The nature of sensitivity in monotone missing not at random models

Peer-reviewed author version

JANSEN, Ivy; HENS, Niel; MOLENBERGHS, Geert; AERTS, Marc; VERBEKE, Geert & Kenward, Michael G. (2006) The nature of sensitivity in monotone missing not at random models. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 50(3). p. 830-858.

DOI: 10.1016/j.csda.2004.10.009
Handle: http://hdl.handle.net/1942/1982

The Nature of Sensitivity in Monotone Missing Not At Random Models

Ivy Jansen [a,*], Niel Hens [a], Geert Molenberghs [a], Marc Aerts [a], Geert Verbeke [b], and Michael G. Kenward [c]

[a] *Center For Statistics, Limburgs Universitair Centrum, Universitaire Campus Building D, 3590 Diepenbeek, Belgium*

[b] *Biostatistical Centre, Katholieke Universiteit Leuven, Belgium*

[c] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, U.K.*

Abstract

Models for incomplete longitudinal data under missingness not at random have gained some popularity. At the same time, cautionary remarks have been issued regarding their sensitivity to often unverifiable modeling assumptions. Consequently, there is evidence for a shift towards using ignorable methodology, supplemented with sensitivity analyses to explore the impact of potential deviations of this assumption in the direction of missingness at random. One such tool is local influence. It is shown that local influence tends to pick up a lot of different anomalies in the data at hand, not just deviations in the MNAR mechanism. This particular behavior is described and insight offered in terms of the non-standard behavior of the likelihood ratio test statistic for MAR missingness versus MNAR missingness within a model of the Diggle and Kenward type.

Key words: Ignorability, Likelihood ratio test, Linear mixed model, Local Influence, Missing at random, Missing not at random, Sensitivity Analysis

1 Introduction

Longitudinal data, while very common in contemporary applications, typically suffer from incompleteness. Over the last century, the focus, when formulating answers to the analysis of such incomplete data, has shifted. Indeed, early

* Corresponding author. Tel. +3211268233. Fax: +3211268299.
Email address: ivy.jansen@luc.ac.be (Ivy Jansen).

work on missing values was largely concerned with overcoming the lack of balance or deviations from the intended study design (Afifi and Elashoff, 1966; Hartley and Hocking, 1971). Later, 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. During the last decade, a multitude of advanced models, allowing for potentially complicated ways in which missingness is influenced by observed and unobserved measurements, have been formulated. To properly describe such methods, let us review the now well established framework 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. The property of ignorability is a convenient one, since it implies that incomplete data, fulfilling this property, can be analyzed as they are, without the need for an explicit missingness model. It has been advocated by many that such an approach ought to be considered way more often than is currently the case (Molenberghs *et al*, 2004; Jansen *et al*, 2004). Examples include the linear and generalized linear mixed-effects models, analyzed using maximum likelihood. Such analyses are less stringent than a complete case analysis or *last observation carried forward*, which have had, and still have, a strong popularity in clinical trial settings. Nevertheless, a shift, away from such simple ad hoc methods, to the more principled (likelihood-based) ignorable analyses, further supported by the availability of the necessary standard statistical software to allow for such analyses in practice, is an important step but it does not solve the issues surrounding incompleteness (Lavori, Dawson, and Shera, 1995; Mallinckrodt *et al*, 2003a,b).

For example, it is possible for the assumption of ignorability not to be true and then one might want to consider more general models. Instances of MNAR models include Diggle and Kenward (1994, DK), Molenberghs, Kenward, and Lesaffre (1997), Van Steen *et al* (2001), and Jansen *et al* (2003). These belong to the so-called selection model family (Little and Rubin, 1987) which factors the joint distribution of the measurement and response mechanisms into the marginal measurement distribution and the response distribution, conditional on the measurements. While such models seem to be the proper answer to the need for more flexible models, a number of criticisms have been formulated. Computational instability is one of them but, more importantly, conclusions based on such models have often been questioned as unreliable because they depend on the specific form assumed for the MNAR process, which can, in

principle, not be verified from the data.

Thus, with the volume of literature on non-random missing data increasing, there has been growing concern as well (Glynn, Laird, and Rubin, 1986), especially for selection models. For example, formal tests for the null hypothesis of random missingness, while technically possible, should be approached with caution. As a result, several authors have advocated to investigate the sensitivity of the results with respect to model assumptions (Little, 1994; Rubin, 1994; Laird, 1994; Molenberghs *et al*, 1999). As a general rule, fitting a MNAR model should be subject to careful scrutiny. First, the impact of the assumed distributional form and the specific model choices on the conclusions, when an MNAR model is fitted, has been shown to be much higher than would be the case if data were complete (Kenward, 1998; Scharfstein, Rotnizky, and Robins, 1999; Kenward, Goetghebeur, and Molenberghs, 2001). Second, one may shift from the classical selection model framework, to which the DK model belongs, to the pattern-mixture model framework (Little, 1993, 1994). The use of pattern-mixture models for sensitivity analysis purposes has been explored by Thijs *et al* (2002). Third, one may want to consider the impact one or a few influential subjects may have on the model parameters. It is natural, at first sight, to make use of the specific influence assessment methodology that has been developed over the years (Cook, 1986; Chatterjee and Hadi, 1988). Applications of local influence analysis to the Diggle and Kenward (1994) model can be found in Verbeke *et al* (2001), Thijs, Molenberghs, and Verbeke (2000), and Molenberghs *et al* (2001). Similar ideas for the context of categorical longitudinal data have been developed in Van Steen *et al* (2001) and Jansen *et al* (2003). Hens *et al* (2004) proposed kernel weighted influence measures.

The original idea behind the use of local influence methods with an eye on sensitivity analysis was to detect observations that had a high impact on the conclusions *due to their aberrant missingness mechanism*. For example, most missing measurements might be MAR, while a few could be MNAR following one or a few deviating mechanisms. However, in most successful applications, where a seemingly MNAR mechanism turned out to be MAR or even MCAR after removing the influential subjects identified upon the use of local influence, the situation turned out to be more complex than anticipated. Indeed, the influential subjects often are influential for other than missingness related features. For example, in the mastitis dataset analyzed by Molenberghs *et al* (2001), the three influential cows had complete data but were identified by an extreme increase between the measurements at two subsequent years. Thijs, Molenberghs, and Verbeke (2000) observed similar behavior when local influence was augmented with global influence (case deletion) as well.

In this paper, we aim to further the study of the method of local influence, not only to better understand its behavior, but also to increase insight in the overall behavior and impact of MNAR mechanisms. This is done using

simulations and general modeling considerations.

In Section 2, a case study is introduced which motivates this work and is used throughout the paper. A general framework to model incomplete longitudinal data is sketched in Section 3, while sensitivity analysis tools, especially local influence, are described in Section 4. This methodology is applied to the rat case study in Section 5. Section 6 is dedicated to the behavior of local influence under standard conditions as well as under a number of anomalous scenarios and the likelihood ratio test statistic for the MNAR parameter in the Diggle and Kenward (1994) model is studied in Section 7.

2 Case Study

The data come from a randomized experiment, designed to study the effect of the inhibition of testosterone production in rats (Department of Orthodontics of the Catholic University of Leuven (K.U.L.) in Belgium; Verdonck *et al* (1997). A total of 50 male Wistar rats have been randomized to either control or one of two treatment groups (low or high dose of the drug Decapeptyl; an inhibitor for the testosterone production). The treatment started at the age of 45 days, and measurements were taken every 10 days, with the first observation taken at the age of 50 days. Our response is a characterization of the height of the skull, taken under anaesthesia. Many rats do not survive anaesthesia implying that for only 22 (44%) rats all 7 designed measurements could have been taken. The investigators' impression is that dropout is independent of the measurements.

The individual profiles are shown in Figure 1. To linearize, we use the logarithmic transformation $t = \ln(1 + (\text{age} - 45)/10)$ for the time scale. Let y_{ij} denote the jth measurement for the ith rat, taken at $t = t_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$. A simple statistical model, as considered by Verbeke *et al* (2001), then assumes that y_{ij} satisfies a model of the form (3.2) with common average intercept β_0 for all three groups, average slopes β_1, β_2 and β_3 for the three treatment groups, respectively, and assuming a so-called compound symmetry covariance structure, i.e., with common variance $\sigma^2 + \tau^2$ and common covariance τ^2.

3 Modeling Incomplete Longitudinal Data

Let us introduce the necessary notation. For subject i and occasion j, define $R_{ij} = 1$ if y_{ij} is observed and 0 otherwise. For convenience, partition $\boldsymbol{y}_i$ into two subvectors such that $\boldsymbol{y}_i^o$ is the vector containing those y_{ij} for which $R_{ij} = 1$

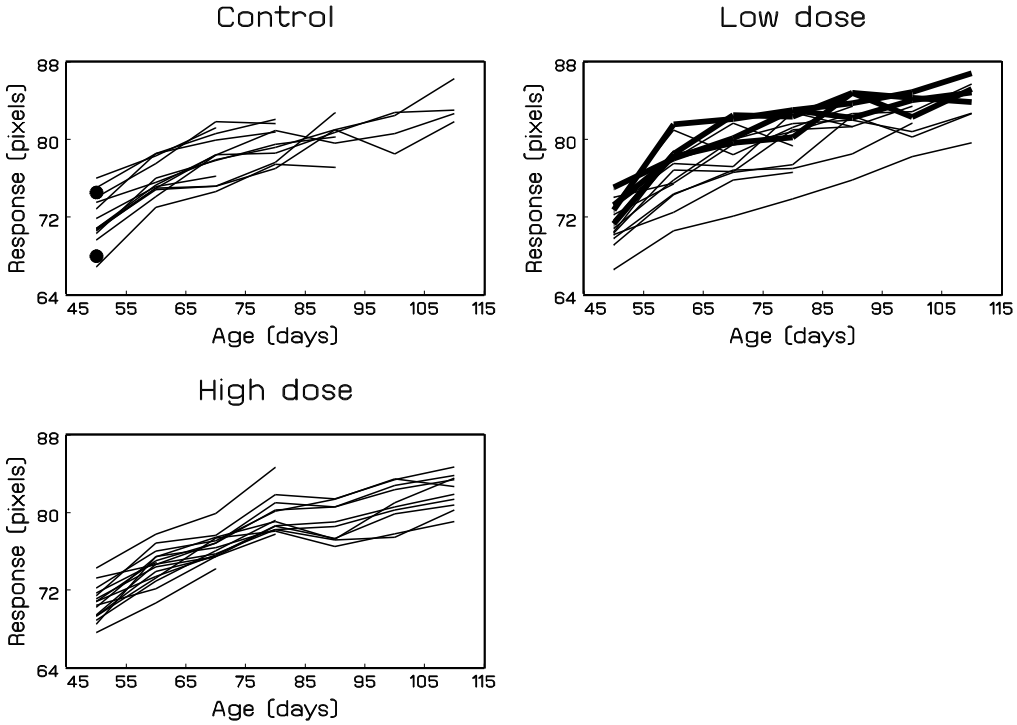

Fig. 1. *Individual growth curves for the three treatment groups separately. Influential subjects are highlighted.*

and $\boldsymbol{y}_i^m$ contains the remaining components. Statistical modeling begins by considering the full data density

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, Z_i, \boldsymbol{\theta}, \boldsymbol{\psi}),$$

where X_i and Z_i are the design matrices for fixed and random effects, respectively, and $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are vectors that parameterize the joint distribution.

A large class of models are based on the *selection model* factorization:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, Z_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}), \tag{3.1}$$

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. It is possible to have additional covariates in the missingness model, but this is suppressed from notation. An alternative taxonomy can be built based on so-called *pattern-mixture models* (Little, 1993, 1994) but these will not be considered within this paper.

Much of the early development of, and debate about, selection models appeared in the econometrics literature in which the tobit model (Heckman, 1976) played a central role. This combines a marginal Gaussian regression model for the response, as might be used in the absence of missing data, with a Gaussian-based threshold model for the probability of a value being missing.

The selection model of Diggle and Kenward (1994) is essentially a variation to this theme, combining the multivariate Gaussian linear model with a logistic dropout model. For the full likelihood analyses, subject-by-subject integration is required in general, unless MAR is assumed. This makes maximization somewhat cumbersome. Diggle and Kenward (1994) used the Nelder and Mead simplex algorithm (Nelder and Mead, 1965). However, such an approach lacks flexibility and is inefficient for high-dimensional problems. Alternatives are the EM algorithm and direct likelihood maximization.

Let us consider the Diggle and Kenward selection model in some more detail. They combine a linear mixed model (Laird and Ware, 1982) for the measurement process with a logistic regression model for the dropout process. The measurement model assumes that the vector $\boldsymbol{y_i}$ of repeated measurements for the ith subject satisfies the linear regression model

$$\boldsymbol{y_i} \sim N(X_i\boldsymbol{\beta}, V_i), \quad i = 1, \ldots, N \tag{3.2}$$

in which $\boldsymbol{\beta}$ is a vector of population-averaged regression coefficients called fixed effects, and where $V_i = Z_i G Z_i' + \Sigma_i$ (Verbeke and Molenberghs, 2000) for positive definite matrices G and Σ_i. The parameters in $\boldsymbol{\beta}$, G, and Σ_i are assembled into $\boldsymbol{\theta}$.

Since no data would be observed otherwise, we assume that the first measurement y_{i1} is obtained for every subject in the study. The model for the dropout process is based on a logistic regression for the probability of dropout at occasion j (let D_i be the occasion at which dropout occurs), given the subject was still in the study up to occasion j. We denote this probability by $g(\boldsymbol{h}_{ij}, y_{ij})$ in which $\boldsymbol{h}_{ij}$ is a subvector of the history $\widetilde{\boldsymbol{h}}_{ij}$, containing all responses observed up to but not including occasion j, as well as covariates. We assume

$$\begin{aligned} \text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] &= \text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{y}_i)\right] \\ &= \boldsymbol{h}_{ij}'\boldsymbol{\psi} + \omega y_{ij} \qquad\qquad i = 1, \ldots, N. \end{aligned} \tag{3.3}$$

In our case $\boldsymbol{h}_{ij}$ will contain the previous measurement $y_{i,j-1}$.

When ω equals zero and the model assumptions made are correct, the dropout model is random, and all parameters can be estimated using standard software since the measurement model and dropout model parameters can then be fitted separately. If $\omega \neq 0$, the dropout process is assumed to be non-random. Earlier, we pointed to the sensitivity of such an approach and a dropout model may be found to be non-random solely since one or a few influential subjects have driven the analysis. This concern will be taken up, as a starting point for the local influence based sensitivity analysis mode, in Section 4.2.

4 Sensitivity Analysis Tools

4.1 A General View On Sensitivity Assessment

We indicated in the introduction and at the end of the previous section that models for incomplete longitudinal data, especially the MNAR based ones, are vulnerable to model-assumption related sensitivity. Even when the linear mixed model would beyond any doubt be the choice of preference to describe the measurement process *should the data be complete*, then the analysis of the actually observed, incomplete version is, in addition, subject to further untestable modeling assumptions.

With the growing volume of MNAR based selection models, including Heckman (1976) and Diggle and Kenward (1994), the need for a careful understanding of such sensitivities, and the development of tools to discern their impact, has been growing as well (Glynn, Laird, and Rubin, 1986). Early, important contributions to sensitivity analysis have been made by Draper (1995), Copas and Li (1997), and Vach and Blettner (1995).

A strong conclusion, arising from most sensitivity analysis work, is that MNAR selection models have to be approached cautiously. This was made clear by several discussants of Diggle and Kenward (1994). This implies, for example, that formal tests for the null hypothesis of MAR versus the alternative of MNAR, should be approached with caution. Verbeke and Molenberghs (2000, Ch. 17) have shown, in the context of an onychomycosis study, that excluding a small amount of measurement error, drastically changes the likelihood ratio test statistics for the MAR null hypothesis. Kenward (1998) revisited the analysis of the mastitis data performed by Diggle and Kenward (1994). In this study, the milk yields of 107 cows were to be recorded during two consecutive years. While data were complete in the first year, 27 measurements were missing in year 2 because these cows developed mastitis and their milk yield was not of use anymore. While in the initial paper there was strong evidence for MNAR, Kenward (1998) showed that removing two out of 107 anomalous profiles, completely removed this evidence. Alternatively, Kenward showed that changing the conditional distribution of the year 2 yield, given the year 1 yield, from a normal to a heavy-tailed t, also led to the same result of no residual evidence for MNAR. This particular conditional distribution is of great importance, because a subject with missing data does not contribute to it, and hence, as we will illustrate later, is a source of sensitivity issues. Thus, fitting a MNAR model should be subject to careful scrutiny.

Such sensitivities to model assumptions have been reported for about two decades. See, for example, Nordheim (1984), Fitzmaurice, Molenberghs, and Lipsitz (1995), Molenberghs *et al* (1999), and Kenward and Molenberghs

(1999). In an attempt to formulate an answer to these concerns, a number of authors have proposed strategies to study sensitivity. We broadly distinguish between two types. A first family of approaches can be termed *substantive-driven* in the sense that they start from particularities of the problem at hand. Kenward's (1998) approach falls within this category. Arguably, such approaches are extremely useful in their own right and in preparation of use of the second family, where what could be termed *general purpose* tools are used.

We could define a sensitivity analysis as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools (such as diagnostic measures). This rather loose and very general definition encompasses a wide variety of useful approaches. The simplest procedure is to fit a selected number of (nonrandom) models which are all deemed plausible or one in which a preferred (primary) analysis is supplemented with a number of variations. The extent to which conclusions (inferences) are stable across such ranges provides an indication about the belief that can be put into them. Variations to a basic model can be constructed in different ways. The most obvious strategy is to consider various dependencies of the missing data process on the outcomes and/or on covariates. Alternatively, the distributional assumptions of the models can be changed.

Several authors have proposed to use local influence tools (Verbeke *et al*, 2001; Thijs, Molenberghs, and Verbeke, 2000; Molenberghs *et al*, 2001; Van Steen *et al*, 2001; Jansen *et al*, 2003), the method that will be considered in detail in the next section. In particular, Molenberghs *et al* (2001) revisited the mastitis example. They were able to identify the same two cows also found by Kenward (1998), in addition to another one. However, it is noteworthy that all three are cows with *complete* information, even though local influence methods were originally intended to identify subjects with other than MAR mechanisms of missingness. Thus, an important question is to what exactly are the sources causing an MNAR model to provide evidence for MNAR against MAR. There is some evidence to believe that a number of outlying aspects, but not necessarily the (outlying) nature of the missingness mechanism in one or a few subjects, is responsible for an apparent MNAR mechanism. In the next section, after briefly introducing local influence, this issue will be taken up.

Of course, it goes without saying that for other than selection models, e.g., pattern-mixture models, different sensitivity analysis tools need to be and have been developed (Thijs *et al*, 2002). These are interesting in their own right but are outside the scope of this paper.

Verbeke *et al* (2001), Thijs, Molenberghs, and Verbeke (2000), and Molenberghs *et al* (2001) investigated sensitivity of estimation of quantities of interest, such as treatment effect, growth parameters, or the dropout model parameters, w.r.t. assumptions regarding the dropout model. To this end, they considered the following perturbed version of dropout model (3.3):

$$\begin{aligned}\text{logit}(g(\boldsymbol{h}_{ij}, y_{ij})) &= \text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{y}_i)\right] \\ &= \boldsymbol{h}'_{ij}\boldsymbol{\psi} + \omega_i y_{ij} \qquad\qquad i = 1, \ldots, N, \end{aligned} \tag{4.1}$$

where the ω_i are local, individual-specific perturbations around a null model. They should not be confused with subject-specific parameters. Our null model will be the MAR model, corresponding to setting $\omega = 0$ in (3.3). Thus, the ω_i are perturbations that will be used only to derive influence measures (Cook, 1986).

Using this proposal, one can study the impact on key model features, induced by small perturbations in the direction, or seemingly so, of MNAR. This can practically be done by constructing local influence measures (Cook, 1986). When small perturbations in a specific ω_i lead to relatively large differences in the model parameters, this suggests that the subject is likely to drive key conclusions. For example, if such a subject would drive the model towards MNAR, then the conditional expectations of the unobserved measurements, given the observed ones, may deviate substantially from the ones under an MAR mechanism (Kenward, 1998). Such an observation is important also for our approach since then the impact (e.g., from influential subjects) on dropout model parameters extends to all functions that include these dropout parameters. One such function is the conditional expectation of the unobserved measurements, given the corresponding dropout pattern : $E(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, D_i, \boldsymbol{\theta}, \boldsymbol{\psi})$. As a consequence, the corresponding measurement model parameters will be affected as well.

We are interested in the influence exerted by the dropout model on the parameters of interest. This can be done, for example, by considering (4.1) as the dropout model. When small perturbations in a specific ω_i lead to relatively large differences in the model parameters, then this suggests that these subjects may have a large impact on the final analysis. However, even though we may be tempted to conclude that such subjects drop out non-randomly, this conclusion is misguided since we are not aiming to detect (groups of) subjects that drop out non-randomly but rather subjects that have a considerable impact on the dropout and measurement model parameters. Indeed, a key observation is that a subject that drives the conclusions towards MNAR may be doing so, not only because its true data generating mechanism is of an MNAR

type, but also for a wide variety of other reasons, such as an unusual mean profile or autocorrelation structure. Earlier analyses have shown that this may indeed be the case. Likewise, it is possible that subjects, deviating from the bulk of the data because they are generated under MNAR, go undetected by this technique. This begs the question that one needs to reflect carefully upon which anomalous features are typically detected and which ones typically go unnoticed.

Let us now introduce the key concepts of local influence (Cook, 1986). Since the resulting influence diagnostics can in many cases be expressed analytically, they often can be decomposed in interpretable components, which yields additional insight. We denote the log-likelihood function corresponding to model (4.1) by $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\gamma}|\omega_i)$, in which $\ell_i(\boldsymbol{\gamma}|\omega_i)$ is the contribution of the i^{th} individual to the log-likelihood, and where $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \boldsymbol{\psi})$ is the s-dimensional vector, grouping the parameters of the measurement model and the dropout model, not including the $N \times 1$ vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)'$ of weights defining the perturbation of the MAR model. It is assumed that $\boldsymbol{\omega}$ belongs to an open subset Ω of $\mathbb{R}^N$. For $\boldsymbol{\omega}$ equal to $\boldsymbol{\omega_0} = (0, 0, \ldots, 0)'$, $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega_0})$ is the log-likelihood function which corresponds to a MAR dropout model.

Let $\widehat{\boldsymbol{\gamma}}$ be the maximum likelihood estimator for $\boldsymbol{\gamma}$, obtained by maximizing $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega_0})$, and let $\widehat{\boldsymbol{\gamma}}_\omega$ denote the maximum likelihood estimator for $\boldsymbol{\gamma}$ under $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega})$. The local influence approach now compares $\widehat{\boldsymbol{\gamma}}_\omega$ with $\widehat{\boldsymbol{\gamma}}$. Strongly different estimates suggest that the estimation procedure is highly sensitive to such perturbations. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\gamma}}_\omega$ and $\widehat{\boldsymbol{\gamma}}$ by the so-called likelihood displacement, defined by $LD(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\gamma}}|\boldsymbol{\omega_0}) - \ell(\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\omega}}|\boldsymbol{\omega})]$. This takes into account the variability of $\widehat{\boldsymbol{\gamma}}$. Indeed, $LD(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega_0})$ is strongly curved at $\widehat{\boldsymbol{\gamma}}$, which means that $\boldsymbol{\gamma}$ is estimated with high precision, and small otherwise. Therefore, a graph of $LD(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ contains essential information on the influence of perturbations. It is useful to view this graph as the geometric surface formed by the values of the $N + 1$ dimensional vector $\boldsymbol{\xi}(\boldsymbol{\omega}) = (\boldsymbol{\omega}', LD(\boldsymbol{\omega}))'$ as $\boldsymbol{\omega}$ varies throughout Ω. Since this *influence graph* can only be depicted when $N = 2$, Cook (1986) proposed to look at local influence, i.e., at the normal curvatures $C_{\boldsymbol{h}}$ of $\boldsymbol{\xi}(\boldsymbol{\omega})$ in $\boldsymbol{\omega}_0$, in the direction of some N dimensional vector $\boldsymbol{h}$ of unit length. Let $\boldsymbol{\Delta}_i$ be the s dimensional vector defined by

$$\boldsymbol{\Delta}_i = \left. \frac{\partial^2 \ell_i(\boldsymbol{\gamma}|\omega_i)}{\partial \omega_i \partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}, \omega_i = 0} \tag{4.2}$$

and define Δ as the $(s \times N)$ matrix with $\boldsymbol{\Delta}_i$ as its i^{th} column. Further, let $\ddot{L}$ denote the $(s \times s)$ matrix of second order derivatives of $\ell(\boldsymbol{\gamma}|\boldsymbol{\omega}_0)$ with respect to $\boldsymbol{\gamma}$, also evaluated at $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}$. Cook (1986) has then shown that $C_{\boldsymbol{h}}$ can be easily calculated by $C_{\boldsymbol{h}} = 2|\boldsymbol{h}'\Delta'\ddot{L}^{-1}\Delta\boldsymbol{h}|$. Obviously, $C_{\boldsymbol{h}}$ can be

calculated for any direction $\boldsymbol{h}$. One evident choice is the vector $\boldsymbol{h_i}$ containing one in the i^{th} position and zero elsewhere, corresponding to the perturbation of the i^{th} weight only. This reflects the influence of allowing the i^{th} subject to drop out non-randomly, while the others can only drop out at random. The corresponding local influence measure, denoted by C_i, then becomes $C_i = 2|\boldsymbol{\Delta}_i' \ddot{L}^{-1} \boldsymbol{\Delta}_i|$. Another important direction is the direction $\boldsymbol{h}_{\max}$ of maximal normal curvature $C_{\max}$. It shows how to perturb the MAR model to obtain the largest local changes in the likelihood displacement. $C_{\max}$ is the largest eigenvalue of $-2\ \Delta'\ \ddot{L}^{-1}\ \Delta$ and $\boldsymbol{h}_{\max}$ is the corresponding eigenvector.

When a subset $\boldsymbol{\gamma}_1$ of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')'$ is of special interest, a similar approach can be used, replacing the log-likelihood by the profile log-likelihood for $\boldsymbol{\gamma}_1$, and the methods discussed above for the full parameter vector directly carry over (Lesaffre and Verbeke, 1998).

4.3 Applied to the Model of DK

Let us focus on (3.2) and (3.3). First note that the dropout mechanism is described by

$$f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = \begin{cases} \displaystyle\prod_{j=2}^{n_i} [1 - g(\boldsymbol{h}_{ij}, y_{ij})] & \text{for a completer } (d_i = n_i + 1), \\ \displaystyle\prod_{j=2}^{d-1} [1 - g(\boldsymbol{h}_{ij}, y_{ij})] g(\boldsymbol{h}_{id}, y_{id}) & \text{for a dropout } (d_i = d \leq n_i), \end{cases}$$

where the g-factors follow from (4.1). The log-likelihood contribution for a complete sequence then is

$$\ell_{i\omega} = \ln f(\boldsymbol{y}_i) + \ln f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi}),$$

where the parameter dependencies are suppressed for notational ease. The density $f(\boldsymbol{y}_i)$ is multivariate normal, following from the linear mixed model. The contribution from an incomplete sequence is more complicated. Its log-likelihood term is

$$\ell_{i\omega} = \ln f(y_{i1}, \ldots, y_{i,d-1}) + \sum_{j=2}^{d-1} \ln[1 - g(\boldsymbol{h}_{ij}, y_{ij})] + \ln \int f(y_{id}|y_{i1}, \ldots, y_{i,d-1}) g(\boldsymbol{h}_{id}, y_{id}) dy_{id}.$$

Further details can be found in Verbeke *et al* (2001). We need expressions for Δ and $\ddot{L}$. Straightforward derivation shows that the columns $\boldsymbol{\Delta}_i$ of Δ are given by

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\right|_{\omega_i=0} = \mathbf{0}, \tag{4.3}$$

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i}\right|_{\omega_i=0} = -\sum_{j=2}^{n_i} \boldsymbol{h}_{ij} y_{ij} g(\boldsymbol{h}_{ij})[1 - g(\boldsymbol{h}_{ij})], \tag{4.4}$$

for complete sequences (no drop out) and by

$$\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\right|_{\omega_i=0} = [1 - g(\boldsymbol{h}_{id})]\frac{\partial \lambda(y_{id}|\boldsymbol{h}_{id})}{\partial \boldsymbol{\theta}}, \tag{4.5}$$

$$\begin{aligned}\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i}\right|_{\omega_i=0} = &-\sum_{j=2}^{d-1} \boldsymbol{h}_{ij} y_{ij} g(\boldsymbol{h}_{ij})[1 - g(\boldsymbol{h}_{ij})] \\ &-\boldsymbol{h}_{id}\lambda(y_{id}|\boldsymbol{h}_{id})g(\boldsymbol{h}_{id})[1 - g(\boldsymbol{h}_{id})],\end{aligned} \tag{4.6}$$

for incomplete sequences. All above expressions are evaluated at $\widehat{\boldsymbol{\gamma}}$, and $g(\boldsymbol{h}_{ij}) = g(\boldsymbol{h}_{ij}, y_{ij})|_{\omega_i=0}$, is the MAR version of the dropout model. In (4.5), we make use of the conditional mean

$$\lambda(y_{id}|\boldsymbol{h}_{id}) = \lambda(y_{id}) + V_{i,21}V_{i,11}^{-1}[\boldsymbol{h}_{id} - \lambda(\boldsymbol{h}_{id})]. \tag{4.7}$$

The variance matrices follow from partitioning the responses as $(y_{i1}, \ldots, y_{i,d-1}|y_{id})'$.

The derivatives of (4.7) w.r.t. the measurement model parameters are

$$\begin{aligned}\frac{\partial \lambda(y_{id}|\boldsymbol{h}_{id})}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_{id} - V_{i,21}V_{i,11}^{-1}X_{i,(d-1)}, \\ \frac{\partial \lambda(y_{id}|\boldsymbol{h}_{id})}{\partial \alpha} &= \left[\frac{\partial V_{i,21}}{\partial \boldsymbol{\alpha}} - V_{i,21}V_{i,11}^{-1}\frac{\partial V_{i,11}}{\partial \boldsymbol{\alpha}}\right] V_{i,11}^{-1}[\boldsymbol{h}_{id} - \lambda(\boldsymbol{h}_{id})]\end{aligned}$$

where $\boldsymbol{x}_{id}'$ is the dth row of X_i, and where $X_{i,(d-1)}$ indicates the first $(d-1)$ rows X_i. Further, $\boldsymbol{\alpha}$ indicates the subvector of covariance parameters within the vector $\boldsymbol{\theta}$.

In practice, the parameter $\boldsymbol{\theta}$ in the measurement model is often of primary interest. Since $\ddot{L}$ is block-diagonal with blocks $\ddot{L}(\boldsymbol{\theta})$ and $\ddot{L}(\boldsymbol{\psi})$, we have that for any unit vector $\boldsymbol{h}$, $C_{\boldsymbol{h}}$ equals $C_{\boldsymbol{h}}(\boldsymbol{\theta}) + C_{\boldsymbol{h}}(\boldsymbol{\psi})$, with

$$C_{\boldsymbol{h}}(\boldsymbol{\theta}) = -2\boldsymbol{h}' \left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\right|_{\omega_i=0}\right]' \ddot{L}^{-1}(\boldsymbol{\theta}) \left[\left.\frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\theta} \partial \omega_i}\right|_{\omega_i=0}\right] \boldsymbol{h} \tag{4.8}$$

$$C_{\boldsymbol{h}}(\boldsymbol{\psi}) = -2\boldsymbol{h}' \left[\left. \frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i} \right|_{\omega_i=0} \right]' \ddot{L}^{-1}(\boldsymbol{\psi}) \left[\left. \frac{\partial^2 \ell_{i\omega}}{\partial \boldsymbol{\psi} \partial \omega_i} \right|_{\omega_i=0} \right] \boldsymbol{h}, \tag{4.9}$$

evaluated at $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}$. It now immediately follows from (4.3) and (4.5) that *direct* influence on $\boldsymbol{\theta}$ only arises from those measurement occasions at which dropout occurs. In particular, from (4.5) it is clear that the corresponding contribution is large only if (1) the dropout probability was small but the subject disappeared nevertheless and (2) the conditional mean 'strongly depends' on the parameter of interest. This implies that complete sequences cannot be influential in the strict sense ($C_i(\boldsymbol{\theta}) = 0$) and that incomplete sequences only contribute, in a direct fashion, at the actual dropout time. However, we make an important distinction between direct and indirect influence. It was shown that complete sequences can have an impact by changing the conditional expectation of the unobserved measurements given the observed ones *and given the dropout mechanism*. Thus, a complete observation which has a strong impact on the *dropout model parameters*, can still drastically change the measurement model parameters and functions thereof.

Expressions (4.8)–(4.9) can be simplified further in specific cases. For example, Verbeke *et al* (2001) considered the compound-symmetric situation. Precisely, they were able to split the overal influence in the approximate sum of three components, describing the mean model parameter $\boldsymbol{\beta}$, the variance components σ^2 and τ^2, and the dropout model parameters $\boldsymbol{\psi}$, respectively:

$$\begin{aligned} C_i^{\text{ap}}(\boldsymbol{\beta}) = {} & 2[1 - g(\boldsymbol{h}_{id})]^2 (\xi_{id}\boldsymbol{x}_{id} + (1-\xi_{id})\boldsymbol{\rho}_{id})' \\ & \times \sigma^2 \left[\sum_{i=1}^{N} \left(\xi_{id} X'_{i(d-1)} X_{i(d-1)} + (1-\xi_{id}) R'_{i(d-1)} R_{i(d-1)} \right) \right]^{-1} \\ & \times (\xi_{id}\boldsymbol{x}_{id} + (1-\xi_{id})\boldsymbol{\rho}_{id}), \end{aligned} \tag{4.10}$$

$$\begin{aligned} C_i^{\text{ap}}(\sigma^2, \tau^2) = {} & 2[1 - g(\boldsymbol{h}_{id})]^2 \xi_{id}^2 (1-\xi_{id})^2 \left[\boldsymbol{h}_{id} - \widetilde{\lambda}(\boldsymbol{h}_{id}) \right]^2 \\ & \times \left(-1, \frac{1}{\tau^2} \right) \ddot{L}^{-1}(\sigma^2, \tau^2) \begin{pmatrix} -1 \\ \frac{1}{\tau^2} \end{pmatrix}, \end{aligned} \tag{4.11}$$

$$\begin{aligned} C_i(\boldsymbol{\psi}) = {} & 2 \left(\sum_{j=2}^{d} \boldsymbol{h}_{ij} y_{ij} v_{ij} \right)' \left(\sum_{i=1}^{N} \sum_{j=2}^{d} v_{ij} \boldsymbol{h}_{ij} \boldsymbol{h}'_{ij} \right)^{-1} \\ & \times \left(\sum_{j=2}^{d} \boldsymbol{h}_{ij} y_{ij} v_{ij} \right), \end{aligned} \tag{4.12}$$

where $R_{i,d-1} = X_{i(d-1)} - \mathbf{1}_{d-1} \widetilde{X_{i(d-1)}}$, $\widetilde{X_{i(d-1)}} = \frac{1}{d-1} \mathbf{1}'_{d-1} X_{i(d-1)}$,

$$[\boldsymbol{h}_{id} - \widetilde{\lambda}(\boldsymbol{h}_{id})] = \frac{1}{d-1} \mathbf{1}'_{d-1} [\boldsymbol{h}_{id} - \lambda(\boldsymbol{h}_{id})],$$

$$\ddot{L}(\sigma^2,\tau^2) = \sum_{i=1}^{N} \frac{d-1}{2(\sigma^2+(d-1)\tau^2)^2} \times \begin{pmatrix} [\sigma^2+(d-1)\tau^2]^2 - \tau^2[2\sigma^2+(d-1)\tau^2] & 1 \\ 1 & (d-1) \end{pmatrix},$$

$d = n_i$ for a complete case and where y_{id} needs to be replaced with

$$\lambda(y_{id}|\boldsymbol{h}_{id}) = \lambda(y_{id}) + (1-\xi_{id})[\boldsymbol{h}_{id} - \widetilde{\lambda}(\boldsymbol{h}_{id})]$$

for incomplete sequences. Further, v_{ij} equals $g(h_{ij})[1-g(h_{ij})]$ which is the variance of the estimated dropout probability under MAR.

5 Analysis and Sensitivity Analysis of the Rat Data

The rat data, which are introduced in Section 2, are analyzed using model (3.2) with the following specific version of dropout model (3.3):

$$\text{logit}\left[\text{pr}(D_i = j|D_i \geq j, \boldsymbol{y}_i)\right] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{5.1}$$

Parameter estimates are shown in Table 1. More details about these estimates and the performance of a local influence analysis can be found in Verbeke *et al* (2001). This section will focus on specific details of this local influence analysis.

Figure 2 displays overall C_i and influences for subvectors θ, β, α, and ψ. In addition, the direction $\boldsymbol{h}_{\max}$ corresponding to maximal local influence is given. Apart from the last one of these graphs, the scales are not unitless and therefore it would be hard to use a common one for all of the panels. This implies that the main emphasis should be on *relative* magnitudes.

The largest C_i are observed for rats #10, #16, #35, and #41, and virtually the same picture holds for $C_i(\boldsymbol{\psi})$. They are highlighted in Figure 1. All four belong to the low dose group. Arguably, their relatively large influence is caused by an interplay of three facts. First, the profiles are relatively high, and hence y_{ij} and h_{ij} in (4.12) are large. Second, since all four profiles are complete, the first factor in (4.12) contains a maximal number of large terms. Third, the computed v_{ij} are relatively large.

Turning attention to $C_i(\boldsymbol{\alpha})$ reveals peaks for rats #5 and #23. Both belong to the control group and drop out after a single measurement occasion. They are highlighted in the first panel of Figure 1. To explain this, observe that the relative magnitude of $C_i(\boldsymbol{\alpha})$, approximately given by (4.12), is determined

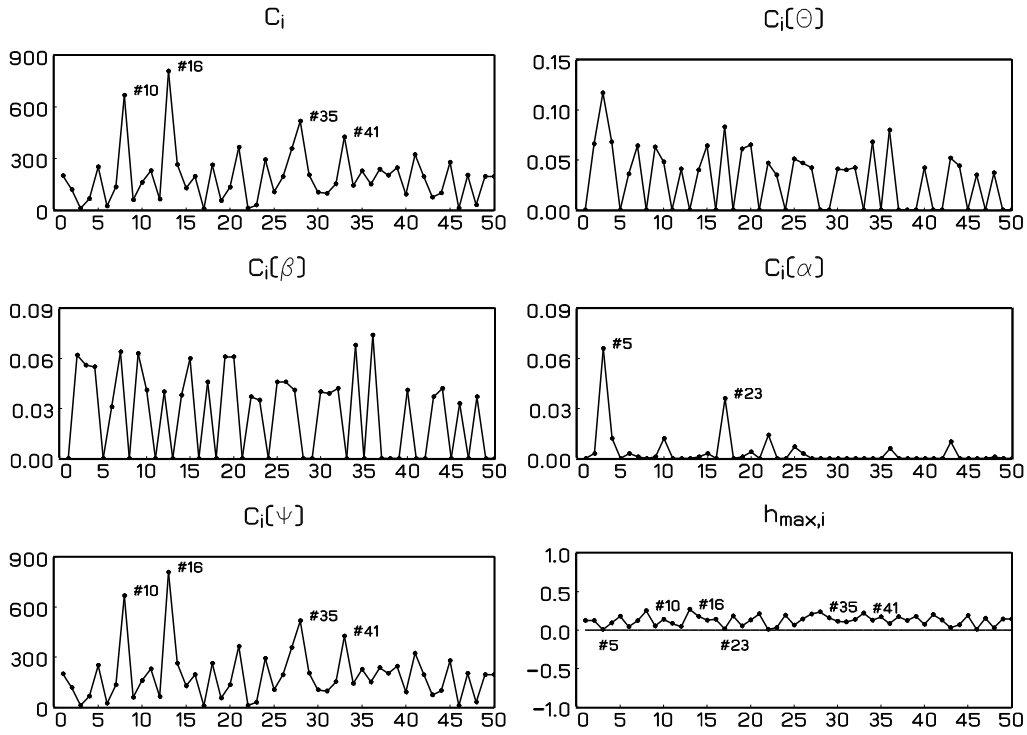

Fig. 2. *Index plots of* C_i, $C_i(\boldsymbol{\theta})$, $C_i(\boldsymbol{\beta})$, $C_i(\boldsymbol{\alpha})$, $C_i(\boldsymbol{\psi})$, *and of the components of the direction* $\boldsymbol{h}_{\max}$ *of maximal curvature.*

by $1 - g(h_{id})$ and $h_{id} - \lambda(h_{id})$. The first term is large when the probability of dropout is small. Now, when dropout occurs early in the sequence, the measurements are still relatively low, implying that the dropout probability is rather small (cf. Table 1). This feature is built into the model by writing the dropout probability in terms of the raw measurements with time-independent coefficients rather than, for example, in terms of residuals. Further, the residual $h_{id} - \lambda(h_{id})$ is large since these two rats are somewhat distant from the group by time mean. A practical implication of this is that the time-constant nature of the dropout model may be unlikely to hold. Therefore, a time-varying version was considered, where the logit of the dropout model takes form $\psi_0 + \psi_1 y_{i,j-1} + \nu_0 t_{ij} + \nu_1 t_{ij} y_{i,j-1}$. There is overwhelming evidence in favor of such a more elaborate MAR model (likelihood ratio statistic of 167.4 on 2 degrees of freedom). Thus, local influence can be used to call into question the posited MAR (and MNAR) models, and to guide further selection of more elaborate, perhaps MAR, models.

Since all deviations are rather moderate, we further explore our approach by considering a second analysis where all responses for rats #10, #16, #35, and #41 have been increased with 20 units. The effect of this distortion will primarily be seen in the variance structure. Precisely, such a change is likely to inflate the random intercept variance, at the expense of the other variance components. In doing so, we will illustrate that (1) such a change is likely to show up in the assessment of the dropout model, underscoring the sensitivity

Table 1
Maximum likelihood estimates (standard errors) of completeley random, random and non-random dropout models, fitted to the rat data set, with and without modification.

Original Data				
Effect	Parameter	MCAR	MAR	MNAR
Measurement model:				
Intercept	β_0	68.61 (0.33)	68.61 (0.33)	68.60 (0.33)
Slope control	β_1	7.51 (0.22)	7.51 (0.22)	7.53 (0.24)
Slope low dose	β_2	6.87 (0.23)	6.87 (0.23)	6.89 (0.23)
Slope high dose	β_3	7.31 (0.28)	7.31 (0.28)	7.35 (0.30)
Random intercept	τ^2	3.44 (0.77)	3.44 (0.77)	3.43 (0.77)
Measurement error	σ^2	1.43 (0.14)	1.43 (0.14)	1.43 (0.14)
Dropout model:				
Intercept	ψ_0	−1.98 (0.20)	−8.48 (4.00)	−10.30 (6.88)
Prev. measurement	ψ_1		0.08 (0.05)	0.03 (0.16)
Curr. measurement	ψ_2			0.07 (0.22)
-2 loglikelihood		1100.4	1097.6	1097.5
Modified Data				
Effect	Parameter	MCAR	MAR	MNAR
Measurement model:				
Intercept	β_0	70.20 (0.92)	70.20 (0.92)	70.25 (0.92)
Slope control	β_1	7.52 (0.25)	7.52 (0.25)	7.42 (0.26)
Slope low dose	β_2	6.97 (0.25)	6.97 (0.25)	6.90 (0.25)
Slope high dose	β_3	7.21 (0.31)	7.21 (0.31)	7.04 (0.33)
Random intercept	τ^2	40.38 (0.18)	40.38 (0.18)	40.71 (8.25)
Measurement error	σ^2	1.42 (0.14)	1.42 (0.14)	1.44 (0.15)
Dropout model:				
Intercept	ψ_0	−1.98 (0.20)	-0.79 (1.99)	2.08 (3.08)
Prev. measurement	ψ_1		−0.015 (0.03)	0.23 (0.15)
Curr. measurement	ψ_2			−0.28 (0.17)
-2 loglikelihood		1218.0	1217.7	1214.8

and that (2) the local influence approach is able to detect such an effect. The parameter estimates for all three models are also shown in Table 1. Clearly, while the fixed-effect parameters remain virtually unchanged, the random intercept parameter has, of course, drastically increased. Likewise, the dropout parameters are affected. In addition, the likelihood ratio statistic for MAR versus MCAR changes from 2.8 to 0.3 and for MNAR versus MAR changes from 0.1 to 2.9. Thus, the evidence has shifted from the first to the second test. While all of these statistics seem to be non-significant, there is an important qualitative effect. Moreover, as discussed in Section 7, the use of the classical χ^2-distribution is very questionable for testing MNAR.

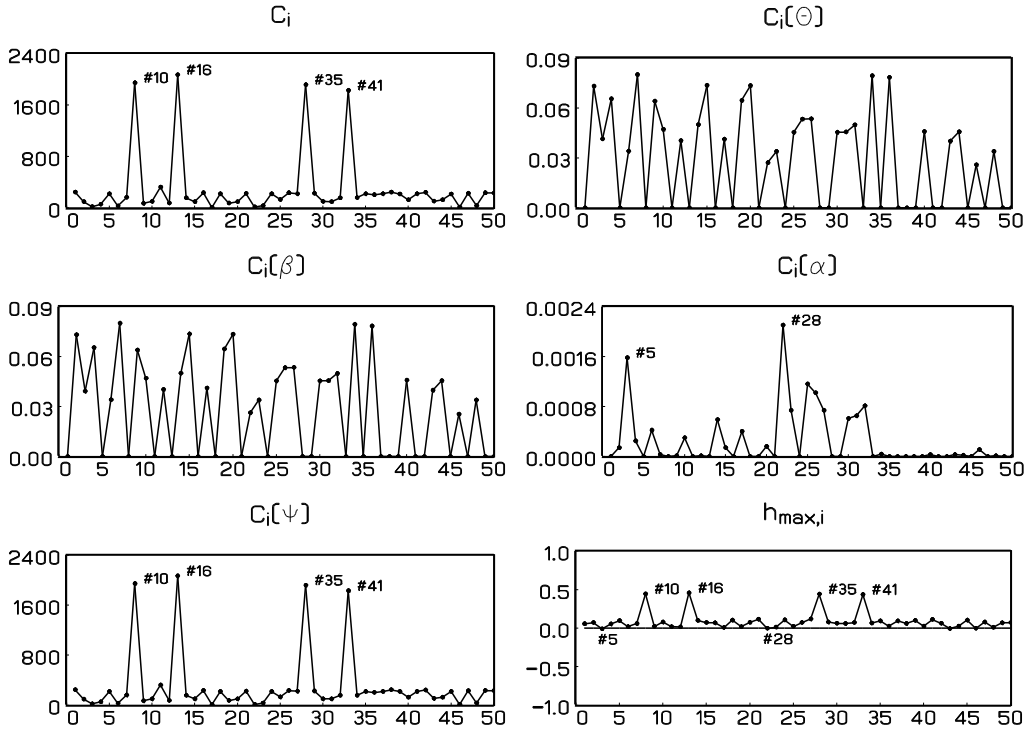

Fig. 3. *Index plots of* C_i, $C_i(\boldsymbol{\theta})$, $C_i(\boldsymbol{\beta})$, $C_i(\boldsymbol{\alpha})$, $C_i(\boldsymbol{\psi})$, *and of the components of the direction* $\boldsymbol{h}_{\max}$ *of maximal curvature, where 4 profiles have been shifted upward.*

In order to check whether these findings are recovered by the local influence approach, let us study Figure 3. In line with the changes in parameter estimates, $C_i(\boldsymbol{\beta})$ shows no peaks in these observations but peaks in $C_i(\boldsymbol{\alpha})$ and $C_i(\boldsymbol{\psi})$ indicate a relatively strong influence from the four extreme profiles.

It will be clear from the above that subjects may turn out to be influential, for reasons different from the nature of the dropout model. Indeed, increasing the profile by 20 units primarily changes the level of the random intercept and ultimately changes the form of the random-effects distribution. Nevertheless, this feature shows in our local influence analysis, where the perturbation is put into the dropout model and not, for example, in the measurement model. This feature requires careful study and will be addressed in the next section.

6 Local Influence Methods and Their Behavior

A number of concerns have been raised, not only about sensitivity, but also about the tools used to assess sensitivity themselves. For example, Verbeke *et al* (2001) noted, based on a case study, that the local influence tool, as defined and implemented in Section 4.2, is able to pick up anomalous features of study subjects that are not necessarily related to the missingness mechanism. In particular, they found that subjects with an unusually high profile, or a

somewhat atypical serial correlation behavior, are detected with the local influence tool. At first sight, this is a little disconcerting, since the ω_i parameter in (4.1) is placed in the dropout model and not in the measurement model, necessitating further investigation regarding which effects are easy or difficult to detect with these local influence methods.

We aim to gain more insight into this phenomenon in a number of ways. To this effect, we undertake a targeted simulation study to explore various sources of influence. First, we took interest in the relative magnitudes of the influence measures to assess how feasible it is to separate influence values that are in line with regular behavior from those that are unduly large. This can be done by proposing a rule of thumb as well as by constructing sampling-based confidence limits and bounds. Second, the impact of one or a few subjects with an anomalous dropout mechanism was explored. Such anomalies are of the type one would intuitively expect to be picked up by the proposed tool. We illustrate that great care is needed. Third, impact due to anomalies in the measurement model was studied. We will show that precisely such anomalies are relatively easily picked up by the tool, in spite of its conception for anomalies in the missingness mechanism. We offer an explanation for why such behaviour is seen, which is then backed up with some complementary considerations in Section 7.

6.1 The Effect of Sample Size

Lesaffre and Verbeke (1998) applied local influence methods to the classical linear mixed-effects model. They introduced ω_i parameters as follows: $\ell = \sum_i \omega_i \ell_i$ where ℓ_i is the log-likelihood contribution for subject i. They were able to show that the sum of the influences is approximately equal to $2N$ with N the sample size. Their result is based on the fact that, in their local influence contributions, Δ_i in (4.2) becomes

$$\Delta_i = \frac{\partial \ell_i}{\partial \boldsymbol{\theta}},$$

so that the entire expression has the flavor of a contribution to the score test. In our case, as can be seen from (4.5) and (4.6), Δ_i is a second rather than a first derivative of the log-likelihood contributions, implying that a, perhaps linear, dependence on the sample size could be envisaged. Such a calibration would be beneficial since it would allow to determine critical values, or at least rules of thumb, to determine what is large enough for a subject's influence to undergo further scrutiny.

To this end, we generated a number of datasets, all under the assumption of MAR and with parameters equal to the ones from the rat example. The only difference between these simulations was the sample size. Selected quantiles

Table 2
Selected quantiles of the local influence measures for datasets of different sample sizes, as obtained from simulations and after fitting a simple empirical model.

	Simulated results						Empirical model		
sample size	50	50	50	500	500	1000	50	500	1000
median	176.7	138.7	146.8	16.5	15.6	7.6	150.0	15.0	7.5
95 percentile	384.5	359.9	317.3	40.6	39.5	18.7	359.4	39.4	20.3
maximum	683.7	674.1	950.7	138.0	137.8	53.6	—	—	—

for sample sizes 50, 500, and 1000 are shown in Table 2. Studying even larger sample sizes would be faced with increasing computation times. This also is the reason for considering a single run at size 1000. While the relationship is less clear, as is to be expected, for the maximum value, an obvious trend is seen in the median values and in the 95th percentile. We indeed notice that the influence for a subject decreases linearly with sample size, and hence the total influence for a dataset is roughly constant. This is confirmed by a simple multiplicative regression model, which yields that the product of the median and the sample size is constant and equal to 7500. Similarly, the product of the 95th percentile and the sample size to the power 0.96 equals 15,367. To ensure calibration at the individual level, one could then multiply all influences by the sample size. This calibration result implies that the rescaled local influence can be used as a rough measure to determine whether large values are present. For example, one could investigate subjects for which the influence exceeds $1/N$ of the calibrated total with a certain amount. However, while useful in its own right, we still do not learn anything about the actual distribution of a local influence profile under the null hypothesis. To gain further insight into this problem, we will derive confidence limits and simultaneous confidence bounds in the next section.

6.2 Pointwise Confidence Limits and Simultaneous Confidence Bounds for the Local Influence Measure

Since for practical purposes only high values of the influence measures are of interest, we will focus on one-sided (upper) limits and bounds. To this end, we simulated 1000 datasets of 50 rats, using the parameters of the MAR model. To have a consistent ordering of the C_i values, not based on the arbitrary order of the rats within the set of data, we sorted them from large to small.

This can be seen from, say, 1000 repetitions of a bootstrap experiment with 50 grid points. At each grid point, the 95% pointwise upper confidence limit then simply is the 95% quantile of the $C_{i,j}$ values at that particular grid point. Construction of the simultaneous confidence bounds is based on Besag *et al* (1995). For each grid point j, order the $C_{i,j}$ values to obtain order statistics $C_{i,j}^{[t]}$ and their corresponding ranks $r_j^{(t)}, t = 1, ..., 1000$. Next, for fixed k, define

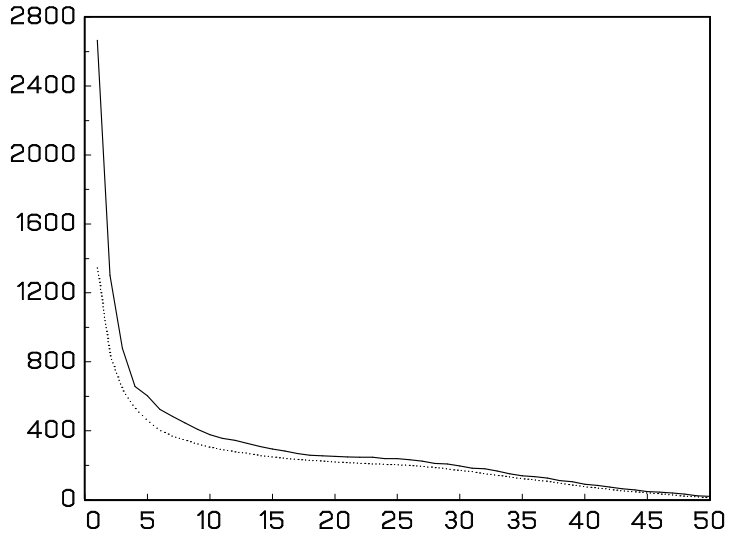

Fig. 4. 95% *pointwise upper confidence limit (dotted) and* 95% *simultaneous upper confidence bound (solid).*

t_k as the k-th order statistic of the set

$$\left\{\max\left(\max_{1\leq j\leq 50} r_j^{(t)}; 1001 - \min_{1\leq j\leq 50} r_j^{(t)}\right); t = 1, ..., 1000\right\}.$$

Then, by construction, the intervals

$$\left\{\left[C_{i,j}^{[1001-t_k]}; C_{i,j}^{[t_k]}\right]; j = 1, ..., 50\right\}$$

have a global confidence level of at least $100(k/1000)\%$. To obtain the 95% simultaneous upper confidence bound, simply take $k = 900$, and restrict consideration to the upper bound $C_{i,j}^{[t_k]}$. A graphical representation of this result is given in Figure 4.

6.3 The Effect of Anomalies in the Missingness Mechanism

To get an idea of the effect of anomalies in the dropout mechanism, a general procedure was followed, as described next. Generate an MNAR dataset, fit these data assuming an MAR mechanism in model (3.3), and use the estimates of those model parameters to generate 1000 datasets, which are then used to construct the pointwise confidence limits and simultaneous confidence bounds as outlined in Section 6.2. Afterwards, add the profile of ordered C_i values from the original MNAR dataset on the graph with the pointwise confidence limits and simultaneous confidence bounds. Several different settings to generate the MNAR dataset were explored, and will be discussed in the remainder of this section.

First, attention is paid to the creation of MNAR based on the model parameters. This was done in the following ways: (1) a set of 50 rats were generated

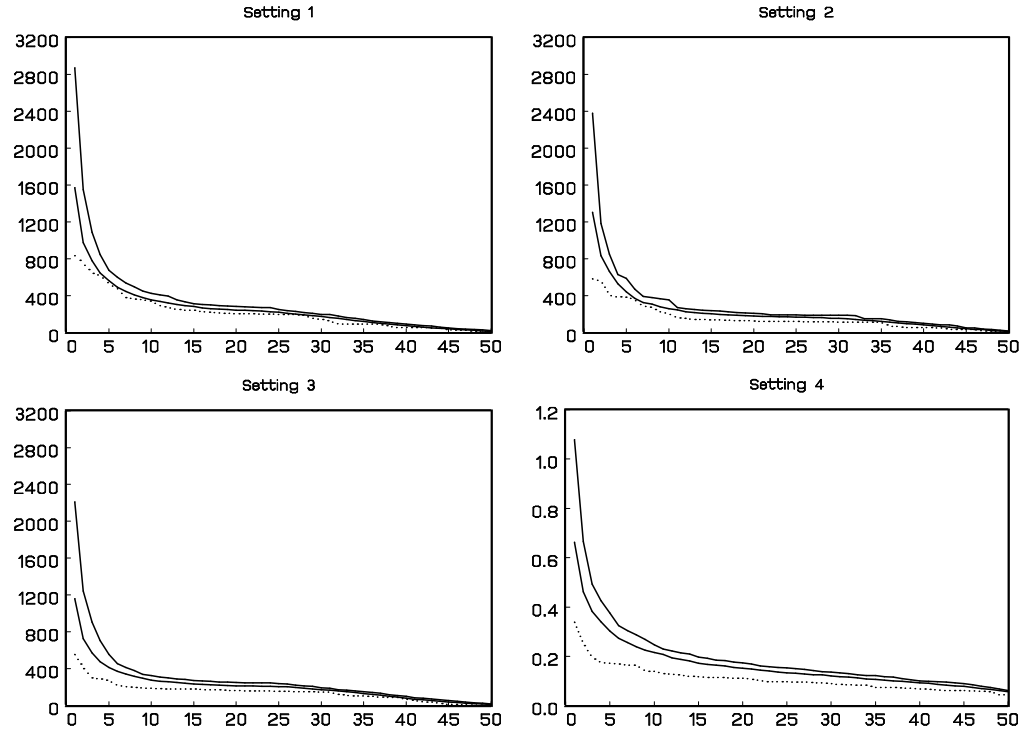

Fig. 5. *Graphical representation of the profiles of different parameter-based MNAR settings (dotted), compared with the* 95% *pointwise upper confidence limit and* 95% *simultaneous upper confidence bound (solid).*

using the MNAR parameters from the original rats data, as presented in the upper part of Table 1; (2) the same parameters were used, except for ψ_2, which was increased to 0.5. (3) only 10% of the data (equivalently to 5 rats) were generated taking $\psi_2 = 0.2$, while for the remaining 90% of the data (45 rats) $\psi_2 = 0$; (4) using the incremental parameterization introduced in Thijs, Molenberghs, and Verbeke (2000), 10% of the rats were generated with $\lambda_2 = 0.2$, and the other 90% with $\lambda_2 = 0$.

All different settings of these simulations were repeated several times, but, since they all gave similar results, and in the interest of space, for each setting only one result is discussed and presented in Figures 5 and 7.

A general trend is observed in settings (1) to (4). The C_i profile of the MNAR dataset crosses neither the 95% pointwise upper confidence limit nor the simultaneous upper confidence bound for large values of C_i. On the other hand, in some settings they cross the 95% pointwise upper confidence limit for small values of C_i (near the end of the profile), but since we are only interested in highly influential subjects, this result is irrelevant for our purposes. Taking a closer look at the rats for whom $\psi_2 \neq 0$ (settings 3 and 4) we note that their C_i values are very small (all within the 10 lowest values). We can therefore conclude that this type of MNAR is not detectable using local influence. Note that the limit and bound for setting (2) is more ragged than for the others.

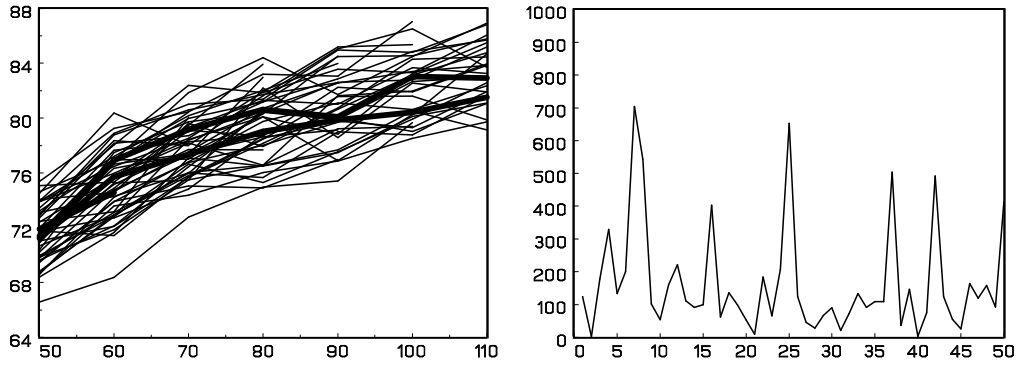

Fig. 6. *Individual growth curves (left panel, subjects 3, 21 and 26 highlighted) and C_i profile (right panel) of the MAR dataset which will be manipulated.*

The reason is that convergence is quite difficult to obtain for this setting, in line with general convergence problems for situations where ψ_2 is substantially different from zero. The scale for setting (4) in Figure 5 is completely different from the scale in the other settings, which is due to the fact that setting (4) considers the effect of the difference between the current and the previous measurement on the dropout process, rather than the raw effect of the current measurement in the other three settings.

Alternatively, in a second round of settings, MNAR was created in a deterministic way. Therefore a dataset is generated using the MAR parameters of the original rats data in Table 1 (the C_i profile of this dataset is shown in Figure 6). Afterwards, the MNAR part was created by manually deleting values from some profiles as follows: (5) all values of skull height from the moment that one of them exceeded 86 mm; (6) all values of skull height from the moment that one of them exceeded 85 mm; (7) second to last values of skull height if the value at age 60 days (2nd value) exceeded 78.83 mm (95^{th} percentile); (8) third to last values of skull height if the value at age 70 days (3rd value) exceeded 80.82 mm (95^{th} percentile).

Our interest is now in seeing how such sets of data give qualitatively different influence graphs than under the original rat dataset. Therefore, we have to proceed somewhat differently from the simulation study done for settings (1)–(4). We now rather directly compare the influence graphs from the four settings (5)–(8) with the original one. Under setting (5), we see that the peak for rat #25, seen in the original analysis, is removed, while others pop up, in the sense that some moderate peaks now become the largest ones. However, no rat really sticks in such a way that further investigation would need to be undetaken. It is noteworthy though, that those whose profiles have been shortened due to the action described under setting (5), have, as a consequence, a smaller influence value. A similar phenomenon has been observed in Jansen *et al* (2003) for categorical responses. Settings (6)–(8) are similar in qualitative terms, even though the phenomena are tiny bit more extreme in setting (6) than in setting (5).

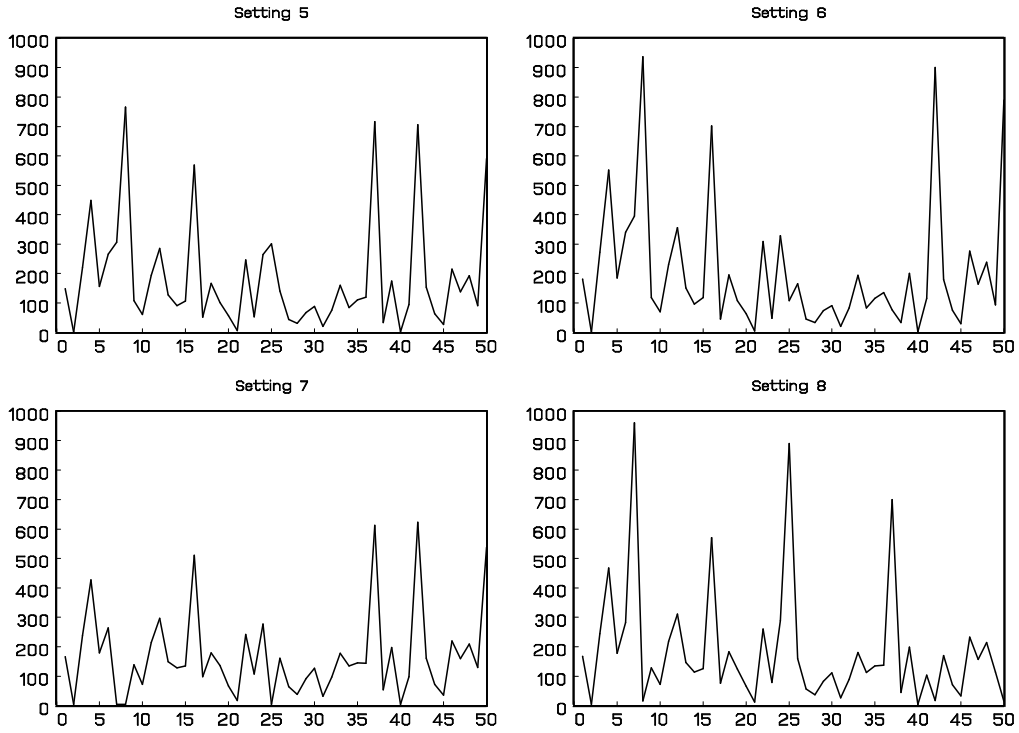

Fig. 7. *Graphical representation of the unordered C_i profiles of different settings with manually created anomalies in the missingness model.*

6.4 The Effect of Anomalies in the Measurement Model

In this section, we shifted attention to anomalies in the measurement model. We generated 4 MAR datasets, each of them with its specific changes to the measurement model for 3 randomly selected rats, namely (1) an increased mean profile by 20 units *after* the dropout probability was calculated, (2) an increased mean profile by 20 units *before* the dropout probability was calculated, (3) an increased variance component by 20 units, and (4) an increased τ^2 (covariance for the compound symmetry) by 20 units. The starting dataset without any changes to the measurement model is the same as was used in Section 6.3, Figure 6.

While settings (3) and (4), focusing on the variance-covariance structure, show virtually no impact (Figure 8), settings (1) and (2) exhibit a dramatic effect. The impact is larger in setting (2) because there also the dropout model is affected. In both settings, rats #3 and #26 clearly stick out, while with differing relative magnitudes. The effect of rat #21 is negligible. These results can be explained by taking a closer look at the individual profiles of those rats. Figure 9 shows that in setting (1) rat #21 has only 2 observations, while rats #3 and #26 have complete profiles. In setting (2), the profile of rat #21 reduces to only one observation, which explains the negligible influence, and the profiles of rats #3 and #26 reduce to 6 and 3 measurements, respectively. Previous conclusions again indicate that shortened profiles tend to give smaller

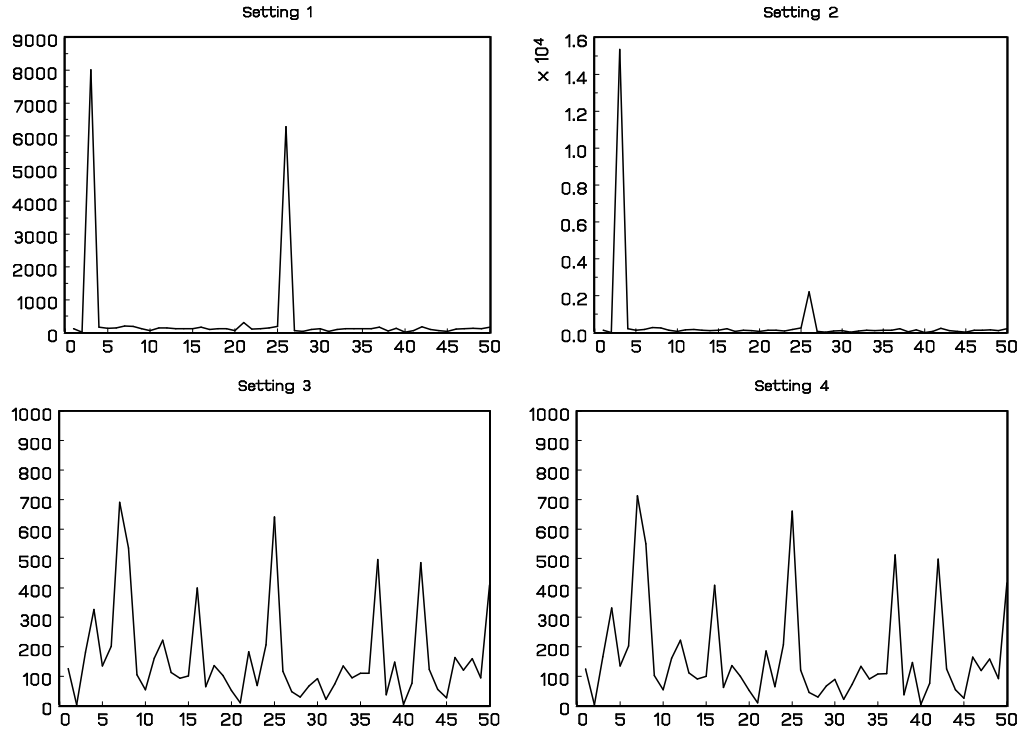

Fig. 8. *Graphical representation of the unordered C_i profiles of different settings with anomalies in the measurement model.*

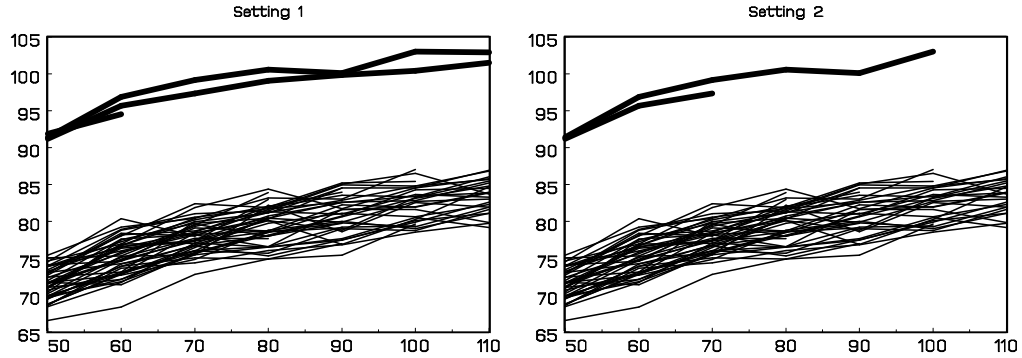

Fig. 9. *Individual growth curves for settings 1 and 2, where the mean profile is increased, before of after the dropout probability was calculated. Subjects 3, 21 and 26 are highlighted.*

influence values (Jansen *et al*, 2003).

6.5 Discussion of Results

These results indicate that there is little or no local influence stemming from having a few subjects that drop out in a non-random way, by setting ψ_2 for these equal to a nonzero value, while there is considerable influence in a number of settings where the measurement model is changed in the sense that a few profiles following a deviating mean-model structure. This indicates that

the non-random parameter ψ_2, rather than capturing true MNAR missingness, has a strong tendency to pick up other deviations, primarily in the measurement model. Many authors have noted that there is very little information in many sets of data for the parameter ψ_2, in addition to the information available for all other parameters. If this were to be true, this ought to show in the behavior of the likelihood ratio test statistic for ψ_2, as well as in the structure of the information matrix for the vector of model parameters. We will explore this further in the next section.

7 Behavior of the Likelihood Ratio Test for MAR versus MNAR

In this section we report a simulation study designed to examine the finite sample behaviour of the likelihood ratio test (LRT) for testing MAR versus MNAR within the selection model framework of Diggle and Kenward (1994). For comparison we also consider the test for MCAR versus MAR. The behaviour of a parametric and a semi-parametric bootstrap approach in this context is also investigated.

It follows from standard theory that the LRT for MCAR versus MAR has an approximate χ^2_1 distribution but the test for MNAR versus MAR is a nonstandard situation. Indeed, Rotnitzky *et al* (2000) have proven that for a similar but simpler setting the limiting distribution is a χ^2-mixture with characteristics governed by the singularity properties of the information matrix. The score equation associated with the MNAR parameter ψ_2 apparently generates a quasi-linear dependence structure in the system of score equations. When reducing the model to a MAR/MCAR model, this dependency disappears. They moreover have shown that convergence to this limiting distribution is extremely slow.

Similar theoretical considerations show that the same phenomena hold for the model of Diggle and Kenward (1994). The slow convergence also raises the question whether, even if known, an asymptotic distribution is of any practical use. This was our motivation to examine the finite sample properties of the LRT in this setting and to investigate whether a bootstrap simulated null distribution, known to be a (slightly) better approximation in several classical settings, could be a useful alternative to χ^2 based distributions.

The simulation settings are as follows: the measurement model (3.2) with $N = 200$, $n_i = 3$, $X_i = 1$ (intercept only), mean vector $\beta = (2, 0, -2)'$ and compound-symmetric covariance structure with common variance equal to 8 and common covariance equal to 6; the missingness model given by (5.1).

As mentioned before, consider the following hypotheses: (1) $\psi_1 = 0$ in model (5.1) with $\psi_2 = 0$, i.e., MCAR vs MAR, (2) $\psi_2 = 0$ in model (5.1) with $\psi_1 \neq 0$,

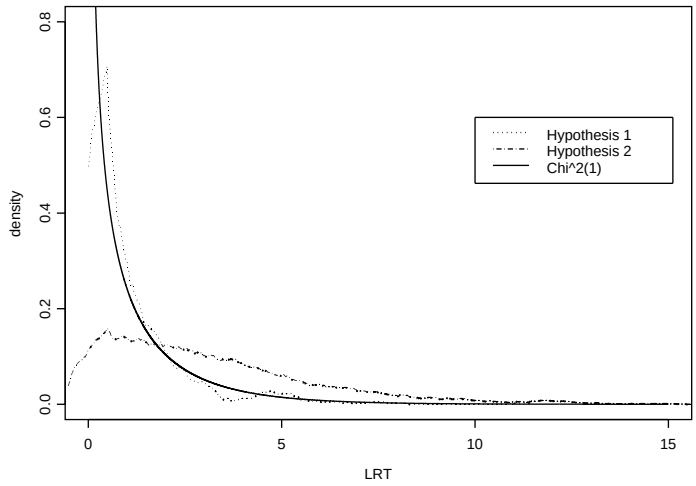

Fig. 10. *Kernel density plots of the simulated null distribution based on 800 samples, under each Hypothesis* $i = 1, 2$*, together with the density of the* χ^2_1*-distribution.*

i.e. MAR vs MNAR. When not put to zero, we took $\psi_0 = -2, \psi_1 = 1, \psi_2 = 2$.

7.1 Simulated Null Distributions

Figure 10, based on 800 samples generated under each of the two null hypotheses, shows the simulated null distribution of the likelihood ratio test.

As expected, the simulated null distribution deviates more pronounced from the χ^2_1 distribution when going from Hypothesis 1 to 2. Indeed, the p-value of the Kolmogorov-Smirnov goodness-of-fit test equals 0.0548 for Hypothesis 1 and 0.0000 for Hypothesis 2. The mean and variance are 0.94 and 1.75 respectively under Hypothesis 1 and 2.54, 8.11 under Hypothesis 2, clearly showing an increase in the values of the LRT. This is also confirmed by the 90, 95 and 99% quantiles as shown in the the bottom lines of Tables 3 and 4. Whereas the theoretical results of Rotnitzky *et al* (2000) indicate that in this setting the asymptotic distribution is stochastically smaller than a χ^2_1 distribution, i.e., a mixture with a χ^2_0, the simulated null distribution is stochastically larger. Probably the slow rate of convergence is causing this opposite behavior. Experimentation with larger sample sizes did not substantially change this result.

This restricted simulation experiment clearly shows that the use of the χ^2_1 distribution, which holds in standard situations, should be discouraged in the settings studied here. There is a need for an alternative approach and in the next section we propose two bootstrap approaches and investigate to which extent they can resolve the aforementioned problems.

7.2 Performance of Bootstrap Approaches

We propose a parametric and a semi-parametric bootstrap likelihood ratio test. In case the asymptotic null distribution does not depend on unknown parameters, the bootstrap is expected to have a smaller asymptotic order of error in level (Efron, 1979; Efron and Tibshirani, 1998; Davison and Hinkley, 1997). Beran (1988) showed that the bootstrap LRT automatically accomplishes the Bartlett adjustment, at least in a standard setting.

7.2.1 Parametric bootstrap

Given the data, a parametric bootstrap procedure for testing Hypotheses 1 or 2 in the selection model can be implemented using the following 4-step algorithm. (1) Fit the initial data under the null and the alternative hypothesis resulting in $(\hat{\theta}_{H_0}, \hat{\psi}_{H_0})$ and $(\hat{\theta}_{H_1}, \hat{\psi}_{H_1})$ respectively, where θ denotes the parameter vector for the measurement part and ψ for the missingness part; compute the LRT for the hypotheses under consideration. (2) Generate a 'bootstrap sample' from the selection model, reflecting the null hypothesis by using the estimates $(\hat{\theta}_{H_1}, \hat{\psi}_{H_0})$. (3) Compute the LRT test for the bootstrap sample. (4) Repeat step 2 and 3 B times and determine the bootstrap p-value as the proportion of bootstrap LRT values larger than its value for the original data from step 1.

Alternatively, step 2 could be based on the estimates $(\hat{\theta}_{H_0}, \hat{\psi}_{H_0})$. But some exploratory simulations showed that both choices resulted in essentially the same p-values. Instead of a p-value, one can also compute critical points of the bootstrap approximate null distribution (90, 95 and 99% quantiles) in step 4 (Tables 3 and 4).

The parametric bootstrap heavily depends on the quality of the estimates $(\hat{\theta}_{H_1}, \hat{\psi}_{H_0})$. In case the initial data are generated under the alternative, one can expect that bias disturbs the procedure, especially for Hypothesis 2. This would lead to the generation of bootstrap data in step 2 which would obey the null constraint but which would be substantially different from the initial data in many other respects. A semi-parametric model based on resampling and less depending on the estimates from the initial sample might perform better.

7.2.2 Semi-parametric Bootstrap

Given the data, a semi-parametric bootstrap procedure for testing Hypotheses 1 or 2 in the selection model can be implemented using the following algorithm. (1) Fit the initial data under the null and the alternative hypothesis resulting in $(\hat{\theta}_{H_0}, \hat{\psi}_{H_0})$ and $(\hat{\theta}_{H_1}, \hat{\psi}_{H_1})$ respectively; compute the LRT for the hypothesis

under consideration. (2) Impute the missing data, conditionally on the observed outcomes at the previous occasion, and based on the probability model for the measurement part (3.2) using the estimate $\hat{\theta}_{H_1}$ (this is a parametric part). (3) Draw (complete) observations from the augmented data set (resulting from step 2), with replacement, yielding a new sample of the same size N (this resampling is the nonparametric part). (4) Observations at time $t \geq 2$ are deleted with a probability according the logistic dropout model (3.3) using the estimate $\hat{\psi}_{H_0}$ (thus reflecting the null hypothesis; this is again a parametric part); this is the final bootstrap sample. (5) Compute the LRT test for the bootstrap sample. (6) Repeat steps 2 and 5 B times and determine the bootstrap p-value as the proportion of bootstrap LRT values larger than its value from the initial data from step 1.

For more details about similar semi-parametric bootstrap implementations in other settings, see Davison and Hinkley (1997).

For Hypothesis i ($i = 1, 2$), two initial data sets were generated under three scenarios: Scenario 1: all N observations generated under Hypothesis i; Scenario 2: all N observations generated under the alternative: $\psi_1 = 1$ for $i = 1$, $\psi_2 = 2$ for $i = 2$; Scenario 3: 10 observations generated under Hypothesis i and 190 observations under the corresponding alternative.

For each Hypothesis i, for each Scenario j and for each initial dataset, $B = 400$ bootstrap samples were generated and bootstrap LRT values were computed. The results (p-values and quantiles) for these 18 combinations are shown in Tables 3 and 4.

Fitting the selection model, obtaining the maximum likelihood estimates and computing the LRT, is a nontrivial iterative computing exercise, not lending itself for intensive simulations. A full simulation study based on, e.g., 100 initial samples was computationally not feasible. The 'optmum' procedure in Gauss 3.2.32 was used. The optimization method used the Broyden-Fletcher-Goldfarb-Shanno procedure to obtain starting values for the Newton-Raphson procedure and it took about one week to obtain the results of one of the 18 combinations. Nevertheless, we think that our limited results do reveal the main characteristics of the performance of both bootstrap procedures.

For Hypothesis 1, Table 3 shows that, for all scenarios, the χ^2_1 approximation and the bootstrap approximation to the null distribution are consistent and in line with our expectations. Note that the results for the two initial data sets under Scenario 3 are not in agreement: one clearly rejects the hypothesis and the other clearly not. Since only 5% of the initial data are generated under the alternative, a less clear rejection pattern is to be expected here.

The results in Table 4 globally show that for testing MAR versus MNAR (Hypothesis 2), also the bootstrap is not able to approximate the true null distri-

Table 3
Hypothesis 1. Critical points based on the parametric and semi-parametric bootstrap procedure (400 bootstrap runs) for two initial data sets. Lower lines show the critical points of the simulated null distribution based on 800 samples, together with those of the χ^2_1 distribution.

		quantiles			p-value
		0.10	0.05	0.01	
Scenario 1	Parametric	3.04	4.17	6.12	0.7556
		2.53	3.35	6.76	0.3566
	Semi-Parametric	2.96	3.83	6.22	0.7890
		2.46	4.16	6.60	0.3616
Scenario 2	Parametric	2.55	3.39	6.36	<0.0025
		2.83	3.68	7.02	<0.0025
	Semi-Parametric	2.41	3.39	6.49	<0.0025
		2.68	3.68	6.37	<0.0025
Scenario 3	Parametric	2.35	3.72	7.91	0.9352
		3.00	3.93	6.48	<0.0025
	Semi-Parametric	2.83	4.13	8.00	0.6085
		2.70	4.40	6.49	<0.0025
simulated H_0		2.23	3.27	6.04	
χ^2_1 distribution		2.71	3.84	6.63	

Table 4
Hypothesis 2. Critical points based on the parametric and semi-parametric bootstrap procedure (400 bootstrap runs) for two initial data sets. Lower lines show the critical points of the simulated null distribution based on 800 samples, together with those of the χ^2_1 distribution.

		quantiles			p-value
		0.10	0.05	0.01	
Scenario 1	Parametric	38.71	42.68	46.21	0.1870
		9.62	12.25	19.77	1.000
	Semi-Parametric	4.86	6.74	10.48	0.0998
		7.40	9.35	14.47	0.2743
Scenario 2	Parametric	22.07	24.84	30.32	0.0349
		9.36	11.39	14.04	0.0025
	Semi-Parametric	12.35	15.63	20.88	0.0050
		17.05	19.75	27.56	0.0224
Scenario 3	Parametric	8.17	10.09	15.02	0.0175
		15.46	17.85	24.38	0.9351
	Semi-Parametric	15.68	19.11	25.58	0.1397
		8.11	10.46	13.55	0.6085
simulated H_0		6.44	9.17	12.10	
χ^2_1 distribution		2.71	3.84	6.63	

bution. Especially the behaviour of the parametric bootstrap is very unstable and variable. The semi-parametric version seems to slightly perform better,

especially for Scenario 1. As the bootstrap is also an asymptotic method, it suffers from the same slow convergence as the χ^2 type distributions.

8 Concluding Remarks

In recent times, models for MNAR missingness have gained in popularity. However, as already noted in the discussion to Diggle and Kenward (1994), it has been made clear at various occasions that caution should be used when interpreting such models, due to the great sensitivity the results exhibit with respect to the model assumptions made. This has led to quite a bit of work on sensitivity analysis. One such tool is local influence but this particular tool itself tends to behave in an, at first sight, non-intuitive fashion. At the same time, there is some confusion about the identifiability issues implied by the Diggle and Kenward (1994) model, since a likelihood ratio and the corresponding p-value can apparently be obtained in a standard way. In this paper, by studying both of these issues, we have provided evidence that they are, in fact, two faces of the same coin.

The behavior of the likelihood ratio statistic for the MNAR parameter ψ_2 is non-standard. The information on ψ_2, available in the data, is very scarce and interwoven with other features of the measurement and dropout model. This translates mathematically into dependent systems of estimating equations and thus singularities in the corresponding information matrix. The rate of convergence to the asymptotic null distribution is extremely slow, implying that also well-established bootstrap methods appear to be deficient.

This implies that there is much less information available, even with increasing sample sizes, than one typically would expect. As a result of this, the ψ_2 parameter is more vulnerable than others for all sorts of deviations in the model, in particular to unusual profiles. This includes profiles with an average away from the bulk of the data, unusual autocorrelation pattern etc.

The bottomline is that local influence tools in the incomplete data context are useful, not to detect individuals that drop out non-randomly, but rather to detect anomalous subjects that lead to a seemingly MNAR mechanism. A careful study of such subjects, combined with appropriate treatment (e.g., correction of errors, removal,...), can lead to a final MAR model, in which more confidence can be put by the researchers, which ultimately is the goal of every sensitivity analysis.

Acknowledgments

Ivy Jansen, Niel Hens, Geert Molenberghs and Marc Aerts 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".

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.

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements, Journal of the American Statistical Association, 83, 687–697.

Besag, J., Green, P., Higdon, D., and Mengersen, K., 1995. Bayesian Computation and Stochastic Systems, Statistical Science, 10, 3–66.

Chatterjee, S. and Hadi, A.S., 1988. Sensitivity Analysis in Linear Regression. New York: John Wiley & Sons.

Cook, R.D., 1986. Assessment of local influence. Journal of the Royal Statistical Society, Series B, 48, 133–169

Copas, J.B. and Li, H.G., 1997. Inference from non-random samples (with discussion). Journal of the Royal Statistical Society, Series B, 59, 55–96.

Davison, A.C. and Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press.

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.

Diggle, P.D. and Kenward, M.G., 1994. Informative dropout in longitudinal data analysis (with discussion). Applied Statistics, 43, 49–93.

Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society, Series B, 57, 45–97.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1–25.

Efron, B. and Tibshirani, R.J., 1998. An Introduction to the Bootstrap. Chapman and Hall/CRC, Boca Raton.

Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R., 1995. Regression models for longitudinal binary responses with informative dropouts. Journal of the Royal Statistical Society, Series B, 57, 691–704.

Glynn, R.J., Laird, N.M., and Rubin, D.B., 1986. Selection Modelling versus mixture modelling with nonignorable nonresponse. In Drawing Inferences from Self-Selected Samples, Ed. H. Wainer, pp. 115–142. New York: Springer Verlag.

Hartley, H.O. and Hocking, R., 1971. The analysis of incomplete data. Biometrics, 27, 783–808.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5, 475–492.

Hens, N., Aerts, M., Molenberghs, G., Thijs, H., and Verbeke, G., 2004. Kernel Weighted Influence Measures. Computational Statistics and Data Analysis, 0, 00–00.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C., 2004. Analyzing Incomplete Binary Longitudinal Clinical Trial Data. Statistical Science, Submitted.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K., 2003. A local influence approach applied to binary data from a psychiatric study. Biometrics, 59, 409–418.

Kenward, M.G., 1998. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. Statistics in Medicine, 17, 2723–2732.

Kenward, M.G, Goetghebeur, E.J.T., and Molenberghs, G., 2001. Sensitivity analysis of incomplete categorical data. Statistical Modelling, 1, 31–48.

Kenward, M.G. and Molenberghs, G., 1999. Parametric models for incomplete continuous and categorical longitudinal studies data. Statistical Methods in Medical Research, 8, 51–83.

Laird, N.M., 1994. Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. Applied Statistics, 43, 84.

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.

Laird, N.M. and Ware, J.H., 1982. Random effects models for longitudinal data. Biometrics, 38, 963–974.

Lesaffre, E. and Verbeke, G., 1998. Local influence in linear mixed models. Biometrics, 54, 570-582.

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. Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. Applied Statistics, 43, 78.

Little, R.J.A. and Rubin, D.B., 1987. Statistical Analysis with Missing Data. New York: Wiley.

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.
Molenberghs, G., Goetghebeur, E., Lipsitz S.R., and Kenward, M.G., 1999. Non-random missingness in categorical data: strengths and limitations. The American Statistician, 53, 110–118.
Molenberghs, G., Kenward, M.G., and Lesaffre, E., 1997. The analysis of longitudinal ordinal data with nonrandom dropout. Biometrika, series 84, 33–44.
Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., and Carroll, R.J., 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics, 0, 00–00.
Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M.G., 2001. Mastitis in dairy cattle: local influence 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.
Nordheim, E.V., 1984. Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome. Journal of the American Statistical Association, 79, 772–780.
Rotnitzky, A., Cox, D.R., Bottai, M., and Robins, J., 2000. Likelihood-based inference with singular information matrix. Bernouilli, 6(2), 243–284.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.
Rubin, D.B., 1994. Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. Applied Statistics, 43, 80–82.
Scharfstein, D.O., Rotnizky, A., and Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). Journal of the American Statistical Association, 94, 1096–1146.
Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D., 2002. Strategies to fit pattern-mixture models. Biostatistics, 3, 245–265.
Thijs, H., Molenberghs, G., and Verbeke, G. (2000) The milk protein trial: influence analysis of the dropout process. Biometrical Journal, 42, 617–646.
Vach, W. and Blettner, M., 1995. Logistic regression with incompletely observed categorical covariates-investigating the sensitivity against violation of the missing at random assumption. Statistics in Medicine, 12, 1315–1330.
Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H., 2001. A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. Statistical Modelling: An International Journal, 1, 125–142.
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, 43–50.
Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J.P., Carels, C., Kuhn, E.R., Darras, V., and de Zegher, F., 1997. Craniofacial growth in the male rat: Effect of endogeneous testosterone. Submitted.