by several factors. First, absence of effect may lead to the decision to quit a trial. Second, occurrence of adverse events or even serious adverse events may have a similar consequence. Third, dropout can be caused by completely trial-unrelated factors, such as patients moving to another town.

The dropout mechanisms that were chosen in our investigation cover a range of situations. In general, dropout is assumed to be higher in those patients who are more severely depressed. This was accommodated for in the 3 scenarios MAR and MNAR. Moreover, placebo dropout may be higher because of interactions between the treating physician and the patient. The first may recognise the absence of characteristic side effects and treatment effect, and inadvertently direct the patient towards quitting the trial. Another possibility is that dropout in the active arm is higher, caused by the additional burden of side effects. However, it is important to realise that any difference in dropout rate between active and placebo arms is observed and subsequent simulation-based investigations can reveal whether these observed differences could lead to biased estimates of treatment effect. If dropout mechanisms are not taken into account, or wrong assumptions are made about its nature, a significant reduction in power or inflated type I error may occur, as shown in figures 4 and 6.

A limitation of the current simulations is that only dropout scenarios depending on

the severity of depression (and a completely random scenario) are investigated. In reality, multiple reasons will cause dropout in any given study, with some patients dropping out because of lack of efficacy (i.e., severely depressed patients) and some dropping out because of adverse events. A realistic clinical trial simulation must therefore consider a combination of dropout mechanisms.

Goodness of fit for longitudinal models

As demonstrated by our analysis, the DREM provides the closest fit to actual patient data, especially for high and low HAMD values. Furthermore, the normalised prediction discrepancy errors (NPDE) show that it is more appropriate to simulate data from this model than from the REM, given that variability is underestimated by the latter. Since one of the frequently used outcome parameters is the percentage of responders (decrease of 50% from baseline HAMD) and the percentage of remitters (HAMD$\leq$7), it is also noteworthy that the percentage of responders is closely resembled by the DREM. The model with a single random effect clearly does not capture this information well. This indicates that the DREM will provide more realistic results than the REM if data analysis of simulated data is based on the percentage responders.

The implementation of the DREM was based on a functional principal component analysis (FPCA), which indicated that such a model may be able to provide a better description of individual patient data. The use of tools provided by this exploratory field may, as shown here, enable the evaluation of more appropriate statistical models. It is important to realise that to be effective, novel statistical methods must account for the variability between patients, which is high and may differ across clinical trials.

LOCF - a statistical blunder

The main finding from the simulations is that LOCF is often biased even under mild dropout scenarios, leading to either inflated type I errors or reduced power, depending on the directionality of the bias. The MMRM/DREM in contrast only show severely inflated type I error under extreme MNAR assumptions, especially when the dropout rate under placebo is higher than under active treatment. The differences between the results of the simulations based on the two studies underline that the risk of increasing type I error increases with higher and unequal dropout rates.

The present simulations show that the assumption that any bias on the part of LOCF is conservative is not justified. Specifically, anti-conservative biases (i.e., favouring active treatment) appear when placebo dropout is higher than the dropout rate under active treatment (figures 3 and 5). An important learning from this exercise is that the direction of the bias is not always predictable, especially under MNAR assumptions. This feature has also been reported elsewhere (Molenberghs *et al.*, 2004; Lane, 2008).

Other authors have used simulations to investigate the bias and type I error of the MMRM and LOCF methods in the presence of various dropout scenarios (Lane, 2008; Mallinckrodt *et al.*, 2001a,b, 2004b). The data in these simulations studies was simulated

base on the MMRM, and not using the DREM, which we have shown to be more appropriate. Furthermore, one author did not investigate the type I error (Lane, 2008) and only showed the power and mean bias, obscuring the actual variability of the bias, whereas others simulated small trials and used parameters that were not based on actual clinical studies (Mallinckrodt *et al.*, 2001b, 2004b). We believe that the methodology used in this paper is more appropriate, as the simulations are based on existing studies and the DREM which provides an adequate description of the data and is more appropriate to simulate from.

A limitation of our approach is that we have not discussed any methods that deal with MNAR patterns. Although methods are available for this problem (Mazumdar *et al.*, 2007), they are not suitable for primary analysis, but may rather provide guidance as to how sensitive results are to the MAR assumptions. We have therefore excluded these methods from our investigations. Yet, they should be used in the context of sensitivity analyses in the analysis of clinical trials (Molenberghs *et al.*, 2004).

Advantages of hierarchical models in a Bayesian framework

The most important advantage of hierarchical models is that individual patient predictions can easily be obtained. Therefore it is easily shown whether or not a model provides a good fit to the data, which may increase the uptake of the model. Also, the availability of individual fits allows for straightforward inclusion of covariates on the subject-level, since correlations between random effects and covariates can be readily explored.

An interesting aspect of the application of the DREM in a Bayesian framework is that the interpretation of the results is straightforward. Rather than a p-value, a Bayesian analysis may be summarised in a posterior probability of inferiority (PPI), i.e., the probability that placebo is more effective than an active treatment. This use of Bayesian methodology is even more interesting because approximation of the denominator degrees of freedom is not required as in the frequentist framework. Lastly, the hierarchical nature of the parameterisation of the DREM is easily accommodated in a Bayesian context. Since future work will encompass the investigation into the feasibility of the application of the model for interim analyses, the availability of the posterior predictive distribution was another important factor in the choice for a Bayesian framework.

A much disputed feature of Bayesian analyses is the obligatory inclusion of prior distributions on each parameter. In the current work, prior distributions have been chosen to be non-informative, such that little information is added to the likelihood of the data. However, especially in the context of interim analyses, it is conceivable that in earlier phases of clinical development informative priors based on historical data are included, allowing prior information to influence the analysis of present studies. In a regulatory context, however, analyses will need to be performed based on data from pivotal trials only.

A disadvantage of the use of MCMC methods is the time it takes to reach convergence and the difficulties that sometimes arise in assessing whether convergence has been

reached. The current model takes approximately two minutes on a standard workstation. Convergence was not an issue in this simple model. We assessed convergence after 8,000 iterations and compared parameter estimates to estimates obtained after 15,000 iterations. Since no difference was observed, 8,000 iterations were considered sufficient and used throughout the simulations.

Conclusions

We strongly advise clinical investigators not to use LOCF analysis for clinical trials in depression. LOCF has either reduced power or an inflated type I error, especially when dropout rates are unequal for active and placebo treatment and total dropout rate is high (as in study 2). The MMRM and DREM control these factors well, as long as the dropout scenario is not extreme. Although there is little difference in bias between the MMRM and the DREM, the latter is preferable because of its ability to describe the data (and especially the variability in the data). In addition, model parameterisation in terms of interindividual variability ensures easier explanation of findings to clinicians and other non-statisticians, who generally make decisions based on statistical analysis.

We have also shown that implementation of the DREM in a Bayesian framework allows straightforward interpretation of the parameters. In future investigations, we will perform clinical trial simulations based on the DREM to explore the relevance of several design factors, and implement the posterior predictive distributions in an interim analysis in the context of adaptive study designs.

REFERENCES

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.

Davis LL, Bartolucci A, and Petty F (2005) Divalproex in the treatment of bipolar depression: A placebo-controlled study. *J Affect Disord* **85**:259–266.

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

Huson LW, Chung J, and Salgo M (2007) Missing data imputation in two phase III trials treating HIV1 infection. *J Biopharm Stat* **17**:159–172.

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.

Kinon BJ, Lipkovich I, Edwards SB, Adams DH, Ascher-Svanum H, and Siris SG (2006) A 24-week randomized study of olanzapine versus ziprasidone in the treatment of schizophrenia or schizoaffective disorder in patients with prominent depressive symptoms. *J Clin Psychopharmacol* **26**:157–162.

Laird NM and Ware JH (1982) Random-effects models for longitudinal data. *Biometrics* **38**:963–974.

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

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, Raskin J, Wohlreich M, Watkin J, and Detke M (2004a) The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF-ANCOVA in eight clinical trials. *BMC Psychiatry* **4**:26–.

Mallinckrodt CH (2006) The test of public scrutiny. *Pharm Stat* **5**:249–252.

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.

Mazumdar S, Liu KS, Houck PR, and Reynolds CF (1999) Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values. *J Psychiatr Res* **33**:87–95.

Mazumdar S, Tang G, Houck PR, Dew MA, Begley AE, Scott J, Mulsant BH, and Reynolds CF (2007) Statistical analysis of longitudinal psychiatric data with dropouts. *J Psychiatr Res* **41**:1032–1041.

Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, and Carroll RJ (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**:445–464.

Molenberghs G and Verbeke G (2004) Meaningful statistical model formulations for repeated measures. *Stat Sin* **14**:989–1020.

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.

Thase ME, Macfadden W, Weisler RH, Chang W, Paulsson B, Khan A, and Calabrese JR (2006) Efficacy of quetiapine monotherapy in bipolar I and II depression - a double-blind, placebo-controlled study (the BOLDER II study). *J Clin Psychopharmacol* **26**:600–609.

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.

Verbeke G and Molenberghs G (2000) *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New-York.

Wisniewski SR, Leon AC, Otto MW, and Trivedi MH (2006) Prevention of missing data in clinical research studies. *Biol Psychiatry* **59**:997–1000.

Wood A, White I, and Thompson S (2004) Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clin Trials* **1**:368–376.

Wood AM, White IR, Hillsdon M, and Carpenter J (2005) Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *Int J Epidemiol* **34**:89–99.