METRON - *International Journal of Statistics*
2005, vol. LXIII, n. 2, pp. 143-176

GEERT VERBEKE – GEERT MOLENBERGHS

Longitudinal and incomplete clinical studies

Summary - Repeated measures are obtained whenever an outcome is measured repeatedly within a set of units. The fact that observations from the same unit, in general, will not be independent poses particular challenges to the statistical procedures used for the analysis of such data. The current paper is dedicated to an overview of frequently used statistical models for the analysis of repeated measurements, with emphasis on model formulation and parameter interpretation.
Missing data frequently occur in repeated measures studies, especially in humans. An important source for missing data are patients who leave the study prematurely, so-called dropouts. When patients are evaluated only once under treatment, then the presence of dropouts makes it hard to comply with the intention-to-treat (ITT) principle. However, when repeated measurements are taken then one can make use of the observed portion of the data to retrieve information on dropouts. Generally, commonly used methods to analyse incomplete longitudinal clinical trial data include complete-case (CC) analysis and an analysis using the last observation carried forward (LOCF). However, these methods rest on strong and unverifiable assumptions about the dropout mechanism. Over the last decades, a number of longitudinal data analysis methods have been suggested, providing a valid estimate for, *e.g.*, the treatment effect under less restrictive assumptions.
We will argue that direct likelihood methods, using all available data, require the relatively weak missing at random assumption only. Finally, since it is impossible to verify that the dropout mechanism is MAR we argue that, to evaluate the robustness of the conclusion, a sensitivity analysis thereby varying the assumption on the dropout mechanism should become a standard procedure when analyzing the results of a clinical trial.

Key Words - Longitudinal data; Marginal models; Missing at random; Mixed models.

1. Introduction

In medical science, studies are often designed to investigate changes in a specific parameter which is measured repeatedly over time in the participating subjects. Such studies are called longitudinal studies, in contrast to cross-

Received October 2004.

sectional studies where the response of interest is measured only once for each individual. As pointed out by Diggle *et al.* (2002) the main advantage of longitudinal studies is that they can distinguish changes over time within individuals (longitudinal effects) from differences among people in their baseline values (cross-sectional effects).

In randomized clinical trials, where the aim is usually to compare the effect of two (or more) treatments at a specific time-point, the need and the advantage of taking repeated measures is at first sight less obvious. Indeed, a simple comparison of the treatment groups at the end of the follow-up period is often sufficient to establish the treatment effect(s) (if any) by virtue of the randomization. However, in some instances, it is important to know how the patients have reached their endpoint, *i.e.*, it is important to compare the average profiles (over time) between the treatment groups. Further, longitudinal studies can be more powerful than studies evaluating the treatments at one single time-point. Finally, follow-up studies often suffer from dropout, *i.e.*, some patients leave the study prematurely, for known or unknown reasons. In such cases, a full repeated measures analysis will help in drawing inferences at the end of the study. Since incompleteness usually occurs for reasons outside of the control of the investigators and may be related to the outcome measurement of interest, it is generally necessary to reflect on the process governing incompleteness. Only in special but important cases is it possible to ignore the missingness process.

When patients are examined repeatedly in a clinical trial, missing data can occur for various reasons and at various visits. When missing data result from patient dropout, the missing data pattern is *monotone* pattern. *Non-monotone* missingness occurs when there are intermittent missing values as well. Our focus will be on dropout.

When referring to the missing-value, or non-response, process we will use the terminology of Little and Rubin (2002). A non-response process is said to be *missing completely a 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. Thus, under ignorable dropout, one can literally ignore the missingness process and nevertheless obtain valid estimates of, say, the treatment. Above definitions are conditional on including the correct set of covariates into the model. An overview of the various mechanisms, and their (non-)ignorability under likelihood, Bayesian, or frequentist inference, is given in Table 1.

TABLE 1: *Overview of missing data mechanisms.*

Acronym	Description	Likelih./Bayesian	Frequentist
MCAR	missing completely at random	ignorable	ignorable
MAR	missing at random	ignorable	non-ignorable
MNAR	missing not at random	non-ignorable	non-ignorable

Let us first consider the case where only one follow-up measurement per patient is made. When dropout occurs, and hence there are no follow-up measurements, one usually is forced to discard such a patient from analysis, thereby violating the intention to treat (ITT) principle which stipulates that all randomized patients should be included in the primary analysis and according to the randomisation scheme. Of course, the effect of treatment can be investigated under extreme assumptions, such as, for example, a worst case and a best case scenario, but such scenarios are most often not really helpful.

Early work regarding missingness focused on the consequences of the induced lack of balance of deviations from the study design (Afifi and Elashoff 1966, Hartley and Hocking 1971). Later, algorithmic developments took place, such as the expectation-maximization algorithm (EM, Dempster, Laird and Rubin 1977) and multiple imputation (Rubin 1987). These have brought likelihood-based ignorable analysis within reach of a large class of designs and models. However, they usually require extra programming in addition to available standard statistical software.

In the meantime, however, clinical trial practice has put a strong emphasis on methods such as *complete case analysis* (CC) and *last observation carried forward* (LOCF) or other simple forms of imputation. Claimed advantages include computational simplicity, no need for a full longitudinal model analysis (*e.g.*, when the scientific question is in terms of the last planned measurement occasion only) and, for LOCF, compatibility with the ITT principle. However, a CC analysis assumes MCAR and the LOCF analysis makes peculiar assumptions on the (unobserved) evolution of the response, underestimates the variability of the response and ignores the fact that imputed values are no real data.

On the other hand, a likelihood-based longitudinal analysis requires only MAR, uses all data (obviating the need for both deleting and filling in data) and is also consistent with the ITT principle. Further, it can be shown that also the incomplete sequences contribute to estimands of interest (treatment effect at the end of the study), even early dropouts. For continuous responses, the linear mixed model is quite popular and is a direct extension of ANOVA and MANOVA approaches, but more broadly valid in incomplete data settings. For categorical responses and count data, so-called marginal (*e.g.*, generalized estimating equations, GEE) and random-effects (*e.g.*, generalized linear mixed-effects models, GLMM) approaches are in use. While GLMM parameters can

be fitted using maximum likelihood, the same is not true for the frequentist GEE method but modifications have been proposed to accommodate the MAR assumption (Robins, Rotnitzky and Zhao 1995).

Finally, MNAR missingness can never be fully ruled out based on the observed data only. It is argued that, rather than going either for discarding MNAR models entirely or for placing full faith on them, a sensible compromise is to make them a component of a sensitivity analysis.

In Section 3, we will first focuss on linear models for Gaussian data. In Section 4, we will discuss models for the analysis of discrete outcomes. Section 5 describes simple methods to deal with incomplete data, while more appropriate methods are described in Section 6. Sensitivity analysis is briefly discussed in Section 8.

2. Case studies

2.1. The Toenail data

As a typical longitudinal example, we consider data from a randomized, double blind, parallel group, multicentre study for the comparison of 2 oral treatments (in the sequel coded as A and B) for toenail dermatophyte onychomycosis (TDO). We refer to De Backer *et al.* (1996) for more details about this study. TDO is a common toenail infection, difficult to treat, affecting more than two percent of the population. Antifungal compounds classically used for treatment of TDO need to be taken until the whole nail has grown out healthy. However, new compounds, have reduced the treatment duration to three months. The aim of the present study was to compare the efficacy and safety of two such new compounds, labelled A and B, and administered during 12 weeks.

In total, 2×189 patients were randomized, distributed over 36 centres. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. As a first response, we consider the unaffected naillength (one of the secondary endpoints in the study), measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in mm. Obviously this response will be related to the toesize. Therefore, we will include here only those patients for which the target nail was one of the two big toenails. This reduces our sample under consideration to 146 and 148 subjects respectively. Individual profiles for 30 randomly selected subjects in each treatment group are shown in Figure 1. Our second outcome will be severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of

severe infections decreased over time, and whether that evolution was different for the two treatment groups. A summary of the number of patients in the study at each time-point, and the number of patients with severe infections is given in Table 2.

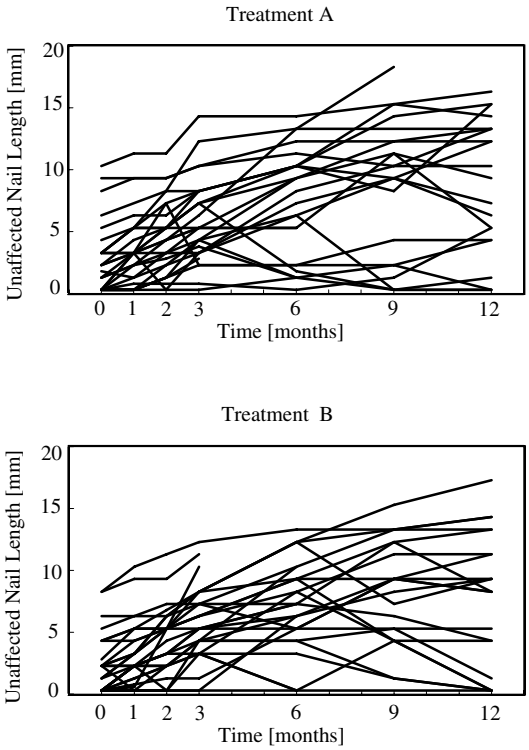

Figure 1. Toenail data: Individual profiles of 30 randomly selected subjects in each treatment arm.

TABLE 2: *Toenail data: Number and percentage of patients with severe toenail infection, for each treatment arm separately.*

	Group A			Group B		
	# severe	# patients	percentage	# severe	# patients	percentage
Baseline	54	146	37.0%	55	148	37.2%
1 month	49	141	34.7%	48	147	32.6%
2 months	44	138	31.9%	40	145	27.6%
3 months	29	132	22.0%	29	140	20.7%
6 months	14	130	10.8%	8	133	6.0%
9 months	10	117	8.5%	8	127	6.3%
12 months	14	133	10.5%	6	131	4.6%

A key issue in the analysis of longitudinal data is that outcome values measured repeatedly within the same subjects tend to be correlated, and this correlation structure needs to be taken into account in the statistical analysis. This is easily seen with paired observations obtained from, *e.g.*, a pre-test/post-test experiment. An obvious choice for the analysis is the paired t-test, based on the subject-specific difference between the two measurements. While an unbiased estimate for the treatment effect can also be obtained from a two-sample t-test, standard errors and hence also p-values and confidence intervals obtained from not accounting for the correlation within pairs will not reflect the correct sampling variability, and hence still lead to wrong inferences. In general, classical statistical procedures assuming independent observations, cannot be used in the context of repeated measurements. In this paper, we will give an overview of the most important models useful for the analysis of clinical trial data, and widely available through commercial statistical software packages.

2.2. Orthdontic Growth data

As an example, we use the orthodontic growth data, introduced by Pothoff and Roy (1964) and used by Jennrich and Schluchter (1986) as well. The data have the typical structure of a clinical trial and are simple yet illustrative. They contain growth measurements for 11 girls and 16 boys. For each subject, the distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14. Figure 2 presents the 27 individual profiles. Little and Rubin [1] deleted 9 of the $[(11+16)\times 4]$ measurements, rendering 9

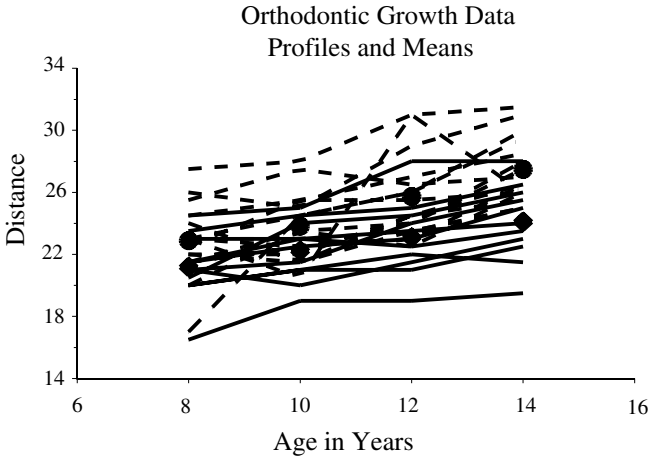

Figure 2. Orthodontic Growth Data. Raw and residual profiles. (Girls are indicated with solid lines. Boys are indicated with dashed lines.)

incomplete subjects which, even though a somewhat unusual practice, has the advantage of allowing a comparison between the incomplete data methods and the analysis of the original, complete data. Deletion is confined to the age 10 measurements and rougly speaking the complete observations at age 10 are those with a higher measurement at age 8. We will put some emphasis on ages 8 and 10, the typical dropout setting, with age 8 fully observed and age 10 partially missing.

3. Linear models for Gaussian data

With repeated Gaussian data, a general, and very flexible, class of parametric models is obtained from a random-effects approach. Suppose that an outcome Y is observed repeatedly over time for a set of persons, and suppose that the individual trajectories are of the type as shown in Figure 3. Obviously, a linear regression model with intercept and linear time effect seems plausible to describe the data of each person separately. However, different persons tend to have different intercepts and different slopes. One can therefore assume that the j^{th} outcome Y_{ij} of subject i ($i = 1, \ldots, N$, $j = 1, \ldots, n_i$), measured at time t_{ij} satisfies $Y_{ij} = \tilde{b}_{i0} + \tilde{b}_{i1} t_{ij} + \varepsilon_{ij}$. Assuming the vector $\tilde{\mathbf{b}}_i = (\tilde{b}_{i0}, \tilde{b}_{i1})'$ of person-specific parameters to be bivariate normal with mean $(\beta_0, \beta_1)'$ and 2×2 covariance matrix D and assuming ε_{ij} to be normal as well, this leads to a so-called linear mixed model. In practice, one will often formulate the model as

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \varepsilon_{ij} \,,$$

with $\tilde{b}_{i0} = \beta_0 + b_{i0}$ and $\tilde{b}_{i1} = \beta_1 + b_{i1}$, and the new random effects $\mathbf{b}_i = (b_{i0}, b_{i1})'$ are now assumed to have mean zero.

The above model can be viewed as a special case of the general linear mixed model which assumes that the outcome vector $\mathbf{Y}_i$ of all n_i outcomes for subject i satisfies

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \varepsilon_{\mathbf{i}} \,, \tag{3.1}$$

in which $\boldsymbol{\beta}$ is a vector of population-average regression coefficients called fixed effects, and where $\mathbf{b}_i$ is a vector of subject-specific regression coefficients. The $\mathbf{b}_i$ are assumed normal with mean vector 0 and covariance D, and they describe how the evolution of the i^{th} subject deviates from the average evolution in the population. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. Note that p and q are the numbers of fixed and subject-specific regression parameters in the model, respectively. The residual components $\varepsilon_{\mathbf{i}}$ are assumed to be independent $N(\mathbf{0}, \Sigma_{\mathbf{i}})$, where Σ_i depends on i only through its dimension n_i. Model (3.1) naturally follows from a so-called two-stage model formulation. First, a linear regression model is specified for

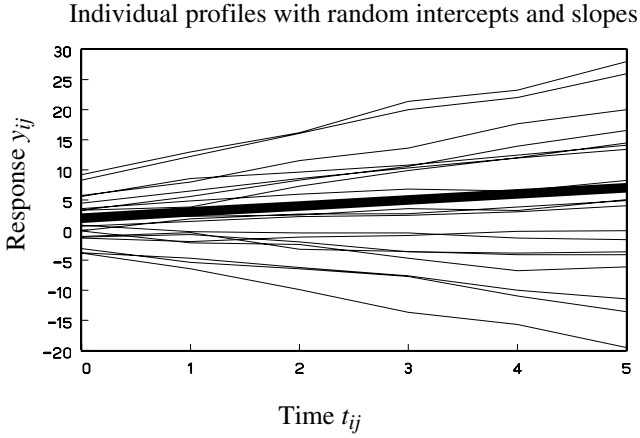

Figure 3. Hypothetical example of continuous longitudinal data which can be well described by a linear mixed model with random intercepts and random slopes. The thin lines represent the observed subject-specific evolutions. The bold line represents the population-averaged evolution. Measurements are taken at six time-points 0, 1, 2, 3, 4, 5.

every subject separately, modelling the outcome variable as a function of time. Afterwards, in the second stage, multivariate linear models are used to relate the subject-specific regression parameters from the first-stage model to subject characteristics such as age, gender, treatment, etc.

Estimation of the parameters in (3.1) is usually based on maximum likelihood (ML) or restricted maximum likelihood (REML) estimation for the marginal distribution of $\mathbf{Y}_i$ which can easily be seen to be

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

Note that model (3.1) implies a model with very specific mean and covariance structures, which may or may not be valid, and hence need to be checked for every specific data set at hand. Note also that, when $\Sigma_i = \sigma^2 I_{n_i}$, with I_{n_i} equal to the identity matrix of dimension n_i, the observations of subject i are independent conditionally on the random effect $\mathbf{b}_i$. The model is therefore called the conditional independence model. Even in this simple case, the assumed random-effects structure still imposes a marginal correlation structure for the outcomes Y_{ij}. Indeed, even if all Σ_i equal $\sigma^2 I_{n_i}$, the covariance matrix in (3.2) is not the identity matrix, illustrating that, marginally, the repeated measurements Y_{ij} of subject i are not assumed to be uncorrelated. Another special case arises when the random effects are omitted from the model. In that case, the covariance matrix of $\mathbf{Y}_i$ is modeled through the residual covariance matrix Σ_i. In the case of completely balanced data, *i.e.*, when n_i is the same for all subjects, and when the measurements are all taken at fixed time-points, one can assume all Σ_i to be equal to a general unstructured covariance matrix Σ,

which results in the classical multivariate regression model. Inference in the marginal model can be done using classical techniques including approximate Wald tests, t-tests, F-tests, or likelihood ratio tests. Finally, Bayesian methods can be used to obtain "empirical Bayes estimates" for the subject-specific parameters $\mathbf{b}_i$ in (3.1). We refer to Henderson *et al.* (1959), Harville (1974, 1976, 1977), Laird and Ware (1982), Verbeke and Molenberghs (2000) for more details about estimation and inference in linear mixed models.

As an illustration, we analyse the unaffected naillength response in the toenail example. The model proposed by Verbeke, Lesaffre and Spiessens (2001) assumes a quadratic evolution for each subject, with subject-specific intercepts, and with correlated errors within subjects. More formally, they assume that Y_{ij} satisfies

$$Y_{ij}(t) = \begin{cases} (\beta_{A0} + b_i) + \beta_{A1} t + \beta_{A2} t^2 + \varepsilon_{\mathbf{i}}(\mathbf{t}), & \text{in group A} \\ (\beta_{B0} + b_i) + \beta_{B1} t + \beta_{B2} t^2 + \varepsilon_{\mathbf{i}}(\mathbf{t}), & \text{in group B,} \end{cases} \tag{3.3}$$

where $t = 0, 1, 2, 3, 6, 9, 12$ is the time in the study, expressed in months. The error components $\varepsilon_{\mathbf{i}}(\mathbf{t})$ are assumed to have common variance σ^2, with correlation of the form $\text{corr}(\varepsilon_{\mathbf{i}}(\mathbf{t}), \varepsilon_{\mathbf{i}}(\mathbf{t} - \mathbf{u})) = \exp(-\varphi \mathbf{u}^2)$ for some unknown parameter φ. Hence, the correlation between within-subject errors is a decreasing function of the time span between the corresponding measurements. Fitted average profiles are shown in Figure 4. An approximate F-test shows that, on average, there is no evidence for a treatment effect ($p = 0.2029$).

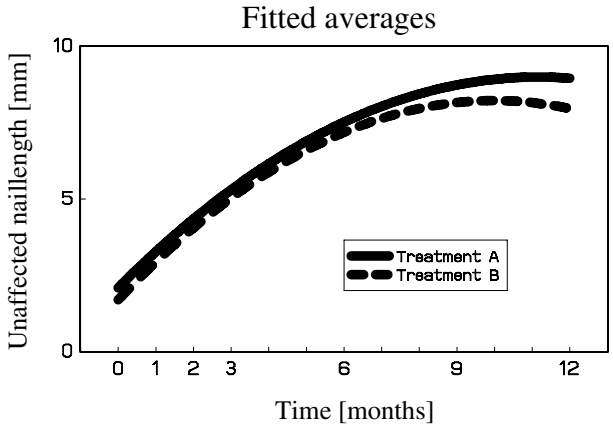

Figure 4. Toenail data: Fitted average profiles based on model (3.3).

Note that, even when interest would only be in comparing the treatment groups after 12 months, this could still be done based on the above fitted model. The average difference between group A and group B, after 12 months, is

given by $(\beta_{A0} - \beta_{B0}) - 12(\beta_{A1} - \beta_{B1}) + 12^2(\beta_{A2} - \beta_{B2})$. The estimate for this difference equals 0.80 mm ($p = 0.0662$). Alternatively, a two-sample t-test could be performed based on those subjects which have completed the study. This yields an estimated treatment effect of 0.77 mm ($p = 0.2584$) illustrating that modelling the whole longitudinal sequence also provides more efficient inferences at specific time-points.

4. Models for discrete outcomes

Whenever discrete data are to be analysed, the normality assumption in the models in the previous section is no longer valid, and alternatives need to be considered. The classical route, in analogy to the linear model, is to specify the full joint distribution for the set of measurements $Y_{ij}, \ldots, Y_{in_i}$ per individual. Clearly, this implies the need to specify all moments up to order n_i. Examples of marginal models can be found in Bahadur (1961), Altham (1978), Efron (1986), Molenberghs and Lesaffre(1994, 1999), Lang and Agresti (1994), and Fahrmeir and Tutz (2001).

Especially for longer sequences and/or in cases where observations are not taken at fixed time-points for all subjects, specifying a full likelihood, as well as making inferences about its parameters, traditionally done using maximum likelihood principles, can become very cumbersome. Therefore, inference is often based on a likelihood obtained from a random-effects approach. Associations and all higher-order moments are then implicitly modelled through a random-effects structure. This will be discussed in Section 4.1. A disadvantage is that the assumptions about all moments are made implicitly, and are very hard to check. As a consequence, alternative methods have been in demand, which require the specification of a small number of moments only, leaving the others completely unspecified. In a large number of cases, one is primarily interested in the mean structure, whence only the first moments need to be specified. Sometimes, there is also interest in the association structure, quantified, for example using odds ratios or correlations. Estimation is then based on so-called generalized estimating equations, and inference no longer directly follows from maximum likelihood theory. This will be explained in Section 4.2. In Section 4.3, both approaches will be illustrated in the context of the toenail data. A comparison of both techniques will be presented in Section 4.4.

4.1. Generalized linear mixed models (GLMM)

As discussed in Section 3, random effects can be used to generate an association structure between repeated measurements. This can be exploited to

specify a full joint likelihood in the context of discrete outcomes. More specifically, conditionally on a vector $\mathbf{b}_i$ of subject-specific regression coefficients, it is assumed that all responses Y_{ij} for a single subject i are independent, satisfying a generalized linear model with mean $\mu_{ij} = h(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i)$ for a pre-specified link function h, and for two vectors $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ of known covariates belonging to subject i aht the j^{th} time point. Let $f_{ij}(y_{ij}|\mathbf{b}_i)$ denote the corresponding density function of Y_{ij}, given $\mathbf{b}_i$. As for the linear mixed model, the random effects $\mathbf{b}_i$ are assumed to be sampled from a normal distribution with mean vector $\mathbf{0}$ and covariance D. The marginal distribution of $\mathbf{Y}_i$ is then given by

$$f(\mathbf{y}_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i \tag{4.4}$$

in which dependence on the parameters $\boldsymbol{\beta}$ and D is suppressed from the notation. Assuming independence accross subjects, the likelihood can easily be obtained, and maximum likelihood estimation becomes available.

In the linear model, the integral in (4.4) could be worked out analytically, leading to the normal marginal model (3.2). In general however, this is no longer possible, and numerical approximations are needed. Broadly, we can distinguish between approximations to the integrand in (4.4), and methods based on numerical integration. In the first approach, Taylor series expansions to the integrand are used, simplifying the calculation of the integral. Depending on the order of expansion and the point around which one expands, slightly different procedures are obtained. We refer to Breslow and Clayton (1993) and to Wolfinger and O'Connell (1993) for an overview of estimation methods. In general, such approximations will be accurate whenever the responses y_{ij} are "sufficiently continuous" and/or if all n_i are sufficiently large. This explains why the approximation methods perform poorly in cases with binary repeated measurements, with a relatively small number of repeated measurements available for all subjects (Wolfinger 1998). Especially in such examples, numerical integration proves very useful. Of course, a wide toolkit of numerical integration tools, available from the optimization literature, can be applied. A general class of quadrature rules selects a set of abscissas and constructs a weighted sum of function evaluations over those. We refer to Hedeker and Gibbons (1994, 1996) and to Pinheiro and Bates (2000) for more details on numerical integration methods in the context of random-effects models.

4.2. Generalized estimating equations (GEE)

Liang and Zeger (1986) proposed so-called generalized estimating equations (GEE) which require only the correct specification of the univariate marginal

distributions provided one is willing to adopt "working" assumptions about the association structure. More specifically, a generalized linear model (McCullagh and Nelder 1989) is assumed for each response Y_{ij}, modelling the mean μ_{ij} as $h(\mathbf{x}'_{ij}\boldsymbol{\beta})$ for a pre-specified link function h, and a vector $\mathbf{x}_{ij}$ of known covariates. In case of independent repeated measurements, the classical score equations for the estimation of $\boldsymbol{\beta}$ are well known to be

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \mu'_i}{\partial \boldsymbol{\beta}} \; V_i^{-1} \; (\mathbf{Y}_i - \mu_i) \;=\; 0 \tag{4.5}$$

where $\mu_i = \mathrm{E}(\mathbf{Y}_i)$ and V_i is a diagonal matrix with $v_{ij} = \mathrm{Var}(Y_{ij})$ on the main diagonal. Note that, in general, the mean-variance relation in generalized linear models implies that the elements v_{ij} also depend on the regression coefficients $\boldsymbol{\beta}$. Generalized estimating equations are now obtained from allowing non-diagonal "covariance" matrices V_i in (4.5). In practice, this comes down to the specification of a "working correlation matrix" which, together with the variances v_{ij} results in a hypothesized covariance matrix V_i for $\mathbf{Y}_i$.

Solving $S(\boldsymbol{\beta}) = 0$ is done iteratively, constantly updating the working correlation matrix using moment-based estimators. Note that, in general, no maximum likelihood estimates are obtained, since the equations are not first-order derivatives of some log-likelihood function for the data under some statistical model. Still, very similar properties can be derived. More specifically, Liang and Zeger (1986) showed that $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed, with mean $\boldsymbol{\beta}$ and with a covariance matrix that can easily be estimated in practice. Hence, classical Wald-type inferences become available. This result holds provided that the mean was correctly specified, whatever working assumptions were made about the association structure. This implies that, strictly speaking, one can fit generalized linear models to repeated measurements, ignoring the correlation structure, as long as inferences are based on the standard errors that follow from the general GEE theory. However, efficiency can be gained from using a more appropriate working correlation model (Mancl and Leroux 1996).

The original GEE approach focusses on inferences for the first-order moments, considering the association present in the data as nuisance. Later on, extensions have been proposed which also allow inferences about higher-order moments. We refer to Prentice (1988), Lipsitz, Laird and Harrington (1991), and Liang, Zeger and Qaqish (1992) for more details on this.

4.3. Application to the toenail data

As an illustration of GEE and GLMM, we analyse the severity of infection binary outcome in the toenail example. We will first apply GEE, based on the

marginal logistic regression model

$$\log\left[\frac{P(Y_i(t)=1)}{1-P(Y_i(t)=1)}\right]=\begin{cases}\beta_{A0}+\beta_{A1}t, & \text{in group A}\\ \beta_{B0}+\beta_{B1}t, & \text{in group B.}\end{cases} \tag{4.5}$$

Furthermore, we use an unstructured 7×7 working correlation matrix. The results are reported in Table 3, and the fitted average profiles are shown in the top graph of Figure 5. Based on a Wald-type test we obtain a significant difference in the average slope between the two treatment groups ($p=0.0158$).

TABLE 3: *Toenail data: Parameter estimates* (*standard errors*) *for a generalized linear mixed model* (GLMM) *and a marginal model* (GEE).

	GLMM	GEE
Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept group A (β_{A0})	−1.63 (0.44)	−0.72 (0.17)
Intercept group B (β_{B0})	−1.75 (0.45)	−0.65 (0.17)
Slope group A (β_{A1})	−0.40 (0.05)	−0.14 (0.03)
Slope group B (β_{B1})	−0.57 (0.06)	−0.25 (0.04)
Random intercepts s.d. (σ)	4.02 (0.38)	

Alternatively, we consider a generalized linear mixed model, modelling the association through the inclusion of subject-specific (random) intercepts. More specifically, we will now assume that

$$\log\left[\frac{P(Y_i(t)=1|b_i)}{1-P(Y_i(t)=1|b_i)}\right]=\begin{cases}\beta_{A0}+b_i+\beta_{A1}t, & \text{in group A}\\ \beta_{B0}+b_i+\beta_{B1}t, & \text{in group B}\end{cases} \tag{4.7}$$

with b_i normally distributed with mean 0 and variance σ^2. The results, obtained using numerical integration methods, are also reported in Table 3. As before, we obtain a significant difference between β_{A1} and β_{B1} ($p=0.0255$).

4.4. Marginal versus hierarchical parameter interpretation

Comparing the GEE results and the GLMM results in Table 3, we observe large differences between the parameter estimates. This suggests that the parameters in both models need to be interpreted differently. Indeed, the GEE approach yields parameters with a population-averaged interpretation. Each regression parameter expresses the average effect of a covariate on the probability of having a severe infection. Results from the generalized linear mixed model however, require an interpretation conditionally on the random effect, *i.e.*, conditionally on the subject. In the context of our toenail example, consider model (4.7) for treatment group A only. The model assumes that the probability of severe infection satisfies a logistic regression model, with the same slope

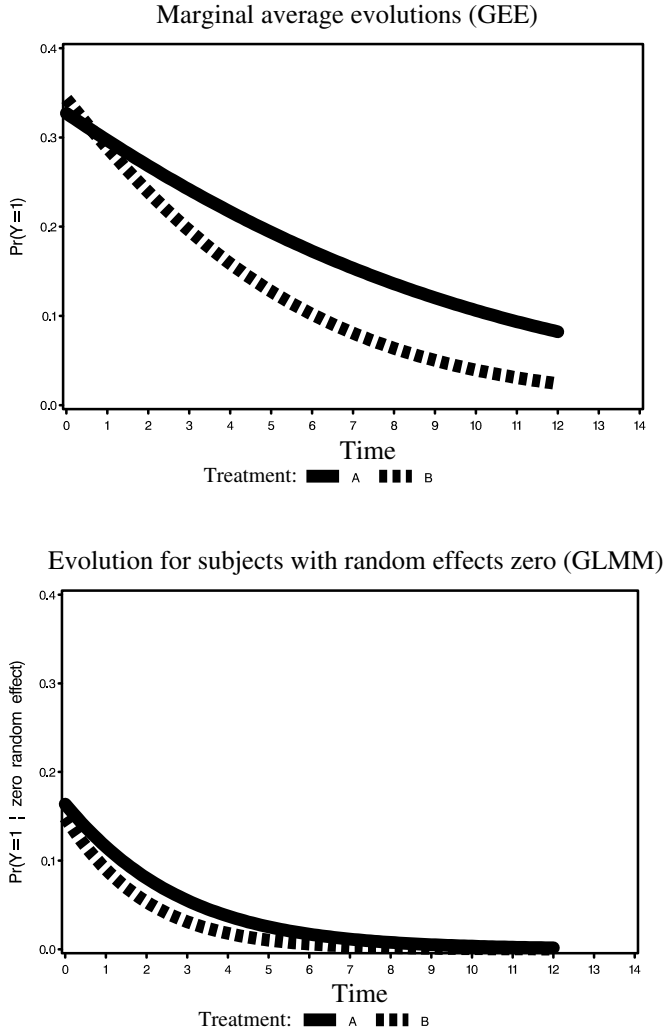

Figure 5. Toenail Data. Treatment-specific evolutions. (a) Marginal evolutions as obtained from the marginal model (4.6) fitted using GEE, (b) Evolutions for subjects with random effects in model (4.7) equal to zero.

for all subjects, but with subject-specific intercepts. The population-averaged probability of severe infection is obtained from averaging these subject-specific profiles over all subjects. This is graphically presented in Figure 6. Clearly, the slope of the average trend is different from the subject-specific slopes, and this effect will be more severe as the subject-specific profiles differ more, *i.e.*, as the random-intercepts variance σ^2 is larger. Formally, the average trend for

group A is obtained as

$$P(Y_i(t) = 1) = E\left[P(Y_i(t) = 1|b_i)\right]$$
$$= E\left[\frac{\exp(\beta_{A0} + b_i + \beta_{A1}t)}{1 + \exp(\beta_{A0} + b_i + \beta_{A1}t)}\right] \neq E\left[\frac{\exp(\beta_{A0} + \beta_{A1}t)}{1 + \exp(\beta_{A0} + \beta_{A1}t)}\right]$$

Hence, the population-averaged evolution is not the evolution for an "average" subject, *i.e.*, a subject with random effect equal to zero. The bottom graph in Figure 5 shows the fitted profiles for an average subject in each treatment group, and these profiles are indeed very different from the population-averaged profiles shown in the top graph of Figure 5 and discussed before. In general, the population-averaged evolution implied by the GLMM is not of a logistic form any more, and the parameter estimates obtained from the GLMM are typically larger in absolute value than their marginal counterparts (Neuhaus, Kalbfleisch and Hauck 1991). However, one should not refer to this phenomenon as bias since the two sets of parameters target at different scientific questions. Note that this difference in parameter interpretation between marginal and random-effects models immediately follows from the non-linear nature, and therefore is absent in the linear mixed model, discussed in Section 3. Indeed, the regression parameter vector $\boldsymbol{\beta}$ in the linear mixed model (3.1)is the same as the regression parameter vector modelling the expectation in the marginal model (3.2).

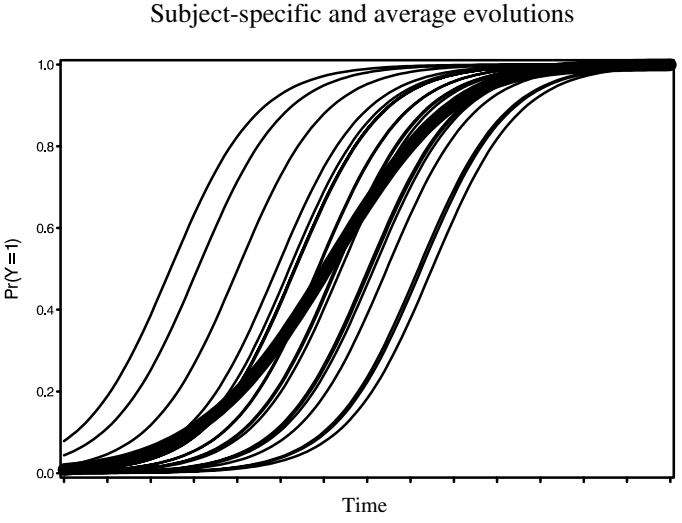

Figure 6. Graphical representation of a random-intercepts logistic model. The thin lines represent the subject-specific logistic regression models. The bold line represents the population-averaged evolution.

5. Methods in common use for incomplete data

Turning to the incomplete data problem, we will focus on two relatively simple methods that have been and still are in extensive use. A detailed account of simple methods to handle missingness is given in Verbeke and Molenberghs (2000).

5.1. Complete case analysis

A *complete case analysis* includes only those cases for analysis, for which all measurements were recorded. This method has obvious advantages. It is very simple to describe and since the data structure is as would have resulted from a complete experiment, standard statistical software can be used without additional work. Further, since the entire estimation is done on the same subset of completers, there is a common basis for inference. Unfortunately, the method suffers from severe drawbacks. First, there is nearly always a substantial loss of information. The impact on precision and power is dramatic. Further, such an analysis will only be representative for patients who remain on study. Of course a complete case analysis could have a role as an auxiliary analysis, especially if a scientific question relates to it. A final important issue about a complete case analysis is that it is only valid when the missingness mechanism is MCAR. However, severe bias can result when the missingness mechanism is MAR but not MCAR. This bias can go both ways, *i.e.*, either overestimating or underestimating the true effect.

5.2. Last observation carried forward

A method that has received a lot of attention (Siddiqui and Ali 1998, Mallinckrodt *et al.* 2003) is *last observation carried forward* (LOCF). As already noted before, in the LOCF method, whenever a value is missing, the last observed value is substituted. For the LOCF approach, the MCAR assumption is necessary but not sufficient for an unbiased estimate. Indeed, it further assumes that subjects' responses would have been constant from the last observed value to the endpoint of the trial. These conditions seldom hold (Verbeke and Molenberghs 2000). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Therefore, carrying observations forward may bias estimates of treatment effects and underestimate the associated standard errors (Verbeke and Molenberghs 2000, Gibbons *et al.* 1993, Heyting, Tolboom and Essers 1992, Lavori *et al.* 1995, Mallinckrodt *et al.* 2001ab). Further, this method artificially increases the amount of information in the data, by treating imputed and actually observed values on equal footing.

Despite its shortcomings, LOCF has been the longstanding method of choice for the primary analysis in clinical trials because of its simplicity, ease of implementation, and the belief that the potential bias from carrying observations forward leads to a "conservative" analysis in comparative trials. An analysis is called conservative when it leads to no treatment difference, while in fact there is treatment difference. However, reports of anti-conservative or liberal behavior of LOCF are common (Kenward *et al.* 2004, Molenberghs *et al.* 2004, Mallinckrodt *et al.* 2004, Little and Yau 1996, Liu and Gould 2002). This means that a LOCF analysis can create treatment effect when none exists. Thus the statement that LOCF analysis has been used to provide a conservative estimate of treatment effect is unacceptable.

Historically, an important motivation behind the simpler methods was their simplicity. Indeed, the main advantage, shared with complete case analysis, is that complete data software can be used. However, with the availability of commercial software tools, such as, for example, the SAS procedures MIXED and NLMIXED and the SPlus and R nlme libraries, this motivation no longer applies.

It is often quoted that LOCF or CC, while problematic for parameter estimation, produce randomization-valid hypothesis testing, but this is questionable. First, in a CC analysis partially observed data are selected out, with probabilities that may depend on post-randomization outcomes, thereby undermining any randomization justification. Second, if the focus is on one particular time point, *e.g.*, the last one scheduled, then LOCF plugs in data. Such imputations, apart from artificially inflating the information content, may deviate in complicated ways from the underlying data (Kenward *et al.* 2004). Third, although the size of a randomization based LOCF test may reach its nominal size under the null hypothesis of no difference in treatment profiles, there will be other regions of the alternative space where the power the LOCF test procedure is equal to its size, which is completely unacceptable.

6. An alternative approach to incomplete data

We will first provide a graphical illustration, using an artificial example, of the various simple methods we have considered and then turn to so-called direct likelihood analysis.

6.1. Illustration of simple methods

Let us take a look at an artificial but insightful example, depicted in Figure 7, which displays the results of the traditional methods, CC and LOCF, next to the result of an MAR method. In this example, the mean response

is supposed to be linear. For both groups (completers and dropouts), the slope is the same, but intercepts differ. Patients with incomplete observations dropped out half way of the study, *e.g.*, because they reached a certain level of the outcome. This is obviously an MAR missingness mechanism. Using a method, valid under the MAR assumption, yields the correct mean profile, being a straight line centered between the mean profiles of the completers and incompleters. If one would perform a CC analysis, the fitted profile will coincide with the mean profile of the complete cases (bold line). Next, under LOCF, data are imputed (dashed line). The resulting fitted profile will be the bold dashed line. Clearly, both traditional methods produce an incorrect result.

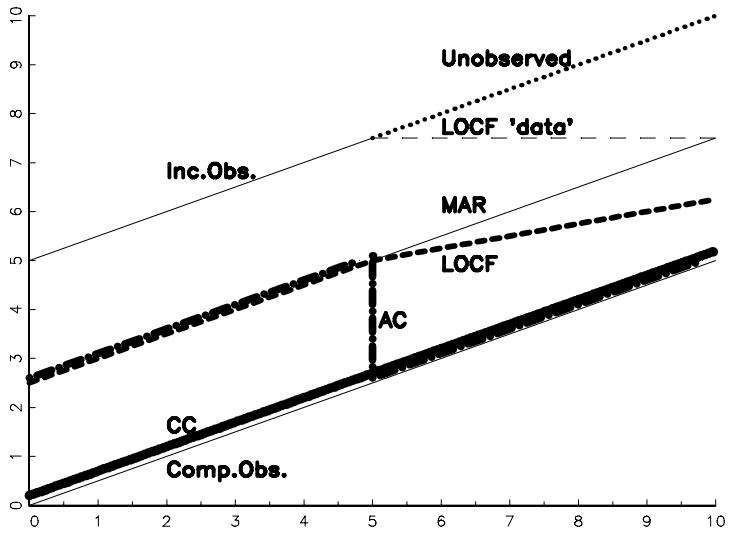

Figure 7. Artificial situation, illustrating the results of the traditional MCAR methods – CC and LOCF – next to the result of the direct likelihood method.

Further, in a traditional available case analysis (AC), one makes use of the information actually available. One such set of estimators could be the treatment-specific mean at a number of designed measurement occasions. With a decreasing sample size over time, means later in time would be calculated using less subjects than means earlier in time. Figure 7 shows a dramatic instance of this, due to the rather extreme nature of this illustrative example. The key message is that such an approach is unable to remove major sources of bias.

6.2. Direct likelihood analysis

For continuous outcomes, Verbeke and Molenberghs (2000) describe likelihood-based mixed-effects models, which are valid under the MAR assumption. Indeed, for longitudinal studies, where missing data are involved, a mixed model only requires that missing data are MAR. As opposed to the traditional techniques, mixed-effects models permit the inclusion of subjects with missing values at some time points (both dropout and intermittent missingness).

This likelihood-based MAR analysis is also termed likelihood-based ignorable analysis, or, as we will be using in the remainder of this entry, a *direct likelihood analysis*. In such a direct likelihood analysis, the observed data are used without deletion nor imputation. In doing so, appropriate adjustments are made to parameters at times when data are incomplete, due to the within-patient correlation.

Thus, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion, such a full longitudinal analysis is a good approach, since the fitted model can be used as the basis for inference at the last occasion.

In many clinical trials the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, which allows the a priori specification of a "saturated" model. For example, a full group-by-time interaction for the fixed effects combined with an unstructured covariance matrix. The direct likelihood analysis is equivalent to a classical MANOVA analysis when data are complete, but more generally valid when they are incomplete. This is a strong answer to the common criticism that a direct likelihood method is making strong assumptions. Indeed, its coincidence with MANOVA for data without missingness shows that the assumptions made are very mild. Therefore, it constitutes a very promising alternative for CC and LOCF. When a relatively large number of measurements is made within a single subject, the full power of random effects modeling can be used (Verbeke and Molenberghs 2000).

The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects, manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values.

A few cautionary remarks are warranted. First, when at least part of the scientific interest is directed towards the nonresponse process, obviously both processes need to be considered. Under MAR, both questions can be answered separately. This implies that a conventional method can be used to study questions in terms of the the outcomes of interest, such as treatment effect and time trend, whereafter a separate model can be considered to study missingness. Second, likelihood inference is often surrounded with references to

the sampling distribution (*e.g.*, to construct measures of precision for estimators and for statistical hypothesis tests, Kenward and Molenberghs 1998). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct, are valid. Thirdly, it may be hard to rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Section 8.

7. Illustration: orthodontic growth data

The simple methods and direct likelihood method from Sections 5 and 6 are now compared using the growth data. For this purpose, a linear mixed model is used, assuming unstructured mean, *i.e.*, assuming a separate mean for each of the eight age×sex combinations, together with an unstructured covariance structure, and using maximum likelihood (ML) as well as restricted maximum likelihood (REML). The mean profiles of the linear mixed model using maximum likelihood for all four data sets, for boys, are given in Figure 8. The girls' profiles are similar and hence not shown.

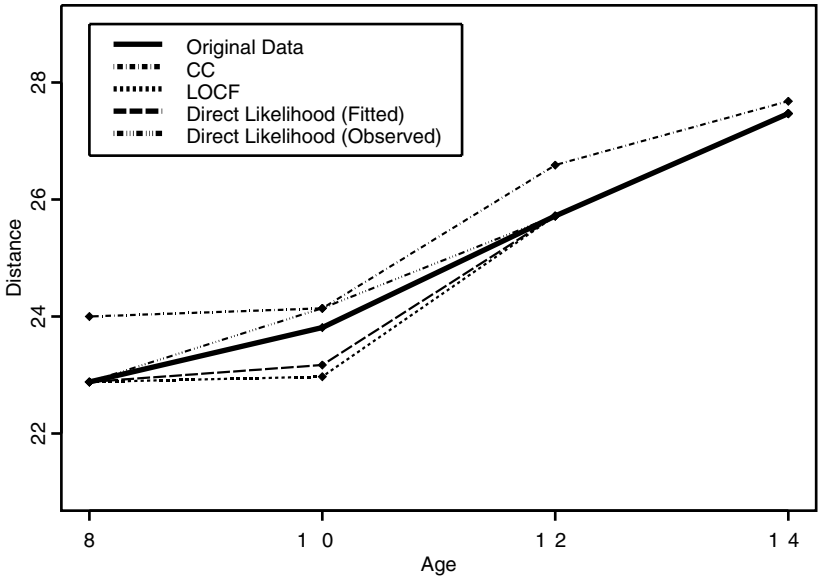

Figure 8. Orthodontic Growth Data. Profiles for the original data, CC, LOCF, and direct likelihood for boys.

Next to this longitudinal approach, we will consider a full MANOVA analysis and a univariate ANOVA analysis, *i.e.*, one per time point. For all of these analyses, Table 4 shows the estimates and standard errors for boys at ages 8

and 10, for the original data and all available incomplete data, as well as for the CC and the LOCF data.

First, we consider the group means for the boys in the original data set in Figure 8, *i.e.*, we observe relatively a straight line. Clearly, there seems to be a linear trend in the mean profile.

In a complete case analysis of the growth data, the 9 subjects which lack one measurement are deleted, resulting in a working data set with 18 subjects. This implies that 27 available measurements will not be used for analysis, a quite severe penalty on a relatively small data set. Observing the profiles for the CC data set in Figure 8, all group means increased relative to the original data set but mostly so at age 8. The net effect is that the profiles overestimate the average length.

For the LOCF data set, the 9 subjects that lack a measurement at age 10 are completed by imputing the age 8 value. It is clear that this procedure will affect the apparently increasing linear trend found for the original data set. Indeed, the imputation procedure forces the means at ages 8 and 10 to be more similar, thereby destroying the linear relationship. Hence, a simple, intuitively appealing interpretation of the trends is made impossible.

TABLE 4: *Orthodontic Growth Data. Comparison of analyses based on means at* (*completely observed age* 8 *and incompletely observed age* 10 *measurement*).

Method	Boys at Age 8	Boys at Age 10
Original Data		
Direct likelihood, ML	22.88 (0.56)	23.81 (0.49)
Direct likelihood, REML	22.88 (0.58)	23.81 (0.51)
MANOVA	22.88 (0.58)	23.81 (0.51)
ANOVA per time point	22.88 (0.61)	23.81 (0.53)
All Available Incomplete Data		
Direct likelihood, ML	22.88 (0.56)	23.17 (0.68)
Direct likelihood, REML	22.88 (0.58)	23.17 (0.71)
MANOVA	24.00 (0.48)	24.14 (0.66)
ANOVA per time point	22.88 (0.61)	24.14 (0.74)
Complete Case Analysis		
Direct likelihood, ML	24.00 (0.45)	24.14 (0.62)
Direct likelihood, REML	24.00 (0.48)	24.14 (0.66)
MANOVA	24.00 (0.48)	24.14 (0.66)
ANOVA per time point	24.00 (0.51)	24.14 (0.74)
Last Observation Carried Forward Analysis		
Direct likelihood, ML	22.88 (0.56)	22.97 (0.65)
Direct likelihood, REML	22.88 (0.58)	22.97 (0.68)
MANOVA	22.88 (0.58)	22.97 (0.68)
ANOVA per time point	22.88 (0.61)	22.97 (0.72)

In case of direct likelihood, we now see two profiles. One for the observed means and one for the fitted means. These two coincide at all ages except age 10. As mentioned earlier, the complete observations at age 10 are those with a higher measurement at age 8. Due to the within-subject correlation, they are the ones with a higher measurement at age 10 as well, and therefore the fitted model corrects in the appropriate direction. The consequences of this are very important. While we are inclined to believe that the fitted means do not follow the observed means all that well, this nevertheless is precisely what we should observe. Indeed, since the observed means are based on a non-random subset of the data, the fitted means take into account all observed data points, as well as information on the observed data at age 8, through the measurements that have been taken for such children, at different time points.

As an aside to this, note that, in case of direct likelihood, the observed average at age 10 coincides with the CC average, while the fitted average does not coincide with anything else. Indeed, if the model specification is correct, then a direct likelihood analysis produces a consistent estimator for the average profile, as if nobody had dropped out. Of course, this effect might be blurred in relatively small data sets due to small-sample variability. Irrespective of the small-sample behavior encountered here, the validity under MAR and the ease of implementation are good arguments that favor this direct likelihood analysis over other techniques.

Let us now compare the different methods by means of Table 4, which shows the estimates and standard errors for boys at age 8 and 10, for the original data and all available incomplete data, as well as for the CC data and the LOCF data.

Table 4 shows some interesting features. In all four cases, a CC analysis gives an upward biased estimate, for both age groups. This is obvious, since the complete observations at age 10 are those with a higher measurement at age 8, as we have seen before. The LOCF analysis gives a correct estimate for the average outcome for boys at age 8. This is not surprising since there were no missing observations at this age. As noted before, the estimate for boys of age 10 is biased downwards. When the incomplete data are analyzed, we see from Table 4 that direct likelihood produces good estimates. The MANOVA and ANOVA per time point analyses give an overestimation of the average of age 10, like in the CC analysis. Further, the MANOVA analysis also yields an overestimation of the average at age 8, again the same as in the CC analysis.

Thus, direct likelihood shares the elegant and appealing features of ANOVA and MANOVA for fully observed data, but is superior with incompletely observed profiles.

8. Sensitivity analysis

When there is residual doubt about the plausibility of MAR, one can conduct a sensitivity analysis. While many proposals have been made, this is still a very active area of research. Obviously, a number of MNAR models can be fitted, provided one is prepared to approach formal aspects of model comparison with due caution. Such analyses can be complemented with appropriate (global and/or local) influence analyses (Verbeke *et al.* 2001). Another route is to construct pattern-mixture models, where the measurement model is considered, conditional upon the observed dropout pattern, and to compare the conclusions with those obtained from the selection model framework, where the reverse factorization is used (Michiels *et al.* 2002, Thijs *et al.* 2002). Alternative sensitivity analyses frameworks are provided by Robins, Rotnitzky, and Scharfstein (1998), Forster and Smith (1998) who present a Bayesian sensitivity analysis, and Raab and Donnelly (1999). A further paradigm, useful for sensitivity analysis, are so-called shared parameter models, where common latent or random effects drive both the measurement process as well as the process governing missingness.

Nevertheless, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data. A discussion of this phenomenon in the survey context has been given in Rubin, Stern, and Vehovar (1995). These authors firstly argue that, in well conducted experiments (some surveys and many confirmatory clinical trials), the assumption of MAR is often to be regarded as a realistic one. Secondly, and very important for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with complete data. Thirdly, while MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on the untestable and often implicit assumptions built in regarding the distribution of the unobserved measurements given the observed ones. The quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. Based on these considerations, we recommend, for primary analysis purposes, the use of ignorable likelihood-based methods or appropriately modified frequentist methods. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis (Verbeke and Molenberghs 2000).

9. Concluding remarks

No doubt repeated measurements occur very frequently in a variety of contexts. This leads to data structures with correlated observations, hence no longer allowing standard statistical modelling assuming independent observa-

tions. Here, we gave a general overview of the main issues in the analysis of repeated measurements, with focuss to a few general classes of approaches often used in practice, and available in many commercially available statistical software packages. A much more complete overview can be found in Diggle *et al.* (2002). Many linear models proposed in the statistical literature for the analysis of continuous data are special cases of linear mixed models discussed in Section 3. We refer to Verbeke and Molenberghs (2000) for more details. We did not discuss non-linear models for continuous data, but the non-linearity implies important numerical and interpretational issues similar to those discussed in Section 4 for discrete data models, and these are discussed in full detail in Davidian and Giltinan (1995) and Vonesh and Chinchilli (1997). An overview of many models for discrete data can be found in Fahrmeir and Tutz (2001). One major approach to the analysis of correlated data is based on random-effects models, both for continuous as well as discrete outcomes. These models are presented in full detail in Pinheiro and Bates (2000).

A variety of models is nowadays available for the analysis of longitudinal data, all posing very specific assumptions. In many other contexts, procedures for model checking or for testing goodness of fit have been developed. For longitudinal data analysis, relatively few techniques are available, and it is not always clear to what extent inferences rely on the underlying parametric assumptions. We refer to Verbeke and Molenberghs (2000) and to Verbeke and Lesaffre (1997) for a selection of available methods for model checking, and for some robustness results, in the context of linear mixed models. Since model checking is far from straightforward, attempts have been made to relax some of the distributional assumptions (Verbeke and Lesaffre 1996).

Regarding incomplete data, a direct likelihood analysis is preferable since it uses all available information, without the need neither to delete nor to impute measurements or entire subjects. It is theoretically justified whenever the missing data mechanism is MAR, which is a more relaxed assumption than MCAR, necessary for simple analyses (CC, LOCF). There is no statistical information distortion, since observations are neither removed (such as in CC analysis) nor added (such as in LOCF analysis). There is software available, such that no additional programming is involved to perform a direct likelihood analysis.

It is very important to realize that, for complete sets of data, direct likelihood, especially with the REML estimation method, is identical to MANOVA (see Table 4). Given the classical robustness of MANOVA, and its close agreement with ANOVA per time point, this provides an extra basis for direct likelihood. Indeed, it is not as assumption-driven as is sometimes believed. This, in addition with the validity of direct likelihood under MAR (and hence its divergence from MANOVA and ANOVA for incomplete data) provides a strong basis for the direct likelihood method.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from Fonds Wetenschappelijk Onderzoek-Vlaanderen Research Project G.0002.98 ôSensitivity Analysis for Incomplete and Coarse Dataö 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.

ALTHAM, P. M. E. (1978) Two generalizations of the binomial distribution, *Applied Statistics*, 27, 162–167.

BAHADUR, R. R. (1961) *A representation of the joint distribution of responses to n dichotomous items. In: Studies in Item Analysis and Prediction,*, H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

BRESLOW, N. E., and CLAYTON, D. G. (1993) Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, 9–25.

DE BACKER, M., DE KEYSER, P., DE VROEY, C., and LESAFFRE, E. (1996) A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250 *mg*/day *vs.* itraconazole 200 *mg*/day–a double-blind comparative trial, *British Journal of Dermatology*, 134, 16–17.

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.

DAVIDIAN, M., and GILTINAN, D. M. (1995) *Nonlinear Models for Repeated Measurement Data*, London: Chapman & Hall.

DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y., and ZEGER, S. L. (2002) *Analysis of Longitudinal Data*, New York: Oxford University Press.

EFRON, B. (1986) Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association*, 81, 709–721.

FAHRMEIR, L., and TUTZ, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models*, Heidelberg: Springer-Verlag.

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

GIBBONS, R. D., HEDEKER, D., ELKIN, I., WATERNAUX, D., KRAEMER, H. C., GREENHOUSE, J. B., SHEA, M. T., IMBER, S. D., SOTSKY, S. M., and WATKINS, J. T. (1993) Some conceptual and statistical issues in analysis of longitudinal psychiatric data, *Archives of General Psychiatry*, 50, 739–750.

HARTLEY, H. O., and HOCKING, R. (1971) The analysis of incomplete data, *Biometrics*, 27, 7783–808.

HARVILLE, D. A. (1974) Bayesian inference for variance components using only error contrasts, *Biometrika*, 61, 383–385.

HARVILLE, D. A. (1976) Extension of the Gauss-Markov theorem to include the estimation of random effects, *The Annals of Statistics*, 4, 384–395.

HARVILLE, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, 72, 320–340.

HEDEKER, D., and GIBBONS, R. D. (1994) A random-effects ordinal regression model for multilevel analysis, *Biometrics*, 50, 933–944.

HEDEKER, D., and GIBBONS, R. D. (1996) MIXOR: A computer program for mixed-effects ordinal regression analysis, *Computer Methods and Programs in Biomedicine*, 49, 157–176.

HENDERSON, C. R., KEMPTHORNE, O., SEARLE, S. R., and VON KROSIG, C. N. (1959) Estimation of environmental and genetic trends from records subject to culling, *Biometrics*, 15, 192–218.

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

JENNRICH, R. I., and SCHLUCHTER, M. D. (1986) Unbalanced repeated measures models with structured covariance matrices, *Biometrics*, 42, 805–820.

KENWARD, M. G., EVANS, S., CARPENTER, J., and MOLENBERGHS, G. (2004) Handling missing responses: time to leave Last Observation Carried Forward (LOCF) behind, Submitted for publication.

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

LAIRD, N. M., and WARE, J. H. (1982) Random effects models for longitudinal data, *Biometrics*, 38, 963–974.

LANG, J. B., and AGRESTI, A. (1994) Simultaneously modeling joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association*, 89, 625–632.

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.

LIANG, K.-Y., and ZEGER, S. L. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, 73, 13–22.

LIANG, K.-Y., ZEGER, S. L., and QAQISH, B. (1992) Multivariate regression analyses for categorical data, *Journal of the Royal Statistical Society, Series B*, 54, 3–40.

LIPSITZ, S. R., LAIRD, N. M. and HARRINGTON, D. P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika*, 78, 153–160.

LITTLE, R. J. A., and RUBIN, D. B. (2002) *Statistical Analysis with Missing Data*, New York: John Wiley & Sons.

LITTLE, R. J. A., and YAU, L. (1996) Intent-to-Treat Analysis in Longitudinal Studies with Drop-Outs, *Biometrics*, 52, 1324–1333.

LIU G., and GOULD A. L. (2002) Comparison of alternative strategies for analysis of longitudinal trials with dropouts, *Journal of Biopharmaceutical Statistics*, 12, 207–26.

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

MALLINCKRODT, C. H., CLARK, W. S., and STACY R. D. (2001A) Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward, *Drug Information Journal*, 35, 4, 1215–1225.

MALLINCKRODT, C. H., CLARK, W. S., and STACY R. D. (2001B) Accounting for dropout bias using mixed-effects models, *Journal of Biopharmaceutical Statistics*, 11, (1 & 2), 9–21.

MALLINCKRODT, C. H., WATKIN, J. G., MOLENBERGHS, G., and CARROLL, R. J. (2004) Choice of the primary analysis in longitudinal clinical trials, *Pharmaceutical Statistics*, 3, 161–169.

MANCL, L. A., and LEROUX, B. G. (1996) Efficiency of regression estimates for clustered data, *Biometrics*, 52, 500–511.

McCullagh, P., and Nelder, J. A. (1989) *Generalized Linear Models*, London: Chapman & Hall.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002) Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout, *Statistics in Medicine*, 21, 1023–1041.

Molenberghs, G., and Lesaffre, E. (1994) Marginal modelling of correlated ordinal data using a multivariate Plackett distribution, *Journal of the American Statistical Association*, 89, 633–644.

Molenberghs, G., and Lesaffre, E. (1999) Marginal modelling of multivariate categorical data, *Statistics in Medicine*, 18, 2237–2255.

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*, 5, 445–464.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review*, 59, 25–30.

Pinheiro, J. C., and Bates, D. M. (2000) *Mixed effects models in S and S-Plus*, New York: Springer-Verlag.

Potthoff, R. F., and Roy, S. N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, 51, 313–326.

Prentice, R. L. (1988) Correlated binary regression with covariates specific to each binary observation, *Biometrics*, 44, 1033–1048.

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

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response, *Journal of the American Statistical Association*, 93, 1321–1339.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, 90, 106–121.

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

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995) Handling "don't know" survey responses: the case of the Slovenian plebiscite, *Journal of the American Statistical Association*, 90, 822–828.

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

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002) Strategies to fit pattern-mixture models, *Biostatistics*, 3, 245–265.

Verbeke, G., and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population, *Journal of the American Statistical Association*, 91, 217–221.

Verbeke, G., and Lesaffre, E. (1997) The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data, *Computational Statistics and Data Analysis*, 23, 541–556.

Verbeke, G., Lesaffre, E., and Spiessens, B. (2001) The practical use of different strategies to handle dropout in longitudinal studies, *Drug Information Journal*, 5, 419–434.

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

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

VONESH, E. F., and CHINCHILLI, V. M. (1997) *Linear and nonlinear models for the analysis of repeated measurements*, New York: Marcel Dekker Inc.

WOLFINGER, R. D. (1998) Towards practical application of generalized linear mixed models. In: B. Marx and H. Friedl (Eds.), *Proceedings of the 13th International Workshop on Statistical Modeling*, pp. 388–395, New Orleans, Louisiana, USA.

WOLFINGER, R., and O'CONNELL, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, 48, 233–243.

GEERT MOLENBERGHS
Center for Statistics
Universiteit Hasselt
Universitaire Campus
B–3590 Diepenbeek (Belgium)
geert.molenberghs@uhasselt.be

GEERT VERBEKE
Biostatistical Centre
Catholic University of Leuven
U.Z. St.-Rafaël
Kapucijnenvoer 35
B–3000 Leuven (Belgium)
geert.verbeke@med.kuleuven.be

GARRETT M. FITZMAURICE

Comment

I congratulate Verbeke and Molenberghs for providing an expository overview of modern methods for the analysis of longitudinal data arising from clinical studies. The authors review regression models for both continuous and discrete outcomes and highlight many of the subtle issues that arise in the analysis of the latter type of outcome variable. Many of the distinctions between the so-called "marginal" and "mixed effects" models for discrete longitudinal data are not always well-understood by statisticians and empirical researchers alike; the authors are to be commended for the clarity with which they have discussed and illustrated the key issues.

A major focus of their article is on the thorny problem of incomplete data, in particular, the monotone missing data patterns produced by dropout in clinical studies. As noted by the authors, a variety of *ad hoc* procedures for handling dropout are widely used. The rationale for many of these procedures is not well-founded and they can result in biased estimates of treatment comparisons. The authors quite rightly point out the limitations of these *ad hoc* techniques. Methods such as "complete-case" (CC) analysis or imputation based on "last observation carried forward" (LOCF), occasionally referred to as "last value carried forward" (LVCF), make strong, and often very unrealistic, assumptions about the responses following dropout. Despite the fact that the shortcomings of these methods are relatively well known, their use in the analysis of clinical studies has persisted. The article by Verbeke and Molenberghs is a very timely reminder of the problems associated with the routine use of these *ad hoc* methods for handling dropout and missingness more generally.

I find myself in wholehearted agreement with the authors when they recommend that the missing at random (MAR) assumption should be at the basis of the default primary analysis of clinical studies. As mentioned by the authors, the MAR assumption can be relaxed in various ways. However, a fundamentally difficult problem arises when the probability of dropout is thought to be related to the specific value that in principle should have been obtained, *i.e.*, when the dropout process is MNAR or *non-ignorable*. As noted by Verbeke

and Molenberghs, the results of analyses based on non-ignorable models for dropout are highly dependent on assumptions that are unverifiable from the data at hand. Consequently, they should be considered a component of a sensitivity analysis rather than a single, final analysis.

It is quite unfortunate that the current status of methods for handling dropout in clinical studies lags so far behind the recent advances in statistical methodology. Despite frequent and well-founded criticisms by statisticians (*e.g.*, Laird, 1988; Heyting *et al.*, 1992; Fitzmaurice, 2003; and many others), LOCF is widely used to handle dropout in clinical studies. Indeed, some regulatory agencies (*e.g.*, the U.S. Food and Drug Administration) seem to actively encourage the continuing use of LOCF as a method for handling dropout, despite all of its obvious shortcomings. One can only hope that the article by Verbeke and Molenberghs, and similar recommendations by others, will quickly remedy this situation.

Because I find myself in substantial agreement with the authors on most of the issues raised in their paper, I will restrict my remaining discussion to one important aspect of the paper that deserves some amplification: the objectives of the analysis of longitudinal clinical studies when there is dropout. When there is dropout, the goals of the analysis need to be clearly specified. Verbeke and Molenberghs have alluded to this in their discussion of the intention-to-treat (ITT) principle. Because many of the currently used methods for handling dropout make unforeseen assumptions about the goals of the analysis, it is worth reviewing the subtle distinctions between two main types of analyses of clinical studies: "pragmatic" and "explanatory" analyses.

Many clinical studies embrace the intention-to-treat (ITT) principle. Broadly speaking, an intention-to-treat analysis follows two main principles: (i) outcome data at all occasions on all subjects randomized to a treatment group should be included in the analysis, including data from those who deviate in any way from the study protocol (*e.g.*, those who dropout), and (ii) the data on all subjects should be analyzed "as randomized" rather than "as treated". That is, if a subject is randomized to one treatment but switches to another treatment prior to completion of the study, that subject is included in the initially assigned treatment group for the purpose of an intention-to-treat analysis. The intention-to-treat analysis is often regarded as a "pragmatic" analysis (Schwartz and Lellouch, 1967), providing an unbiased estimate of the effect of treatment assignment or of the practical "utility" of a treatment after taking into account the "cost" (*e.g.*, withdrawal from treatment due to side effects) of prescribing the treatment. A pragmatic analysis addresses the following scientific question: "What is the effect of starting on one particular treatment rather than another?" In contrast, an "explanatory" analysis (Schwartz and Lellouch, 1967), often referred to as an "as treated" analysis, focuses on what is thought to be the true underlying effects of the different treatments (*e.g.*, the effects

due to the biological or pharmacological properties of drug regimens). An explanatory analysis addresses the following scientific question: "What is the effect of a particular treatment *if somehow* all subjects randomized could be persuaded to remain on their treatment assignment throughout the duration of the study?" Regardless of the relative merits of pragmatic and explanatory analyses, a pragmatic analysis is usually one of the required analyses that are requested by the regulatory agencies (*e.g.*, the U.S. Food and Drug Administration) for approval of new therapies or treatments. Consequently, clinical study investigators often have the goal of producing a pragmatic analysis of their data.

When there is dropout, and further repeated measures of the response can be obtained following dropout (or, at least, on a random sample of those who dropout), both pragmatic and explanatory analyses can be conducted. Furthermore, the results from pragmatic and explanatory analyses can be compared. However, when no further repeated measures of the response are obtained following dropout, many of the methods widely used for analyzing longitudinal data provide either an explanatory analysis or are somewhat ambiguous regarding the type of analysis that is being conducted. Consequently, the results of the analysis do not necessarily match the intended goal of the study.

To highlight the main differences between the pragmatic and explanatory analysis, some additional notation is required. Let Y_{ij}^* denote the response for the i^{th} subject at the j^{th} occasion *assuming that the subject remains on the assigned treatment throughout the duration of the study*. Note that this will be "counter-factual" if the i^{th} subject is a dropout. Let Y_{ij} denote the actual response for the i^{th} subject at the j^{th} occasion. Note that if a subject remains on his or her treatment assignment throughout the study, then $Y_{ij} = Y_{ij}^*$. To understand the key differences between the pragmatic and explanatory analysis, we assume in what follows that further repeated measures of the response can be obtained following dropout. The goal of a pragmatic analysis is to make inferences about $\mu_i = E(Y_i|X_i)$. In a longitudinal clinical study the pragmatic analysis compares the treatment groups, as randomized, in terms of the average rate of change in the outcome at all occasions, regardless of whether measures were made when a subject was on or off the study protocol. Note that the pragmatic analysis follows the intention-to-treat principle by including all repeated measures on all individuals, without excluding any subjects or any outcomes. In contrast, the goal of an explanatory analysis is to make inferences about $\mu_i^* = E(Y_i^*|X_i)$. Note that the explanatory analysis includes only outcome data from individuals prior to dropout; the outcomes following dropout need to be imputed in some manner, with the method of imputation conditioning on the treatment group to which an individual was randomized rather than the treatments actually received following dropout. Since the explanatory analysis does not include "off-treatment" outcomes when they are available, it is in

violation of the intention-to-treat principle of including outcome data from all subjects at all possible occasions.

Next, consider some of the widely used *ad hoc* methods for handling dropout discussed by Verbeke and Molenberghs. In the complete-case (CC) analysis it is somewhat unclear what is the goal of the analysis since the referent population can be somewhat ambiguous. However, if it can be assumed that dropout is an MCAR process (*i.e.*, the "study completers" are a random subset of the sample), then the complete-case analysis provides an explanatory, rather than a pragmatic, analysis. With an *ad hoc* imputation method such as LOCF, it is unclear whether it represents an attempt to impute Y_{ij}^{*} or Y_{ij}; as a result, it is ambiguous whether the goal of the analysis is explanatory or pragmatic. More principled methods for handling dropout, *e.g.*, imputations based on propensity scores, the direct likelihood approach advocated by Verbeke and Molenberghs, and inverse probability weighting methods (*e.g.*, Heyting *et al.*, 1992; Robins, Rotnitzky and Zhao, 1995), effectively impute Y_{ij}^{*} since these methods condition on the treatment to which the subject was randomized rather than the treatments actually received. As a result, even these more principled methods provide an explanatory, rather than a pragmatic, analysis.

In summary, when there is dropout in longitudinal clinical studies, it is not generally appreciated that most of the commonly used methods for handling dropout, including the direct likelihood approach recommended by the authors, come closest to providing an inherently explanatory analysis. This fact appears to have escaped the attention of many of the regulatory agencies that require a pragmatic analysis as part of the approval process for new therapies or treatments.

Finally, careful consideration of the objectives of the analysis of longitudinal clinical studies raises important implications for study design. In particular, when the main analytic goal is to produce a pragmatic analysis, studies should be designed to take further measures of the outcomes following dropout, if not on all subjects who dropout, then at least on a random subsample. The additional outcome data following dropout can then be used for imputation of the incomplete data, and both pragmatic and explanatory analysis can be conducted. For example, Hogan and Laird (1996) describe a pattern-mixture model-based approach for conducting a pragmatic analysis when further measures of the outcomes are available on a random sample of subjects who dropout. An appealing aspect of their model is that it can be used to produce both pragmatic and explanatory analyses. When it is not feasible to obtain repeated measures of the outcome following dropout, then information on the treatments actually received following dropout should be collected. The latter information can then be used in an "imputation model" that conditions on the treatments that were received rather than the treatment to which a subject was randomized. Little and Yau (1996) refer to these as "as treated" imputations

rather than "as randomized" imputations. If information on the treatments actually received following dropout cannot be collected, an imputation model needs to be adopted that conditions on the treatments that were assumed to have been received (Little and Yau, 1996). Given these imputed values, a pragmatic analysis can then proceed by conducting analyses, applied to the observed and imputed data, that compare the treatments *as randomized*. As noted by Little and Yau (1996), in a pragmatic analysis the critical distinction is between the model used for imputation and the model used for analysis. The imputation model conditions on the treatments actually received (or makes assumptions about the treatments actually received in the context of a sensitivity analysis) rather than on the treatment randomized; while the analysis model is based on the treatments as randomized, rather than on the treatments actually received.

Once again, I congratulate Verbeke and Molenberghs for their interesting overview of recent methods for the analysis of longitudinal data arising from clinical studies and for highlighting the shortcomings of many *ad hoc*, but widely used, methods for handling dropout. The methods for handling dropout in current use lag far behind the advances in statistical methodology over the last 25 years. Verbeke and Molenberghs have made sound recommendations for the proper handling of missing data in the analysis of a longitudinal clinical study; their article should be required reading for all statisticians and regulators involved in the analysis of such a study.

ACKNOWLEDGMENTS

The author is grateful for the support provided by grant GM 29745 from the U.S. National Institutes of Health.

REFERENCES

FITZMAURICE, G. M. (2003) Methods for handling dropouts in longitudinal clinical trials, *Statistica Neerlandica*, 57, 75–99.

HEYTING, A., TOLBOOM, J. T., and ESSERS, J. G. (1992) Statistical handling of drop-outs in longitudinal clinical trials, *Statistics in Medicine*, 11, 2043–2061.

HOGAN, J. W., and LAIRD, N. M. (1996). Intention-to-treat analyses for incomplete repeated measures data, *Biometrics*, 52, 1002–1017.

LAIRD, N. M. (1988) Missing data in longitudinal studies, *Statistics in Medicine*, 7, 305–315.

LITTLE, R. J. A., and YAU, L. (1996) Intent-to-treat analysis for longitudinal studies with drop-outs, *Biometrics*, 52, 1324–1333.

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

SCHWARTZ, D., and LELLOUCH, L. (1967) Explanatory and pragmatic attitudes in therapeutical trials, *Journal of Chronic Diseases*, 20, 637–648.

GARRETT M. FITZMAURICE
Division of General Medicine
Brigham and Women's Hospital
1620 Tremont Street
Boston, MA 02120 (U.S.A.)
and
Department of Biostatistics
Harvard School of Public Health
655 Huntington Avenue
Boston MA 02115 (U.S.A.)