Epidemiology
Open Access
ISSN: 2161-1165

Satty, Epidemiol 2016, 6:1
DOI: 10.4172/2161-1165.1000221

Research Article Open Access

An Analysis of Selection Models for Incomplete Longitudinal Clinical Trials Due to Dropout: An Application to Multi-centre Trial Data

Ali Satty*

Faculty of Mathematical Sciences and Statistics, Department of Statistics and Actuarial Science, Elneelain University, Khartoum, Sudan

Abstract

A common problem encountered in statistical analysis is that of missing data, which occurs when some variables have missing values in some units. The present paper deals with the analysis of longitudinal continuous measurements with incomplete data due to non-ignorable dropout. In repeated measurements data, as one solution to a problem, the selection model assumes a mechanism of outcome-dependent dropout and jointly both the measurement together with dropout process of repeated measures. We consider the construction of a particular type of selection model that uses a logistic regression model to describe the dependency of dropout indicators on the longitudinal measurement. We focus on the use of the Diggle-Kenward model as a tool for assessing the sensitivity of a selection model in terms of the modeling assumptions. Our main objective here is to investigate the influence on inference that might be exerted on the considered data by the dropout process. We restrict attention to a model for repeated Gaussian measures, subject to potentially non-random dropout. To investigate this, we carry out an application for analyzing incomplete longitudinal clinical trial with dropout by using a practical example in the form of a multi-centre clinical trial data.

Keywords: Incomplete longitudinal data; Selection model; Diggle and Kenward model; Dropout; Missing not at random

Introduction

A typical characteristic of longitudinal studies is that study subjects are measured over repeated time intervals. The dropout of subjects along the time scale is common. The dropout process is assumed to be stochastic in nature and generally dependent upon the observed or unobserved outcomes. It also may depend upon covariates, such as the treatment arm an individual is allocated to. The dropout may be regarded as a "failure" outcome in certain limited settings. Of prime concern to this study, is the more general situation that characterizes the statistical behavior of the original outcome, while dropout is treated as a "nuisance" occurrence that must be tolerated. As a result of this, the distinction between the outcome and the dropout processes needs to be simultaneously maintained. Rubin, Little and Rubin [1,2] introduce different mechanisms for denoting dropout or non-response. A dropout, or non-response, process is said to be missing completely at random (MCAR) if the non-response process is a random event independent of both unobserved and observed outcomes, missing at random (MAR), if conditional upon the observed outcomes, the non-response process is independent of the unobserved outcomes, and missing not at random (MNAR) when the non-response process depends only upon the unobserved outcomes. In the context of likelihood and Bayesian inferences, and when the parameters describing the measurement process are functionally independent of the describing the non-response process, MCAR and MAR are ignorable, while a non-random process is non-ignorable [1,2]. When data are MNAR, the missingness cannot be ignored from the analysis. In this case, the longitudinal measurement process and the missingness indicators may be considered jointly [3].

It is possible to consider more general models when one assumes random missingness mechanism to be untrue [4]. Examples on work of MNAR modeling include [5]. These belong to the so-called selection models family [2]. A selection model factors the joint distribution of the measurement and dropout mechanism into two parts, that is, a marginal measurement model that describes the distribution of the underlying complete data, and a dropout mechanism that describes the distribution of the missing data indicators, conditional upon the complete data. For more details, see, for example Diggle and Kenward [5]. This is intuitively appealing since the marginal measurement distribution would be of interest also with complete data [3]. Furthermore, the missing data mechanisms (MCAR, MAR and MNAR) are most easily developed within the selection setting. However, it is often argued, especially within the context of non-random missingness model, that selection models, although identifiable, should be approached with caution [6]. Indeed, one has to make untestable assumptions about the missing data process. Selection models originated from the ref. [7]. The theoretical translation from the model [7] to selection model [5] have been addressed [5,8]. Consider a selection model for the study of a longitudinal measurement when data are MNAR by letting the probability of dropout depend on the possibly unobserved measurements. They use a linear mixed model for the longitudinal measurement and logistic regression model for the dropout process to describe the dependency between dropout indicators and measurements. The dropout indicators are used to indicate participant dropout. However, the intermittent missing data is assumed to be missed at random, and it can be ignored in the model. For alternatives for the missing data processes [3] an earlier work on the selection model analysis is given by refs. [6,7]. Selection models that are applied to the regression analysis of categorical variables with outcome subject to non-ignorable non-response are applied by [9,10] used a selection perspective for the conditional expectation model in a semi-parametric approach. For the ignorable non-response hypothesis, [11] proposed a general class of selection models under non-monotone missing data pattern. In the case of the selection models for repeated measurements, sensitivity of the conclusions to the assumptions about the dropout mechanism has been illustrated by Kenward [12]. A semi-

***Corresponding author:** Ali Satty, Faculty of Mathematical Sciences and Statistics, Department of Statistics and Actuarial Science, Elneelain University, Khartoum, Sudan, E-mail: alisatty1981@gmail.com

Received July 13, 2015; **Accepted** January 19, 2016; **Published** January 26, 2016

Citation: Satty A (2016) An Analysis of Selection Models for Incomplete Longitudinal Clinical Trials Due to Dropout: An Application to Multi-centre Trial Data. Epidemiol 6: 221. doi:10.4172/2161-1165.1000221

parametric approach of missing data mechanism is proposed [13] in order to avoid the impact of the parametric missing data specification in a selection model perspective. With regard to the nonmonotone pattern, selection models have been extended [14]. In addition to Troxel's work, within the selection model framework, models have been proposed for non-monotone pattern as well, for instance, see [15]. In the context of categorical and other types of measure, in many examples, see [16,17], the selection models were also developed. Additionally, a number of proposals have been made for non-Gaussian outcomes, see [18]. Further details in selection models can be found in refs. [18-21].

This paper deals with the analysis of longitudinal data when there is non-ignorable dropout. We illustrate this analysis by considering the problem of missing data that occurs with a continuous outcome. We focus on the use of the Diggle and Kenward [5] model as a tool for assessing the sensitivity of a selection to the modeling assumptions. We restrict attention to a model for repeated Gaussian measures, subject to where dropout possibly depends upon missing outcomes, i.e., MNAR. A monotone missing pattern has been constructed in the model. Similar to Diggle and Kenward [5], a selection model is specified that uses a logistic regression model to describe the dependency of missing data indicators upon the longitudinal response. In the current application, we modify the analysis software to accommodate the case of more than two treatment arms as a computational extension. Our main objective here is to investigate the influence that might be exerted on the considered data by the dropout process. In order to investigate our objective, we carry out an application for analyzing incomplete longitudinal data with dropout. We outline the fitting of the selection model which is based on the linear mixed model for the measurement process as well as a logistic regression for dropout process. The model was fitted using standard statistical software (SAS version 9.2, IML macro). This is done by using a practical example in the form of a multi-centre clinical trial data. The remainder of the article is organized as follows: the data setting and modeling framework are introduced in Section 2. In Section 3, a background for the selection model is provided, followed by descriptions of the selection model based on Diggle and Kenward model frameworks as well as detailed discussion of the linear mixed model and dropout model. In Section 4, we present an application including a description of the data set used in the analysis. The results of the estimation of the model are then described in Section 5. We conclude with a discussion of the results in Section 6.

Modeling Longitudinal Data with Dropout

To introduce some necessary notation, we follow the terminology provided by Molenberghs and Kenward [3,8] based on the standard modeling frameworks of [1,22]. So, assume that for each independent subject i=1,..., N in the study a sequence of responses Yij is designed to be measured at a fixed set of occasions j=1,..., n. The outcomes are grouped into a vector $Yi=(Y_{i1},..., Y_{in})^t$. It is often necessary to split the outcome vector Y_i into two sub-vectors, Y^o_i and Ymi, indicating the observed and missing components, respectively. Additionally, one can define an indicator R_{ij}, for each occasion j as follows: $R_{ij}=1$, if Y_{ij} is observed and $R_{ij}=0$ if not. The indicators of missing data (R_{ij}) can be grouped into a vector Ri which is of parallel structure to Y_i. The processes generating the vectors Y_i and R_i are referred to as the measurement and missing data processes, respectively. We now pay attention to the dropout setting which is a particular case of monotone pattern of missingness in which a missing value whenever it occurs to any subject in the sequence of repeated measurements of the outcome is never followed by any observed measurement on that subject. Alternatively, when dropout occurs, one could use a scalar variable D_i called the dropout indicator, rather than the missing data indicator R_i, defined as $D_i=1+P_{nij}=1\ R_{ij}$, indicating the occasion at which dropout occurs. Next, we consider the density of the full data (Y_i,R_i), denoted by

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi), \tag{1}$$

Where X_i and W_i are covariate matrices for the measurement and missing data mechanism, respectively, and the parameter vectors θ and ψ describe the measurement and missingness processes, respectively. The taxonomy, constructed by refs. [1,22], is based on the following factorization

$$f(y_i, r_i \mid X_i, Wi, \theta, \psi)=f(y_i \mid X_i, \theta)f(r_i \mid y_i, W_i, \psi), \tag{2}$$

Where the first and second factors denote the marginal density of the measurement process and the density of the missing data process, conditional upon the outcomes, respectively. Factorization (2) forms the basis of selection modeling as the second factor corresponds to the self-selection of individuals into observed and missing groups. Using the reversed factorization, an alternative taxonomy which can be considered is called pattern mixture models. They have the following form

$$f(y_i, r_i \mid X_i, Wi, \theta, \psi)=f(y_i \mid r_i, X_i, \theta)f(r_i, W_i, \psi). \tag{3}$$

In fact, equation (3) can be described as a mixture of different populations, characterized by the observed missing data pattern. An initial attention of these models were provided by ref. [2,6], while further attention later was provided by many authors, see, for example [23]. As we mentioned above, Rubin's taxonomy [1,22] of missing data process is based on the second factor of equation (2), thus within the selection modeling framework

$$f(r_i \mid y_i, W_i, \psi)=f(r_i \mid y^o_i, y^m_i, W_i, \psi). \tag{4}$$

In equation (4), the covariates for the measurement process are assumed measured but suppressed for simplicity sake. The form in equation (4) can be discussed as follows: when the missingness process is independent of the responses, i.e.,

$$f(r_i \mid y_i, W_i, \psi)=f(r_i \mid W_i, \psi). \tag{5}$$

Then the process corresponds to the case of missing completely at random (MCAR). If the missingness process is only independent of the unobserved responses Y m, but depends on the observed responses Y o, consequently, assuming the form

$$f(r_i \mid y_i, W_i, \psi)=f(r_i \mid y^o_i, W_i, \psi). \tag{6}$$

Then the process corresponds to the case of missing at random (MAR). Finally, when the missingness process depends on the missing data Y^m_i, the process corresponds to the case of missing not at random (MNAR). As pointed out by Rubin, Little and Rubin [1,2], when MAR mechanism holds, the parameters θ and ψ are functionally independent. In practice, the likelihood of interest then depends upon the factor $f(y^o_i \mid \theta)$. For this reason, when using a likelihood based analysis under the MAR assumption, the missing value mechanism is sometimes said to be "ignorable". By contrast, if the likelihood of interest only depends upon the factor $f(y^m_i \mid \theta)$, then this is referred to as "non-ignorable" setting. Therefore, when ignorability holds, likelihood-based and Bayesian inferences are valid [8,18].

A Selection Models for Non-ignorable Dropout

In the framework of the selection models, it is not always reasonable to assume that MAR holds, and a wide range modeling approaches for MNAR data have been proposed. One such is the model proposed by

Diggle and Kenward [5] for continuous outcomes with dropout. In this section, we first describe the Diggle and Kenward [5] selection model for continuous longitudinal data. We then discuss in detail the linear mixed model and the dropout model.

Diggle and Kenward's (1994) model for continuous longitudinal outcomes

A model for longitudinal Gaussian data with non-random dropout has been proposed by Diggle and Kenward [5]. Their model assumes that the missingness mechanism is MNAR which combines the multivariate normal model for longitudinal Gaussian data with a logistic regression for the dropout process. From the notation presented in Section (2) recall that for subject *i, i=1,..., N,* a sequence of responses Y_{ij} is designed to be measured at time points t_{ij}, *j=1,..., n*, resulting in a vector of observed outcomes $Y_i=(Y_{i1},..., Y_{ini})^t$ of measurements for each subject. Note that although n measurements per subject were planned the vector Y_i is of size $n_i < n$ because of missing observation. In the case of dropout, the complete Y_i is only partially observed. If we let Di be the occasion where dropout occurs, then $D_i > 1$, and Y_i can be partitioned into the $(D_i - 1)$-dimensional observed component Y^o_i and the $(n_i - D_i+1)$-dimensional missing component Y^m_i. If no dropout occurs, we let $D_{i=} n_i+1$, and Y_i equal Y^o_i. For the ith subject, the observed data is (Y^o_i, di), thus, the likelihood contribution is proportional to the marginal density function.

$$f(\mathrm{y}_i, d_i \mid \theta,\psi) = \int f(\mathrm{y}_i, d_i \mid \theta,\psi) dy_i^m = \int f(\mathrm{y}_i \mid \theta) f(d_i \mid \mathrm{y}_i,\psi) dy_i^m. \quad (7)$$

In equation (7), a marginal model for Y_i can be combined with a model for the dropout process, conditional upon the measurement, and the measurement process model, including the vectors of unknown parameters, θ and ψ, respectively. More formally, we denote the conditional probability of dropout by g_j (y_{ij}, h_{ij}) at time *j* given the response at *time j*, and $h_{ij}=(y_{i1},..., y_{ij-1})$ which denotes a possibly observed history of subject *i* until time t_j, *j*−1. According to Diggle and Kenward [5], the dropout process allows the conditional probability for dropout at occasion *j*, given that the subject was still observed at the previous occasion, to depend upon the history *hij* and the possibly unobserved current outcome $y_{ij,}$ but not upon future outcomes y_{ik}, *k>j*. Now, for calculating the dropout probability for each occasion, we use the conditional probabilities $P(D_i=j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)$ which can be expressed as follows:

$$P(D_i=j \mid Y_i)=p(D_i=j \mid h_{ij}, y_{ij}, \psi)$$

$$\mathrm{P(Di=j \mid yi,)=P(Di=j \mid hij, yij,)}$$

$$= \begin{cases} P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi) & j = 2 \\ P(D_i = j \mid D_i \geq j, h_{ij}, y_{ij}, \psi) & \\ \times \prod_{k=2}^{j-1} [1 - \mathrm{P}(D_i = k \mid D_i \geq k, h_{ik}, y_{ik}, \psi)] & j = 3,..., n_i \\ \prod_{k=2}^{n_i} [1 - \mathrm{P}(D_i = k \mid D_i \geq k, h_{ik}, y_{ik}, \psi)] & j = n_i + 1, \end{cases} \quad (8)$$

Assuming no missing values at occasion *j=1*. As mentioned above Diggle and Kenward [5] combine a multivariate normal for the measurement process together with a logistic model for the dropout process. To obtain parameter and precision estimates from the combined measurement/dropout model, they use maximum likelihood that involves marginalization over the unobserved components, i.e., Y^m_i. In fact, under repeated measurements for the i^{th} subject, the measurement model assumes that the vector Y_i satisfies the linear regression model $Y_i \sim N(X_i\beta, V_i)$, where *i=1,...N*, in which β is a vector of population-averaged regression coefficients. Further Verbeke and Molenberghs [8] pointed out that the matrix V_i can be left unstructured or assumed to be of a specific form, for example, resulting from a linear mixed model, a factor-analytic structure, or spatial covariance structure. As Molenberghs and Kenward [3], there is some advantages to using an unstructured covariance matrix. Following, we introduce the measurement and dropout models that can be combined for the dropout process.

Measurement model: For continuous outcomes proposed linear mixed-effects models, and they can be written as follows

$$Y_i=X_i\beta+Z_ib_i+S_i+\varepsilon_i, \quad (9)$$

Where *Yi* is the ni-dimensional response vector for subject *i*, $1 \leq i \leq N$, *N* is the number of subjects, X_i and Z_i are $(n_i \times p)$ *and* $(n_i \times q)$ known design matrices, β is the p-dimensional vector containing the fixed effects, $b_i \sim N(0, G)$ is the q-dimensional vector containing the random effects. The residual components $\varepsilon_i \sim N(0, \sigma^2 I_n)$, and $b_1,..., b_n$, $\varepsilon_1,..., \varepsilon_n$ are assumed to be independent. The serial correlation is captured by the realization of a Gaussian stochastic process, S_i which is assumed to follow a $N(0, \tau^2 H_i)$ law. Here, the serial covariance matrix *Hi* dependent upon *i* through the number *n* of observations and through the time points t_{ij} at which measurements are taken. Using autocorrelation function $\rho(t_{ij} - t_{ik})$, the structure of the matrix H_i is determined. A first simplifying assumption is that H_i depends upon the time interval between two measurements Y_{ij} and $Y_{ik,}$ i.e., $\rho(t_{ij} - t_{ik})=\rho(u)$, where $u=|t_{ij} - t_{ik}|$ represents the time lag. The autocorrelation function decreases such that $\rho(0)=1$ and $\rho(u) \to 0$ as $u \to \infty$. Finally, *G* is a general $(q \times q)$ covariance matrix with its *(i, j)* element given by $d_{ij}=d_{ji}$. The random effects in model (9) stem from heterogeneity between subjects, in the sense that various aspects of their behavior may exhibit inter-subject random variation. It follows from model (9) that, given the random effect bi, Yi is normally distributed with mean vector $X_i\beta+Z_ib_i$ and covariance matrix V_i. Thus, after integrating over random effects, inference for the marginal distribution of the outcome Y_i, can be written as follow

$$Y_i \sim N(X_i\beta, V_i), \quad (10)$$

Where $V_i=Z_iGZ^t+\sigma^2 \boldsymbol{I}_{ni}+\tau^2 H_i$ is a $(n_i \times n_i)$ covariance matrix which combine both the measurement error and serial components. On the other hand, to include various fixed effects, a random intercept, and allowing Gaussian serial correlation, a linear mixed model is used for the measurement model process. In this case the covariance matrix V_i becomes

$$V_i=dJ_n+\sigma^2 I_i+\tau^2 H_i, \quad (11)$$

Where J_{ni} is an $(n_i \times n_i)$ matrix with all elements equal to 1, I_i is the $(n_i \times n_i)$ identity matrix, and H_i is determined through the autocorrelation function ρ^{uijk}, where μ_{jk} the Euclidean distance between t_{ij} and t_{ik}, thus

$$H_i = \begin{pmatrix} 1 & p^{u12} & \cdots & p^{u1n} \\ p^{u12} & 1 & \cdots & p^{u2n} \\ \vdots & \vdots & \ddots & \vdots \\ p^{u1n} & p^{u2n} & \cdots & 1 \end{pmatrix} \quad (12)$$

Where $\sigma^2>0$ and $0 \leq \rho \leq 1$. The covariance structure V_i in equation (11) combines both serial autocorrelation and a shared random effect variance in the estimation. The main problem with this approach, which is due to Diggle and Kenward [5], is that it assumes stationary. In practice, if times of measurement are common, the unstructured matrices can be used (aside from very small trials) and for unbalanced times, a random coefficient model.

Dropout model: As noted previously, we focus only on incompleteness due to dropout, and thus we assume that the first measurement Y_{i1} is measured for all subjects in the study. In agreement with notation introduced in Section 2, the selection model arises when the joint likelihood of the measurement process and the dropout process is factorized as follows

$$f(y_i, r_i \mid X_i, W_i, \theta, \psi)=f(y_i \mid X_i, \theta)f(r_i \mid y_i, W_i, \psi). \quad (12)$$

We use the linear mixed-effects model introduced in equation (9) to model the measurements process, together with a logistics regression to describe the dropout process. According to Diggle and Kenward [5], the model for dropout process is based on a logistics regression for the conditional probability of dropout at occasion j, given the subject is still in the study. Again, the $g_i(y_{ij}, h_{ij})$ denotes this probability of dropout at time j, in which $h_{ij}=(Y_{i1}, Y_{i2},..., Y_{ij-1})$ is a vector possibly containing all observed measurements up to including occasion $j-1$, as well as relevant covariates, in the conditional probability of dropout model. Modeling the dropout mechanism may be simplified by allowing dropout to depend upon the current measurement and immediately preceding measurement only with corresponding regression coefficients, i.e., ψ_1 and ψ_2. A commonly used version of such a logistic dropout model is

$$logit[P(D_i=j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)]=\psi_0+\psi_c W_i+\psi_1 y_{i,j-1}+\psi_2 y_{ij}, \quad (13)$$

Where ψ_0 and ψ_c denote the intercept and the vector of parameters for covariates W_i, respectively. The model in equation (13) contains special cases corresponding to MAR and MCAR mechanism that can be obtained from $\psi_2=0$ or $\psi_1=\psi_2=0$, respectively. As pointed out by Diggle and Kenward [5,8], a likelihood ratio test (LRT) can be used to compare the model fit under a model that assumes the missing data due to dropout are MCAR versus MAR, that is, the LRT for MCAR versus MAR has an approximate χ^2 distribution. The LRT statistic is used to test the hypothesis of $\psi_2=0$ (i.e., MAR), where dropout is no longer dependent upon the current measurement, and similarly to test the hypothesis of $\psi_1=\psi_2=0$ (i.e., MCAR), where dropout is assumed to be at random, the dropout therefore, does not depend upon the outcome altogether. However, the use of the LRT is inappropriate for hypothesis test for MNAR versus MAR when all the other modeling assumptions hold, due to the fact that the behavior of the LRT statistic for the MNAR parameter ψ_2 is non-standard, since the availability of the information on ψ_2 is very rare and interwoven with other features of both measurement and dropout models [4]. In addition Little and Rotnitzky et al. [24,25] illustrated that the limiting distribution is a χ^2 mixture with characteristics controlled by the singular information matrix. Therefore, for the ψ_2 associated with MNAR model, the score equation creates a quasi-linear dependence structure in the system of score equations. This issue is studied in detail by Jansen et al. [4], while, in the context of an onychomycosis study [26] have stated that excluding a small amount of measurement error can change drastically the LRT statistic for the MAR null hypothesis, see also for example [8]. In practice, such a distinction (MAR/MNAR) can only be made using untestable modeling assumptions such a distributional form, see, [12]. This problem is really laid bare in Molenberghs et al. [27] which show that the formal-based distinction between MAR and MNAR is not possible as for any MNAR model there exists an MAR model that fits the data equally well. The similarity of the MAR and MNAR models with respect to fitting to the observed data, may present different predictions of the unobserved outcomes, conditional upon the observed ones. Hence, it is broadly agreed that the role of such MNAR models is in sensitivity analysis, which is if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed. Further detail on the precise nature of sensitivity analysis can be found in Molenberghs and Kenward [3,18].

Application to the Multi-centre Trial Data

Below we describe the data set that is used in the analysis as well as the application schemes that are used in the analysis of the selection models based on Diggle and Kenward [5] approach. In terms of the application of the statistical techniques considered in this study, we use the statistical software, SAS programme.

Data set - multicentre trial data

The example that is used here concerns the analysis of repeated measures designs and demonstrates how to investigate a specific scenario based on dealing with longitudinal data that has a nonignorable dropout mechanism. The data is based on experiments that rely on the split-plot design assumptions. Such experiments which include repeated measures designs have structures that involve more than one size of experimental unit. In this case, a subject is measured over time where time is one of the factors in the treatment structure of the experiment. By measuring the subject at several different time occasions, the subject is essentially being (split) into parts (time intervals), and the response for each part is measured. The larger experimental unit is the subject or the collection of time intervals which constitute a cluster. The smaller unit is the interval of time during which the subject is exposed to a treatment or an interval just between time measurements. The only departure from the classical split-plot assumptions is because in this case the subplot treatments (time intervals) are not randomized. The data used is from a multi-centre experiment data which is a typical longitudinal example. The data used here is described and reported in Milliken and Johnson [28]. This example considers an experiment that involve three drugs where each subject was measured repeatedly at three different time points $(j=1, 2, 3)$, where the outcome is described only as a measure of a continuous blood component. The data were collected (Table 1).

By three different investigators (or in three different centres) and contains fifty-one patients. There are seventeen patients assigned to each drug. All of the fifty-one patients were observed at the first occasion, whereas only eight and ten patients were not seen at the third occasion and at both the second and third occasions, respectively. In Table 1, we present the numbers of dropouts by time, centre and drug. The dropouts occur for all drugs and centres. It is clear that drug2 contains more percentages of missing values. The observed data for all subjects are shown in Figure 1. The main purpose of this experiment has been to estimate the effects of the drugs on the blood component over time, as well as to investigate the relationship between drugs and blood component. In this study, we restrict attention to the influence that might be caused on these effects by the dropout mechanisms as well as to examine which dropout mechanism better describe the data. The full results of the analysis of this trial using a likelihood based linear mixed models approach have been reported elsewhere by Milliken and Johnson [28] (Figure 1).

Time	Centre-R			Centre-S			Centre-T		
	Drug 1	Drug 2	Drug 3	Drug 1	Drug 2	Drug 3	Drug 1	Drug 2	Drug 3
1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	2	0	3	3
3	2	3	0	1	0	2	3	4	3
Total	3	4	0	1	0	4	3	7	6
Total	7			5			16		

Table 1: Numbers of dropouts in the multi-centre trial.

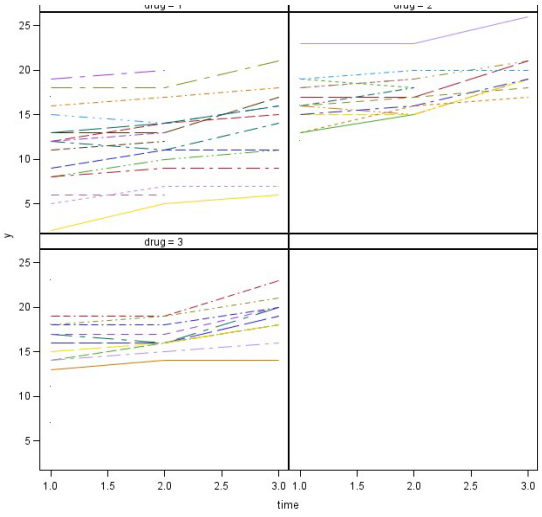

Figure 1: Multi-centre data. Observed data for all subjects.

Diggle-Kenward model applied to the multi-centre trial data

To apply the selection models due to Diggle-Kenward model based on continuous longitudinal data, in the current computations, we modified the SAS macro that was reported in Dmitrienko et al. [29] that maximizes the log-likelihood for the model using PROC IML to the case of three drugs as opposed to most application which are based on two drugs. We carried out an application to the above modeling strategy to the multi-centre data as earlier described. We fit the Diggle and Kenward model in accordance with the MCAR, MAR and MNAR assumptions to our own data set. The three post-baseline visits correspond to the measurements taken at times 1, 2 and 3. In the linear mixed model in equation (9), we allow the inclusion of a variety of fixed effects, a random intercept, and Gaussian serial correlation. Furthermore, the dropout model in equation (13) is considered, assuming that the dropout does not depend upon the covariates. Apart from the explicit MCAR, MAR, and MNAR versions of this model, we will also conduct an ignorable analysis (that is, an analysis based on the measurement model only, ignoring the dropout model). Firstly, we fit a linear mixed model (LMM) of the form in equation (9) in order to obtain initial values for the parameters estimation of the measurement model. Assuming that the first measurement Y_{i1} is observed for every subject in the study. We thus assume a linear time trend of the response within each drug group. This implies that each profile can be described using two parameters, namely the intercept and a slope. The error matrix is chosen to be of the form (11). Since the multi-centre trial data contains fifty-one subjects $(i=1,...,51)$ observed at three time points $(j=1, 2, 3)$ for three drugs $(p=1, 2, 3)$, the model can be written as follows

$$Y_{ijp}=\beta_0+A_p+\beta_1 T_j+\beta_{1p}(T A)_{jp}+\varepsilon_{ijp}, \quad (14)$$

where Y_{ijp} is the blood component of subject i at time j on drug p, A_p denotes the p_{th} drug effect, T_j denotes the j_{th} measurement time effect, $(AT)_{jp}$ denotes the interaction effect between time and drug, and $\varepsilon_{ip} \sim N(0, V_i)$, where $V_i=dJ_3+\sigma^2 I_3+\tau^2 H_i$, with

$$H_i=\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.$$

Using the set to zero constraint $(A_1=0)$, β_0 is the intercept for the drug1 group, $(\beta_0+\alpha_2)$ is the intercept for the drug2 group, and for drug3, the intercept is $(\beta_0+\alpha_3)$, where α denotes the fixed effects. These are, respectively, referred to as β_{01}, β_{02} *and* β_{03} in the results, as we will see in Tables 2-4. The slopes are β_1, $(\beta_1+\beta_{12})$ and $(\beta_1+\beta_{13})$ for drug1, drug2, and drug3, respectively, referred to as β_{11}, β_{12} and β_{13} in the results presented in Tables 3 and 4. The SAS PROC MIXED with REPEATED statement can be used to obtain the initial values. In conforming to the model introduced in equation (13), we use the following logistic regression model for the dropout model probabilities.

$$logit[P(D_i=j \mid D_i \geq j, h_{ij}, y_{ij}, \psi)]=\psi_0+\psi_1 y_{i,j-1}+\psi_2 y_{ij},\ j=1, 2, 3, \quad (15)$$

Where ψ_1 and ψ_2 denote the logistic regression coefficients for current and immediately previous observations, respectively, and j denotes the time points. In practice, the combined model for measurement and dropout can be fitted to the data using a generic function maximization routine in the maximum likelihood [12]. In doing so Diggle and Kenward [5] used the simplex algorithm of [30] to maximize the log-likelihood. However, for the same purpose, we use another optimization method that is available in SAS software, so-called Newton-Raphson ridge optimization. For more detail of this method, see [29]. Therefore, we use SAS IML macro which maximizes the likelihood for the model, so as to fit the selection models for the dropout process. The results of initial values for the parameter estimates of the logistic dropout model can be obtained as in Table 2 [29].

Results

Next, we introduce the results of the application that was discussed earlier. The initial values for the parameters of the linear mixed model are listed in Table 3. The results of maximum likelihood for the parameter estimates (standard errors) from the measurement model, as well as the results of the variance model under the three missingness mechanisms are presented in Table 4. Examining these (Table 3).

Results, we see that as expected, the parameters estimation and corresponding standard errors of the fixed effects of the measurement model and the variance model were the same under ignorability, MCAR and MAR mechanisms. This confirms what is expected in theory, see, [3], for example. We now study factors that influence dropout. As discussed above we fit the three dropout models in turn, under the mechanisms MCAR $(\psi_1=\psi_2=0)$, MAR $(\psi_2=0)$, and MNAR, respectively. Table 5 shows the results of the three dropout models that were considered. Here, the evidence for the MNAR setting is only border line. Thus, under the MNAR assumption, the maximum likelihood estimates for $\psi 1$ *(-0.29)* and $\psi 2$ *(0.30)* were more or less equal, but with opposite signs, pointing to a relationship between the incremental change and probability of dropout. This finding agrees with the theoretical findings of [3], noting that the dropout often depends upon the increment $y_{ij}-y_{i,j-1}$. This can be justified by the fact that two subsequent measurements are usually positively correlated [3,31]. Furthermore, as can be seen in the dropout model, the parameter estimate $(\psi_2=0.30)$ in our model is

	parameter		
Dropout mechanism	ψ_0	ψ_1	ψ_2
MCAR	1		
MAR	ψ_0 ,*MCAR*	1	
MNAR	ψ_0 ,*MAR*	ψ_0 ,*MAR*	1

Table 2: Initial values for the parameters of the dropout model.

Effect	Parameter	Estimate	Rounded to initial value
Fixed-effects parameters			
Drug 1 intercept	β_{01}	13.9102	13.91
Drug 2 intercept	β_{02}	-3.6667	-3.67
Drug 3 intercept	β_{03}	0.6853	0.69
Drug 1 slope	β_{11}	1.198	1.2
Drug 2 slope	β_{12}	1.5146	1.51
Drug 3 slope	β_{13}	1.3481	1.35
Variance parameters			
Random-intercept variance	d	8.9976	9
Serial process variance	τ^2	3.4068	3.41
Serial process correlation	ρ	1	1
Measurement error variance	σ^2	0.7423	0.74
p-value			
Drug 1 effect		0.0061	
Drug 2 effect		0.6004	

Table 3: Multi-centre data. Parameter estimates of the linear mixed model, used as initial values for the Diggle-Kenward model.

Effect	Parameter	MCAR	MAR	MNAR
Measurement model				
Drug 1 intercept	β_{01}	13.91 (0.92)	13.91 (0.92)	13.90 (0.92)
Drug 2 intercept	β_{02}	-3.67 (1.30)	-3.67 (1.30)	-3.71 (1.30)
Drug 3 intercept	β_{03}	0.69 (1.30)	0.69 (1.30)	0.61 (1.32)
Drug 1 slope	β_{11}	1.20 (0.17)	1.20 (0.17)	1.24 (0.17)
Drug 2 slope	β_{12}	1.51 (0.19)	1.51 (0.19)	1.60 (0.20)
Drug 3 slope	β_{13}	1.35 (0.18)	1.35 (0.18)	1.38 (0.18)
Variance model				
Random-intercept variance	d	8.99 (2.63)	8.99 (2.63)	8.99 (2.63)
Serial process variance	τ^2	3.41 (0.56)	3.41 (0.56)	3.35 (0.57)
Serial process correlation	ρ	1.00 (0.00)	1.00 (0.00)	1(0.01)
Measurement error variance	σ^2	0.74 (0.12)	0.74(0.12)	0.76 (0.12)
-2£		596.99	591.43	595.56

Table 4: Multi-centre data: Maximum likelihood for the parameter estimates (standard errors) under MCAR, MAR, and MNAR assumptions without covariate in the dropout model

Parameter	Dropout Mechanism		
	MCAR	MAR	MNAR
ψ_0	-1.41 (0.31)	-1.52 (0.85)	-2.03 (0.89)
ψ_1		0.01 (0.12)	-0.29 (0.40)
ψ_2			0.30 (0.41)

Table 5: Dropout model: Comparison of the Parameter estimates (standard errors) for MCAR, MAR and MNAR models.

positive, indicating a strong association between the dropout and the increment in the outcome variable (blood component) between two successive times. In addition, as mentioned previously, the maximum likelihood estimates of ψ_1 and ψ_2 have different signs, and furthermore, although there is a strong positive association between ψ_1 and ψ_2, the likelihood based 95% confidence interval for these two parameters (ψ_1, ψ_2) is largely contained in the negative-positive quadrant, that is, the intervals for the parameter space *where* $\psi_1 < 0$ and $\psi 2 > 0$. The full dropout model estimated from the MNAR process is as follows:

$$logit[P\ (dropout)]=-2.03 - 0.29y_{i,j-1}+0.30y_{ij}. \quad (16)$$

One of our interests is to investigate whether the dropout process is MAR or MCAR, in other words, whether or not $\psi_1=\psi_2=0$ in equation (15). The likelihood ratio test is used to compare

The maximum likelihood parameter estimates and minus twice the maximized log-likelihood from the MCAR, MAR and MNAR models appears in Table 4. Comparing the log-likelihood estimates from the MAR and MCAR models, we see that the likelihood ratio for the null hypothesis $\psi_1=\psi_2=0$ is 596.99-591.43=5.56 which is significant with $p<0.01$ on 1 degree of freedom. The test suggests that an MAR dropout process cannot be ruled out, i.e., there is evidence in favour of the MAR, that is, dropouts are not completely at random in the context of the assumed model. Further, the test also support MAR over MNAR as the LRT statistics is 1.43 which is not significant. However, great care has to be taken regarding the sensitivity of the MNAR model to modeling assumptions fit here. From the dropout model in equation (15), it is possible to extend the model by using Table 4.

More observed outcomes. According to Diggle and Kenward [5,32], the dropout in the non-ignorable models tends to depend upon the increment (i.e., the difference between the current and previous measurements, $yij - yi,j-1$). Including this effect implies a switch to the MAR framework. Some insight into this fitted model can be obtained by rewriting it in terms of the increment. In our case, we obtain the following

$$logit[P\ (dropout)]=-2.03+0.30(y_{ij} - y_{i,j-1})+0.01y_{i,j-1}, \quad (17)$$

which indicate that dropout is related to the increment $y_{ij} - y_{i,}j-1$, rather than to any of the actual observations y_{ij} or $y_{i,j-1}$, and such that individuals that improve most (large increments) are very likely to drop out from the study. On the other hand, it is useful also to rewrite this with respect to the increment and the sum of the successive measurements. Thereby, by rewriting equation (15), the fitted dropout model equals

$$logit[P\ (D_i=j \mid D_i \geq j, h_{ij}, y_{ij}, v)]=v0+v1(y_{i,j+}y_{i,j-1})+v2(y_{ij} - y_{i,j-1}),\ j=1, 2, 3, \quad (18)$$

Where $v_1=(\psi_1+\psi_2)/2$ and $v_2=(\psi_1 - \psi_2)/2$. The parameters v_1 and v_2 represent dependence on level and increment in the outcome (blood component), and these quantities are likely to be much less strongly correlated than $y_{i,j}$ and $y_{i,j-1}$. Thus from the fitted MNAR model in equation (18), we have

$$logit[P(dropout)] = -2.03 + 0.005(y_{i,j+} y_{i,j} - 1) - 0.295(y_{ij} - y_{i,j} - 1), \quad (19)$$

Which is to say that the probability of dropout increases with larger negative increments. In the other words, those patients who showed or would have shown a greater decrease in the overall level of the blood component from the previous time have a higher probability of dropout. This is said, given the fact that those patients who have a large improvement compared with the previous time and, a sudden shift in profile, are more likely to drop out of the study Table 5.

In terms of the significance of the drug effects, the corresponding p-values are displayed in Table 6. The p-values of the drug effects at the first point in time do not change much, it being significant in all three cases. However, for all cases, the p-values of the drug2 effects were not statistically significant. It is clear from the different dropout models that the drug effects do not differ to a large extent, the impact caused by drugs might be only on the dropout rate through their effects on the blood component. This is similar to the results from Diggle and Kenward [5] which stated that the drug effects should be made directly into the dropout model, either by using it as constants or allowing the relationship between dropout and outcome to differ between the drugs (Table 6).

Discussion

In this paper we have discussed the performance of the selection models based on Diggle-Kenward approach in terms of the analysis of longitudinal continuous measurements with incomplete data when there are dropouts missing not at random. We considered the use of the Diggle and Kenward [5] model as a tool to assess the sensitivity of a selection model with regard to the modeling assumptions. A model for repeated Gaussian measures, subject to a possibly MNAR assumption were considered. However, a monotone missing pattern was constructed in the model, that is, if a subject's observation was missing for a particular time point, and then all subsequent data for that subject was also to be deleted. Similar to Diggle and Kenward [5], a selection models is specified that uses a logistic regression model to describe the dependency of missing data indicators on the longitudinal measurement. In particular, we have investigated the influence on inference that might be caused of the data by the dropout process. In doing so, we carried out an application for analyzing incomplete longitudinal data with dropout. The model was fitted by using an example from a multi-centre clinical trial data.

The application notably reveals that dropout increases with one element, i.e., large increments. This implied an occurrence of unfavorable values at the previous time. In fact, this case is, in practical terms, very common in fitting selection models of [5,8,32]. Our findings were similar to those of [8,33] in that the example followed in the study yielded parameter estimates for the dropout model that present different signs for current and previous observations, indicating the relationships between incremental changes and the probability of dropping out. The results further suggest that there is an evidence in favour of the prevalence of an MAR process rather than an MCAR process in the context of the assumed model. However, [5,8,18] advise one to take care in interpreting the evidence for such conclusions, using only the data under analysis.

p-value	MCAR	MAR	MNAR
Drug 1 effect	0.0062	0.0061	0.002
Drug 2 effect	0.6002	0.6006	0.601

Table 6: Multi-centre data: p-values for drug effects under MCAR, MAR, and MNAR assumptions.

On the other hand, when all the other modeling assumptions can be guaranteed to hold, the use of the LRT, in a well-defined sense, is inappropriate for hypothesis test for MNAR versus MAR [4]. This is certainly true for the model based on Diggle and Kenward [5] who investigated the tests of MAR null hypothesis against MNAR, but it is important to note that their tests are conditional on the alternative model holding. In practice, such a distinction can only be made using untestable modeling assumptions such a distributional form, see [12]. This problem is really laid bare in Verbeke et al. [26] which showed that for any MNAR model there exist an MAR model that fits the data equally well. Further, they stated that it is not possible to use fit of an MNAR model for or against an MAR model, unless one puts a priori belief in the posited MNAR model. In other words, as the original MNAR model, the MAR model can give the same estimates of predictions to the observed data, and depending on the same parameter vector. This in line with previous study conducted by Gill et al. [34]. For more discussions of examination the differences between an MNAR model and its MAR counterpart, we recommend [12,26] articles. Hence, it is broadly agreed that the role of such MNAR models is in sensitivity analysis that is if the assumptions are changed, the conclusions from the primary (typically MAR) analysis are also changed, as the nature of sensitivity comes due to the non-verifiability in the MNAR model from the data.

Finally, in line with previous studies, for example, [3,8,31,35], the selection model of Diggle and Kenward is viewed as a member of the sensitivity analysis framework. An alternative approach to modeling incomplete longitudinal data under a non-ignorable assumption has frequently been proposed in the literature are the pattern mixture models [23]. There is also what is known as (influence tools) to deal with incomplete longitudinal data with nonignorable missingness and these are useful for detecting subjects that cause non-ignorable dropout, as well as other subjects that lead to non-random missingness. Here, we note that the scope of this study is limited to selection models based on Diggle-Kenward model, the other approaches are not included in this article. On the other hand, in order to assess sensitivity it is useful to obtain further insight into the data by comparing both the selection and the pattern mixture models, for instance, see, [31,35]. While it is not the focus of our current study, sensitivity analyses are an important issue of modeling incomplete longitudinal data and should be routinely conducted. To this end, special attention should go to the comparisons between the various sensitivity analysis frameworks.

Conclusions

Our findings from a non-systematic review for this study revealed that a majority of published literature on epidemiologic studies that used multivariable regression models have not mentioned anything related to testing for statistical interactions, effect modification, or heterogeneity of effect. Although calculation and interpretation of interactive effects are more difficult these are essential if the effects are interactive or synergic. We recommend inclusion of interaction terms that are clinically significant even if the interaction effects are not statistically significant. The failure to identify interactive effects in regression models could lead to significant bias, misinterpretation of the results, and in some instances to incorrect public health interventions with potential adverse implications.

Acknowledgements

We gratefully acknowledge the support we received from Geert Molenberghs (Interuniversity Institute for Biostatistics and Statistical Bioinformatics - Universiteit Hasselt) for providing the PROC IML code from which the current code is based. We are also thankful to Milliken, G. A. (Kansas State University) for the kind permission to use his data.

References

1. Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.
2. Little RJA, Rubin DB (1987) Statistical analysis with missing data. New York: John Wiley and Sons.
3. Molenberghs G, Kenward MG (2007) Missing data in clinical studies. West Sussex, England: John Wiley.
4. Jansen I, Hens N, Molenberghs G, Aerts M, Verbeke G, et al. (2006) The nature of sensitivity in missing not at random models. Computational Statistics and Data Analysis 50: 830-858.
5. Diggle PJ, Kenward MG (1994) Informative drop-out in longitudinal data analysis (with discussion). Applied Statistics 43: 49-93.
6. Glynn RJ, Laird NM, Rubin DB (1986) Selection modelling versus mixture modelling with nonignorable nonresponse. In drawing inferences from self-selected samples, Wainer H (ed.). New York: Springer.
7. Heckman JJ (1976) The common structure of statistical models of trucation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5: 475-492.
8. Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. New York: Springer.
9. Baker SG, Laird NM (1988) Regression analysis for categorical variables with outcomes subject to non-ignorable non-response. Journal of the American Association, 83: 62-69.
10. Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89: 846-866.
11. Gill RD, van der Laan MJ, Robins JM (1997) Coarsening at random: characterizations, conjectures and counterexamples. In Proc. 255-294.
12. Kenward MG (1998) Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. Stat Med 17: 2723-2732.
13. Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable dropout using semiparametric non-response models (with discussion). Journal of the American Statistical Association 94: 1096-1146.
14. Troxel AB, Harrington DP, Lipsitz SR (1998) Analysis of longitudinal data with non-ignorable non-monotone missing values. Applied Statistics 47: 425-438.
15. Jansen I, Molenberghs G (2008) A flexible marginal modeling strategy for non-monotone missing data. Journal of the Royal Statistical Society 171: 347-373.
16. Fitzmaurice GM, Molenberghs G, Lipsitz SR (1995) Regression models for longitudinal binary responses with informative dropouts. Journal of the Royal Statistical Society 57: 691-704.
17. Nordheim EV (1984) Inference from nonrandomly missing categorical data: an example from a genetic study on Turners syndrome. Journal of the American Statistical Association 79: 772-780.
18. Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. New York: Springer.
19. Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90: 106-121.
20. Rotnitzky A, Robins J (1997) Analysis of semi-parametric regression models with non-ignorable non-response. Stat Med 16: 81-102.
21. Robins JM, Rotnitzky A, Scharfstein DO (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response. Journal of the American Statistical Association 93: 1321-1339.
22. Little RJA, Rubin DB (2002) Statistical analysis with missing data. New York: John Wiley and Sons.
23. Little RJA (1993) Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88: 125-134.
24. Little RJA (1994) A class of pattern-mixture models for normal incomplete data. Biometrika 81: 471-483.
25. Rotnitzky A, Cox DR, Bottai M, Robins J (2000) Likelihood-based inference with singular information matrix. Bernoulli 6: 243-284.
26. Verbeke G, Lesaffre E, Spiessens B (2001) The practical use of different strategies to handle dropout in longitudinal studies. Drug Infromation Journal 35: 419-434.
27. Molenberghs G, Beunckens C, Sotto C, Kenward M (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society 70: 371-388.
28. Milliken GA, Johnson DE (2009). Analysis of messy data. Design Experiments, volume 1. Second Edition. Chapman and Hall/CRC.
29. Dmitrienko A, Offen WW, Faries D, Chuang-Stein C, Molenberghs G (2005) Analysis of clinical trial data using the SAS system, Cary, NC: SAS Publishing.
30. Nelder JA, Mead R (1965) A simple method for function minimisation. The Computer Journal 7: 303-313.
31. Kenward MG, Molenberghs G (1999) Parametric models for incomplete continuous and categorical longitudinal data. Stat Methods Med Res 8: 51-83.
32. Molenberghs G, Kenward MG, Lesaffre E (1997) The analysis of longitudinal ordinal data with non-random dropout. Biometrika 84: 33-44.
33. Molenberghs G, Kenward MG, Goetghebeur E (2001) Sensitivity analysis for incomplete contigency tables: The solvenian plebiscite case. Applied Statistics 50: 15-29.
34. Gill RD, van der Laan MJ, Robins JM (1997) Coarsening at random: characterizations, conjectures and counterexamples. In Proc. 255-294.
35. Molenberghs G, Thijs H, Kenward MG, Verbeke G (2003) Sensitivity analysis of continuous incomplete longitudinal outcomes. Statistica Neerlandica 57: 122-135.