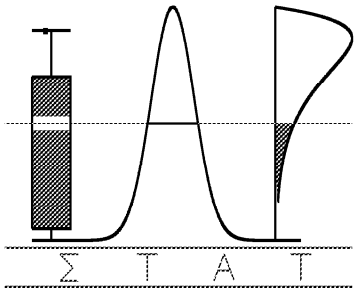
T E C H N I C A L

R E P O R T

0229

Analyzing Incomplete Longitudinal Clinical Trial Data

G. Molenberghs, H. Thijs, R.J. Carroll, M.G. Kenward



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Analyzing Incomplete Longitudinal Clinical Trial Data

Geert Molenberghs and Herbert Thijs

Biostatistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium.
Email: geert.molenberghs@luc.ac.be and herbert.thijs@luc.ac.be

Michael G. Kenward

London School of Hygiene & Tropical Medicine, London, U.K.

Raymond J. Carroll

Department of Statistics, Texas A& M University, College Station, TX, U.S.A.

Abstract

Using standard missing data taxonomy, largely due to Rubin and co-workers, it is argued 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, are poorly principled and restrictive. Given the availability of flexible software for analyzing longitudinal sequences of unequal length, it is argued there is not even a computational reason for not shifting to a likelihood-based ignorable analysis. Such analyses are valid under the much weaker assumption of MAR. While the occurrence of MNAR missingness cannot be ruled out, it is argued that such analyses are, themselves, surrounded with problems and therefore, rather than either forgetting about them or blindly shifting to them, their optimal place is within a sensitivity analyses. The concepts developed here are exemplified using data from three clinical trials, where it is shown that shifting the analysis method may have an impact on the conclusions of the study.

Some Key Words: 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.

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, Laird, and Rubin 1977), and data imputation and augmentation procedures (Rubin 1987), combined with powerful computing resources have largely provided a solution to this aspect of the problem. There remains the very difficult and important question of assessing the impact of missing data on subsequent statistical inference.

Certain important concepts are now in common use in the missing value literature. 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.

Many 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 since the marginal measurement distribution would be of interest also with complete data. Little and Rubin's taxonomy is most easily developed in the selection setting. Parameterizing and making inference about the effect of treatment and its evolution over time is straightforward in the selection model context.

In the specific case of a clinical trial setting, standard methodology used to analyze longitudinal data subject to non-response is mostly based on the MCAR assumption and most often include such methods as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation. This is often even done without questioning the possible influence of these assumptions on the final results. As will be discussed in subsequent sections, this is unfortunate since (1) the assumption of MCAR is very strong and (2) such methods as LOCF, CC, and imputation make additional, strong, and often very unrealistic assumptions. On the other hand, under MAR, valid inference can be obtained through a likelihood-based analysis, without the need for modeling the dropout process. As a consequence, one can simply use, for example, linear or generalized linear mixed models (Verbeke and Molenberghs 2000), without additional complication or effort. We will argue that such an analysis not only enjoys much wider validity than the simple methods but in addition is simple to conduct, *without additional data manipulation* using such tools as, for example, the SAS procedures MIXED or NLMIXED. Thus, clinical trial practice should shift away from the *ad hoc* methods and focus on likelihood-based ignorable analyses instead.

This is not to forget that the reasons for dropout in realistic settings are varied and it is therefore difficult to fully justify on a priori grounds the assumption of MAR. For example, the rate of and the reasons for dropout varied considerably across 11 clinical trials of similar

design, of the same drug in the same indication. In one study, completion rates were 80% for drug and placebo. In another study, two thirds of the patients on drug completed the study, while only one third did so on placebo. In 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 to adverse events was more common. At first sight, this calls for a further shift towards MNAR models. However, some careful considerations have to be made. First, MNAR models raise a number of non-trivial issues: irrespective of the MNAR route taken, e.g., a parametric model of the type of Diggle and Kenward (1994), or a semiparametric approach such as in Robins, Rotnitzky, and Zhao (1998), assumptions will be required that cannot be assessed from the data under analysis. Hence in this setting there cannot be anything that could be termed a definitive analysis. This limits the use of MNAR methods in a regulatory framework. Second and often overlooked, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated. This is very important in settings such as those mentioned above where dropout frequency and patterns vary across otherwise similar studies. While the *rate* of dropout and its dependence on treatment arm does not provide evidence against even MCAR, its variation from trial to trial indicates that one should not assume all studies are alike. A discussion of this phenomenon in the survey context has been given in Rubin, Stern, and Vehovar (1995). These authors argue that, in well conducted experiments (some surveys and many confirmatory clinical trials), the assumption of MAR is often to be regarded as a realistic one. Third, and very important for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with complete data. Based on these considerations, we recommend, for primary analysis purposes, the use of ignorable likelihood-based methods. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis, within which MNAR models can play a major role, together

with, for example, pattern-mixture models (Verbeke and Molenberghs 2000, Ch. 18–20).

A three-trial case study is introduced in Section 2. A thorough discussion on the problems associated with simple method is presented in Section 3. Particular emphasis is given to available case methods, since they allow easy transition from simple methods (requiring MCAR) to likelihood-based analyses, ignorable under MAR (Section 4). A detailed treatment of likelihood-based analyses is given in Section 5) and the case study is analyzed in Section 6. A perspective on sensitivity analysis is sketched in Section 7.

2 Case Study

The ideas developed in this paper are motivated from and applied to three sets of data, coming from three clinical trials (enrolling 167, 342, and 713 patients, respectively). The Hamilton Depression Rating Scale ($HAMD_{17}$) is used to measure the depression status of the patients. For each patient, a baseline assessment is 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. Individual profiles, for each study, are shown in Figure 1. The primary contrast is between A1 and C for studies 1 and 2, whereas in study 3 one is interested in A *versus* B. Emphasis is on the difference between arms at the end of the study.

3 Simple Methods

As suggested in the introduction, missing data nearly always entail problems for the practicing statistician. First, inference will often be invalidated when the observed measurements do not constitute a simple random subset of the complete set of measurements. Second, even when correct inference follows, it is not always easy to trick standard software into operation on a ragged data structure. Even in the simple case of a one-way ANOVA design and under an MCAR mechanism operating, problems occur since missingness destroys the balance between the sizes of the subsamples. This implies that a slightly more complicated least squares analysis has to be invoked. Of course, a regression module for the latter analysis is included in most statistical software packages. The trouble is that the researcher has to know which tool to choose for particular classes of incomplete data. Little and Rubin (1987) give an extensive treatment of methods for the analysis incomplete data. Some of these methods were proposed three quarters of a century ago. Examples are Yates' (1933) iterated ANOVA and Bartlett's (1937) ANCOVA procedures to analyze incomplete ANOVA designs. The former method is an early example of the Expectation-Maximization (EM) algorithm.

We will briefly review a number of techniques that are valid when the measurement and missing data processes are independent (MCAR) and their parameters are separated. One of our main points is that many of these methods are used in situations where the MCAR assumption, as well as the additional assumptions behind each of the techniques separately, are not tenable. This should be seen as bad practice since it will often lead to biased estimates, invalid tests and hence erroneous conclusions. Ample detail and illustrations of several problems are provided in Verbeke and Molenberghs (1997). A complete case analysis removes incomplete cases whereas imputation methods (such as LOCF) fill in missing values. In both cases the effect is a "rectangular" data matrix, but at high cost. Regarding impu-

tation, 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 properly acknowledges the uncertainty stemming from filling in missing values rather than observing them (Rubin 1987, Schafer 1997). A third family of methods is based on analyzing the data as they are. A well-known example is given by a so-called *available case* analysis. In Section 4, it will be shown how a transition can be made from MCAR to MAR by replacing a frequentist MCAR analysis with a likelihood-based one. The latter is the basis for Section 5.

The impact and *danger* of using simple methods is illustrated, using a relatively simple case study, in Verbeke and Molenberghs (1997). The more desirable solutions, such as ignorable likelihood analyses, are taken up at length in Verbeke and Molenberghs (2000).

3.1 Complete Case Analysis

A complete case analysis includes only those cases for analysis, for which all measurements were recorded. This method has obvious advantages. It is very simple to describe and since the data structure is as would have resulted from a complete experiment, standard statistical software can be used. Further, since the entire estimation is done on the same subset of completers, there is a common basis for inference, unlike for the available case methods.

Unfortunately, the method suffers from severe drawbacks. First, 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 is dramatic. Even though the

reduction of the number of complete cases will be less dramatic in realistic settings where the missingness indicators are correlated, the effect just sketched will often undermine a complete case analyses. In addition, 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. As mentioned earlier, the MCAR assumption is much more restrictive than the MAR assumption.

3.2 Simple Forms of Imputation

An alternative way to obtain a data set on which complete data methods can be used is based on filling in rather then deletion. The principle of imputation is easy. The observed values are used to impute values for the missing observations. There are several ways to use the observed information. First, one can use information on the same subject (e.g., last observation carried forward). Second, information can be borrowed from other subjects (e.g., mean imputation). Finally, both within and between subject information can be used (e.g., conditional mean imputation, hot deck imputation). A standard reference is Little and Rubin (1987).

However, great care has to be taken with 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 method could work for a linear model with one fixed effect and one error term, but that it generally does not for hierarchical models, split-plot designs, and

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 would be biased. Second, even for a correct imputation model, the uncertainty resulting from missingness is masked. 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. Methods such as the last observation carried forward require additional and often unrealistically strong assumptions.

The main advantage, shared with complete case analysis, is that complete data software can be used. With the availability of such software like the SAS procedures MIXED and NLMIXED, it is no longer necessary to restrict oneself to complete data software.

3.3 Last Observation Carried Forward

In the LOCF method, whenever a value is missing, the last observed value is substituted. The technique can be applied to both monotone and nonmonotone missing data. It is typically applied to settings where incompleteness is due to attrition. Very strong and often unrealistic assumptions have to be made to ensure validity of this method. First, one has to believe that a subjects' measurement stays at the same level from the moment of dropout onward (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. Second, this method 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 equal footing.

Verbeke and Molenberghs (1997, Ch. 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 severely affected by application of this technique. A similar conclusion, based on the case study, is reached in Section 4.

3.4 Imputing Unconditional Means

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 equally damaging as LOCF.

3.5 Buck's Method: Conditional Mean Imputation

This method (Buck 1960, Little and Rubin 1987) is technically hardly more complex than mean imputation. Let us describe it for 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

$$\boldsymbol{Y}_i^m | \boldsymbol{y}_i^o \sim N(\boldsymbol{\mu}^m + \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o), \boldsymbol{\Sigma}^{mm} - \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om}). \tag{3.1}$$

Superscripts o and m refer to "observed" and "missing" components, respectively. The second step calculates the conditional mean from this regression and substitutes it for the 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 for MCAR mechanisms. Little and Rubin (1987) added that the method is valid under certain types of MAR mechanism. 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 underestimated. There is a connection between *the concept* of conditional mean imputation and a likelihood-based ignorable analysis, as described in Section 5.

4 Available Case Methods

Available case methods (Little and Rubin 1987) use as much of the data as possible. Consider a single multivariate normal sample, based on $i = 1, \ldots, N$ subjects, for which $j = 1, \ldots, n$ assessment occasions are planned, producing measurements Y_{ij}. Let us describe estimation of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The jth component μ_j of the mean vector and the jth diagonal variance element σ_{jj} are estimated using all cases that are observed on the jth variable, disregarding their response status at the other measurement occasions. The (j,k)th element $(j \neq k)$ of the covariance matrix is computed using all cases that are observed on both the jth and the kth variable. This method is more efficient than the complete case method. The number of components of the outcome vector has no direct effect on the sample available for a particular mean or covariance component.

The method is valid only under MCAR. In this respect, it is no fundamental improvement over a complete case analysis. An additional issue is that, although more information is used and a consistent estimator is obtained under MCAR, it is not guaranteed that the covariance matrix is positive (semi-)definite. While this is a small sample problem, for samples with a large number of variables and/or with fairly high correlations between pairs of outcomes, this nuisance feature is likely to occur.

Although a complete case analysis is possible for virtually every statistical method and single imputation is also fairly generally applicable, extending an available case analysis beyond such simple settings as multivariate means and covariances is tedious.

5 Likelihood-based Ignorable Analysis

Let us assume MAR holds. Below and based on arguments laid out in Rubin (1976) and Little and Rubin (1986) it will be argued formally that likelihood based inference is valid, whenever the mechanism is MAR and provided the technical condition holds that the parameters describing the nonresponse mechanism are distinct from the measurement model parameters (Little and Rubin 1987). 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. We will now further qualify the extent of this statement; a few cautionary remarks apply.

First, when at least part of the scientific interest is directed towards the nonresponse process, obviously both processes need to be considered. Still, 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 precision 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 the parametric assumptions are correct, are valid. Third, it may be hard to fully rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Section 7. In preparation of the case study analysis, let us formalize these concepts.

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 $f(\boldsymbol{y}_i, d_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi})$, where X_i, Z_i, and W_i are covariate matrices. We will use the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ to describe the measurement and missingness processes, respectively.

The taxonomy of Rubin (1976) and Little and Rubin (1987), informally described in the introduction, is based on the selection model factorization:

$$f(\boldsymbol{y}_i, d_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, Z_i, \boldsymbol{\theta}) f(d_i | \boldsymbol{y}_i, W_i, \boldsymbol{\psi}), \tag{5.2}$$

where the first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. The (hypothetical) full data likelihood contribution for subject i assumes the form

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} | X_i, Z_i, W_i, \boldsymbol{y}_i, d_i) \propto f(\boldsymbol{y}_i, d_i | X_i, Z_i, \boldsymbol{\theta}, \boldsymbol{\psi}). \tag{5.3}$$

Since inference has to be based on what is observed, the full data likelihood L^* has to be replaced by the observed data likelihood L:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | X_i, Z_i, W_i, \boldsymbol{y}_i, d_i) \propto f(\boldsymbol{y}_i^o, d_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}),$$

with

$$\begin{aligned} f(\boldsymbol{y}_i^o, d_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i, d_i | X_i, Z_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, Z_i, \boldsymbol{\theta}) f(d_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, W_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m. \end{aligned}$$

Under an MAR process, we obtain

$$\begin{aligned} f(\boldsymbol{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, Z_i, \boldsymbol{\theta}) f(d_i | \boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= f(\boldsymbol{y}_i^o | X_i, Z_i, \boldsymbol{\theta}) f(d_i | \boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}), \end{aligned} \tag{5.4}$$

i.e., the likelihood factorizes into two components of the same functional form as the general factorization (5.2) of the complete data. If further $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint in the sense that the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is the product of the individual parameter spaces, the so-called separability condition, then inference can be based on the marginal observed data density only.

Turning to the measurement model, assume the measurements are continuous and it is deemed sensible to consider a linear mixed-effects model with serial correlation:

$$\boldsymbol{Y}_i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i, \tag{5.5}$$

(Verbeke and Molenberghs 2000) where $\boldsymbol{Y}_i$ is the n dimensional response vector for subject i, $1 \leq i \leq N$, 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, $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2 I_{n_i})$ is a n dimensional vector of measurement error components, and $\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N}$ 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 $\rho(t_{ij}-t_{ik})$. A first simplifying assumption is that it depends only on the time interval between two measurements Y_{ij} and Y_{ik}, i.e., $\rho(t_{ij}-t_{ik}) = \rho(|t_{ij}-t_{ik}|)$, where $u = |t_{ij} - t_{ik}|$ denotes time lag. This function decreases such that $\rho(0) = 1$ and $\rho(+\infty) = 0$. 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

$$\boldsymbol{Y}_i \;\sim\; N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{5.6}$$

Here, $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n \times n)$ covariance matrix grouping 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. The model for the dropout process is based on, for example, a logistic regression for the probability of dropout at occasion j, given 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

$$\text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = \text{logit}\left[\text{pr}(D_i = j|D_i \geq j, \boldsymbol{y}_i)\right] = \boldsymbol{h}_{ij}\boldsymbol{\psi} + \omega y_{ij}, \qquad i = 1, \ldots, N. \tag{5.7}$$

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 (5.7) provides the building blocks for the dropout process $f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi})$.

Note that, under an ignorable likelihood analysis, the dropout model (5.7) does not need to be specified and it is sufficient to specify (5.5). Such an ignorable linear mixed model specification is termed MMRM by Mallinckrodt *et al* (2001ab). Precisely, MMRM is a

particular form of a linear mixed model, relevant for acute phase confirmatory clinical trials, fitting within the ignorable likelihood paradigm.

6 Analysis of Case Study

The primary null hypothesis (difference between the treatment and placebo in mean change of the HAMD17 total score at endpoint) will be tested using a model of the type (5.5). The model will include the fixed categorical effects of treatment, investigator, visit, and treatment-by-visit interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-visit interaction. The following covariance structures will be used to estimate the within-patient errors in preliminary analyses: unstructured, auto-regressive, compound symmetric, and simple structures, with heterogeneous variances by visit. The covariance structure leading to the best fit, as determined by Akaike's information criterion, will be considered the primary analysis. Satterthwaite's approximation will be used to estimate denominator degrees of freedom. The significance of differences in least-square means will be based on Type III tests. Analyses will be 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. In practice, one commonly combines LOCF and CC with an analysis at the last occasion only. This will be looked at further in Section 6.2, where it will be shown that even in such a context it is possible and better to adopt an MAR strategy as well. In addition, the impact of focusing on two *versus* all treatment arms will be discussed.

6.1 Longitudinal Analysis

When comparing the three strategies, there is little or no difference between the three methods, neither for the treatment main effect, nor for the treatment by time interaction (p values are reported in Table 2). 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).

Profiles for each of the treatment arm by investigator combinations, for a selected set of baseline values shows very little difference (plots not shown). Two specific features of the mean structure are the time trends and the treatment effects (over time). Let us discuss these in turn. The time trends for all studies are displayed in Figure 2. Both LOCF and CC are different from MAR, with a larger difference for CC. This is due to the selective effect of CC; this same effect carries over, in part, to the LOCF method. The effect is strongest in the third study. There is also some deviation of the treatment effect as opposed to MAR (Figure 3). The most striking feature is the qualitative difference between the three studies. 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, somewhat in contrast with the often claimed conservatism of LOCF. So, one might want to argue that LOCF is only sometimes valid. However, there are two considerations. First, under the conditions where LOCF would be valid, so is MAR. Further, due to the aberrant behavior in other key aspects (variance and correlation structures), one has to be extremely careful and conservative behavior is not guaranteed.

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 homogenous or classical compound symmetry (CS) and hence a random-intercept structure. LOCF on the other hand even 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, one should be very careful.

The fitted correlations are given in Table 3. Clearly, CC and MAR produce virtually the same correlation. However, the correlation coefficient estimated under LOCF is much higher. This is entirely due to the fact that after dropout, a constant value is imputed for the remainder of the study period, thereby increasing the correlation. 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 truly longitudinal aspects of the mean structure. For example, (standard errors of) time trends and interactions of time with covariates (such as treatment effect) can be affected. Precisely, if the 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 three structures are affected. This is in line with earlier analyses such as in Verbeke and Molenberghs (1997, 2000). It is important to note that, generally, the directionality of the errors made (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 relevant from a clinical trial point of view.

6.2 Single Time Point and All *Versus* Two Treatment Arms

In practice, a primary endpoint as, for example, specified in the protocol of a clinical trial, may focus on one particular measurement occasion. Explicitly, one is often interested in the treatment difference at the last assessment time. Thus, LOCF, CC, and MAR approaches can be studied towards this specific goal. The conduct of an LOCF and CC analysis is straightforward. In the first case, the profiles are extended from dropout until the last assessment. In the second case, incomplete profiles are simply deleted. In both cases, a rectangular set of data results and consequently studying the treatment effect at the last endpoint is very simple. Still, there are a few choices to make. First, one can include all treatments into the model *versus* only the two arms of interest. Second, the comparison can be based on a full linear mixed model *versus* a simple model for the last time point only (e.g., based on a two-sample t test or on change from baseline).

For MAR, by its very nature, one still explicitly wants to consider the incomplete profiles, to use the information contained in these for the correct estimation of effects at later, incomplete times. Thus, one considers the full linear mixed model, while it can be argued that this approach is slightly more complicated than CC and LOCF, it ought to be clear that such an approach is not excessively complicated and certainly within the realm of standard statistical modeling technology. The linear mixed model can be parameterized such that the parameter of interest is displayed as a fixed effect. Alternatively, additional contrast statements can be added to the program.

In summary, the analyses considered are classified along the following three dimensions.

Method for handling missingness. As before, we consider complete cases (CC), last observation carried forward (LOCF), and missing at random (MAR).

Model. Either the full linear mixed model is used, from which then the appropriate test is derived, or the (preprocessed) data at the final evaluation time are used in a t test, or change from baseline. Preprocessing refers to either deletion (CC) or imputation (LOCF). Since, in the case of MAR, no processing is done but the data are used as observed, effecting a t test does not allow exploiting the power of MAR, and hence no t test is used in this case.

Treatments included. Either all treatments are included or just the two treatments of primary interest. This choice has an effect on the p value in the linear mixed model case. Indeed, take for example the covariance structure. Model-based smoothing of the covariance takes place either on two arms or on all arms. Hence, due to correlation between model parameters, the estimated treatment effects and also the p values derived thereof, might change. 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 such is not the purpose of the current report and does not substantially affect the conclusions thereof.

Table 4 shows a summary of the results, in terms of p values. As far as study 3 goes, the relatively large sample size implies that 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 leads to a completely different picture, with clearly non-significant results. For study 2, the mixed models lead to small difference, with a noticeable shift towards borderline significance for MAR with all treatments. An endpoint analysis shows, again, results that are strongly different (non-significant) as opposed to the mixed models.

Once again, the results are borderline. However, 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 on the one hand, and under MAR only the mixed analysis makes sense on the other hand.

These results, in conjunction with those of Section 6.1, underscore the problems with LOCF and CC. By selecting a subset (CC), a different type of patients might be retained in the treated versus the untreated arm. This can be explained by, for example, a difference in therapeutic effect, a difference in side effects, or a combination thereof. Exactly as with CC, the difference of completers versus incomplete observations can cause distortions within an LOCF analysis. In addition to differences in sets to which the technique 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.

In conclusion, there is very little justification for LOCF and CC analyses. Historically, the most important justification came from simplicity. Currently, with the availability of commercial software tools, such as the SAS procedures MIXED and NLMIXED, this justification no longer holds. Arguably, an MAR analysis is the preferred choice. Of course, the correctness of an 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 hold. These and related issues are briefly discussed in the next section (see also Verbeke and Molenberghs 2000).

7 Sensitivity Analysis

Even though the assumption of likelihood ignorability encompasses the MAR and not only 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). Diggle and Kenward (1994) fitted models to the full data using the simplex algorithm (Nelder and Mead 1965). A module for the linear mixed model with dropout is implemented in the OSWALD software, written for S-Plus (Smith, Robertson, and Diggle 1996). It is based on an extension of the Diggle and Kenward (1994) model, as described by models (5.5) and (5.7). The result of fitting these models to studies 1–3, using GAUSS code developed by the authors, is presented in Table 5. Along with the effect of treatment, time, the interaction between time and treatment, and baseline value were included into the model. The model for dropout is based on (5.7) 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 4, 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 becomes prohibitive. This should be seen as another advantage of a likelihood-based ignorable analysis. 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. However, as pointed out in the introduction and by several authors

(discussion to Diggle and Kenward 1994, Verbeke and Molenberghs 2000, Ch. 18), one has to be extremely careful with interpreting evidence for or against MNAR in a selection model context. 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, we see, in line with several other observations (Molenberghs *et al* 2001, Verbeke *et al* 2001) 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 compliment to selection models (Thijs *et al* 2001, 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.

8 Discussion

In this paper, we have shown 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 much more sensible to refer to linear mixed models, in combination with the assumption of MAR. Such an analysis is stable, providing 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, Kenward, and Lesaffre (1997), Verbeke *et al* 2001, Molenberghs *et al* 2001). Moreover, such analyses can be conducted without any problem 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

that such an analysis (1) can be specified a priori in a protocol without any difficulty, (2) is consistent with the intention to treat (ITT) principle, even when only the measurement at the last occasion is of interest, and (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, the test for the treatment effect at the last occasion neglects sequences with dropout. However, as Little and Rubin (1987, Ch. 6) showed, likelihood based estimation of means in an incomplete multivariate (normal) setting involves adjustment in terms of the conditional expectation of the unobserved measurements given the observed ones. Such an adjustment is similar in spirit to Buck's method of conditional imputation, but without the problems associated to explicit single imputation. An illustration of this is provided in Verbeke and Molenberghs (1997, p. 229) where it is shown that for a saturated normal model fit to an incomplete sample, observed and expected means do not coincide, precisely as a result of the aforementioned adjustment. Thus, a likelihood based ignorable analysis (such as MMRM) should be seen as a major improvement over LOCF, appropriately using all information on all patients (consistent with ITT), without the risk of distorting key model features such as mean profile, relative importance of the components of variability, etc. Regarding the third issue, the case study produced smaller p values under MAR than under LOCF (Table 4). Reversely, 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 way too often.

When there is residual doubt about the plausibility of MAR, one can conduct a sensitivity analysis. While many proposals have been made, this is still 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 sensitivity analyses frameworks are provided by Robins, Rotnitzky, and Scharfstein (1998), Forster and Smith (1997) who present a Bayesian sensitivity analysis, and Raab and Donnelly (1999).

Acknowledgments

Geert Molenberghs and Herbert Thijs 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". Ray Carroll's research was supported by a grant from the National Cancer Institute (CA57030). We are thankful to Eli Lilly for kind permission to use their data. We are also indebted to Dr. Craig Mallinckrodt for helpful discussions.

s

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.

Bartlett, M.S. (1937) Some examples of statistical methods of research in agriculture and applied botany. *Journal of the Royal Statistical Society, Series B*, **4**, 137–170.

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.

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. *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*, W.G. Madow, I. Olkin, and D.B. Rubin (Eds.). New York: Academic Press, 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**, 7783–808.

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

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: John Wiley & Sons.

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**, 4, 1215–1225.

Mallinckrodt, C.II., Clark, W.S., and Stacy R.D. (2001b) Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, **11**, (1 & 2), 9–21.

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., 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.

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.

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

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

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 & Hall.

Smith, D.M., Robertson, B., and Diggle, P.J. (1996) *Object-oriented Software for the Analysis of Longitudinal Data in S.* Technical Report MA 96/192. Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2001) Strategies to fit pattern-mixture models. *Biostatistics*, **2**, 000 000.

Verbeke, G. and Molenberghs, G. (1997) *Linear Mixed Models in Practice: A SAS-Oriented Approach.* Lecture Notes in Statistics 126. New York: Springer-Verlag.

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

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.

Yates, F. (1933) The analysis of replicated experiments when the field results are incomplete. *Empirical Journal of Experimental Agriculture*, **1**, 129–142.

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

Table 2: *Analysis of case study. Initial models. p values for (treatment main effect, treatment by time interaction).*

Method	Study 1	Study 2	Study 3
LOCF	(0.421, 0.565)	(0.406, 0.231)	(0.964,<0.001)
CC	(0.322, 0.684)	(0.254, 0.399)	(0.918,<0.001)
MAR	(0.288, 0.510)	(0.191, 0.138)	(0.476,<0.001)

Table 3: *Fitted within-patient correlation coefficients.*

Method	Study 1	Study 2	Study 3
LOCF	0.65	0.54	0.74
CC	0.57	0.37	0.57
MAR	0.53	0.39	0.60

Table 4: *Additional analyses of case study. 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

Table 5: Fitting 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	$-1.98(0.65)$	$-2.04(0.64)$		

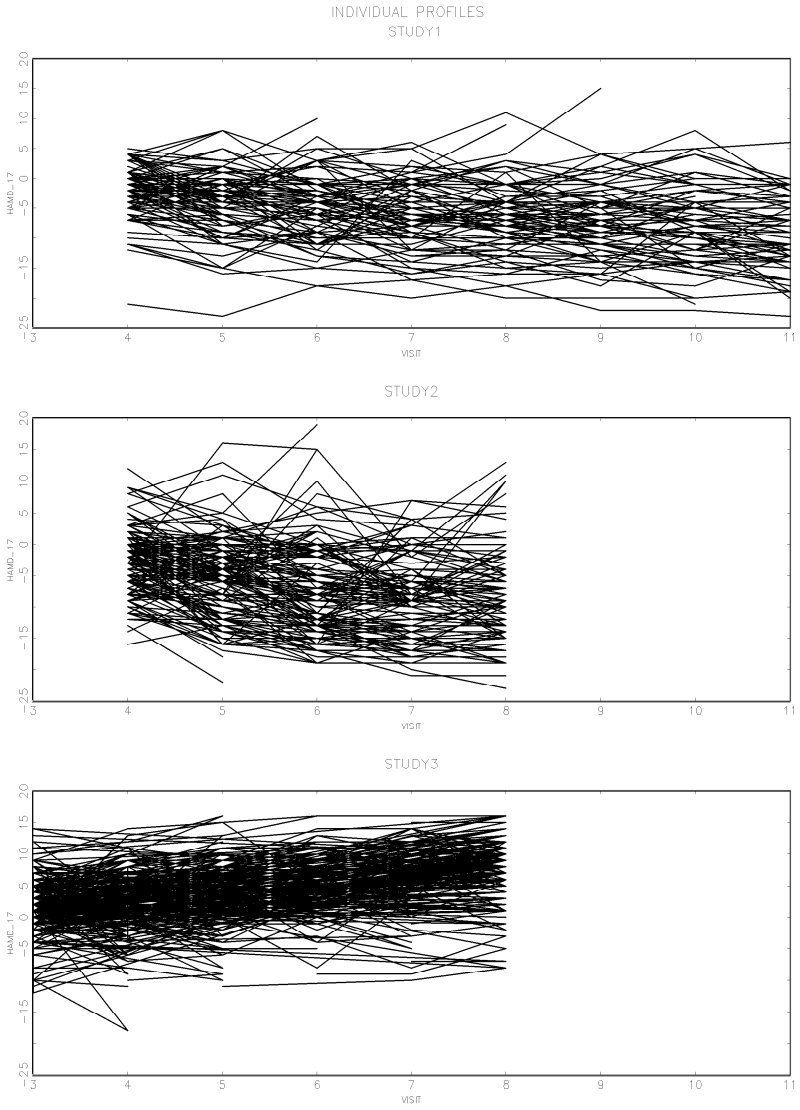


Figure 1: *Individual profiles for each of the three studies.*

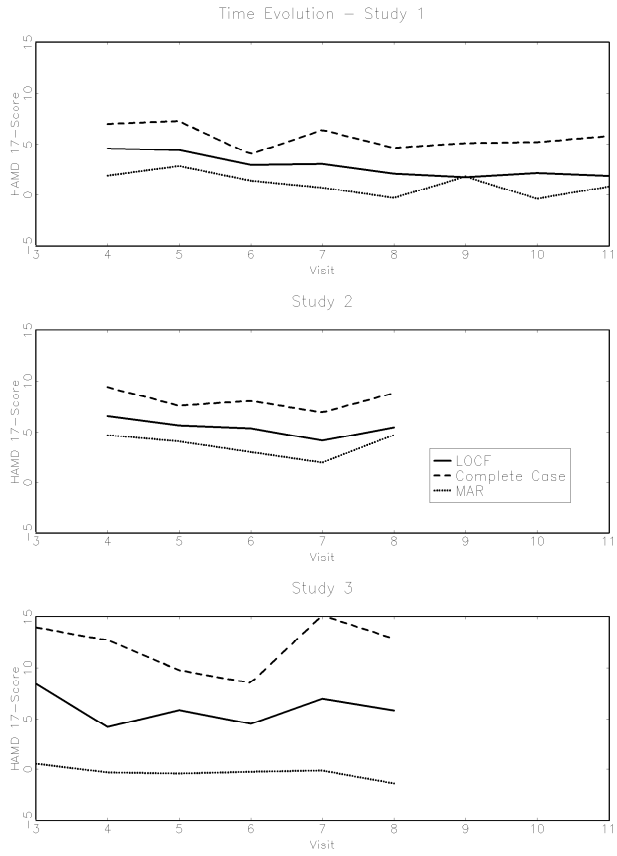


Figure 2: *Summary of All Placebo Time Evolutions.*

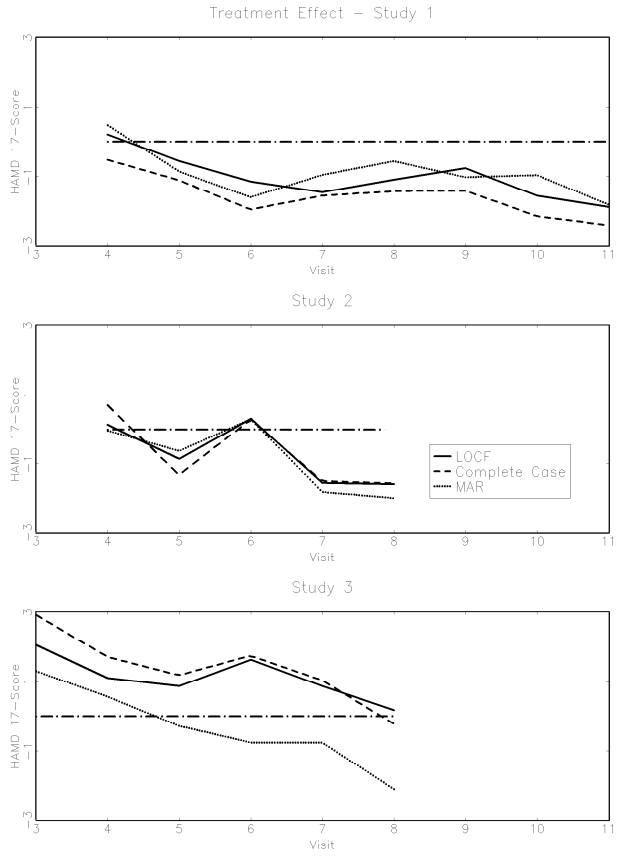


Figure 3: *Summary of All Treatment Effects.*

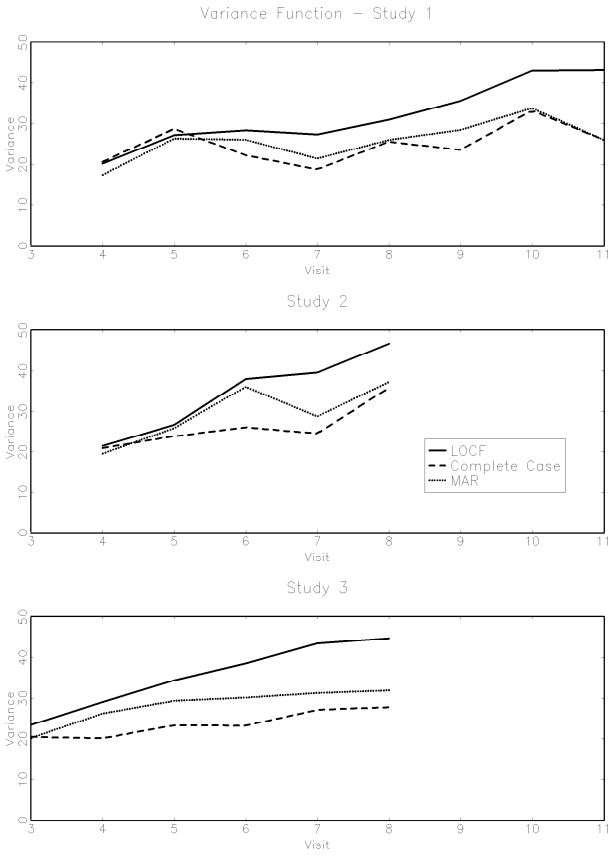


Figure 4: *Variance Functions.*