Biostatistics (2004), **5**, 3, *pp.* 445–464
doi: 10.1093/biostatistics/kxh001

Analyzing incomplete longitudinal clinical trial data

GEERT MOLENBERGHS*, HERBERT THIJS, IVY JANSEN, CAROLINE BEUNCKENS
Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium
geert.molenberghs@luc.ac.be

MICHAEL G. KENWARD
London School of Hygiene and Tropical Medicine, London, UK

CRAIG MALLINCKRODT
Eli Lilly and Company, Indianapolis, IN, USA

RAYMOND J. CARROLL
Department of Statistics, Texas A&M University, College Station, TX, USA

SUMMARY

Using standard missing data taxonomy, due to Rubin and co-workers, and simple algebraic derivations, it is argued that some simple but commonly used methods to handle incomplete longitudinal clinical trial data, such as complete case analyses and methods based on last observation carried forward, require restrictive assumptions and stand on a weaker theoretical foundation than likelihood-based methods developed under the missing at random (MAR) framework. Given the availability of flexible software for analyzing longitudinal sequences of unequal length, implementation of likelihood-based MAR analyses is not limited by computational considerations. While such analyses are valid under the comparatively weak assumption of MAR, the possibility of data missing not at random (MNAR) is difficult to rule out. It is argued, however, that MNAR analyses are, themselves, surrounded with problems and therefore, rather than ignoring MNAR analyses altogether or blindly shifting to them, their optimal place is within sensitivity analysis. The concepts developed here are illustrated using data from three clinical trials, where it is shown that the analysis method may have an impact on the conclusions of the study.

Keywords: Complete case analysis; Ignorability; Last observation carried forward; Missing at random; Missing completely at random; Missing not at random.

1. INTRODUCTION

In a longitudinal clinical trial, each unit is measured on several occasions. It is not unusual in practice for some sequences of measurements to terminate early for reasons outside the control of the investigator, and any unit so affected is called a dropout. It might therefore be necessary to accommodate dropout in the modeling process.

*To whom corespondence should be addressed.

Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Afifi and Elashoff, 1966; Hartley and Hocking, 1971). More recently, general algorithms such as expectation-maximization (EM) (Dempster *et al.*, 1977), and data imputation and augmentation procedures (Rubin, 1987), combined with powerful computing resources have largely solved the computational difficulties. There remains the difficult and important question of assessing the impact of missing data on subsequent statistical inference.

When referring to the missing-value, or non-response, process we will use terminology of Little and Rubin (1987, Chapter 6). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable*, while a non-random process is non-ignorable.

Numerous missing data methods are formulated as selection models (Little and Rubin, 1987) as opposed to pattern-mixture modeling (PMM; Little, 1993, 1994). A selection model factors the joint distribution of the measurement and response mechanisms into the marginal measurement distribution and the response distribution, conditional on the measurements. This is intuitively appealing because the marginal measurement distribution would be of interest with complete data. Little and Rubin's taxonomy is most easily developed in the selection model setting. Parametrizing and making inference about treatment effects and their evolution over time is straightforward in the selection model context.

In many clinical trial settings, the standard methodology used to analyze incomplete longitudinal data is based on such methods as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation. This is often done without questioning the possible influence of these assumptions on the final results, even though several authors have written about this topic. A relatively early account is given in Heyting *et al.* (1992). Mallinckrodt *et al.* (2003a,b) and Lavori *et al.* (1995) propose direct-likelihood and multiple-imputation methods, respectively, to deal with incomplete longitudinal data. Siddiqui and Ali (1998) compare direct-likelihood and LOCF methods.

As will be discussed in subsequent sections, it is unfortunate that such a strong emphasis is placed on methods like LOCF and CC in clinical trial settings, since they are based on strong and unrealistic assumptions. Even the strong MCAR assumption does not suffice to guarantee that an LOCF analysis is valid. In contrast, under the less restrictive assumption of MAR, valid inference can be obtained through a likelihood-based analysis without modeling the dropout process. One can then use linear or generalized linear mixed models (Verbeke and Molenberghs, 2000), without additional complication or effort. We will argue that such an analysis is more likely to be valid, and even easier to implement than LOCF and CC analyses.

Nevertheless, approaches based on MNAR need to be considered. In practical settings, the reasons for dropout are varied and it may therefore be difficult to justify the assumption of MAR. For example, in 11 clinical trials of similar design, considered by Mallinckrodt *et al.* (2003b), with the same drug and involving patients with the same disease state, the rate of and the reasons for dropout varied considerably. In one study, completion rates were 80% for drug and placebo. In another study, two-thirds of the patients on drug completed all visits, while only one-third did so on placebo. In yet another study, 70% finished on placebo but only 60% on drug. Reasons for dropout also varied, even within the drug arm. For example, at low doses more patients on drug dropped out due to lack of efficacy whereas at higher doses dropout due to adverse events was more common. At first sight, this calls for a further shift towards MNAR models. However, caution ought to be used since no modeling approach, whether MAR or MNAR, can recover the lack of information due to incompleteness of the data.

Table 1. *Overview of number of patients and post baseline visits per study*

	Number of patients	Post-baseline visits
Study 1	167	4–11
Study 2	342	4–8
Study 3	713	3–8

First, if MAR can be guaranteed to hold, a standard analysis would follow. However, only rarely is such an assumption known to hold (Murray and Findlay, 1988). Nevertheless, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data (Mallinckrodt *et al.*, 2001a,b). A discussion of this phenomenon in the survey context has been given in Rubin *et al.* (1995). These authors argue that, in rigidly controlled experiments (some surveys and many clinical trials), the assumption of MAR is often reasonable. Second, and very importantly for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with complete data. Third, while MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on untestable and often implicit assumptions regarding the distribution of the unobserved measurements given the observed measurements. The quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. This point is irrespective of the MNAR route taken, whether a parametric model of the type of Diggle and Kenward (1994) is chosen, or a semiparametric approach such as in Robins *et al.* (1998). Hence, in incomplete-data settings, a definitive MNAR analysis does not exist. We therefore argue that clinical trial practice should shift away from the *ad hoc* methods and focus on likelihood-based ignorable analyses instead. The cost involved in having to specify a model will likely be small to moderate in realistic clinical trial settings. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis, within which MNAR models and pattern-mixture models can play a major role (Verbeke and Molenberghs, 2000, Chapter 18–20).

A three-trial case study is introduced in Section 2. The general data setting is introduced in Section 3, as well as a formal framework for incomplete longitudinal data. A discussion on the problems associated with simple methods is presented in Section 4. In Section 5, using algebraic derivations, we explore the origins of the asymptotic bias in LOCF, complete-case and likelihood-based ignorable analyses. The case study is analyzed in Section 6. A perspective on sensitivity analysis is sketched in Section 7.

2. Case studies

The ideas developed in this paper are motivated from, and applied to, data from three clinical trials of anti-depressants. The three trials contained 167, 342, and 713 patients with post-baseline data, respectively (Mallinckrodt *et al.*, 2003b). The Hamilton Depression Rating Scale ($HAMD_{17}$) was used to measure the depression status of the patients. For each patient, a baseline assessment was available. Post-baseline visits differ by study (Table 1).

For blinding purposes, therapies are recoded as A1 for primary dose of experimental drug, A2 for secondary dose of experimental drug, and B and C for non-experimental drugs. The treatment arms across the three studies are as follows: A1, B, and C for study 1; A1, A2, B, and C for study 2; A1 and B for study 3. The primary contrast is between A1 and C for studies 1 and 2, whereas in study 3 one is interested in A *versus* B.

In this case study, emphasis is on the difference between the treatment arms in mean change of the

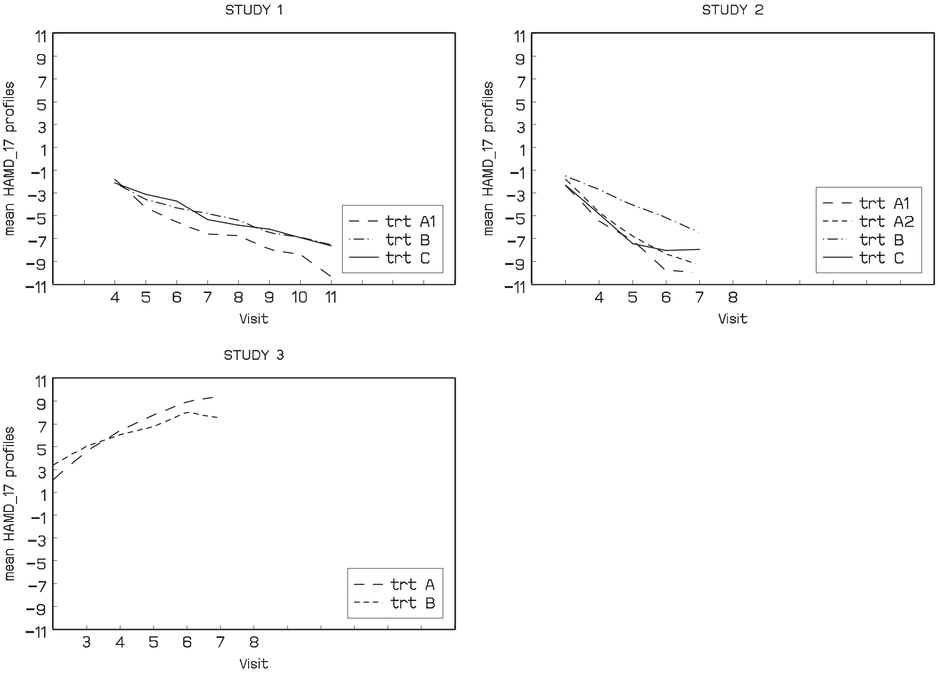

Fig. 1. Mean profiles for each of the three studies.

$HAMD_{17}$ score at the endpoint. For each study, mean profiles within each treatment arm are given in Figure 1. However, as time evolves, more and more patients drop out, resulting in fewer observations for later visits. Indeed, a graphical representation of dropout, per study and per arm, is given in Figure 2. Due to this fact, Figure 1 might be misleading if interpreted without acknowledging the diminishing basis of inference.

3. DATA SETTING AND MODELING FRAMEWORK

Assume that for subject $i = 1, \ldots, N$ in the study a sequence of responses Y_{ij} is designed to be measured at occasions $j = 1, \ldots, n$. The outcomes are grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$. In addition, define a dropout indicator D_i for the occasion at which dropout occurs and make the convention that $D_i = n + 1$ for a complete sequence. It is often necessary to split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components respectively.

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured, but have been suppressed from notation for simplicity.

Most strategies used to analyze such data are, implicitly or explicitly, based on two choices.

Model for measurements. A choice has to be made regarding the modeling approach to the measurements. Several views are possible.

View 1. One can choose to analyze the entire longitudinal profile, irrespective of whether interest focuses on the entire profile (e.g. difference in slope between groups) or on a specific time point (e.g.

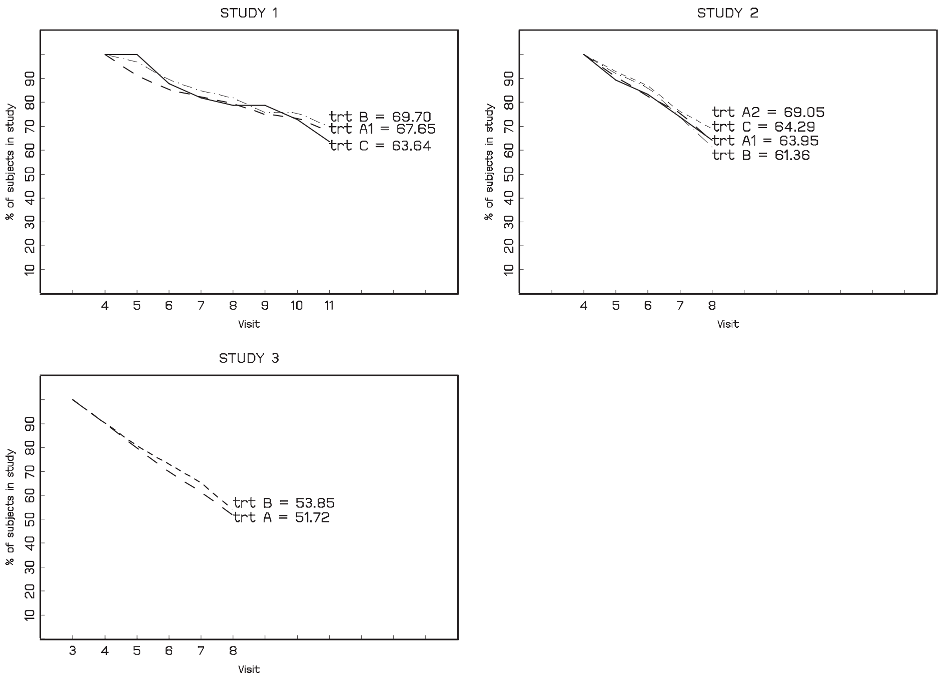

Fig. 2. Evolution of dropout per study and per treatment arm. Treatment arms of primary interest, are shown in bolder typeface.

the last planned occasion). In the latter case, one would make inferences about such an occasion using the posited model.

View 2. One states the scientific question in terms of the outcome at a well-defined point in time. Several choices are possible:

View 2a. The scientific question is defined in terms of the *last planned occasion*. In this case, one can either accept the dropout as it is or use one or other strategy (e.g. imputation) to incorporate the missing outcomes.

View 2b. One can choose to define the question and the corresponding analysis in terms of the *last observed measurement*.

While Views 1 and 2a necessitate reflection on the missing data mechanism, View 2b avoids the missing data problem because the question is couched completely in terms of observed measurements. Thus, under View 2b, an LOCF analysis might be acceptable, provided it matched the scientific goals, but is then better described as a Last Observation analysis because nothing is carried forward. Such an analysis should properly be combined with an analysis of time to dropout, perhaps in a survival analysis framework. Of course, an investigator should reflect very carefully on whether View 2b represents a relevant and meaningful scientific question (see also Shih and Quan, 1997).

Method for handling missingness. A choice has to be made regarding the modeling approach for the missingness process. Under certain assumptions this process can be ignored (e.g. a likelihood-based ignorable analysis). Some simple methods, such as a complete case analysis and LOCF, do not explicitly address the missingness process either.

We first describe the measurement and missingness models in turn, then formally introduce and comment on ignorability.

The measurement model will depend on whether or not a full longitudinal analysis is done. When the focus is on the last observed measurement or on the last measurement occasion only, one typically opts for classical two- or multi-group comparisons (t test, Wilcoxon, etc.). When a longitudinal analysis is deemed necessary, the choice depends on the nature of the outcome. For continuous outcomes, such as in our case studies, one typically assumes a linear mixed-effects model, perhaps with serial correlation:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{W}_i + \varepsilon_i, \tag{3.1}$$

(Verbeke and Molenberghs, 2000) where $\boldsymbol{Y}_i$ is the n-dimensional response vector for subject i, $1 \leqslant i \leqslant N$,$N$ is the number of subjects, X_i and Z_i are $(n \times p)$ and $(n \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p-dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$ is the q-dimensional vector containing the random effects, $\varepsilon_i \sim N(\boldsymbol{0}, \sigma^2 I_{n_i})$ is a n-dimensional vector of measurement error components, and $\boldsymbol{b_1}, \dots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \dots, \boldsymbol{\varepsilon_N}$ are assumed to be independent. Serial correlation is captured by the realization of a Gaussian stochastic process, $\boldsymbol{W}_i$, which is assumed to follow a $N(\boldsymbol{0}, \tau^2 H_i)$ law. The serial covariance matrix H_i only depends on i through the number n of observations and through the time points t_{ij} at which measurements are taken. The structure of the matrix H_i is determined through the autocorrelation function $\rho(t_{ij} - t_{ik})$. This function decreases such that $\rho(0) = 1$ and $\rho(u) \to 0$ as $u \to \infty$. Finally, D is a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$. Inference is based on the marginal distribution of the response $\boldsymbol{Y_i}$ which, after integrating over random effects, can be expressed as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_i D Z_i' + \Sigma_i). \tag{3.2}$$

Here, $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n \times n)$ covariance matrix combining the measurement error and serial components.

Assume that incompleteness is due to dropout only, and that the first measurement Y_{i1} is obtained for everyone. A possible model for the dropout process is a logistic regression for the probability of dropout at occasion j, given that the subject is still in the study. We denote this probability by $g(\boldsymbol{h}_{ij}, y_{ij})$ in which $\boldsymbol{h}_{ij}$ is a vector containing all responses observed up to but not including occasion j, as well as relevant covariates. We then assume that $g(\boldsymbol{h}_{ij}, y_{ij})$ satisfies

$$\operatorname{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = \operatorname{logit}\left[\operatorname{pr}(D_i = j | D_i \geqslant j, \boldsymbol{y}_i)\right] = \boldsymbol{h}_{ij}\boldsymbol{\psi} + \omega y_{ij}, \qquad i = 1, \dots, N, \tag{3.3}$$

(Diggle and Kenward, 1994). When ω equals zero, the dropout model is MAR, and all parameters can be estimated using standard software since the measurement model, for which we use a linear mixed model, and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If $\omega \neq 0$, the posited dropout process is MNAR. Model (3.3) provides the building blocks for the dropout process $f(d_i | \boldsymbol{y}_i, \boldsymbol{\psi})$.

Rubin (1976) and Little and Rubin (1987) have shown that, under MAR and the condition that parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are functionally independent, likelihood-based inference remains valid when the missing data mechanism is ignored (see also Verbeke and Molenberghs, 2000). Practically speaking, the likelihood of interest is then based upon the factor $f(\boldsymbol{y}_i^o | \boldsymbol{\theta})$. This is called *ignorability*. The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects, manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values. Note that the estimands are the parameters of (3.2), which is a model for complete data, corresponding to what one would expect to see in the absence of dropouts.

A few cautionary remarks are warranted. First, when at least part of the scientific interest is directed towards the nonresponse process, obviously both processes need to be considered. Under MAR, both processes can be modeled and parameters estimated separately. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g. to construct measures of precision for estimators and for statistical hypothesis tests; Kenward and Molenberghs, 1998). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct, are valid. Thirdly, it may be hard to rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Section 7. Fourthly, such an analysis can proceed only under View 1, i.e. a full longitudinal analysis is necessary, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion. In the latter case, the fitted model can be used as the basis for inference at the last occasion. A common criticism is that a model needs to be considered, with the risk of model misspecification. However, it should be noted that in many clinical trial settings the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, allowing the a priori specification of a saturated model (e.g. full group by time interaction model for the fixed effects and unstructured variance–covariance matrix). Such an ignorable linear mixed model specification is termed MMRM (mixed-model random missingness) by Mallinckrodt *et al.* (2001a,b). Thus, MMRM is a particular form of a linear mixed model, fitting within the ignorable likelihood paradigm. Such an approach is a promising alternative to the often used simple methods such as complete-case analysis or LOCF. These will be described in the next section and further studied in subsequent sections.

4. Simple methods

We will briefly review a number of relatively simple methods that still are commonly used. For the validity of many of these methods, MCAR is required. For others, such as LOCF, MCAR is necessary but not sufficient. The focus will be on the complete case method, for which data are removed, and on imputation strategies, where data are filled in. Regarding imputation, one distinguishes between single and multiple imputation. In the first case, a single value is substituted for every ‘hole’ in the data set and the resulting data set is analyzed as if it represented the true complete data. Multiple imputation acknowledges the uncertainty stemming from filling in missing values rather than observing them (Rubin, 1987; Schafer, 1997). LOCF will be discussed within the context of imputation strategies, although LOCF can be placed in other frameworks as well.

A *complete case analysis* includes only those cases for which all measurements were recorded. This method has obvious advantages. It is simple to describe and almost any software can be used since there are no missing data. Unfortunately, the method suffers from severe drawbacks. Firstly, there is nearly always a substantial loss of information. For example, suppose there are 20 measurements, with 10% of missing data on each measurement. Suppose, further, that missingness on the different measurements is independent; then, the estimated percentage of incomplete observations is as high as 87%. The impact on precision and power may be dramatic. Even though the reduction of the number of complete cases will be less severe in settings where the missingness indicators are correlated, this loss of information will usually militate against a CC analysis. Secondly, severe bias can result when the missingness mechanism is MAR but not MCAR. Indeed, should an estimator be consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR. A CC analysis can be conducted when Views 1 and 2 of Section 3 are adopted. It obviously is not a reasonable choice with View 2b.

An alternative way to obtain a data set on which complete data methods can be used is to fill in rather than delete (Little and Rubin, 1987). Concern has been raised regarding imputation strategies. Dempster and Rubin (1983) write: ‘The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.’ For example, Little and Rubin (1987) show that the application of imputation could be considered acceptable in a linear model with one fixed effect and one error term, but that it is generally not acceptable for hierarchical models, split-plot designs, repeated measures with a complicated error structure, random-effects, and mixed-effects models.

Thus, the user of imputation strategies faces several dangers. First, the imputation model could be wrong and, hence, the point estimates biased. Second, even for a correct imputation model, the uncertainty resulting from missingness is ignored. Indeed, even when one is reasonably sure about the mean value the unknown observation *would have had*, the actual stochastic realization, depending on both the mean and error structures, is still unknown. In addition, most methods require the MCAR assumption to hold while some even require additional and often unrealistically strong assumptions.

A method that has received considerable attention (Siddiqui and Ali, 1998; Mallinckrodt *et al.*, 2003a,b) is *last observation carried forward* (LOCF). In the LOCF method, whenever a value is missing, the last observed value is substituted. The technique can be applied to both monotone and nonmonotonic missing data. It is typically applied in settings where incompleteness is due to attrition.

LOCF can, but should not necessarily, be regarded as an imputation strategy, depending on which of the views of Section 3 is taken. The choice of viewpoint has a number of consequences. First, when the problem is approached from a missing data standpoint, one has to think it plausible that subjects’ measurements do not change from the moment of dropout onwards (or during the period they are unobserved in the case of intermittent missingness). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Secondly, LOCF shares with other single imputation methods that it artificially increases the amount of information in the data, by treating imputed and actually observed values on an equal footing. This is especially true if a longitudinal view is taken. Verbeke and Molenberghs (1997, Chapter 5) have shown that all features of a linear mixed model (group difference, evolution over time, variance structure, correlation structure, random effects structure, ...) can be affected. A similar conclusion, based on the case study, is reached in Section 6.

Thus, scientific questions with which LOCF is compatible will be those that are phrased in terms of the last obtained measurement (View 2b). Whether or not such questions are sensible should be the subject of scientific debate, which is quite different from a *post hoc* rationale behind the use of LOCF. Likewise, it can be of interest to model the complete cases separately and to make inferences about them. In such cases, a CC analysis is of course the only reasonable way forward. This is fundamentally different from treating a CC analysis as one that can answer questions about the randomized population as a whole.

We will briefly describe two other imputation methods. The idea behind *unconditional mean imputation* (Little and Rubin, 1987) is to replace a missing value with the average of the observed values on the same variable over the other subjects. Thus, the term *unconditional* refers to the fact that one does not use (i.e. condition on) information on the subject for which an imputation is generated. Since values are imputed that are unrelated to a subject’s other measurements, all aspects of a model, such as a linear mixed model, are typically distorted (Verbeke and Molenberghs, 1997). In this sense, unconditional mean imputation can be as damaging as LOCF.

Buck’s method or *conditional mean imputation* (Buck, 1960; Little and Rubin, 1987) is similar in complexity to mean imputation. Consider, for example, a single multivariate normal sample. The first step is to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ from the complete cases, assuming that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a subject with missing components, the regression of the missing components ($\boldsymbol{Y}_i^m$)

on the observed ones ($\boldsymbol{y}_i^o$) is

$$\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o \sim N(\boldsymbol{\mu}^m + \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o), \boldsymbol{\Sigma}^{mm} - \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om}). \tag{4.1}$$

The second step calculates the conditional mean from the regression of the missing components on the observed components, and substitutes the conditional mean for the corresponding missing values. In this way, ‘vertical’ information (estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) is combined with ‘horizontal’ information ($\boldsymbol{y}_i^o$). Buck (1960) showed that under mild conditions, the method is valid under MCAR mechanisms. Little and Rubin (1987) added that the method is also valid under certain MAR mechanisms. Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant across missingness patterns. Again, this method shares with other single imputation strategies that, although point estimation may be consistent, the precision will be overestimated. There is a connection between *the concept* of conditional mean imputation and a likelihood-based ignorable analysis, in the sense that the latter analysis produces expectations for the missing observations that are formally equal to those obtained under a conditional mean imputation. However, in likelihood-based ignorable analyses, no explicit imputation takes place, hence the amount of information in the data is not overestimated and important model elements, such as mean structure and variance components, are not distorted.

Historically, an important motivation behind the simpler methods was their simplicity. Currently, with the availability of commercial software tools such as, for example, the SAS procedures MIXED and NLMIXED and the SPlus and R nlme libraries, this motivation no longer applies. Arguably, a MAR analysis is the preferred choice. Of course, the correctness of a MAR analysis rests upon the truth of the MAR assumption, which is, in turn, never completely verifiable. Purely resorting to MNAR analyses is not satisfactory either since important sensitivity issues then arise. These and related issues are briefly discussed in the next section (see also Verbeke and Molenberghs, 2000).

It is often quoted that LOCF or CC, while problematic for parameter estimation, produce randomization-valid hypothesis testing, but this is questionable. First, in a CC analysis partially observed data are selected out, with probabilities that that may depend on post-randomization outcomes, thereby undermining any randomization justification. Second, if the focus is on one particular time point, e.g. the last one scheduled, then LOCF plugs in data. Such imputations, apart from artificially inflating the information content, may deviate in complicated ways from the underlying data (see next section). In contrast, a likelihood-based MAR analysis uses all available data, with the need for neither deletion nor imputation, which suggests that a likelihood-based MAR analysis would usually be the preferred one for testing as well. Third, although the size of a randomization based LOCF test may reach its nominal size under the null hypothesis of no difference in treatment profiles, there will be other regions of the alternative space where the power of the LOCF test procedure is equal to its size, which is completely unacceptable.

5. Bias in LOCF, CC, and ignorable likelihood methods

Using the simple but insightful setting of two repeated follow-up measures, the first of which is always observed while the second can be missing, we establish some properties of the LOCF and CC estimation procedures under different missing data mechanisms, against the background of a MAR process operating. In this way, we bring LOCF and CC within a general framework that makes clear their relationships with more formal modeling approaches, enabling us to make a coherent comparison among the different approaches. The use of a moderate amount of algebra leads to some interesting conclusions.

Let us assume each subject i is to be measured on two occasions $t_i = 0, 1$. Subjects are randomized to one of two treatment arms: $T_i = 0$ for the standard arm and 1 for the experimental arm. The probability of an observation being observed on the second occasion ($D_i = 2$) is p_0 and p_1 for treatment groups 0 and 1, respectively. We can write the means of the observations in the two dropout groups as follows:

$$\text{dropouts } D_i = 1 : \beta_0 + \beta_1 T_i + \beta_2 t_i + \beta_3 T_i t_i, \tag{5.1}$$

$$\text{completers } D_i = 2 : \gamma_0 + \gamma_1 T_i + \gamma_2 t_i + \gamma_3 T_i t_i. \tag{5.2}$$

The true underlying population treatment difference at time $t_i = 1$, as determined from (5.1)–(5.2), is equal to

$$\begin{aligned}\Delta_{\text{true}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1 + \beta_2 + \beta_3) \\ & -[p_0(\gamma_0 + \gamma_2) + (1 - p_0)(\beta_0 + \beta_2)].\end{aligned} \tag{5.3}$$

If we use LOCF as the estimation procedure, the expectation of the corresponding estimator equals

$$\begin{aligned}\Delta_{\text{LOCF}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1) \\ & -[p_0(\gamma_0 + \gamma_2) + (1 - p_0)\beta_0].\end{aligned} \tag{5.4}$$

Alternatively, if we use CC, the above expression changes to

$$\Delta_{\text{CC}} = \gamma_1 + \gamma_3. \tag{5.5}$$

In general, these are both biased estimators.

We will now consider the special but important cases where the true missing data mechanisms are MCAR and MAR, respectively. Each of these will impose particular constraints on the β and γ parameters in model (5.1)–(5.2). Under MCAR, the β parameters are equal to their γ counterparts and (5.3) simplifies to

$$\Delta_{\text{MCAR,true}} = \beta_1 + \beta_3 \equiv \gamma_1 + \gamma_3. \tag{5.6}$$

Suppose we apply the LOCF procedure in this setting, the expectation of the resulting estimator then simplifies to

$$\Delta_{\text{MCAR,LOCF}} = \beta_1 + (p_1 - p_0)\beta_2 + p_1\beta_3. \tag{5.7}$$

The bias is given by the difference between (5.6) and (5.7):

$$B_{\text{MCAR,LOCF}} = (p_1 - p_0)\beta_2 - (1 - p_1)\beta_3. \tag{5.8}$$

While of a simple form, we can learn several things from this expression by focusing on each of the terms in turn. First, suppose $\beta_3 = 0$ and $\beta_2 \neq 0$, implying that there is no differential treatment effect between the two measurement occasions but there is an overall time trend. Then, the bias can go in either direction depending on the sign of $p_1 - p_0$ and the sign of β_2. Note that $p_1 = p_0$ only in the special case that the dropout rate is the same in both treatment arms. Whether or not this is the case has no impact on the status of the dropout mechanism (it is MCAR in either case, even though in the second case dropout is treatment-arm dependent), but is potentially very important for the bias implied by LOCF. Second, suppose $\beta_3 \neq 0$ and $\beta_2 = 0$. Again, the bias can go in either direction depending on the sign of β_3, i.e. depending on whether the treatment effect at the second occasion is larger or smaller than the treatment effect at the first occasion. In conclusion, even under the strong assumption of MCAR, we see that the bias in the LOCF estimator typically does not vanish and, even more importantly, the bias can be positive or negative and can even induce an apparent treatment effect when one does not exist.

In contrast, as can be seen from (5.5) and (5.6), the CC analysis is unbiased.

Let us now turn to the MAR case. In this setting, the constraint implied by the MAR structure of the dropout mechanism is that the conditional distribution of the second observation given the first is the same in both dropout groups (Molenberghs *et al.*, 1998). Based on this result, the expectation of the second observation in the standard arm of the dropout group is

$$E(Y_{i2}|D_i = 1, T_i = 0) = \gamma_0 + \gamma_2 + \sigma(\beta_0 - \gamma_0) \tag{5.9}$$

where $\sigma = \sigma_{21}\sigma_{11}^{-1}$, σ_{11} is the variance of the first observation in the fully observed group and σ_{12} is the corresponding covariance between the pair of observations. Similarly, in the experimental group we obtain

$$E(Y_{i2}|D_i = 1, T_i = 1) = \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 + \sigma(\beta_0 + \beta_1 - \gamma_0 - \gamma_1). \tag{5.10}$$

The true underlying population treatment difference (5.3) then becomes

$$\Delta_{\text{MAR,true}} = \gamma_1 + \gamma_3 + \sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \tag{5.11}$$

In this case, the bias in the LOCF estimator can be written as

$$\begin{aligned} B_{\text{MAR,LOCF}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1) \\ & - p_0(\gamma_0 + \gamma_2) - (1 - p_0)\beta_0 - \gamma_1 - \gamma_3 \\ & - \sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \end{aligned} \tag{5.12}$$

Again, although involving more complicated relationships, it is clear that the bias can go in either direction, thus contradicting the claim often put forward that the bias in LOCF leads to conservative conclusions. Further, it is far from clear what conditions need to be imposed in this setting for the corresponding estimator to be either unbiased or conservative.

The bias in the CC estimator case takes the form

$$B_{\text{MAR,CC}} = -\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \tag{5.13}$$

Even though this expression is simpler than in the LOCF case, it is still true that the bias can operate in either direction.

Thus, in all cases, LOCF typically produces bias of which the direction and magnitude depend on the true but unknown treatment effects. Hence, caution is needed when using this method. In contrast, an ignorable likelihood based analysis, as outlined in Section 4, provides a consistent estimator of the true treatment difference at the second occasion under both MCAR and MAR. While this is an assumption, it is rather a mild one in contrast to the stringent conditions required to justify the LOCF method, even when the qualitative features of the bias are considered more important than the quantitative ones. Note that the LOCF method is not valid even under the strong MCAR condition, whereas the CC approach is valid under MCAR.

6. Analysis of case studies

We now analyze the three clinical trials, introduced in Section 2. The primary null hypothesis (zero difference between the treatment and placebo in mean change of the HAMD17 total score at endpoint) is tested using a model of the type (3.1). The model includes the fixed categorical effects of treatment, investigator, time, and treatment by time interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-time interaction. In line with the protocol design, we use the heterogeneous compound symmetric covariance structure. Satterthwaite's approximation will be used to estimate denominator degrees of freedom. The significance of differences in least-square means is based on Type III tests. These examine the significance of each partial effect, that is, the significance of an effect with all the other effects in the model. Analyses are implemented using the SAS procedure MIXED.

Given this description, the effect of simple approaches, such as LOCF and CC, *versus* MAR, can be studied in terms of their impact on various linear mixed model aspects (fixed effects, variance structure, correlation structure). It will be shown that the impact of the simplifications can be noticeable. This is the subject of Section 6.1, dedicated to View 1. Section 6.2 focuses on Views 2a and 2b, where the last planned occasion and the last measurement obtained are of interest, respectively. In addition, we consider the issues arising when switching from a two-treatment arm to an all-treatment arm comparison.

6.1 *View 1: longitudinal analysis*

For each study in this longitudinal analysis, we will only consider the treatments that are of direct interest. This means we estimate the main difference between these treatments (treatment main effect) as well as the difference between both over time (treatment by time interaction). Treatment main effect estimates and standard errors, p values for treatment main effect and treatment by time interaction, and estimates for the within-patient correlation are reported in Table 2. When comparing LOCF, CC, and MAR, there is little difference between the three methods, in either the treatment main effect or the treatment by time interaction. Nevertheless, some important differences will be established between the strategies in terms of other model aspects. These will be seen to be in line with the reports in Verbeke and Molenberghs (1997, 2000).

Two specific features of the mean structure are the time trends and the treatment effects (over time). We discuss these in turn. The placebo time trends as well as the treatment effects (i.e. differences between the active arms and the placebo arms) are displayed in Figure 3. Both LOCF and CC are different from MAR, with a larger difference for CC. The effect is strongest in the third study. It is striking that different studies lead to different conclusions in terms of relative differences between the approaches. While there is a relatively small difference between the three methods in Study 2 and a mild one for Study 1, for Study 3 there is a strong separation between LOCF and CC on the one hand, and MAR on the other hand. Importantly, the *average* effect is smaller for MAR than for LOCF and CC. This result is in agreement with the proofs in Section 5, which showed that the direction of the bias on LOCF is in fact hard to anticipate.

The variance–covariance structure employed is heterogeneous compound symmetry (CSH), i.e. a common correlation and a variance specific to each measurement occasion. The latter feature allows us to plot the fitted variance function over time. This is done in Figure 4. It is very noticeable that MAR and CC produce a relatively similar variance structure, which tends to rise only mildly. LOCF on the other hand, deviates from both and points towards a (linear) increase in variance. If further modeling is done, MAR and CC produce homogeneous or classical compound symmetry (CS) and hence a random-intercept structure. LOCF on the other hand, suggests a random-slope model. The reason for this discrepancy is that an incomplete profile is completed by means of a flat profile. Within a pool of linearly increasing or decreasing profiles, this leads to a progressively wider spread as study time elapses. Noting that the fitted variance function has implications for the computation of mean-model standard errors, the potential for misleading inferences is clear.

The fitted correlations are given in Table 2. Clearly, CC and MAR produce virtually the same correlation. However, the correlation coefficient estimated under LOCF is much stronger. This is entirely due to the fact that after dropout, a constant value is imputed for the remainder of the study period, thereby

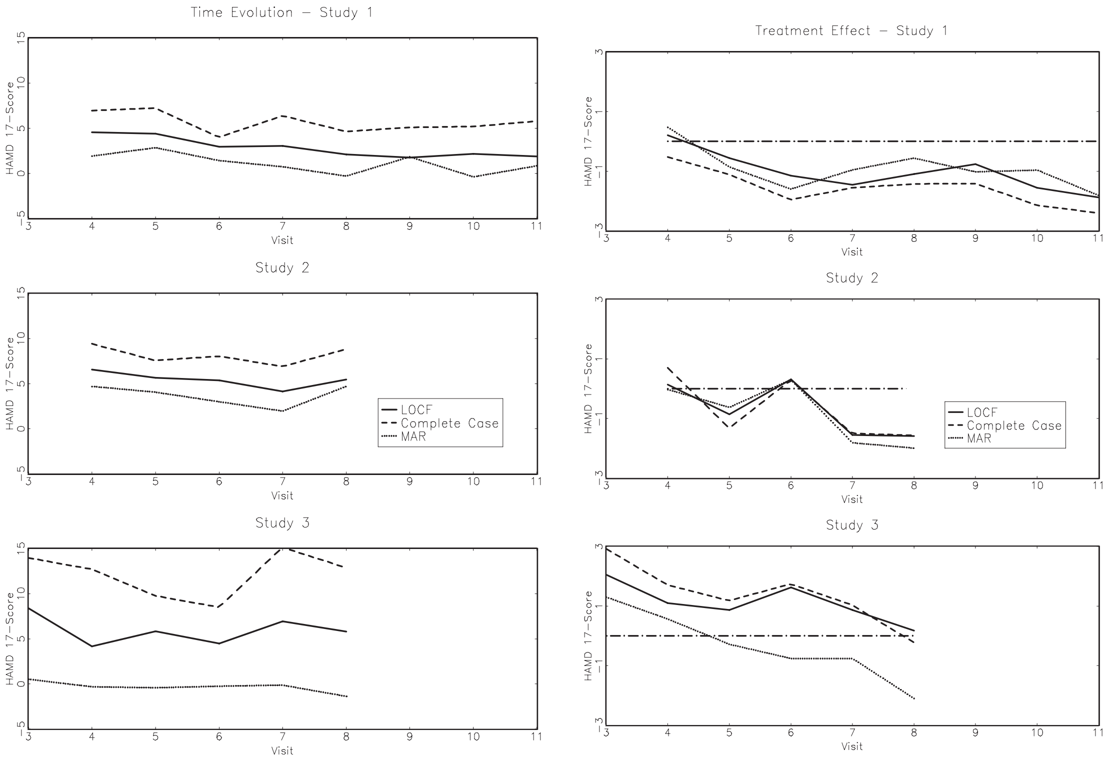

Fig. 3. Summary of all placebo time evolutions (left hand panels) and all treatment effects (right hand panels).

increasing the correlation between the repeated measurements. Of course, the problem is even more severe than shows from this analysis since, under LOCF, a constant correlation structure can be changed into one which progressively strengthens as time elapses. It should be noted that the correlation structure has an impact on all longitudinal aspects of the mean structure. For example, estimates and standard errors of time trends and estimated interactions of time with covariates can all be affected. In particular, if the estimated correlation is too high, the time trend can be ascribed a precision which is too high, implying the potential for a *liberal* error.

In conclusion, all aspects of the linear mixed models (mean structure, variance structure, correlation structure) may be influenced by the method of analysis. This is in line with results reported in Verbeke and Molenberghs (1997, 2000). It is important to note that, generally, the direction of the errors (conservative or liberal) is not clear *a priori*, since different distortions (in mean, variance, or correlation structure) may counteract each other. We will now study a number of additional analyses that are extremely relevant from a clinical trial point of view.

6.2 *Views 2a and 2b and all-* versus *two-treatment arms*

When emphasis is on the last measurement occasion, LOCF and CC are straightforward to use. When the last observed measurement is of interest, the corresponding analysis is not different from the one obtained under LOCF. In these cases, a t test will be used. Note that it is still possible to obtain inferences from a full linear mixed-effects model in this context. While this seems less sensible, since one obviously would

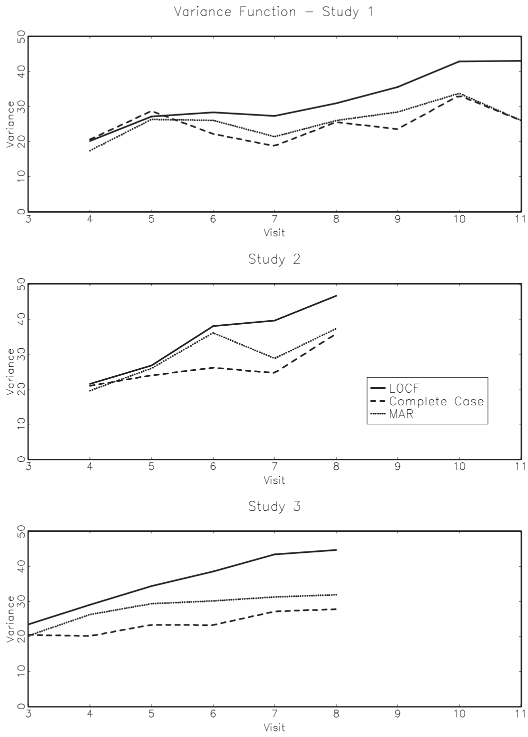

Fig. 4. Variance functions per study and per method.

Table 2. *Analysis of case study. View 1. Treatment effects (standard errors), p values for treatment main effect and for treatment by time interaction, and within-patient correlation coefficients*

		Treatment	*p* value	Within-patient
Study	Method	effect (s.e.)	(effect, interaction)	correlation
1	LOCF	−1.60(1.40)	(0.421, 0.565)	0.65
	CC	−1.96(1.38)	(0.322, 0.684)	0.57
	MAR	−1.81(1.24)	(0.288, 0.510)	0.53
2	LOCF	−1.61(1.05)	(0.406, 0.231)	0.54
	CC	−1.97(1.16)	(0.254, 0.399)	0.37
	MAR	−2.00(1.12)	(0.191, 0.138)	0.39
3	LOCF	1.12(0.71)	(0.964,<0.001)	0.74
	CC	1.75(0.77)	(0.918,<0.001)	0.57
	MAR	2.10(0.69)	(0.476,<0.001)	0.60

get distorted estimates of such longitudinal characteristics as time evolution, etc., we nevertheless add these for the sake of comparison. However, it should be understood that the t test analysis is more in line with clinical trial practice.

For MAR, by its very nature, one is drawn to consider the incomplete profiles, to use the information contained in these for the correct estimation of effects at later times, where there may be some missingness.

Table 3. *Analysis of case study. Views 2a and 2b. p values are reported. ('mixed' refers to the assessment of treatment at the last visit based on a linear mixed model)*

Method	Model	Data used	Study 1	Study 2	Study 3
CC	mixed	All treatments	0.076	0.055	0.001
		Two treatments	0.070	0.088	0.001
CC	*t* test	All treatments	0.092	0.156	0.017
		Two treatments	0.092	0.156	0.017
LOCF	mixed	All treatments	0.053	0.052	0.001
		Two treatments	0.056	0.082	0.001
	t test	All treatments	0.246	0.172	0.120
		Two treatments	0.246	0.172	0.120
MAR	mixed	All treatments	0.052	0.048	0.001
		Two treatments	0.047	0.077	0.001

Thus, one has to consider the full linear mixed model. To this end, the MMRM approach has been developed (Mallinckrodt *et al.*, 2001a,b).

An important issue that occurs whenever there are more than two treatment arms is whether one uses all treatments or only the two of interest. This choice has an effect on the *p* value in the linear mixed model case. Consider, for example, the covariance structure. Model-based smoothing of the covariance structure takes place either on two arms or on all arms. Hence, due to correlations between model parameters, the estimated treatment effects and also the resulting *p* values might change. Generally, one might argue that efficiency can be gained by using all treatment arms, but this comes at the cost of an increased risk of misspecification. This risk can be avoided by assuming a treatment-arm specific covariance matrix in conjunction with a treatment-arm specific mean evolution. For the *t* tests, however, there is no change. Of course, one might entertain the possibility of correcting for multiple comparisons when more than two arms are involved, but this is not the purpose of the current paper and does not substantially affect our conclusions.

Table 3 summarizes results in terms of *p* values. In study 3, which has a relatively large sample size, all *p* values indicate a significant difference with, very importantly, the sole exception of the *t* tests under LOCF. This re-emphasizes the problems with the LOCF method as discussed in Section 6.1. In studies 1 and 2, more subtle differences are observed.

For study 1, we have the following conclusions. All mixed models lead to borderline differences: LOCF and CC are not significant, MAR is borderline (depending on the number of treatments included). An endpoint analysis (i.e. using the last available measurement) leads to a completely different picture, with clearly non-significant results. For study 2, the mixed models lead to small differences, with a noticeable shift towards borderline significance for MAR with all treatments. An endpoint analysis shows, again, results that are notably different (non-significant) from the mixed models.

If the *t* tests under LOCF and CC are compared with the mixed analysis of MAR, studies 1 and 2 show dramatic differences. Such a comparison is not contrived since the *t* tests for LOCF and CC are well in line with common data-analytic practice and under MAR only the mixed analysis makes sense.

These results, in conjunction with those of Section 6.1, underscore the limitations of LOCF and CC. By selecting a subset (CC), a different type of patient might be retained in the treated versus the untreated arm. This can be explained by a difference in therapeutic effect, a difference in side effects or a combination thereof. As with CC, the difference of complete versus incomplete observations can cause distortions within an LOCF analysis. In addition to differences in sets to which the techniques are applied, there are further distortions which take place, in the mean structure, the variance structure and the correlation structure. These effects may counteract and/or strengthen each other, depending on the situation.

Table 4. *Analysis of case study. Fitted MAR and MNAR models to the case study data. Columns MAR and MNAR report twice the negative likelihood. The resulting likelihood ratio is given in the column labeled* χ^2

	MAR	MNAR		
Study	-2 likelihood		χ^2	p
1	2005.89	2004.99	0.90	0.32
2	2330.06	2320.41	9.65	0.0019
3	10234.53	10199.05	35.48	<0.0001
	Treat. effect (s.e.)			
1	$-1.58(1.14)$	$-1.55(1.10)$		
2	$-1.84(1.07)$	$-1.64(1.07)$		
3	1.98(0.65)	2.04(0.64)		

In conclusion, use of likelihood-based ignorable methods is more justifiable than LOCF and CC.

7. SENSITIVITY ANALYSIS

Although the assumption of likelihood ignorability encompasses both MAR and the more stringent and often implausible MCAR mechanisms, it is difficult to exclude the option of a more general nonrandom dropout mechanism. One solution is to fit an MNAR model as proposed by Diggle and Kenward (1994) who fitted models to the full data using the simplex algorithm (Nelder and Mead, 1965). The result of fitting these models to studies 1–3, using GAUSS code developed by the authors, is presented in Table 4. The effects of treatment, time, the interaction between time and treatment, and baseline value were all included in the model. The model for dropout is based on (3.3) and includes the effect of the previous outcome (MAR), with in addition the effect for current, possibly unobserved outcome in the MNAR case.

Note that the results are not directly comparable to those reported in Table 3, where inference is based on the last measurement, but rather to the treatment main effect results reported in Table 2. The model considered here is somewhat simpler than the model considered in Section 6.1, since fitting such a complicated model in the MNAR case may become computationally prohibitive. Note that studies 1–3 show a dramatically different picture in terms of evidence for MNAR, with apparently no, fairly strong, and very strong evidence for MNAR, respectively. However, as pointed out in the introduction and by several authors (discussion to Diggle and Kenward, 1994; Verbeke and Molenberghs, 2000, Chapter 18), one has to be extremely careful with interpreting evidence for or against MNAR using only the data under analysis.

A sensible compromise between blindly shifting to MNAR models or ignoring them altogether, is to make them a component of a sensitivity analysis. In that sense, it is important to consider the effect on key parameters such as treatment effect. Here, in line with several other observations (Molenberghs *et al.*, 2001; Verbeke *et al.*, 2001) we see that the impact on the treatment effect parameter is extremely small, providing additional support for the use of likelihood-based ignorable models. One such route for sensitivity analysis is to consider pattern-mixture models as a complement to selection models (Thijs *et al.*, 2002; Michiels *et al.*, 2002). Further routes to explore sensitivity are based on global and local influence methods (Verbeke *et al.*, 2001). A more extensive case study on the advantages and problems related to several sensitivity analysis is a topic of ongoing research.

The same considerations can be made when compliance data are available. In such a case, arguably a definitive analysis would not be possible and it might be sensible to resort to sensitivity analysis ideas (Cowles *et al.*, 1996).

8. Discussion

In this paper, we have used both formal derivations and case studies to show that there is little justification for analyzing incomplete data from longitudinal clinical trials by means of such simple methods as LOCF and CC. This is true even if a single point in time (e.g. the last measurement occasion) is of primary interest. It is more sensible to use linear mixed models in combination with the assumption of MAR. Such an approach, tailored to the needs of clinical trials, has been proposed by Mallinckrodt *et al.* (2001a,b). This type of analysis is stable and provides sensible assessments of important aspects such as treatment effect and time evolution, even if the assumption of MAR is violated in favor of MNAR. This is in line with analyses conducted by Diggle and Kenward (1994), Molenberghs *et al.* (1997, 2001) and Verbeke *et al.* (2001). Moreover, such analyses can be conducted routinely using standard statistical software such as the SAS procedures MIXED and NLMIXED.

A related and, for the regulatory clinical trial context, very important set of assertions is the following: (1) an ignorable likelihood analysis can be specified a priori in a protocol without any difficulty; (2) an ignorable likelihood analysi is consistent with the intention to treat (ITT) principle, even when only the measurement at the last occasion is of interest; (3) the difference between an LOCF and an ignorable likelihood analysis can be both liberal and conservative. The first is easy to see since, given ignorability, formulating a linear mixed model for either complete or incomplete data involves exactly the same steps. Let us expand on the second issue. It is often believed that when the last measurement is of interest a test for the treatment effect at the last occasion neglects sequences with dropout, even when such sequences contain post-randomization outcomes. As a result, it is often asserted that to be consistent with ITT some form of imputation, based on an incomplete patient's data, e.g. using LOCF, is necessary. However, as Little and Rubin (1987, Chapter 6) showed, likelihood based estimation of means in an incomplete multivariate setting involves adjustment in terms of the conditional expectation of the unobserved measurements given the observed ones. Thus, a likelihood based ignorable analysis (such as MMRM) should be seen as a proper way to accommodate information on a patient with post-randomization outcomes, even when such a patient's profile is incomplete. This fact, in conjunction with the use of treatment allocation as randomized rather than as received, shows that MMRM is fully consistent with ITT. Regarding the third issue, the case study produced smaller p values under MAR than under LOCF (Table 3). Conversely, consider a situation where the treatment difference increases over time, reaches a maximum around the middle of the study period, with a decline thereafter until complete disappearance at the end of the study. Suppose further that the bulk of dropout occurs around the middle of the study. Then, an endpoint analysis based on MAR will produce the correct nominal level, whereas LOCF might reject the null hypothesis too often. When considering LOCF, we often have in mind examples in which the disease shows progressive improvement over time. However, when the goal of a treatment is maintenance of condition in a progressively worsening disease state, LOCF can exaggerate the treatment benefit. For example, in Alzheimer's disease the goal is to prevent the patient from worsening. Thus, in a one-year trial where a patient on active treatment drops out after one week, carrying the last value forward implicitly assumes no further worsening. This is obviously not conservative.

Note that the inadequacy of LOCF, especially when conceived as a single imputation method, will vary across types of disease. LOCF is particularly inappropriate if either the effect of treatment is expected to change over time or there are secular trends. Thus, it would do slightly better in diseases where the treatment induces a steady but reversible response, such as asthma or rheumatism (Senn *et al.*, 2000).

When there is residual doubt about the plausibility of MAR, one can conduct a sensitivity analysis. This is a very active area of research. Obviously, a number of MNAR models can be fitted, provided one is prepared to approach formal aspects of model comparison with due caution. Such analyses can be complemented with appropriate (global and/or local) influence analyses. Another route is to construct pattern-mixture models and to compare the conclusions with those obtained from the selection model framework. Alternative frameworks for sensitivity analyses are provided by Robins *et al.* (1998) and Forster and Smith (1998), who present a Bayesian sensitivity analysis, and Raab and Donnelly (1999).

ACKNOWLEDGEMENTS

The first four authors gratefully acknowledge support from *Fonds Wetenschappelijk Onderzoek-Vlaanderen* Research Project G.0002.98 'Sensitivity Analysis for Incomplete and Coarse Data' and from Belgian IUAP/PAI network 'Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data'. Raymond Carroll's research was supported by a grant from the National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30–ES09106). We are thankful to Eli Lilly for kind permission to use their data.

REFERENCES

AFIFI, A. AND ELASHOFF, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association* **61**, 595–604.

BUCK, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B* **22**, 302–306.

COWLES, M. K., CARLIN, B. P. AND CONNETT, J. E. (1996). Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association* **91**, 86–98.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

DEMPSTER, A. P. AND RUBIN, D. B. (1983). Overview. In Madow, W. G., Olkin, I. and Rubin, D. B. (eds), *Incomplete Data in Sample Surveys,* Vol. II: Theory and Annotated Bibliography. New York: Academic, pp. 3–10.

DIGGLE, P. J. AND KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49–93.

FORSTER, J. J. AND SMITH, P. W. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society, Series B* **60**, 57–70.

HARTLEY, H. O. AND HOCKING, R. (1971). The analysis of incomplete data. *Biometrics* **27**, 7783–7808.

HEYTING, A., TOLBOOM, J. AND ESSERS, J. (1992). Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine* **11**, 2043–2061.

KENWARD, M. G. AND MOLENBERGHS, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science* **12**, 236–247.

LAVORI, P. W., DAWSON, R. AND SHERA, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine* **14**, 1913–1925.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

MALLINCKRODT, C. H., CLARK, W. S. AND STACY, R. D. (2001a). Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal* **35**, 1215–1225.

MALLINCKRODT, C. H., CLARK, W. S. AND STACY, R. D. (2001b). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics* **11**, 9–21.

MALLINCKRODT, C. H., CLARK, W. S., CARROLL, R. J. AND MOLENBERGHS, G. (2003a). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics* **13**, 179–190.

MALLINCKRODT, C. H., SANGER, T. M., DUBE, S., DEBROTA, D. J., MOLENBERGHS, G., CARROLL, R. J., ZEIGLER POTTER, W. M. AND TOLLEFSON, G. D. (2003b). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* **53**, 754–760.

MICHIELS, B., MOLENBERGHS, G., BIJNENS, L., VANGENEUGDEN, T. AND THIJS, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine* **21**, 1023–1041.

MOLENBERGHS, G., KENWARD, M. G. AND LESAFFRE, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* **84**, 33–44.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G. AND DIGGLE, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica* **52**, 153–161.

MOLENBERGHS, G., VERBEKE, G., THIJS, H., LESAFFRE, E. AND KENWARD, M. G. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis* **37**, 93–113.

MURRAY, G. D. AND FINDLAY, J. G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statististics in Medicine* **7**, 941–946.

NELDER, J. A. AND MEAD, R. (1965). A simplex method for function minimisation. *The Computer Journal* **7**, 303–313.

RAAB, G. M. AND DONNELLY, C. A. (1999). Information on sexual behaviour when some data are missing. *Applied Statistics* **48**, 117–133.

ROBINS, J. M., ROTNITZKY, A. AND SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93**, 1321–1339.

ROBINS, J. M., ROTNITZKY, A. AND ZHAO, L.‘P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

RUBIN, D. B., STERN, H. S. AND VEHOVAR, V. (1995). Handling “don’t know” survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* **90**, 822–828.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

SENN, S. J., STEVENS, L. AND CHATURVEDI, N. (2000). Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine* **19**, 861–877.

SHIH, W. J. AND QUAN, H. (1997). Testing for treatment differences with dropouts present in clinical trials—A composite approach. *Statistics in Medicine* **16**, 1225–1239.

SIDDIQUI, O. AND ALI, M. W. (1998). A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics* **8**, 545–563.

THIJS, H., MOLENBERGHS, G., MICHIELS, B., VERBEKE, G. AND CURRAN, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3**, 245–265.

VERBEKE, G. AND MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*, Lecture Notes in Statistics, 126. New York: Springer.

VERBEKE, G. AND MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

VERBEKE, G., MOLENBERGHS, G., THIJS, H., LESAFFRE, E. AND KENWARD, M. G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics* **57**, 7–14.

[*Received April 25, 2003; first revision October 30, 2003; second revision December 23, 2003; accepted for publication January 19, 2004*]