Made available by Hasselt University Library in https://documentserver.uhasselt.be

Random effects models for longitudinal data

Peer-reviewed author version

VERBEKE, Geert; MOLENBERGHS, Geert & Rizopoulos, Dimitris (2010) Random effects models for longitudinal data. In: van Montfort, Kees & Oud, Johan H.L. & Satorra, Albert (Ed.) Longitudinal Research with Latent Variables Longitudinal Research with Latent Variables, p. 37-69.

Handle: http://hdl.handle.net/1942/10773

Random Effects Models for Longitudinal Data

Geert Verbeke[1,2] **Geert Molenberghs**[2,1] **Dimitris Rizopoulos**[3]

[1]I-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium

[2]I-BioStat, Universiteit Hasselt, Agoralaan, B-3590 Diepenbeek, Belgium

[3] Department of Biostatistics, Erasmus University Medical Center,
NL-3000 CA Rotterdam, the Netherlands

1 Introduction

Repeated measures are obtained whenever an outcome is measured repeatedly within a set of units. An array of examples is presented in Section 2. 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. In Section 3, an overview is presented of the most commonly used method for both Gaussian and non-Gaussian repeated measures.

Given their ubiquity, it is not surprising that methodology for repeated measures has emerged in a variety of fields. For example, Laird and Ware (1982) proposed the so-called linear mixed-effects models in a biometric context, whereas Goldstein (1979) proposed what is termed multilevel modeling in the framework of social sciences. Though the nomenclature is different, the underlying idea is the same: hierarchical data are modeled by introducing random coefficients, constant within a given level but changing across levels. Let us provide two examples. In a longitudinal context, where data are hierarchical because a given subject is measured repeatedly over time, a random effect is one that remains constant within a patient but changes across patients. A typical example of a multilevel setting consists of school children that are nested within classes which are, in turn, nested within schools. Random effects are then introduced to capture class-level as well as school-level variability. Examples abound in other fields as well. Methodology has been developed for continuous, Gaussian data, as well as for non-Gaussian settings, such as binary, count, and ordinal data. Overviews can be found in Verbeke and Molenberghs (2000) for the Gaussian case and in Molenberghs and Verbeke (2005) for the non-Gaussian setting.

In addition, a number of important contemporary extensions and issues will be discussed.

First, it is not uncommon for multiple repeated measures sequences to be recorded and analyzed simultaneously, leading to so-called multivariate longitudinal data. This poses specific methodological and computational challenges, especially when the problem is high-dimensional. An overview is presented in Section 4.

Second, it is quite common for longitudinal data to be collected in conjunction with time-to-event outcomes. An overview is presented in Section 5. Broadly, there are three main situations where this can occur: (a) The emphasis can be on the survival outcome with the longitudinal outcome(s)

acting as a covariate process; (b) interest can be on both simultaneously, such as in the evaluation of surrogate markers in clinical studies, with a longitudinal marker for a time-to-event outcome; (c) the survival process can act, either in discrete or continuous time, as a dropout process on the longitudinal outcome.

The above considerations lead us to include a third main theme, surrogate marker evaluation, in Section 6, and a fourth and final theme, incomplete data, in Section 7.

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 1.

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,

Table 1: *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%

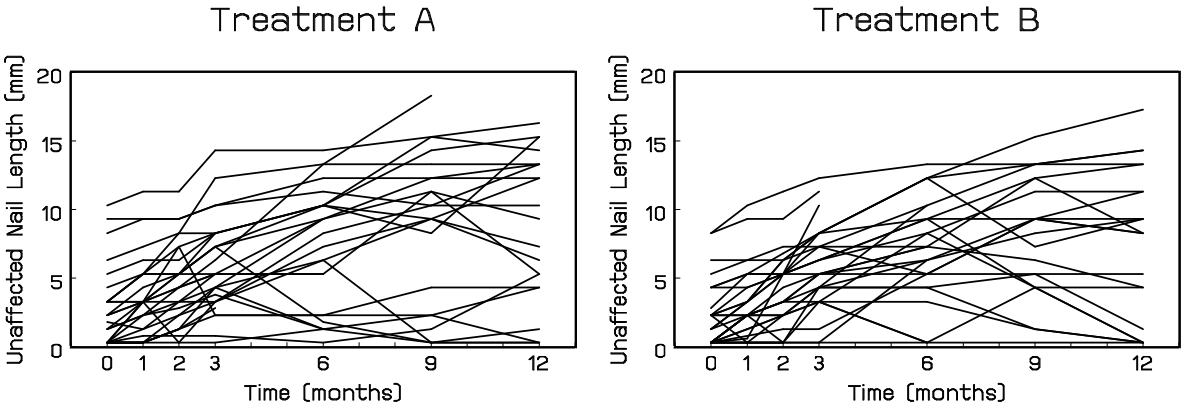

Figure 1: *Toenail data: Individual profiles of 30 randomly selected subjects in each treatment arm.*

classical statistical procedures assuming independent observations, cannot be used in the context of repeated measurements. In this chapter, 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 Hearing Data

In a hearing test, hearing threshold sound pressure levels (dB) are determined at different frequencies to evaluate the hearing performance of a subject. A hearing threshold is the lowest signal intensity a subject can detect at a specific frequency. In this study, hearing thresholds measured at eleven different frequencies (125Hz, 250Hz, 500Hz, 750Hz, 1000Hz, 1500Hz, 2000Hz, 3000Hz, 4000Hz, 6000Hz and 8000Hz), obtained on 603 male participants from the Baltimore Longitudinal Study of Aging (BLSA, Shock *et al.* 1984), are considered. Hearing thresholds are measured at the left as well as at the right ear, leading to 22 outcomes measured repeatedly over time. The number of visits per subject varies from 1 to 15 (a median follow-up time of 6.9 years). Visits are unequally spaced. The age at first visit of the participants ranges from 17.2 to 87 years (with a median age at first visit of 50.2 years). Analyses of the hearing data collected in the BLSA study

can be found in Brant and Fozard (1990), Morrell and Brant (1991), Pearson *et al.* (1995), Verbeke and Molenberghs (2000), and Fieuws and Verbeke (2006). It is well known that the hearing performance deteriorates as one gets older, which will be reflected by an increase in hearing threshold over time. The aim of our analysis will be to investigate whether this interaction between time and age is frequency related. Also of interest is to study the association between evolutions at different frequencies. Both questions can only be answered using a joint model for all 22 outcomes.

2.3 Liver Cirrhosis Data

As an illustrative example for the joint modeling of longitudinal and time-to-event data we consider data on 488 patients with histologically verified liver cirrhosis, collected in Copenhagen from 1962 to 1969 (Andersen *et al.*, 1993). Liver cirrhosis is the condition in which the liver slowly deteriorates and malfunctions due to chronic injury. From the 488 patients, 251 were randomly assigned to receive prednisone and 237 placebo. Patients were scheduled to return at 3, 6, and 12 months, and yearly thereafter, and provide several biochemical values related to liver function. Our main research question here is to test for a treatment effect on survival after adjusting for one of these markers namely, the prothrombin index, which is indicative of the severity of liver fibrosis. Since the prothrombin levels are in fact the output of a stochastic process generated by the patients and is only available at the specific visit times the patients came to the study center, it constitutes a typical example of time-dependent covariate measured intermittently and with error.

2.4 Orthodontic Growth Data

Consider the orthodontic growth data, introduced by Potthoff 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 (2002) deleted 9 of the $[(11+16)\times 4]$ measurements, rendering 9 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.

2.5 Age-related Macular Degeneration Trial

These data arise from a randomized multi-center clinical trial comparing an experimental treatment (interferon-α) to a corresponding placebo in the treatment of patients with age-related macular degeneration. In this chapter we focus on the comparison between placebo and the highest dose (6 million units daily) of interferon-α (Z), but the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with macular degeneration progressively lose vision. In the trial, the patients' visual acuity was assessed

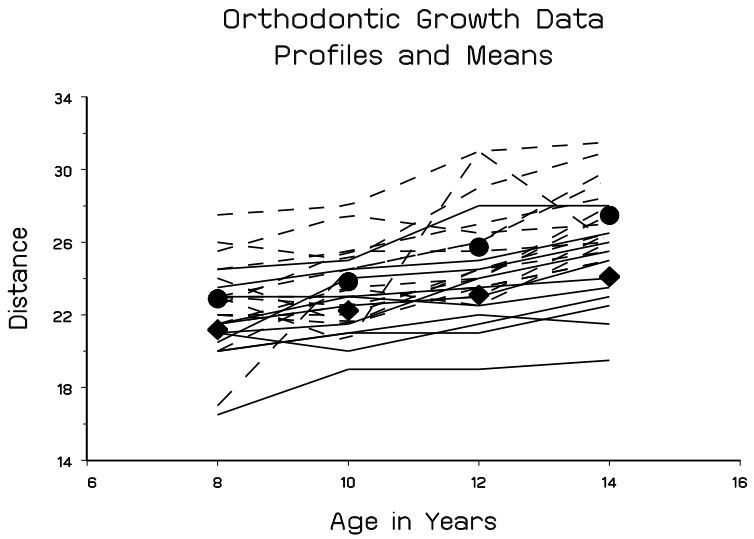

Figure 2: *Orthodontic Growth Data. Raw and residual profiles. (Girls are indicated with solid lines. Boys are indicated with dashed lines.)*

Table 2: *The Age-related Macular Degeneration Trial. Mean (standard error) of visual acuity at baseline, at 6 months and at 1 year according to randomized treatment group (placebo* versus *interferon-α).*

Time point	Placebo	Active	Total
Baseline	55.3 (1.4)	54.6 (1.3)	55.0 (1.0)
6 months	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
1 year	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

at different time points (4 weeks, 12 weeks, 24 weeks, and 52 weeks) through their ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The raw patient's visual acuity is the total number of letters correctly read. In addition, one often refers to each line with at least 4 letters correctly read as a 'line of vision.'

Table 2 shows the visual acuity (mean and standard error) by treatment group at baseline, at 6 months, and at 1 year. Visual acuity can be measured in several ways. First, one can record the number of letters read. Alternatively, dichotomized versions (at least 3 lines of vision lost, or at least 3 lines of vision lost) can be used as well. Therefore, these data will be useful to illustrate methods for the joint modeling of continuous and binary outcomes, with or without taking the longitudinal nature into account. In addition, though there are 190 subjects with both month 6 and month 12 measurements available, the total number of longitudinal profiles is 240, but for only 188 of these have the four follow-up measurements been made.

Thus indeed, 50 incomplete subjects could be considered for analysis as well. Both intermittent missingness as well as dropout occurs. An overview is given in Table 3. Thus, 78.33% of the profiles are complete, while 18.33% exhibit monotone missingness. Out of the latter group, 2.5%

Table 3: *The Age-related Macular Degeneration Trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.*

Measurement occasion					
4 wks	12 wks	24 wks	52 wks	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

or 6 subjects have no follow-up measurements. The remaining 3.33%, representing 8 subjects, have intermittent missing values. Thus, as in many of the examples seen already, dropout dominates intermediate patterns as the source of missing data

3 Modeling Tools for Longitudinal Data

In many branches of 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-sectional studies where the response of interest is measured only once for each individual. As pointed out by Diggle *et al.* (2002) one of the main advantages 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, for example, where the aim usually is to compare the effect of two (or more) treatments at a specific time-point, the need and 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 necessary to compare the average profiles (over time) between the treatment groups. Furthermore, longitudinal studies can be more powerful than studies evaluating the treatments at one single time-point. Finally, follow-up studies more often than not 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. Given that 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, 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. We will return to the missing data issue in Section 7. We are now in a position to discuss first a key modeling tool for Gaussian longitudinal data, where after we will switch the the non-Gaussian case.

3.1 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 people, and suppose that the individual trajectories are of the type 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 people tend to have different intercepts and different slopes. One can therefore assume that the jth 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 $\tilde{b}_i = (\tilde{b}_{i0}, \tilde{b}_{i1})^\top$ of person-specific parameters to be bivariate normal with mean $(\beta_0, \beta_1)^\top$ 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 $b_i = (b_{i0}, b_{i1})^\top$ are now assumed to have mean zero.

The above model is a special case of the general linear mixed model which assumes that the outcome vector Y_i of all n_i outcomes for subject i satisfies

$$Y_i \;=\; X_i \beta \;+\; Z_i b_i \;+\; \varepsilon_i, \tag{3.1}$$

in which β is a vector of population-average regression coefficients, called fixed effects, and where b_i is a vector of subject-specific regression coefficients. The b_i are assumed normal with mean vector $\mathbf{0}$ and covariance D, and they describe how the evolution of the ith 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 ε_i are assumed to be independent $N(0, \Sigma_i)$, where Σ_i depends on i only through its dimension n_i.

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 Y_i which can easily be seen to be

$$Y_i \;\sim\; N(X_i \beta, Z_i D Z_i^\top + \Sigma_i). \tag{3.2}$$

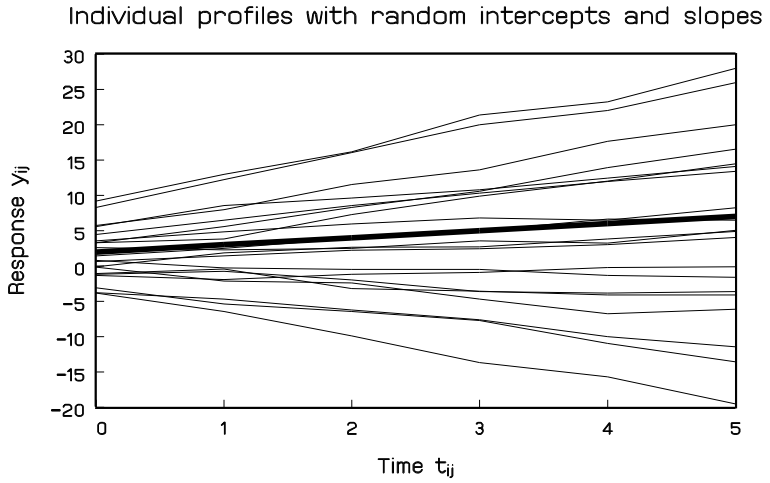

Figure 3: *Hypothetical example of continuous longitudinal data which can be well described by a linear mixed model with random intercepts and random slopes. The thin lines represent the observed subject-specific evolutions. The bold line represents the population-averaged evolution. Measurements are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

Note that model (3.1) implies a model with very specific mean and covariance structures, which may or may not be valid, and hence needs 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 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 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 b_i in (3.1). We refer to Henderson *et al.* (1959), Harville (1974, 1976, 1977), Laird and Ware (1982), Verbeke and Molenberghs (2000), and Fitzmaurice, Laird, and Ware (2004) for more details about estimation and inference in linear mixed models.

3.2 Models for Discrete Outcomes

Whenever discrete data are to be analyzed, 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 modeled through a random-effects structure. This will be discussed in Section 3.2.1. A disadvantage is that the assumptions about all moments are made implicitly, and therefore 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 3.2.2. In Section 3.3, both approaches will be illustrated in the context of the toenail data. A comparison of both techniques will be presented in Section 3.2.3.

3.2.1 Generalized Linear Mixed Models (GLMM)

As discussed in Section 3.1, 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 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 $\mu_{ij} = g({x_{ij}}^\top \beta + {z_{ij}}^\top b_i)$ for a pre-specified link function $g(\cdot)$, and for two vectors x_{ij} and z_{ij} of known covariates belonging to subject i aht the jth time point. Let $f_{ij}(y_{ij}|b_i)$ denote the corresponding density function of Y_{ij}, given b_i. As for the linear mixed model, the random effects b_i are assumed to be sampled from a normal distribution with mean vector 0 and covariance D. The marginal distribution of Y_i is then given by

$$f(y_i) \quad = \quad \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i) f(b_i) db_i, \tag{3.3}$$

in which dependence on the parameters β 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 (3.3) 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 (3.3), 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), Wolfinger and O'Connell (1993), Molenberghs and Verbeke (2005), and Fitzmaurice, Laird, and Ware (2004) 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.

3.2.2 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}, modeling the mean μ_{ij} as $g({x_{ij}}^\top \beta)$ for a pre-specified link function $g(\cdot)$, and a vector x_{ij} of known covariates. In case of independent repeated measurements, the classical score equations for the estimation of β are well known to be

$$S(\beta) \;=\; \sum_i \frac{\partial \mu_i^\top}{\partial \beta}\, V_i^{-1}\, (Y_i - \mu_i) \;=\; 0, \tag{3.4}$$

where $\mu_i = \mathrm{E}(Y_i)$ 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 β. Generalized estimating equations are now obtained from allowing non-diagonal 'covariance' matrices V_i in (3.4). 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 Y_i.

Solving $S(\beta) = 0$ 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. Still, very similar properties can be derived. More specifically, Liang and Zeger (1986) showed that $\widehat{\beta}$ is asymptotically normally distributed, with mean β 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 focuses 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.

3.2.3 Marginal versus Hierarchical Parameter Interpretation

Comparing the GEE results and the GLMM results in Table 4, we observe large differences between the corresponding parameter estimates. This suggests that the parameters in both models have a different interpretation. 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 (3.7) for treatment group A only. The model assumes that the probability of severe infection satisfies a logistic regression model, with the same slope 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 4. 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

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

Hence, the population-averaged evolution is not the evolution for an 'average' subject, i.e., a subject with random effect equal to zero. The right hand graph in Figure 6 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 left hand graph of Figure 6 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 given that the two sets of parameters target at different scientific questions. Observe that this difference in parameter interpretation between marginal and random-effects models immediately follows from their non-linear nature, and therefore is absent in the linear mixed model, discussed in Section 3.1. Indeed, the regression parameter vector β in the linear mixed model (3.1) is the same as the regression parameter vector modeling the expectation in the marginal model (3.2).

3.3 Analysis of Toenail Data

As an illustration, we analyze unaffected nail length 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

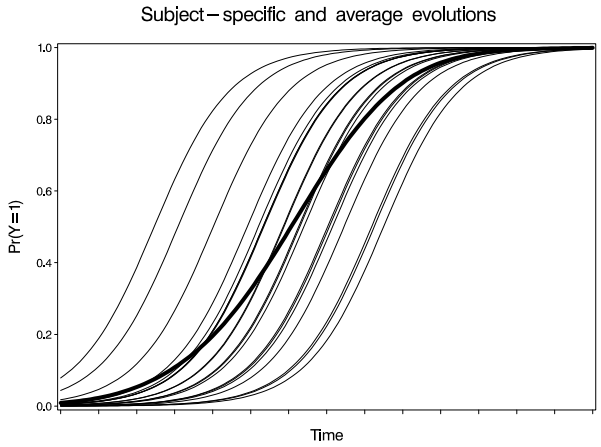

Figure 4: *Graphical representation of a random-intercepts logistic model. The thin lines represent the subject-specific logistic regression models. The bold line represents the population-averaged evolution.*

assume that Y_{ij} satisfies

$$Y_{ij}(t) \quad = \quad \begin{cases} (\beta_{A0} + b_i) + \beta_{A1} t + \beta_{A2} t^2 + \varepsilon_i(t), & \text{in group A} \\ (\beta_{B0} + b_i) + \beta_{B1} t + \beta_{B2} t^2 + \varepsilon_i(t), & \text{in group B,} \end{cases} \tag{3.5}$$

where $t = 0, 1, 2, 3, 6, 9, 12$ is the number of months since randomization. The error components $\varepsilon_i(t)$ are assumed to have common variance σ^2, with correlation of the form $\text{corr}(\varepsilon_i(t), \varepsilon_i(t-u)) = \exp(-\varphi u^2)$ 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 5. An approximate F-test shows that, on average, there is no evidence for a treatment effect ($p = 0.2029$).

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 that have completed the study. This yields an estimated treatment effect of 0.77 mm ($p = 0.2584$) illustrating that modeling the whole longitudinal sequence also provides more efficient inferences at specific time-points.

As an illustration of GEE and GLMM, we analyze the binary outcome 'severity of infection' in the toenail study. We will first apply GEE, based on the marginal logistic regression model

$$\log\left[\frac{P(Y_i(t) = 1)}{1 - P(Y_i(t) = 1)}\right] = \begin{cases} \beta_{A0} + \beta_{A1} t, & \text{in group A} \\ \beta_{B0} + \beta_{B1} t, & \text{in group B.} \end{cases} \tag{3.6}$$

Furthermore, we use an unstructured 7×7 working correlation matrix. The results are reported in Table 4, and the fitted average profiles are shown in the top graph of Figure 6. Based on a Wald-type test we obtain a significant difference in the average slope between the two treatment groups ($p = 0.0158$).

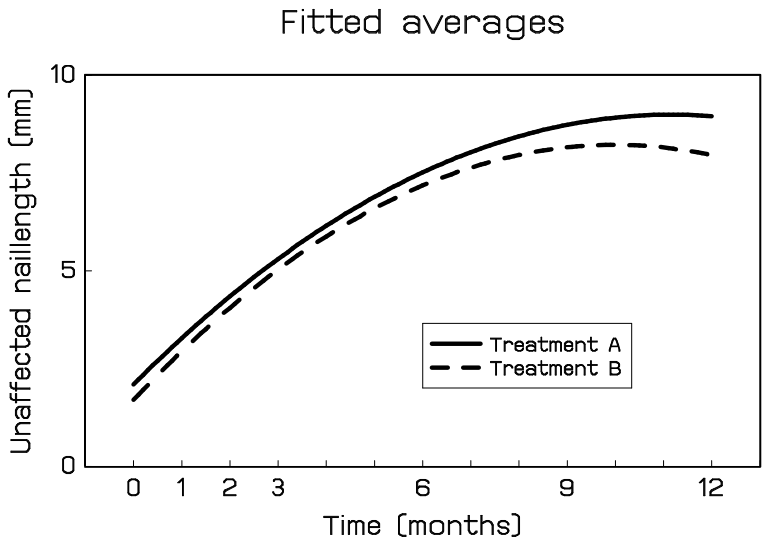

Figure 5: *Toenail data: Fitted average profiles based on model (3.5).*

Table 4: *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, modeling the association through the inclusion of subject-specific, i.e., random, intercepts. More specifically, we will now assume that

$$\log\left[\frac{P(Y_i(t)=1|b_i)}{1-P(Y_i(t)=1|b_i)}\right] = \begin{cases} \beta_{A0}+b_i+\beta_{A1}t, & \text{in group A} \\ \beta_{B0}+b_i+\beta_{B1}t, & \text{in group B} \end{cases} \tag{3.7}$$

with b_i normally distributed with mean 0 and variance σ^2. The results, obtained using numerical integration methods, are also reported in Table 4. As before, we obtain a significant difference between β_{A1} and β_{B1} ($p = 0.0255$).

4 Multivariate Longitudinal Data

So far, we have considered a single, repeatedly measured outcome. However, often one observes more than one outcome at the same time, which is essentially known as multivariate outcomes. These can all be of the same data type, e.g., all Gaussian or all binary, or of a mixed type, e.g., when the outcome vector is made up of continuous and binary components. Statistical problems where various outcomes of a mixed nature are observed have been around for about a half century

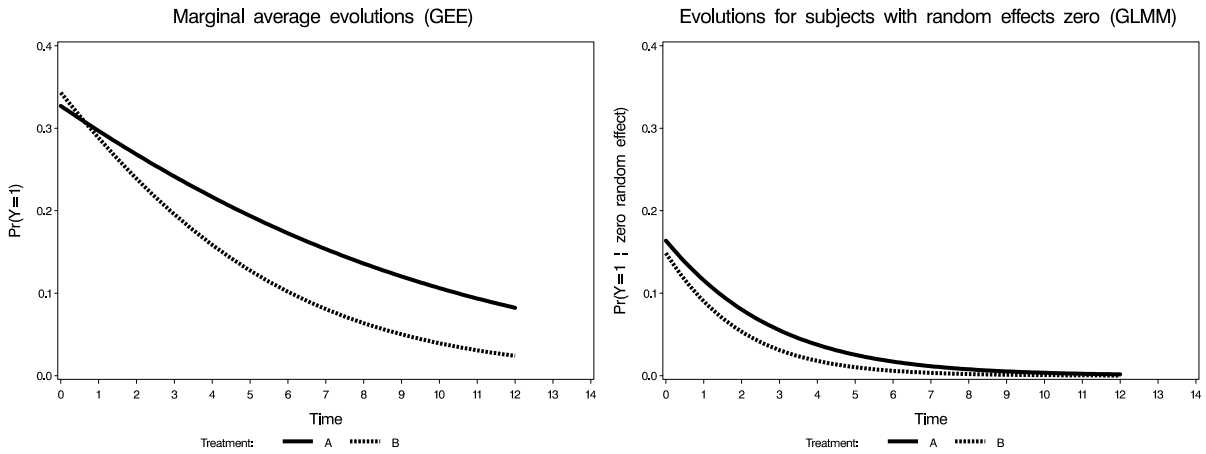

Figure 6: *Toenail Data. Treatment-specific evolutions. (a) Marginal evolutions as obtained from the marginal model (3.6) fitted using GEE, (b) Evolutions for subjects with random effects in model (3.7) equal to zero.*

and are rather common at present. Many research questions can often only fully be addressed in a joint analysis of all outcomes simultaneously. For example, the association structure can be of direct scientific relevance.

It is definitely possible for all of these features to occur simultaneously, whereby a multivariate outcome vector, possible of a mixed nature, is measured repeatedly over time. An array of research questions can then be addressed in this way. A possible question might be how the association between outcomes evolves over time or how outcome-specific evolutions are related to each other (Fieuws and Verbeke, 2004). Another example is discriminant analysis based on multiple, longitudinally measured, outcomes. Third, interest may be in the comparison of average trends for different outcomes. As an example, consider testing the difference in evolution between many outcomes or joint testing of a treatment effect on a set of outcomes. All of these situations require a joint model for all outcomes.

Let us focus, for a moment, on the combined analysis of a continuous and a discrete outcome. There then broadly are three approaches. The first one postulates a marginal model for the binary outcome and then formulates a conditional model for the continuous outcome, given the categorical one. For the former, one can use logistic regression, whereas for the latter conditional normal models are a straightforward choice, i.e., a normal model with the categorical outcome used as a covariate (Tate 1954). The second family starts from the reverse factorization, combining a marginal model for the continuous outcome with a conditional one for the categorical outcome. Conditional models have been discussed by Cox and Wermuth (1992, 1994a, 1994b), Krzanowski (1988), and Little and Schluchter (1985). Schafer (1997) presents a so-called *general location model* where a number of continuous and binary outcomes can be modeled together. The third model family directly formulates a joint model for the two outcomes. In this context, one often starts from a bivariate continuous variable, one component of which is explicitly observed and the other one observed in

dichotomized, or generally discretized, version only (Tate 1955). Molenberghs, Geys, and Buyse (2001) presented a model based on a Plackett-Dale approach, where a bivariate Plackett distribution is assumed, of which one margin is directly observed and the other one only after dichotomization. General multivariate exponential family based models have been proposed by Prentice and Zhao (1991), Zhao, Prentice, and Self (1992), and Sammel, Ryan, and Legler (1997).

Of course, these developments have not been limited to bivariate joint outcomes. One can obviously extend these ideas and families to a multivariate continuous outcome and/or a multivariate categorical outcome. For the first and second families, one then starts from conditional and marginal multivariate normal and appropriately chosen multinomial models. Such a model within the first family has been formulated by Olkin and Tate (1961). Within the third family, models were formulated by Hannan and Tate (1965) and Cox (1974) for a multivariate normal with a univariate bivariate or discrete variable.

As alluded to before, apart from an extension from the bivariate to the multivariate case, one can introduce other hierarchies as well. We will now assume that each of the outcomes may be measured repeatedly over time, and there could even be several repeated outcomes in both the continuous and the categorical subgroup. A very specific hierarchy stems from clustered data, where a continuous and a categorical, or several of each, are observed for each member of a family, a household, a cluster, etc. For the specific context of developmental toxicity studies, often conducted in rats and mice, a number of developments have been made. An overview of such methods, together with developments for probit-normal and Plackett-Dale based models, was presented in Regan and Catalano (2002). Catalano and Ryan (1992) and Fitzmaurice and Laird (1995) propose models for a combined continuous and discrete outcome, but differ in the choice of which outcome to condition on the other one. Both use generalized estimating equations to allow for clustering. Catalano (1997) extended the model by Catalano and Ryan (1992) to accommodate ordinal variables. An overview can be found in Aerts *et al* (2002).

Regan and Catalano (1999a) proposed a probit-type model to accommodate joint continuous and binary outcomes in a clustered data context, thus extending the correlated probit model for binary outcomes (Ochi and Prentice 1984) to incorporate continuous outcomes. Molenberghs, Geys, and Buyse (2001) used a Plackett latent variable to the same effect, extending the bivariate version proposed by Molenberghs, Geys, and Buyse (2001). Estimation in such hierarchical joint models can be challenging. Regan and Catalano (1999a) proposed maximum likelihood, but considered GEE as an option too (Regan and Catalano 1999b). Geys, Molenberghs, and Ryan (1999) made use of pseudo-likelihood. Ordinal extensions have been proposed in Regan and Catalano (2000).

Thus, many applications of this type of joint models can already be found in the statistical literature. For example, the approach has been used in a non-longitudinal setting to validate surrogate endpoints in meta-analyses (Buyse *et al.* 2000, Burzykowski *et al.* 2001) or to model multivariate clustered data (Thum 1997). Gueorguieva (2001) used the approach for the joint modeling of a continuous and a binary outcome measure in a developmental toxicity study on mice. Also in a longitudinal setting, Chakraborty *et al.* (2003) obtained estimates of the correlation between blood and semen HIV-1 RNA by using a joint random-effects model. Other examples with longitudinal

studies can be found in MacCallum *et al.* (1997), Thiébaut *et al.* (2002) and Shah, Laird, and Schoenfeld (1997). All of these examples refer to situations where the number of different outcomes is relatively low. Although the model formulation can be done irrespective of the number of outcomes to be modeled jointly, standard fitting procedures, such as maximum likelihood estimation, is only feasible when the dimension is sufficiently low or if one is willing to make a priori strong assumptions about the association between the various outcomes. An example of the latter can be found in situations where the corresponding random effects of the various outcomes are assumed to be perfectly correlated (Oort 2001, Sivo 2001, Roy and Lin 2000, and Liu and Hedeker 2006). Fieuws and Verbeke (2006) have developed a model-fitting procedure that is applicable, irrespective of the dimensionality of the problem. This is the route that will be followed in the next sections.

4.1 A Mixed Model for Multivariate Longitudinal Outcomes

A flexible joint model that can handle any number of outcomes measured longitudinally, without any restriction to the nature of the outcomes can be obtained by modeling each outcome separately using a mixed model (linear, generalized linear, or non-linear), by assuming that, conditionally on these random effects, the different outcomes are independent, and by imposing a joint multivariate distribution on the vector of all random effects. This approach has many advantages and is applicable in a wide variety of situations. First, the data can be highly unbalanced. For example, it is not necessary that all outcomes are measured at the same time points. Moreover, the approach is applicable for combining linear mixed models, non-linear mixed models, or generalized linear mixed models. The procedure also allows the combination of different types of mixed models, such as a generalized linear mixed model for a discrete outcome and a non-linear mixed model for a continuous outcome.

Let m be the dimension of the problem, i.e., the number of outcomes that need to be modeled jointly. Further, let Y_{rij} denote the jth measurement taken on the ith subject, for the rth outcome, $i = 1, \ldots, N$, $r = 1, \ldots, m$, and $j = 1, \ldots, n_{ri}$. Note that we do not assume that the same number of measurements is available for all subjects, nor for all outcomes. Let Y_{ri} be the vector of n_{ri} measurements taken on subject i, for outcome r. Our model assumes that each Y_{ri} satisfies a mixed model. Let $f_{ri}(y_{ri}|b_{ri}, \theta_r)$ be the density of Y_{ri}, conditional on a q_r-dimensional vector b_{ri} of random effects for the rth outcome on subject i. The vector θ_r contains all fixed effects and possibly also a scale parameter needed in the model for the rth outcome. Note that we do not assume the same type of model for all outcomes: A combination of linear, generalized linear, and non-linear mixed models is possible. It is also not assumed that the same number q_r of random effects is used for all m outcomes. Finally, the model is completed by assuming that the vector b_i of all random effects for subject i is multivariate normal with mean zero and covariance D, i.e.,

$$b_i = \begin{pmatrix} b_{1i} \\ b_{2i} \\ \vdots \\ b_{mi} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} & \cdots & D_{1m} \\ D_{21} & D_{22} & \cdots & D_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ D_{m1} & D_{m2} & \cdots & D_{mm} \end{pmatrix} \right].$$

The matrices D_{rs} represent the covariances between b_{ri} and b_{si}, $r, s = 1, \ldots, m$. Finally, D is the

matrix with blocks D_{rs} as entries.

A special case of the above model is the so-called shared-parameter model, which assumes the same set of random effects for all outcomes. This clearly can be obtained as a special case of the above model by assuming perfect correlation between some of the random effects. The advantage of such shared-parameter models is the relatively low dimension of the random-effects distribution, when compared to the above model. The dimension of the random effects in shared parameter models does not increase with the number of outcomes to be modeled. In the above model, each new outcome added to the model introduces new random effects, thereby increasing the dimension of b_i. Although the shared-parameter models can reasonably easy be fitted using standard software, this is no longer the case for the model considered here. Estimation and inference under the above model will require specific procedures, which will be discussed in Section 4.2. A disadvantage of the shared-parameter model is that it is based on much stronger assumptions about the association between the outcomes, which may not be valid, especially in high-dimensional settings as considered in this chapter. Note also that, joining valid univariate mixed models does not necessarily lead to a correct joint model. Fieuws and Verbeke (2004) illustrate this in the context of linear mixed models for two continuous outcomes. It is shown how the joint model may imply association structures between the two sets of longitudinal profiles that may strongly depend on the actual parameterization of the individual models and that are not necessarily valid.

4.2 A Pairwise Model-fitting Approach

Whereas the modeling approach from the previous setting is rather versatile, it might become computationally cumbersome for high-dimensional applications. It is therefore useful to consider the approach of Fieuws and Verbeke (2006), when a large number of repeated sequences are to be analyzed simultaneously. The general idea is that all parameters in the full multivariate model can be identified from all pairwise models, i.e., all bivariate models for each pair of outcomes. Therefore, using pseudo-likelihood ideas, also termed pairwise or composite likelihood (Molenberghs and Verbeke 2005), fitting the full model is replaced by maximum likelihood estimation of each bivariate model separately. This can be done using standard statistical software. Afterwards, all results are appropriately combined, and Wald-type inferences become available from noticing that the pairwise fitting approach is equivalent to maximizing the sum of all the log-likelihoods from all fitted pairs. This sum can be interpreted as a pseudo-log-likelihood function, and inferences then immediately follow from the general pseudo-likelihood theory, as will now be explained in the following sections.

4.2.1 Pairwise Fitting

Let Ψ^* be the vector of all parameters in the multivariate joint mixed model for $(Y_1, Y_2, \ldots, Y_m)$. The pairwise fitting approach starts from fitting all $m(m-1)/2$ bivariate models, i.e., all joint models for all possible pairs

$$(Y_1, Y_2), (Y_1, Y_3), \ldots, (Y_1, Y_m), (Y_2, Y_3), \ldots, (Y_2, Y_m), \ldots, (Y_{m-1}, Y_m)$$

of the outcomes $Y_1, Y_2, \ldots, Y_m$. Let the log-likelihood function corresponding to the pair (r, s) be denoted by $\ell(y_r, y_s|\Psi_{rs})$, and let Ψ_{rs} be the vector containing all parameters in the bivariate model for pair (r, s).

Let Ψ now be the stacked vector combining all $m(m-1)/2$ pair-specific parameter vectors Ψ_{rs}. Estimates for the elements in Ψ are obtained by maximizing each of the $m(m-1)/2$ log-likelihoods $\ell(y_r, y_s|\Psi_{rs})$ separately. It is important to realize that the parameter vectors Ψ and Ψ^* are not equivalent. Indeed, some parameters in Ψ^* will have a single counterpart in Ψ, e.g., the covariances between random effects of different outcomes. Other elements in Ψ^* will have multiple counterparts in Ψ, e.g., fixed effects from one single outcome. In the latter case, a single estimate for the corresponding parameter in Ψ^* is obtained by averaging all corresponding pair-specific estimates in $\widehat{\Psi}$. Standard errors of the so-obtained estimates clearly cannot be obtained from averaging standard errors or variances. Indeed, two pair-specific estimates corresponding to two pairwise models with a common outcome are based on overlapping information and hence correlated. This correlation should also be accounted for in the sampling variability of the combined estimates in $\widehat{\Psi}^*$. Correct asymptotic standard errors for the parameters in $\widehat{\Psi}$, and consequently in $\widehat{\Psi}^*$, can be obtained from pseudo-likelihood ideas.

4.2.2 Inference for Ψ

Fitting all bivariate models is equivalent to maximizing the function

$$p\ell(\Psi) \equiv p\ell(y_{1i}, y_{2i}, \ldots, y_{mi}|\Psi) = \sum_{r<s} \ell(y_r, y_s|\Psi_{rs}), \tag{4.8}$$

ignoring the fact that some of the vectors Ψ_{rs} have common elements, i.e., assuming that all vectors Ψ_{rs} are completely distinct. The function in (4.8) can be considered a pseudo-likelihood function, maximization of which leads to so-called pseudo-likelihood estimates, with well-known asymptotic statistical properties. We refer to Arnold and Strauss (1991) and Geys, Molenberghs, and Ryan (1997) for more details. Our application of pseudo-likelihood methodology is different from most other applications in the sense that the same parameter vector is usually present in the different parts of the pseudo-likelihood function. Here, the set of parameters in Ψ_{rs} is treated pair-specific, which allows separate maximization of each term in the pseudo log-likelihood function (4.8). In Section 4.2.3, we will account for the fact that Ψ_{rs} and $\Psi_{rs'}$, $s \neq s'$, are not completely distinct, as they share the parameters referring to the rth outcome.

It now follows directly from the general pseudo-likelihood theory that $\widehat{\Psi}$ asymptotically satisfies

$$\sqrt{N}(\widehat{\Psi} - \Psi) \quad \approx \quad N(0, I_0^{-1} I_1 I_0^{-1})$$

in which $I_0^{-1} I_1 I_0^{-1}$ is a 'sandwich-type' robust variance estimator, and where I_0 and I_1 can be constructed using first- and second-order derivatives of the components in (4.8). Strictly speaking, I_0 and I_1 depend on the unknown parameters in Ψ, but these are traditionally replaced by their estimates in $\widehat{\Psi}$.

4.2.3 Combining Information: Inference for Ψ^*

In a final step, estimates for the parameters in Ψ^* can be calculated, as suggested before, by taking averages of all the available estimates for that specific parameter. Obviously, this implies that $\widehat{\Psi}^* = A^\top \widehat{\Psi}$ for an appropriate weight matrix A. Hence, inference for the elements in $\widehat{\Psi}^*$ will be based on

$$\sqrt{N}(\widehat{\Psi}^* - \Psi^*) = \sqrt{N}(A^\top \widehat{\Psi} - A^\top \Psi) \approx N(0, A^\top I_0^{-1} I_1 I_0^{-1} A).$$

It can be shown that pseudo-likelihood estimates are less efficient than the full maximum likelihood estimates (Arnold and Strauss 1991). However, these results refer to efficiency for the elements in Ψ, not directly to the elements in Ψ^*. In general, the degree of loss of efficiency depends on the context, but Fieuws and Verbeke (2006) have presented evidence for only very small losses in efficiency in the present context of the pairwise fitting approach for multivariate random-effects models.

4.3 Analysis of the Hearing Data

Let $Y_{r,i}(t)$ denote the rth hearing threshold for subject i taken at time t, $r = 1, \ldots, 11$ for the right ear, and $r = 12, \ldots, 22$ for the left ear. Morrell and Brant (1991), and Pearson *et al.* (1995) have proposed the following linear mixed model to analyze the evolution of the hearing threshold for a single frequency:

$$Y_{r,i}(t) = (\beta_{r,1} + \beta_{r,2}\text{Age}_i + \beta_{r,3}\text{Age}_i^2 + a_{r,i}) + (\beta_{r,4} + \beta_{r,5}\text{Age}_i + b_{r,i})t + \beta_{r,6}V_i(t) + \varepsilon_{r,i}(t). \quad (4.9)$$

The time t is expressed in years from entry in the study and Age_i equals the age of subject i at the time of entry in the study. The binary time-varying covariate V_i represents a learning effect from the first to the subsequent visits. Finally, the $a_{r,i}$ are random intercepts, the $b_{r,i}$ are the random slopes for time, and the $\varepsilon_{r,i}$ represent the usual error components. The regression coefficients $\beta_{r,1}, \ldots, \beta_{r,6}$ are fixed, unknown parameters. The 44 random effects $a_{1,i}, a_{2,i}, \ldots, a_{22,i}, b_{1,i}, b_{2,i}, \ldots, b_{22,i}$ are assumed to follow a joint zero-mean normal distribution with covariance matrix D. At each time point t, the error components $\varepsilon_{1,i}, \ldots, \varepsilon_{22,i}$ follow a 22-dimensional zero-mean normal distribution with covariance matrix R. The total number of parameters in D and R equals $990 + 253 = 1243$.

We applied the pairwise approach to fit model (4.9) to the Hearing data introduced in Section 2.2. As discussed before, one of the key research questions is whether the deterioration of hearing ability with age is different for different frequencies, because this would yield evidence for selective deterioration. Formally, this requires testing the null-hypotheses $H_0 : \beta_{1,5} = \beta_{2,5} = \ldots = \beta_{11,5}$ for the right side, and $H_0 : \beta_{12,5} = \beta_{13,5} = \ldots = \beta_{22,5}$ for the left side. Figure 7 shows all estimates $\widehat{\beta}_{r,5}$ with associated 95% confidence intervals, for the left and right ear separately. We clearly observe an increasing trend implying that age accelerates hearing loss, but that this is more severe for higher frequencies. Wald-type tests indicate that these estimates are significantly different between the outcomes, at the left side ($\chi^2_{10} = 90.4$, $p < 0.0001$) as well as at the right side ($\chi^2_{10} = 110.9$, $p < 0.0001$).

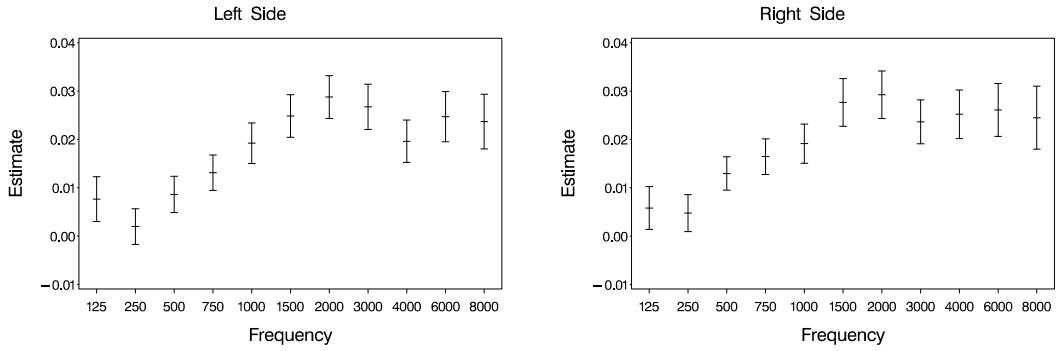

Figure 7: *Hearing data. Estimates* $\widehat{\beta}_{r,5}$ *with associated 95% confidence intervals, for the measurements from left and right ear separately.*

4.4 Some Reflections

The advantage of this technique is that all implied univariate models belong to the well-known mixed model family. This implies that one can first model each outcome separately (with separate data exploration and model building), before joining the univariate models into the full multivariate model. Moreover, the parameters in the multivariate model keep their interpretation from the separate univariate models. Finally, this approach is sufficiently flexible to allow for different types of models for the different outcomes (linear, non-linear, generalized linear).

A disadvantage of the approach is that, when the number of outcomes becomes large, the dimension of the random effects can become too large to fit the full multivariate model using standard software for mixed models. Using results of Fieuws and Verbeke (2006), and Fieuws *et al.* (2006), we have explained how all parameters in the multivariate model can be estimated from fitting the model to all possible pairs of outcomes. Inferences follow from pseudo-likelihood theory. Although the estimates obtained from the pairwise approach do not maximize the full multivariate likelihood, they still have similar asymptotic properties, with no or only marginal loss of efficiency when compared to the maximum likelihood estimates. It should be emphasized that we do not advocate fitting multivariate models in order to gain efficiency for parameters in single univariate models. As long as no inferences are needed for combinations of parameters from different outcomes, and if no outcomes share the same parameters, univariate mixed models are by far the preferred tools for the analysis.

Fitting of the models can usually be done using standard software for the linear, non-linear, and generalized linear mixed models. However, calculation of the standard errors requires careful data manipulation. An example using the SAS software can be obtained from the authors, and will soon be available from the authors' website. In case all univariate mixed models are of the linear type (e.g., our model for the Hearing data), a SAS macro can be used.

5 Joint Models for Longitudinal and Time-to-Event Data

As we have seen earlier in this chapter, it is very common in longitudinal studies to collect measurements on several types of outcomes. In this section we focus on settings in which the outcomes recorded on the subjects simultaneously include a set of repeated measurements and the time at which an event of particular interest occurs, for instance, death, development of a disease or dropout from the study. Typical areas where such studies are encountered encompass HIV/AIDS and cancer studies. In HIV studies, seropositive patients are monitored until they develop AIDS or die, and they are regularly measured for the condition of their immune system using markers such as the CD4 lymphocyte count, the estimated viral load, or whether viral load is below detectable limits. Similarly, in cancer trials the event outcome is death or metastasis, while patients also provide longitudinal measurements of antibody levels or of other markers of carcinogenesis, such as the prostate specific antigen levels for prostate cancer.

Depending on the research questions, these two outcomes can be analyzed either separately or jointly. Here, we will focus on situations in which a joint analysis is required. This is typically the case when interest is on the event time and one wishes to account for the effect of the longitudinal outcome as a time-dependent covariate. Traditional approaches for analyzing time-to-event data, such as the partial likelihood for the Cox proportional hazards models, assume that the time-dependent covariate is a predictable process; that is, the value of this covariate at time point t is not affected by the occurrence of an event at time point u, with $t > u$ (Therneau and Grambsch, 2000, Sect. 1.3). For instance, age can be included as predictable time-dependent covariate in a standard analysis, because if we know the age of a subject at baseline, we can 'predict' her age at every time point without error. However, the type of time-dependent covariates encountered in longitudinal studies are often not predictable. In particular, they are the output of a stochastic process generated at the level of the subject, and it is directly related to the failure mechanism. The stochastic nature of these covariates complicates matters in two ways. First, we do not actually observe the 'true' values for these covariates, owing to the fact that the longitudinal responses usually contain measurement error. Second, we are only able to observe the, error-contaminated, values intermittently at the specific time points at which we have collected measurements and not at any time point t. These special features complicate analysis with the traditional partial likelihood approaches (Tsiatis, DeGruttola, and Wolfsohn 1995, Wulfsohn and Tsiatis 1997). Hence, to produce valid inferences, a model for the joint distribution of the longitudinal and survival outcomes is required instead.

Early attempts to tackle such problems considered using the last available value of the longitudinal outcome for each subject as a representative value for the complete longitudinal history. This method is also known as 'Last Value or Last Observation Carried Forward' (LVCF or LOCF, Molenberghs and Kenward 2007). Even though the simplicity of such an approach is apparent, Prentice (1982) showed that it leads to severe bias in the estimation of the model parameters. Later approaches (Self and Pawitan, 1992; Tsiatis, DeGruttola, and Wulfsohn 1995) focused on joint models with a survival sub-model for the time-to-event and a longitudinal sub-model for the longitudinal process, in which so-called two-stage procedures have been proposed to derive

estimates of the model parameters. In particular, at a first stage, the longitudinal model is estimated ignoring the survival outcome, and at the second stage a survival model is fitted using the subject-specific predictions of time-dependent covariates based on the longitudinal model. Such approaches were shown to reduce bias compared to the naive LVCF without completely eliminating it. This persistent bias prompted a turn of focus to full maximum likelihood methods. A fully parametric approach was proposed by DeGruttola and Tu (1994) who postulated a log-normal sub-model for the time-to-event and a linear mixed model for the longitudinal responses, respectively. Later, Wulfsohn and Tsiatis (1997) extended this work by assuming a relative risk model for the survival times with an unspecified baseline risk function. Excellent overviews of the joint modeling literature are given by Tsiatis and Davidian (2004) and Yu *et al.* (2004). In the rest of this section we will present the basics of the joint modeling framework and provide a perspective on its features.

5.1 Joint Modeling Framework

To introduce joint models for longitudinal and time-to-event data, we need to adapt and extend the notation introduced so far in this chapter. In particular, for the time-to-event outcome we denote by T_i the observed failure time for the ith subject ($i = 1, \ldots, n$), which is taken as the minimum of the true event time T_i^* and the censoring time C_i, i.e., $T_i = \min(T_i^*, C_i)$. Furthermore, we define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function that takes the value 1 if the condition $T_i^* \leq C_i$ is satisfied, and 0 otherwise. Thus, the observed data for the time-to-event outcome consist of the pairs $\{(T_i, \delta_i), i = 1, \ldots, n\}$. For the longitudinal responses, we let $y_i(t)$ to denote the value of the longitudinal outcome at time point t for the ith subject. However, we do not actually observe $y_i(t)$ at all time points but only at very specific occasions t_{ij} at which measurements were taken. Thus, the observed longitudinal data consist of the measurements $y_{ij} = \{y_i(t_{ij}), j = 1, \ldots, n_i\}$. As noted above, this feature of the longitudinal outcome is one of the main reasons why it cannot be simply included as a standard time-dependent covariate in a survival model.

In survival analysis, relative risk models have traditionally been used to quantify effects of both time-independent and time-dependent covariates on the risk of an event (Therneau and Grambsch, 2000). In our setting, we introduce the term $m_i(t)$ that denotes the *true* and *unobserved* value of the longitudinal outcome at time t, which is included as a time-dependent covariate in a relative risk model:

$$\begin{aligned} h_i(t \mid \mathcal{M}_i(t), w_i) &= \lim_{dt \to 0} P\{t \leq T_i^* < t + dt \mid T_i^* \geq t, \mathcal{M}_i(t), w_i\}/dt \\ &= h_0(t) \exp\{\gamma^\top w_i + \alpha m_i(t)\}, \end{aligned} \tag{5.10}$$

where $\mathcal{M}_i(t) = \{m_i(u), 0 \leq u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t, $h_0(\cdot)$ denotes the baseline risk function, and w_i a vector of baseline covariates, such as a treatment indicator, history of diseases, etc., with a corresponding vector of regression coefficients γ. Similarly, parameter α quantifies the effect of the underlying longitudinal outcome to the risk for an event.For instance, in the AIDS example introduced in Section 5, α measures the effect of the number of CD4 cells to the risk for death. An important note regarding Model

(5.10) is that the risk for an event at time t is assumed to depend on the longitudinal history $\mathcal{M}_i(t)$ only through the current value of the time-dependent covariate $m_i(t)$; on the contrary, survival probabilities depend on the whole history via:

$$\mathcal{S}_i(t \mid \mathcal{M}_i(t), w_i) = P(T_i^* > t \mid \mathcal{M}_i(t), w_i) = \exp\Big(-\int_0^t h_0(s)\exp\{\gamma^\top w_i + \alpha m_i(s)\}\, ds\Big), \qquad (5.11)$$

which implies that a correct specification of $\mathcal{M}_i(t)$ is required to produce valid estimates of $\mathcal{S}_i(t \mid \mathcal{M}_i(t), w_i)$. To complete the specification of the survival model, we need to specify the baseline risk function. Within the joint modeling framework, $h_0(t)$ is typically left unspecified (Wulfsohn and Tsiatis 1997). However, Hsieh, Tseng, and Wang (2006) have recently noted that leaving this function completely unspecified leads to an underestimation of the standard errors of the parameter estimates. In particular, problems arise stemming from the fact that the non-parametric maximum likelihood estimate for this function cannot be obtained explicitly under the random-effects structure. To avoid this problem, we could either opt for a standard survival distribution on the one hand, such as the Weibull or Gamma distributions, or for more flexible models on the other, in which $h_0(t)$ is sufficiently well approximated using step functions or spline-based approaches.

So far, in the definition of the survival model we have assumed that the true underlying longitudinal covariate $m_i(t)$ is available at any time point t. Nevertheless, longitudinal information is actually collected intermittently for each subject at a few time points t_{ij}. Therefore, our aim is to estimate $m_i(t)$ and successfully reconstruct the complete longitudinal history, using the available measurements $y_{ij} = \{y_i(t_{ij}), j = 1, \ldots, n_i\}$ of each subject and a set of modeling assumptions. For the remainder of this section, we will focus on normal data and postulate a linear mixed effects model, as in Section 3.1, to describe the subject-specific longitudinal evolutions. Here we make explicit the model's time-dependent nature,

$$y_i(t) = m_i(t) + \varepsilon_i(t) = x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t), \;\; \varepsilon_i(t) \sim N(0, \sigma^2), \qquad (5.12)$$

where β denotes the vector of the unknown fixed effects parameters, $x_i(t)$ and $z_i(t)$ denote row vectors of the design matrices for the fixed and random effects, respectively, and $\varepsilon_i(t)$ is the measurement error term, which is assumed independent of b_i, and with variance σ^2. As we have seen above, the survival function is a function of the complete longitudinal history, and therefore, it is important to adequately specify $x_i(t)$ and $z_i(t)$ to capture interesting characteristics of the data and produce a good estimate of $\mathcal{M}_i(t)$. For instance, in applications in which subjects show highly non-linear longitudinal trajectories, it is advisable to consider flexible representations for $x_i(t)$ and $z_i(t)$ using a possibly high-dimensional vector of functions of time t, expressed in terms of high-order polynomials or splines (Ding and Wang 2008, Brown, Ibrahim, and DeGruttola 2005). An alternative approach is to consider correlated error terms. Joint models with such error structures have been proposed by Wang and Taylor (2001), who postulated an integrated Ornstein-Uhlenbeck process, and by Henderson, Diggle, and Dobson (2000), who considered a latent Gaussian stochastic process shared by both the longitudinal and event processes. We should note however that there is a conflict for information between the random-effects structure and a measurement error structure

that assumes correlated errors, given that both aim at modeling the marginal correlation in the data. Thus, depending on the features of the data at hand, it is advisable to either opt for an elaborate random-effects structure (using e.g., splines in the design matrix $z_i(t)$) or for correlated error terms, but not for both. For an enlightening discussion on the philosophical differences between these two approaches, we refer to Tsiatis and Davidian (2004, Sect. 2.2). Finally, a suitable distributional assumption for the random-effects component is required to complete the specification of the joint model. So far, in this chapter, we have relied on standard parametric assumptions for this distribution, with a typical choice being the multivariate normal distribution with mean zero and covariance matrix D. However, within the joint modeling framework and mainly for two reasons, there is the concern that relying on standard distributions may influence the derived inferences. First, the random effects have a more prominent role in joint models, because on the one hand they capture the correlations between the repeated measurements in the longitudinal outcome and on the other they associate the longitudinal outcome with the event process. Second, joint models belong to the general family of shared parameter models, and correspond to a non-random dropout mechanism. We wil return to this in Section 7. As is known from the missing-data literature, handling dropout can be highly sensitive to modeling assumptions. These features motivated Song, Davidian, and Tsiatis (2002) to explore the need for a more flexible model for the distribution of the random effects, especially in the joint modeling framework. However, the findings of these authors suggested that parameter estimates and standard errors were rather robust to misspecification. This feature has been further theoretically corroborated by Rizopoulos, Verbeke, and Molenberghs (2008), who showed that, as the number of repeated measurements per subject n_i increases, misspecification of the random-effects distribution has a minimal effect in parameter estimators and standard errors.

5.2 Likelihood and Estimation

The main estimation methods that have been proposed for joint models are (semi-parametric) maximum likelihood (Hsieh, Tseng, and Wang 2006, Henderson, Diggle, and Dobson 2000, Wulfsohn and Tsiatis 1997) and Bayes using MCMC techniques (Chi and Ibrahim 2006, Brown and Ibrahim 2003, Wang and Taylor 2001, Xu and Zeger 2001). Moreover, Tsiatis and Davidian (2001) have proposed a conditional score approach in which the random effects are treated as nuisance parameters, and they developed a set of unbiased estimating equations that yields consistent and asymptotically normal estimators. Here, we review the basics of the maximum likelihood method for joint models as one of the more traditional approaches.

Maximum likelihood estimation for joint models is based on the maximization of the log-likelihood corresponding to the joint distribution of the time-to-event and longitudinal outcomes $\{T_i, \delta_i, y_i\}$. To define this joint distribution, we will assume that the vector of time-independent random effects b_i underlies both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process. Formally, we have that,

$$f(T_i, \delta_i, y_i \mid b_i; \theta) \;=\; f(T_i, \delta_i \mid b_i; \theta)\, f(y_i \mid b_i; \theta), \tag{5.13}$$

$$f(y_i \mid b_i; \theta) \quad = \quad \prod_j f\{y_i(t_{ij}) \mid b_i; \theta\}, \tag{5.14}$$

where θ is the parameter vector, y_i is the $n_i \times 1$ vector of longitudinal responses of the ith subject, and $f(\cdot)$ denotes an appropriate probability density function. Under this conditional independence assumption we can now define separate models for the longitudinal responses and the event time data by conditioning on the shared random effects. Under the modeling assumptions presented in the previous section and the conditional independence assumptions (5.13) and (5.14), the joint likelihood contribution for the ith subject can be formulated as

$$f(T_i, \delta_i, y_i; \theta) = \int f(T_i, \delta_i \mid b_i; \theta) \Big[\prod_j f\{y_i(t_{ij}) \mid b_i; \theta\} \Big] f(b_i; \theta)\, db_i, \tag{5.15}$$

where the likelihood of the survival part is written as

$$f(T_i, \delta_i \mid b_i; \theta) = \{h_i(T_i \mid b_i; \theta)\}^{\delta_i} \mathcal{S}_i(T_i \mid b_i; \theta), \tag{5.16}$$

with $h_i(\cdot)$ and $\mathcal{S}_i(\cdot)$ are given by (5.10) and (5.11), respectively, $f\{y_i(t_{ij}) \mid b_i; \theta\}$ is the univariate normal density for the longitudinal responses, and $f(b_i; \theta)$ is the multivariate normal density for the random effects. A further implicit assumption in the above definition of the likelihood is that both the censoring mechanism and the visiting process (i.e., the stochastic mechanism that generates the time points at which the longitudinal measurements are collected) are non-informative, and thus they can be ignored. This non-informativeness assumption is similar in spirit to the missing at random (MAR) assumption in the missing data framework (see also Section 7), and in particular, it is assumed that the probabilities of visiting and censoring at time point t depend only on the observed longitudinal history but not on the event times and future longitudinal measurements themselves. As observed longitudinal history we define all available information for the longitudinal process prior to time point t, i.e., $\mathcal{Y}_i(t) = \{y_i(u), 0 \leq u < t\}$; note that this is different from $\mathcal{M}_i(t)$, which denotes the history of the true unobserved longitudinal outcome $m_i(t)$. In practice, this assumption is valid when the decision on whether a subject withdraws from the study or appears at the study center for the scheduled visit to provide a longitudinal measurement at time t, depends only on $\mathcal{Y}_i(t)$ (and possibly on baseline covariates), but there is no additional dependence on future longitudinal responses and the underlying random effects b_i. Unfortunately, the observed data do not often contain enough information to corroborate these assumptions, and therefore, it is essential to use external information from subject-matter experts as to their validity.

Maximization of the log-likelihood function corresponding to (5.15) with respect to θ is a computationally challenging task, because it requires a combination of numerical integration and optimization algorithms. Numerical integration is required, owing to the fact that neither the integral with respect to the random effects in (5.15), nor the integral of the risk function in (5.11) allow for an analytical solution, except in very special cases. Standard numerical integration techniques, such as Gaussian quadrature and Monte Carlo have been successfully applied in the joint modelling framework (Song, Davidian, and Tsiatis 2002, Henderson, Diggle, and Dobson 2000, Wulfsohn and Tsiatis 1997). Furthermore, Rizopoulos, Verbeke, and Lesaffre (2009b) have recently discussed the use of Laplace approximations for joint models, that can be especially useful in high-dimensional

Table 5: *Liver cirrhosis data. Parameter estimates with standard errors in parenthesis. For the longitudinal process 'a:b' denotes the interaction term between covariates 'a' and 'b'. For the random effects σ_{b1} denotes the standard deviation of the random intercepts term, σ_{b2} the standard deviation of the random slopes term, ρ_{b12} the correlation between the random intercepts and random slopes, and σ the measurement error standard deviation.*

	Survival Process		Longitudinal Process		Variance Comp.	
Model	Parameter	Estimate (s.e.)	Effect	Est. (s.e.)	Param.	Est.
Naive	prednisone	0.054 (0.130)				
Cox	prothrombin	−0.032 (0.003)				
Joint	prednisone	−0.214 (0.140)	intercept	70.49 (1.36)	σ_{b1}	18.51
Model	prothrombin	−0.040 (0.004)	prednisone	11.10 (1.96)	σ_{b2}	4.22
			baseline	−1.49 (1.35)	ρ_{b12}	0.04
			baseline:prednisone	−11.20 (1.89)	σ	16.86
			time	0.40 (0.39)		
			time:prednisone	−1.05 (0.68)		

random-effects settings (e.g., when splines are used in random-effects design matrix). For the maximization of the approximated log-likelihood the EM algorithm has been traditionally used in which the random effects are treated as 'missing data'. The main motivation for using this algorithm is the closed-form M-step updates for certain parameters of the joint model. However, a serious drawback of the EM algorithm is its linear convergence rate that results in slow convergence especially near the maximum. Nonetheless, Rizopoulos, Verbeke, and Lesaffre (2009b) have noted that a direct maximization of the observed data log-likelihood, using for instance, a quasi-Newton algorithm (Lange 2004), requires very similar computations to the EM algorithm. Therefore hybrid optimization approaches that start with EM and then continue with direct maximization can be easily employed.

5.3 Analysis of Liver Cirrhosis Data

To illustrate the virtues of the joint modeling approach, we will start with a 'naive' analysis, in which we ignore the special characteristics of the prothrombin index and we fit a Cox model that includes treatment indicator and prothrombin as an ordinary time-dependent covariate. The results are presented in Table 5. We observe that, after adjusting for prothrombin in the Cox model, there is no statistical evidence for a treatment effect. We proceed by specifying and fitting a joint model that explicitly postulates a linear mixed effects model for the prothrombin index. In particular, in the longitudinal sub-model, we include fixed effects of time, treatment, and an indicator for the baseline measurement at $t = 0$, as well as the interactions of treatment with time and treatment with the baseline indicator. In the random-effects design matrix, we include an intercept and a time term. For the survival sub-model and similarly to the Cox model above we include the treatment effect and as time-dependent covariate the true underlying effect of prothrombin as estimated from

the longitudinal model. The baseline risk function is assumed piecewise constant

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q),$$

where $0 = v_0 < v_1 < \cdots < v_Q$ denotes a split of the time scale, with v_Q being larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$. For the internal knots $v_1, \ldots, v_{Q-1}$ we use equally spaced percentiles of the observed survival times T_i. The parameter estimates and standard errors from the joint model fit are also shown in Table 5. For the treatment effect, we arrive at a similar conclusion as with the standard analysis, that is, there is no clear evidence that prednisone decreases the risk for an event. However, a comparison between the standard time-dependent Cox model with the joint model reveals some interesting features. In particular, we observe that the estimated treatment effect from the joint model is much bigger in size and on the opposite direction compared to the time-dependent Cox model, with a standard error of the same magnitude in both models. Similarly, the effect of the prothrombin index from the joint model is about 2.5 standard errors larger compared to the same effect from the Cox model. These comparisons convincingly demonstrate the degree of attenuation in the regression coefficients of the standard analysis due to the measurement error in the prothrombin levels.

5.4 Some Reflections

Joint modeling of longitudinal and time-to-event data is one of the most rapidly evolving areas of current biostatistics research, with several extensions of the standard joint model that we have presented here already proposed in the literature. These include, among others, handling multiple failure types (Elashoff and Li 2008), considering categorical longitudinal outcomes (Faucett, Schenker, and Elashoff 1998), assuming that several longitudinal outcomes affect the time-to-event (Chi and Ibrahim 2006, Brown and Ibrahim 2003), replacing the relative risk model by an accelerated failure time model (Tseng, Hsieh, and Wang 2005), and associating the two outcomes via latent classes instead of random effects (Proust-Lima *et al.* 2009, Lin *et al.* 2002). Even though there has been considerable work on such extensions, little attention has been given to the development of diagnostic and model-assessment tools for these models. The main problem of using standard diagnostic tools, such as residuals, is the nonrandom dropout caused by the occurrence events. To this end, Dobson and Henderson (2003) defined residuals conditional on the dropout times and recommended plotting these residuals per dropout pattern. Another, more recent proposal by Rizopoulos, Verbeke, and Molenberghs (2009a) takes dropout into account by multiply imputing the longitudinal responses that would have been observed had the event not occurred, and use afterwards standard residuals plots.

Finally, one of the main practical limitations for joint modeling finding its way into the tool box of modern statisticians was the lack of free and reliable software. The R package **JM**[1] has been developed to fill this gap to some extent. **JM** has a user-friendly interface to fit joint models and also provides several supporting functions that extract or calculate various quantities based on the

[1] **JM** can be freely down loaded from the CRAN website at `http://cran.r-project.org`.

fitted model (e.g., residuals, fitted values, empirical Bayes estimates, various plots, and others). At `http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:jm`, more information can be found.

6 The Use in Surrogate Markers

Over the years, longitudinal data models, survival analysis tools, and the combination thereof, have been used in the so-called validation of surrogate endpoints in clinical studies. Reviews can be found in Burzykowski, Molenberghs, and Buyse (2005), Molenberghs *et al.* (2008, 2009). We provide a bird's eye perspective on these developments and their extensions towards information theory.

The field is interesting in its own right, because the use of surrogate endpoints in the development of new therapies has always been very controversial, partly owing to a number of unfortunate historical instances where treatments showing a highly positive effect on a surrogate endpoints were ultimately shown to be detrimental to the subjects' clinical outcome, and conversely, some instances of treatments conferring clinical benefit without measurable impact on presumed surrogates (Fleming and DeMets 1996). For example, in cardiovascular disease, the unsettling discovery that the two major anti arrhythmic drugs encanaide and flecanaide reduced arrhythmia but caused a more than 3-fold increase in overall mortality stressed the need for caution in using non-validated surrogate markers in the evaluation of the possible clinical benefits of new drugs (CAST 1989). On the other hand, the dramatic surge of the AIDS epidemic, the impressive therapeutic results obtained early on with zidovudine, and the pressure for an accelerated evaluation of new therapies, have all led to the use of CD4 blood count and later of viral load as endpoints that replaced time to clinical events and overall survival (DeGruttola and Tu 1994), in spite of serious concerns about their limitations as surrogate markers for clinically relevant endpoints (Lagakos and Hoth 1992). Loosely speaking, a *surrogate endpoint* is a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit, harm, or lack thereof.

One important reason for the present interest in surrogate endpoints is the advent of a large number of biomarkers that closely reflect the disease process. An increasing number of new drugs have a well-defined mechanism of action at the molecular level, allowing drug developers to measure the effect of these drugs on the relevant biomarkers (Ferentz 2002). There is increasing public pressure for new, promising drugs to be approved for marketing as rapidly as possible, and such approval will have to be based on biomarkers rather than on some long-term clinical endpoint (Lesko and Atkinson 2001). If the approval process is shortened, there will be a corresponding need for earlier detection of safety signals that could point to toxic problems with new drugs. It is a safe bet, therefore, that the evaluation of tomorrow's drugs will be based primarily on biomarkers, rather than on the longer-term, harder clinical endpoints that have dominated the development of new drugs until now. It is therefore imperative to use *validated* surrogates, though one needs to reflect on the precise meaning and extent of validation (Schatzkin and Gail 2002).

6.1 A Meta-analytic Framework for Normally Distributed Outcomes

Several methods have been suggested for the formal evaluation of surrogate markers, some based on a single trial with others, currently gaining momentum, of a meta-analytic nature. The first formal single trial approach to validate markers is due to Prentice (1989), who gave a definition of the concept of a surrogate endpoint, followed by a series of operational criteria. Freedman, Graubard, and Schatzkin (1992) augmented Prentice's hypothesis-testing based approach, with the estimation paradigm, through the so-called *proportion of treatment effect explained*. In turn, Buyse and Molenberghs (1998) added two further measures: the *relative effect* and the *adjusted association*. All of these proposals are hampered by the fact that they are single-trial based, in which there evidently is replication at the patient level, but not at the level of the trial.

Although the single trial based methods are relatively easy in terms of implementation, they are surrounded with the difficulties stated before. Therefore, several authors, such as Daniels and Hughes (1997), Buyse *et al.* (2000), and Gail *et al.* (2000) have introduced the meta-analytic approach. This section briefly outlines the methodology.

The meta-analytic approach was formulated originally for two continuous, normally distributed outcomes, and extended in the meantime to a large collection of outcome types, ranging from continuous, binary, ordinal, time-to-event, and longitudinally measured outcomes (Burzykowski, Molenberghs, and Buyse 2005). First, we focus on the continuous case, where the surrogate and true endpoints are jointly normally distributed.

The method is based on the linear mixed model of Section 3.1. Both a fixed-effects and a random-effects view can be taken. Let T_{ij} and S_{ij} be the random variables denoting the true and surrogate endpoints for the jth subject in the ith trial, respectively, and let Z_{ij} be the indicator variable for treatment. First, consider the following fixed-effects models:

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \tag{6.17}$$

$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \tag{6.18}$$

where μ_{Si} and μ_{Ti} are trial-specific intercepts, α_i and β_i are trial-specific effects of treatment Z_{ij} on the endpoints in trial i, and ε_{Si} and ε_{Ti} are correlated error terms, assumed to be zero-mean normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}. \tag{6.19}$$

In addition, we can decompose

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{Si} \\ m_{Ti} \\ a_i \\ b_i \end{pmatrix}, \tag{6.20}$$

where the second term on the right hand side of (6.20) is assumed to follow a zero-mean normal

distribution with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}. \tag{6.21}$$

A classical hierarchical, random-effects modeling strategy results from the combination of the above two steps into a single one:

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \tag{6.22}$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \tag{6.23}$$

Here, μ_S and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random intercepts, and a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects $(m_{Si}, m_{Ti}, a_i, b_i)$ are assumed to be mean-zero normally distributed with covariance matrix (6.21). The error terms ε_{Sij} and ε_{Tij} follow the same assumptions as in the fixed effects models.

After fitting the above models, surrogacy is captured by means of two quantities: trial-level and individual-level coefficients of determination. The former quantifies the association between the treatment effects on the true and surrogate endpoints at the trial level, while the latter measures the association at the level of the individual patient, after adjustment for the treatment effect. The former is given by:

$$R^2_{\text{trial}} = R^2_{b_i|m_{Si},a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^\top \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \tag{6.24}$$

The above quantity is unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, lies within the unit interval.

Apart from estimating the strength of surrogacy, the above model can also be used for prediction purposes. To this end, observe that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0|m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^\top \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \tag{6.25}$$

$$\text{Var}(\beta + b_0|m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^\top \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \tag{6.26}$$

A prediction can be made using (6.25), with prediction variance (6.26). Of course, one has to properly acknowledge the uncertainty resulting from the fact that parameters are not known but merely estimated.

Though the above hierarchical modeling is elegant, it often poses a considerable computational challenge (Burzykowski, Molenberghs, and Buyse 2005). To address this problem, Tibaldi *et al.* (2003) suggested several simplifications.

Table 6: *Examples of possible surrogate endpoints in various diseases (Abbreviations: AIDS = acquired immune deficiency syndrome; ARMD = age-related macular degeneration; HIV = human immunodeficiency virus).*

Disease	Surrogate Endpoint	Type	Final Endpoint	Type
Resectable solid tumor	Time to recurrence	Censored	Survival	Censored
Advanced cancer	Tumor response	Binary	Time to progression	Censored
Osteoporosis	Bone mineral density	Longitudinal	Fracture	Binary
Cardiovascular disease	Ejection fraction	Continuous	Myocardial infraction	Binary
Hypertension	Blood pressure	Longitudinal	Coronary heart disease	Binary
Arrhythmia	Arrhythmic episodes	Longitudinal	Survival	Censored
ARMD	6-month visual acuity	Continuous	24-month visual acuity	Continuous
Glaucoma	Intraoccular pressure	Continuous	Vision loss	Censored
Depression	Biomarkers	Multivariate	Depression scale	Continuous
HIV infection	CD4 counts + viral load	Multivariate	Progression to AIDS	Censored

6.2 Non-Gaussian Endpoints

Statistically speaking, the surrogate endpoint and the clinical endpoint are realizations of random variables. As will be clear from the formalism in Section 6.1, one is in need of the joint distribution of these variables. The easiest, but not the only, situation is where both are Gaussian random variables, but one also encounters binary (e.g., CD4+ counts over 500/mm3, tumor shrinkage), categorical (e.g., cholesterol levels <200 mg/dl, 200-299 mg/dl, 300+ mg/dl, tumor response as complete response, partial response, stable disease, progressive disease), censored continuous (e.g., time to undetectable viral load, time to cardiovascular death), longitudinal (e.g., CD4+ counts over time, blood pressure over time), and multivariate longitudinal (e.g., CD4+ and viral load over time jointly, various dimensions of quality of life over time) endpoints. The models used to validate a surrogate for a clinical endpoint will depend on the type of variables observed in the problem at hand. Table 6 shows some examples of potential surrogate endpoints in various diseases. In what follows, we will briefly discuss the settings of binary endpoints, failure-time endpoints, the combination of an ordinal and a survival endpoint, and longitudinal endpoints.

6.2.1 Binary Endpoints

Renard *et al.* (2002) have shown that extension to this situation is easily done using a latent variable formulation. That is, one posits the existence of a pair of continuously distributed latent variable responses $(\widetilde{S}_{ij},\ \widetilde{T}_{ij})$ that produce the actual values of (S_{ij}, T_{ij}). These unobserved variables are assumed to have a joint normal distribution and the realized values follow by double dichotomization. On the latent-variable scale, we obtain a model similar to (6.17)–(6.18) and in the matrix (6.19) the variances are set equal to unity in order to ensure identifiability. This leads to the following model:

$$\begin{cases} \Phi^{-1}(P[S_{ij}=1|Z_{ij}, m_{S_i}, a_i, m_{T_i}, b_i]) &= \mu_S + m_{S_i} + (\alpha + a_i)Z_{ij}, \\ \Phi^{-1}(P[T_{ij}=1|Z_{ij}, m_{S_i}, a_i, m_{T_i}, b_i]) &= \mu_T + m_{T_i} + (\beta + b_i)Z_{ij}, \end{cases}$$

where Φ denotes the standard normal cumulative distribution function. Renard *et al.* (2002) used pseudo-likelihood methods to estimate the model parameters. Similar ideas have been used in the case one of the endpoints is continuous, with the other one binary or categorical (Burzykowski, Molenberghs, and Buyse 2005, Ch. 6).

6.2.2 Two Failure-time Endpoints

Assume now that S_{ij} and T_{ij} are failure-time endpoints. Model (6.17)–(6.18) is replaced by a model for two correlated failure-time random variables. Burzykowski *et al.* (2001) used copulas to this end (Clayton 1978, Hougaard 1986). Precisely, one assumes the joint survivor function of (S_{ij}, T_{ij}) is written as:

$$F(s, t) = P(S_{ij} \geq s, T_{ij} \geq t) = K_\xi\{F_{Sij}(s), F_{Tij}(t)\}, \quad s, t \geq 0, \tag{6.27}$$

where (F_{Sij}, F_{Tij}) denote marginal survivor functions and K_ξ is a copula, i.e., a distribution function on $[0, 1]^2$ with ξ taking values on the real line.

When the hazard functions are specified, estimates of the parameters for the joint model can be obtained using maximum likelihood. Shih and Louis (1995) discuss alternative estimation methods. The association parameter is generally hard to interpret. However, it can be shown (Genest and McKay 1986) that there is a link with Kendall's τ:

$$\tau = 4 \int_0^1 \int_0^1 K_\xi(u, v) K_\xi(du, dv) - 1,$$

providing an easy measure of surrogacy at the individual level. At the second stage R^2_{trial} can be computed based on the pairs of treatment effects estimated at the first stage.

6.2.3 An Ordinal Surrogate and a Survival Endpoint

Assume that T is a failure-time random variable and S is a categorical variable with K ordered categories. To propose validation measures, similar to those introduced in the previous section, Burzykowski *et al.* (2004) also used bivariate copulas, combining ideas of Molenberghs, Geys, and Buyse (2001) and Burzykowski *et al.* (2001). One marginal distribution is a proportional odds logistic regression, while the other is a proportional hazards model. The Plackett copula (Dale 1986) was chosen to capture the association between both endpoints. The ensuing global odds ratio is relatively easy to interpret.

6.2.4 Longitudinal Endpoints

Most of the previous work focuses on univariate responses. Alonso *et al* (2003) showed that going from a univariate setting to a multivariate framework represents new challenges. The R^2 measures proposed by Buyse *et al* (2000), are no longer applicable. Alonso *et al* (2003) based their calculations of surrogacy measures on a two-stage approach rather than a full random-effects approach. They assume that information from $i = 1, \ldots, N$ trials is available, in the ith of which, $j = 1, \ldots, n_i$ subjects are enrolled and they denoted the time at which subject j in trial i is measured as t_{ijk}. If

T_{ijk} and S_{ijk} denote the associated true and surrogate endpoints, respectively, and Z_{ij} is a binary indicator variable for treatment then along the ideas of Galecki (1994), they proposed the following joint model, at the first stage, for both responses

$$\begin{cases} T_{ijk} = \mu_{Ti} + \beta_i Z_{ij} + g_{Tij}(t_{ijk}) + \varepsilon_{Tijk}, \\ \\ S_{ijk} = \mu_{Si} + \alpha_i Z_{ij} + g_{Sij}(t_{ijk}) + \varepsilon_{Sijk}, \end{cases} \tag{6.28}$$

where μ_{Ti} and μ_{Si} are trial-specific intercepts, β_i and α_i are trial-specific effects of treatment Z_{ij} on the two endpoints and g_{Tij} and g_{Sij} are trial-subject-specific time functions that can include treatment-by-time interactions. They also assume that the vectors, collecting all information over time for patient j in trial i, $\widetilde{\varepsilon}_{T_{ij}}$ and $\widetilde{\varepsilon}_{S_{ij}}$ are correlated error terms, following a mean-zero multivariate normal distribution with covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma_{TSi}^\top & \Sigma_{SSi} \end{pmatrix} = \begin{pmatrix} \sigma_{TTi} & \sigma_{TSi} \\ \sigma_{TSi} & \sigma_{SSi} \end{pmatrix} \otimes R_i. \tag{6.29}$$

Here, R_i is a correlation matrix for the repeated measurements.

If treatment effect can be assumed constant over time, then (6.24) can still be useful to evaluate surrogacy at the trial level. However, at the individual level the situation is totally different, the R^2_{ind} no longer being applicable, and new concepts are needed.

Using multivariate ideas, Alonso *et al* (2003) proposed the *variance reduction factor* (VRF) to capture individual-level surrogacy in this more elaborate setting. They quantified the relative reduction in the true endpoint variance after adjustment by the surrogate as

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})}, \tag{6.30}$$

where $\Sigma_{(T|S)_i}$ denotes the conditional variance-covariance matrix of $\widetilde{\varepsilon}_{T_{ij}}$ given $\widetilde{\varepsilon}_{S_{ij}}$: $\Sigma_{(T|S)i} = \Sigma_{TTi} - \Sigma_{TSi}\Sigma_{SSi}^{-1}\Sigma_{TSi}^\top$. Here, Σ_{TTi} and Σ_{SSi} are the variance-covariance matrices associated with the true and surrogate endpoint respectively and Σ_{TSi} contains the covariances between the surrogate and the true endpoint. Alonso *et al* (2003) showed that the VRF_{ind} ranges between zero and one, and that $VRF_{\text{ind}} = R^2_{\text{ind}}$ when the endpoints are measured only once.

An alternative proposal is

$$\theta_p = \sum_i \frac{1}{Np_i} \text{tr}\left\{\left(\Sigma_{TTi} - \Sigma_{(T|S)i}\right)\Sigma_{TTi}^{-1}\right\}. \tag{6.31}$$

Structurally, both VRF and θ_p are similar, the difference being the reversal of summing the trace and calculating the ratio. In spite of this strong structural similarity the VRF is not symmetric in S and T and it is only invariant with respect to linear orthogonal transformations, whereas θ_p is both symmetric and invariant with respect to the broader class of linear bijective transformations.

A common problem of all previous proposals is that they are strongly based on the normality assumption and extensions to non-normal settings are difficult. To overcome this limitation, Alonso *et al* (2005), introduced a new parameter, the so-called R^2_Λ, to evaluate surrogacy at the individual

level when both responses are measured over time or in general when multivariate or repeated measures are available

$$R^2_\Lambda = \frac{1}{N}\sum_i (1-\Lambda_i), \tag{6.32}$$

where: $\Lambda_i = \dfrac{|\Sigma_i|}{|\Sigma_{TTi}|\,|\Sigma_{SSi}|}$. This parameter not only allows the detection of more general patterns of association but can also be extended to more general settings that those defined by the normal distribution. They proved that R^2_Λ ranges between zero and one, and that in the cross-sectional case $R^2_\Lambda = R^2_{\text{ind}}$. These authors have shown that $R^2_\Lambda = 1$ whenever there is a deterministic relationship between two linear combinations of both endpoints, allowing the detection of strong associations in cases where the VRF or θ_p would fail in doing so.

6.3 Towards a Unified Approach

The longitudinal method of the previous section, while elegant, hinges upon normality. First using the likelihood reduction factor (Section 6.3.1) and then an information-theoretic approach (Section 6.3.2), extension, and therefore unification, will be achieved.

6.3.1 The Likelihood Reduction Factor

Estimating individual-level surrogacy, as the previous developments clearly show, has frequently been based on a variance-covariance matrix coming from the distribution of the residuals. However, if we move away from the normal distribution, it is not always clear how to quantify the association between both endpoints after adjusting for treatment and trial effect. To address this problem, Alonso *et al* (2004) considered the following generalized linear models in the ith trial

$$g_T(T_{ij}) = \mu_{T_i} + \beta_i Z_{ij}, \tag{6.33}$$

$$g_T(T_{ij}) = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \tag{6.34}$$

The longitudinal case would be covered by considering particular functions of time in (6.33) and (6.34). Consider G^2_i as the log-likelihood ratio test statistics to compare (6.33) with (6.34) in trial i, and quantify the association between both endpoints at the individual level using a scaled likelihood reduction factor (LRF)

$$\text{LRF} = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G^2_i}{n_i}\right). \tag{6.35}$$

Alonso *et al.* (2004) established a number of properties for LRF, in particular its ranging in the unit interval, and its reduction to R^2_Λ in the longitudinal and to R^2_{ind} in the cross-sectional case.

6.3.2 An Information-theoretic Unification

This proposal avoids the needs for a joint, hierarchical model, and allows for unification across different types of endpoints. The entropy of a random variable (Shannon 1948), a good measure of

randomness or uncertainty, is defined in the following way for the case of a discrete random variable Y, taking values $\{k_1, k_2, \ldots, k_m\}$, and with probability function $P(Y = k_i) = p_i$:

$$H(Y) = \sum_i p_i \log\left(\frac{1}{p_i}\right). \tag{6.36}$$

The differential entropy $h_d(X)$ of a continuous variable X with density $f_X(x)$ and support S_{f_X} equals

$$h_d(Y) = -E[\log f_X(X)] = -\int_{S_{f_X}} f_X(x) \log f_X(x) dx. \tag{6.37}$$

The joint and conditional (differential) entropies are defined in an analogous fashion. Defining the information of a single event as $I(A) = \log p_A$, the entropy is $H(A) = -I(A)$. No information is gained from a totally certain event, $p_A \approx 1$, so $I(A) \approx 0$), while an improbable event is informative. $H(Y)$ is the average uncertainty associated with P. Entropy is always non-negative, satisfies $H(Y|X) \leq H(Y)$ for any pair of random variables, with equality holding under independence, and is invariant under a bijective transformation (Cover and Tomas 1991). Differential entropy enjoys some but not all properties of entropy: it can be infinitely large, negative, or positive, and is coordinate dependent. For a bijective transformation $Y = y(X)$, it follows $h_d(Y) = h_d(X) - \mathrm{E}_Y\left(\log\left|\frac{dx}{dy}(y)\right|\right)$.

We can now quantify the amount of uncertainty in Y, expected to be removed if the value of X were known, by $I(X, Y) = h_d(Y) - h_d(Y|X)$, the so-called *mutual information*. It is always non-negative, zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y, and $I(X, X) = h_d(X)$. The mutual information measures the information of X, shared by Y.

We will now introduce the entropy-power (Shannon 1948) for comparison of continuous random variables. Let X be a continuous n-dimensional random vector. The entropy-power of X is

$$\mathrm{EP}(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}. \tag{6.38}$$

The differential entropy of a continuous normal random variable is $h(X) = \frac{1}{2}\log\left(2\pi\sigma^2\right)$, a simple function of the variance and, on the natural logarithmic scale: $\mathrm{EP}(X) = \sigma^2$. In general, $\mathrm{EP}(X) \leq \mathrm{Var}(X)$ with equality if and only if X is normally distributed.

We can now define an information-theoretic measure of association (Schemper and Stare 1996):

$$R_h^2 = \frac{\mathrm{EP}(Y) - \mathrm{EP}(Y|X)}{\mathrm{EP}(Y)}, \tag{6.39}$$

which ranges in the unit interval, equals zero if and only if (X, Y) are independent, is symmetric, is invariant under bijective transformation of X and Y, and, when $R_h^2 \to 1$ for continuous models, there is usually some degeneracy appearing in the distribution of (X,Y). There is a direct link between R_h^2 and the mutual information: $R_h^2 = 1 - e^{-2I(X,Y)}$. For Y discrete: $R_h^2 \leq 1 - e^{-2H(Y)}$, implying that R_h^2 then has an upper bound smaller than 1; we then redefine

$$R_{h\max}^2 = \frac{R_h^2}{1 - e^{-2H(Y)}},$$

reaching 1 when both endpoints are deterministically related.

We can now redefine surrogacy, while preserving previous proposals as special cases. While we will focus on individual-level surrogacy, all results apply to the trial level too. Let $Y = T$ and $X = S$ be the true and surrogate endpoints, respectively. We consider S a good surrogate for T at the individual (trial) level, if a "large" amount of uncertainty about T (the treatment effect on T) is reduced when S (the treatment effect on S) is known. Equivalently, we term S a good surrogate for T at the individual level, if our lack of knowledge about the true endpoint is substantially reduced when the surrogate endpoint is known.

A meta-analytic framework, with N clinical trials, produces N_q different R^2_{hi}, and hence we propose a meta-analytic R^2_h:

$$R^2_h = \sum_{i=1}^{N_q} \alpha_i R^2_{hi} = 1 - \sum_{i=1}^{N_q} \alpha_i e^{-2I_i(S_i,T_i)},$$

where $\alpha_i > 0$ for all i and $\sum_{i=1}^{N_q} \alpha_i = 1$. Different choices for α_i lead to different proposals, producing an uncountable family of parameters. This opens the additional issue of finding an *optimal* choice. In particular, for the cross-sectional normal-normal case, Alonso and Molenberghs (2007) have shown that $R^2_h = R^2_{\text{ind}}$. The same holds for R^2_Λ, defined in (6.28) for the longitudinal case. Finally, when the true and surrogate endpoints have distributions in the exponential family, then LRF $\xrightarrow{P} R^2_h$ when the number of subjects per trial goes to infinity.

6.3.3 Fano's Inequality and the Theoretical Plausibility of Finding a Good Surrogate

Fano's inequality shows the relationship between entropy and prediction:

$$\mathrm{E}\left[(T - g(S))^2\right] \geq \mathrm{EP}(T)(1 - R^2_h) \tag{6.40}$$

where $\mathrm{EP}(T) = \dfrac{1}{2\pi e} e^{2h(T)}$. Note that nothing has been assumed about the distribution of our responses and no specific form has been considered for the prediction function g. Also, (6.40) shows that the predictive quality strongly depends on the characteristics of the endpoint, specifically on its power-entropy. Fano's inequality states that the prediction error increases with $\mathrm{EP}(T)$ and therefore, if our endpoint has a large power-entropy then a surrogate should produce a large R^2_h to have some predictive value. This means that, for some endpoints, the search for a good surrogate can be a dead end street: the larger the entropy of T the more difficult it is to predict. Studying the the power-entropy before trying to find a surrogate is therefore advisable.

7 Incomplete Data

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

Table 7: *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

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 7.

Let us first consider the case where one 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, fore 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.

7.1 Direct Likelihood Analysis

For continuous outcomes, Verbeke and Molenberghs (2000) describe likelihood-based mixed-effects models, in the spirit of Section 3.1, that 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 section, 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. A direct-likelihood analysis is equivalent to a classical MANOVA analysis when data are complete. 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. However, when data are incomplete, one should be aware that MANOVA and comparisons per time point are only valid under MCAR and less efficient compared to a likelihood analysis; this was also noted in Section 3.3, where the t-test for treatment differences at month 12 for the toenail data was found less efficient than the linear mixed effects model. On the other hand, under MAR, both MANOVA and comparisons per time point will not only be less efficient, but more importantly, they will produced biased results, because they do not take into account that the observed data no longer constitute a random sample from the target population. Therefore, the full likelihood analysis constitutes a very promising alternative to CC and LOCF. When a relatively large number of measurements is made within a single subject, the full power of random-effects modeling can

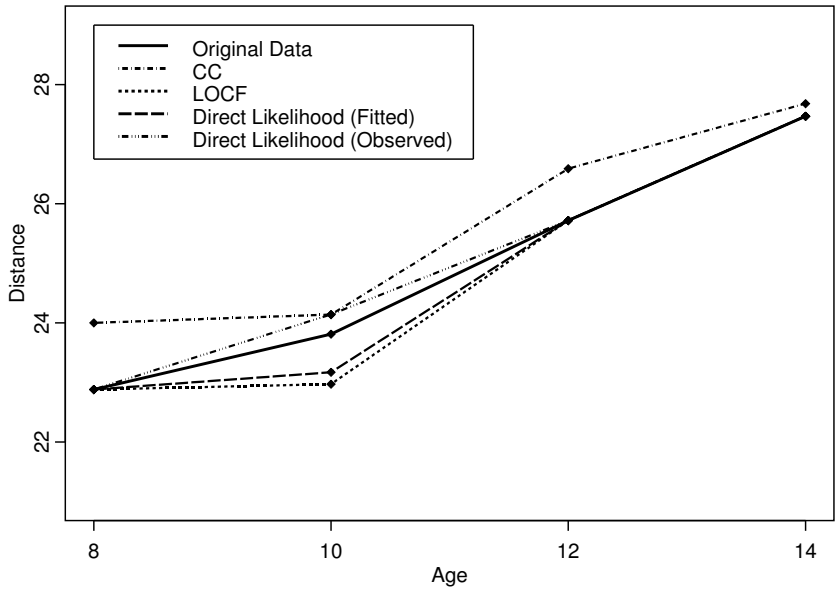

Figure 8: *Orthodontic Growth Data. Profiles for the original data, CC, LOCF, and direct likelihood for boys.*

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 7.4.

7.2 Illustration: Orthodontic Growth Data

The simple methods and direct likelihood method are now compared using the growth data of Section 2.4. 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.

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 8 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.

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 8, 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 8 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

Table 8: *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)

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 8 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.

7.3 Incompleteness and Estimating Equations

7.3.1 Weighted Generalized Estimating Equations

As Liang and Zeger (1986) pointed out, GEE-based inferences are valid only under MCAR, due to the fact that they are based on frequentist considerations. An important exception, mentioned by

these authors, is the situation where the working correlation structure (discussed in the previous section), happen to be correct, since then the estimates and model-based standard errors are valid under the weaker MAR. This is because then, the estimating equations can be interpreted as likelihood equations. In general, of course, the working correlation structure will not be correctly specified. The ability to do so is the core motivation of the method, and therefore Robins, Rotnitzky, and Zhao (1995) proposed a class of *weighted estimating equations* to allow for MAR, extending GEE.

The idea is to weight each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he dropped out. This can be calculated, for example, as

$$\begin{aligned}\nu_{id_i} \equiv P[D_i = d_i] &= \prod_{k=2}^{d_i-1} (1 - P[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1]) \times \\ &\quad P[R_{id_i} = 0 | R_{i2} = \ldots = R_{i,d_i-1} = 1]^{I\{d_i \leq T\}}.\end{aligned}$$

Recall that we partitioned Y_i into the unobserved components Y_i^m and the observed components Y_i^o. Similarly, we can make the exact same partition of μ_i into ${\mu_i}^m$ and ${\mu_i}^o$. In the weighted GEE approach, which is proposed to reduce possible bias of $\hat{\beta}$, the score equations to be solved when taking into account the correlation structure are:

$$\begin{aligned}S(\beta) &= \sum_{i=1}^{N} \frac{1}{\nu_{id_i}} \frac{\partial \mu_i}{\partial \beta^\top} (A_i^{1/2} C_i A_i^{1/2})^{-1} (y_i - \mu_i) \;=\; 0 \\ &= \sum_{i=1}^{N} \sum_{d=2}^{n+1} \frac{I(D_i = d)}{\nu_{id}} \frac{\partial \mu_i}{\partial \beta^\top}(d) (A_i^{1/2} C_i A_i^{1/2})^{-1}(d) (y(d) - \mu_i(d)) \;=\; 0, \qquad (7.41)\end{aligned}$$

where $y_i(d)$ and $\mu_i(d)$ are the first $d-1$ elements of y_i and μ_i respectively. We define $\frac{\partial \mu_i}{\partial \beta^\top}(d)$ and $(A_i^{1/2} C_i A_i^{1/2})^{-1}(d)$ analogously.

It is worthwhile to note that the recently proposed so-called *doubly robust* methods (van der Laan and Robins 2002) is more efficient and robust to a wider class of deviations. However, it is harder to implement than the original proposal.

An alternative mode of analysis, generally overlooked but proposed by Schafer (2003), would consist in multiply imputing the missing outcomes using a parametric model, e.g., of a random-effects or conditional type, followed by conventional GEE and conventional multiple-imputation inference on the so-completed sets of data. This approach is discussed in Beunckens, Sotto, and Molenberghs (2007).

7.3.2 Analysis of the Age-related Macular Degeneration Trial

We compare analyses performed on the completers only (CC), on the LOCF imputed data, as well as on the observed data. For the observed, partially incomplete data, GEE is supplemented with WGEE. The GEE analyses are reported in Table 9. A working exchangeable correlation matrix is considered. The model has four intercepts and four treatment effects. Precisely, the marginal

Table 9: *Age-related Macular Degeneration Trial. Parameter estimates (model-based standard errors; empirically corrected standard errors) for the marginal models: GEE on the CC and LOCF population, and on the observed data. In the latter case, also WGEE is used.*

Effect	Par.	CC	LOCF	Observed data	
				Unweighted	WGEE
Int.4	β_{11}	-1.01(0.24;0.24)	-0.87(0.20;0.21)	-0.87(0.21;0.21)	-0.98(0.10;0.44)
Int.12	β_{21}	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.21;0.21)	-1.78(0.15;0.38)
Int.24	β_{31}	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.22;0.22)	-1.11(0.15;0.33)
Int.52	β_{41}	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.72(0.25;0.39)
Tr.4	β_{12}	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.28;0.28)	0.80(0.15;0.67)
Tr.12	β_{22}	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.29;0.29)	1.87(0.19;0.61)
Tr.24	β_{32}	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.73(0.20;0.52)
Tr.52	β_{42}	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.74(0.31;0.52)
Corr.	ρ	0.39	0.44	0.39	0.33

Table 10: *Age-related Macular Degeneration Trial. Parameter estimates (standard errors) for a logistic regression model to describe dropout.*

Effect	Parameter	Estimate (s.e.)
Intercept	ψ_0	0.14 (0.49)
Previous outcome	ψ_1	0.04 (0.38)
Treatment	ψ_2	-0.86 (0.37)
Lesion level 1	ψ_{31}	-1.85 (0.49)
Lesion level 2	ψ_{32}	-1.91 (0.52)
Lesion level 3	ψ_{33}	-2.80 (0.72)
Time 2	ψ_{41}	-1.75 (0.49)
Time 3	ψ_{42}	-1.38 (0.44)

regression model takes the form

$$\text{logit}[P(Y_{ij} = 1|T_i)] = \beta_{j1} + \beta_{j2}T_i,$$

where $j = 1, \ldots, 4$ refers to measurement occasion, T_i is the treatment assignment for subject $i = 1, \ldots, 240$ and Y_{ij} is the indicator for whether or not 3 lines of vision have been lost for subject i at time j. The advantage of having separate treatment effects at each time is that particular attention can be given at the treatment effect assessment at the last planned measurement occasion, i.e., after one year. From Table 9 it is clear that the model-based and empirically corrected standard errors agree extremely well. This is due to the unstructured nature of the full time by treatment mean structure. However, we do observe differences in the WGEE analyses. Not only are the parameter estimates mildly different between the two GEE versions, there is a dramatic difference between the model-based and empirically corrected standard errors. Nevertheless, the two sets of empirically corrected standard errors agree very closely, which is reassuring.

When comparing parameter estimates across CC, LOCF, and observed data analyses, it is clear that LOCF has the effect of artificially increasing the correlation between measurements. The effect is

mild in this case. The parameter estimates of the observed-data GEE are close to the LOCF results for earlier time points and close to CC for later time points. This is to be expected, as at the start of the study the LOCF and observed populations are virtually the same, with the same holding between CC and observed populations near the end of the study. Note also that the treatment effect under LOCF, especially at 12 weeks and after 1 year, is biased downward in comparison to the GEE analyses. To properly use the information in the missingness process, WGEE can be used. To this end, a logistic regression for dropout, given covariates and previous outcomes, needs to be fitted. Parameter estimates and standard errors are given in Table 10. Intermittent missingness will be ignored. Covariates of importance are treatment assignment, the level of lesions at baseline (a four-point categorical variable, for which three dummies are needed), and time at which dropout occurs. For the latter covariates, there are three levels, since dropout can occur at times 2, 3, or 4. Hence, two dummy variables are included. Finally, the previous outcome does not have a significant impact, but will be kept in the model nevertheless. In spite of there being no strong evidence for MAR, the results between GEE and WGEE differ quite a bit. It is noteworthy that at 12 weeks, a treatment effect is observed with WGEE which goes unnoticed with the other marginal analyses. This finding is mildly confirmed by the random-intercept model, when the data as observed are used.

7.4 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).

7.5 The Link Between Joint Modeling and Incomplete Data

In Section 5, the main research interest was in the time-to-event outcome, and we have motivated joint modeling approaches in order to adequately take into account in our analysis the effect of a time-dependent covariate measured with error. However, joint modeling may be also required when interest is in the longitudinal outcome. In particular, the occurrence of events causes dropout due to the fact that no longitudinal measurements are usually available at and after the event (e.g., death). As we have seen earlier in this section, if the probability of dropout depends on unobserved longitudinal components, i.e., is MNAR, then the dropout process must be explicitly taken into account in order to produce valid inferences for the longitudinal model. One of the modeling frameworks that has been proposed in the missing data literature to handle nonrandom dropout is the shared parameter models (Wu and Carroll 1988, Follmann and Wu 1995). These models posit a survival sub-model for the time-to-dropout and a mixed effects sub-model for the longitudinal responses, and therefore, they belong in fact to same family as the joint model (5.15). When approached from the missing-data point of view, the basic assumption behind these models is that the probability of dropout at time t depends on values of the longitudinal outcome at both past and future time points, through a set of random effects. To show this more clearly, we define for each subject the observed and missing part of the longitudinal response vector. The observed part $y_i^o = \{y_i(t_{ij}) : t_{ij} < T_i, j = 1, \ldots, n_i\}$ contains all observed longitudinal measurements of the ith subject before the observed event time, whereas the missing part $y_i^m = \{y_i(t_{ij}) : t_{ij} \geq T_i, j = 1, \ldots, n_i'\}$ contains the longitudinal measurements that would have been taken until the end of the study, had the event not occurred. Under these definitions, we can derive the dropout mechanism, which is the conditional distribution of the time-to-dropout given the complete vector of longitudinal responses (y_i^o, y_i^m),

$$f(T_i^* \mid y_i^o, y_i^m; \theta) = \int f(T_i^* \mid b_i; \theta) f(b_i \mid y_i^o, y_i^m; \theta)\, db_i, \tag{7.42}$$

which states that the time-to-dropout depends on y_i^m through the posterior distribution of the random effects $f(b_i \mid y_i^o, y_i^m; \theta)$. In practice, this implies that such models are most meaningful when subjects who experience the event sooner, are the ones that show steeper evolutions in their longitudinal profiles.

8 Software Considerations

Let us provide a brief overview of useful software tools, relative to the methodology described and exemplified in this chapter.

Linear mixed models can be fitted using the SAS procedures MIXED, GLIMMIX, and NLMIXED, and the R packages nlme and lme4.

Generalized linear mixed models have been implemented in the SAS procedures GLIMMIX and NLMIXED; they can also be fitted using the R packages lme4, glmmML, MCMCglmm among others.

GEE can be fitted using the SAS procedure GENMOD and the R packages gee and geepack.

When incomplete data are analyzed using multiple imputation, the SAS procedures MI and MIANALYZE apply. Likewise, a suite of R functions is available in packages mice, mitools and Hmisc. For direct-likelihood analysis, simply the aforementioned SAS and R tools apply. Weighted estimating equations require user-defined software.

User-defined software is also needed for the validation of surrogate markers, for high-dimensional data, and for joint modeling.

The authors and their collaborators have developed a variety of software tools, made available via their web sites.

9 Concluding Remarks

Models for the analysis of longitudinal and otherwise hierarchical data are omnipresent these days throughout empirical research. Indeed, models and analysis techniques for longitudinal data, be it for Gaussian or non-Gaussian outcomes, are showing up in biometry, medical statistics, epidemiology, psychometry, econometrics, social science, and survey applications. The models are appealing for the intuition behind their formulation. Inferential apparatus is now well developed, and many methods have been implemented in standard software packages.

In this chapter, we have presented basic methodology for Gaussian and non-Gaussian longitudinal data, including the linear and generalized linear mixed model and generalized estimating equations. We also placed a strong emphasis on the use of these methods in conjunction with a time-to-event outcome, also known as joint modeling. Furthermore, we have indicated how models for longitudinal data are playing a role in the validation of surrogate markers.

Finally, we have placed some emphasis on the problem of incomplete data, and how likelihood-based or Bayesian analysis of incomplete longitudinal data can be performed easily when data are not fully observed, given the missing data are missing at random. Related to this, we have indicated how the joint modeling framework can play a role when the missing data are not missing at random.

Acknowledgments

The authors gratefully acknowledge support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002) *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

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.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2003) Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal* **45**, 931–945.

Alonso, A. and Molenberghs, G. (2007) Surrogate marker evaluation from an information theory perspective. *Biometrics*, **63**, 180–186.

Alonso, A., Molenberghs, G., Geys, H., and Buyse, M. (2005) A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine* **25**, 205–211.

Alonso, A. Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J., and Buyse, M. (2004) Prentice's approach and the meta analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics*, **60**, 724–728.

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

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer.

Arnold, B.C. and Strauss, D. (1991) Pseudolikelihood estimation: some examples, *Sankhya: The Indian Journal of Statistics - Series B*, **53**, 233–243.

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.

Beunckens, C., Sotto, C., and Molenberghs, G. (2007) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, **00**, 000–000.

Brant, L.J. and Fozard, J.L. (1990) Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging. *Journal of the Acoustical Society of America*, **88**, 813–820.

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

Brown, E. and Ibrahim, J. (2003) A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221–228.

Brown, E., Ibrahim, J., and DeGruttola, V. (2005) A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, **61**, 64–73.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004) The validation of surrogate endpoints using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society, Series A*, **167**, 103–124.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001) Validation of surrogate endpoints in multiple randomized clinical trials with failure time end points. *Applied Statistics*, **50**, 405–422.

Buyse, M. and Molenberghs, G. (1998) The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000) The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infraction. *New England Journal of Medicine*, **321**, 406–412.

Catalano, P.J. (1997) Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, **16**, 883–900.

Catalano, P.J. and Ryan, L.M. (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651–658.

Chakraborty, H., Helms, R.W., Sen, P.K., and Cohen, M.S. (2003) Estimating correlation by using a general linear mixed model: Evaluation of the relationship between the concentration of HIV-1 RNA in blood and semen. *Statistics in Medicine*, **22**, 1457–1464.

Chi, Y.-Y. and Ibrahim, J. (2006) Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **62**, 432–445.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

Cover, T. and Tomas, J. (1991) *Elements of Information Theory.* New York: Wiley.

Cox, N.R. (1974) Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, **30**, 171–178.

Cox, D.R. and Wermuth, N. (1992) Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441–461.

Cox, D. R. and Wermuth, N. (1994a) A note on the quadratic exponential binary distribution. *Biometrika*, **81**, 403–408.

Cox, D.R. and Wermuth, N. (1994b) *Multivariate Dependencies: Models, Analysis and Interpretation.* London: Chapman & Hall.

Dale, J.R. (1986) Global cross ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Daniels, M.J. and Hughes, M.D. (1997) Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1515–1527.

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

DeGruttola, V. and Tu, X. (1994) Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

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

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data.* New York: Oxford University Press.

Ding, J. and Wang, J.-L. (2008) Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**, 546–556.

Dobson, A. and Henderson, R. (2003) Diagnostics for joint longitudinal and dropout time modeling. *Biometrics*, **59**, 741–751.

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

Elashoff, R., Li, G., and Li, N. (2008) A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, **64**, 762–771.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models.* Heidelberg: Springer.

Faucett, C., Schenker, N., and Elashoff, R. (1998) Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association*, **93**, 427–437.

Ferentz, A.E. (2002). Integrating pharmacogenomics into drug development. *Pharmacogenomics*, **3**, 453–467.

Fieuws, S. and Verbeke, G. (2004) Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. *Statistics in Medicine*, **23**, 3093–3104.

Fieuws, S. and Verbeke, G. (2006) Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Fieuws, S., Verbeke, G., Boen, F., and Delecluse, C. (2006) High-dimensional multivariate mixed models for binary questionnaire data. *Applied Statistics*, **55**, 1–12.

Fitzmaurice, G.M. and Laird, N.M. (1995) Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845–852.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004) *Applied Longitudinal Analysis*. New York: John Wiley & Sons.

Fleming, T.R. and DeMets, D.L. (1996) Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine*, **125**, 605–613.

Folk, V.G. and Green, B.F. (1989) Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, **13**, 373–389.

Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

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.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992) Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., Carroll, R.J. (2000) On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.

Galecki, A. (1994) General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: theory and methods*, **23**, 3105–3119.

Genest, C. and McKay, J. (1986) The joy of copulas: bivariate distributions with uniform marginals. *American Statistician*, **40**, 280–283.

Geys, H., Molenberghs, G., and Ryan, L.M. (1997) Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Geys, H., Molenberghs, G., and Ryan, L. (1999) Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 734–745.

Gueorguieva, R. (2001) A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, **1**, 177–193.

Goldstein, H. (1979) *The Design and Analysis of Longitudinal Studies.* London: Academic Press.

Hartley, H.O. and Hocking, R. (1971) The analysis of incomplete data. *Biometrics*, **27**, 7783–7808.

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, R., Diggle, P. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.

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.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

Hsieh, F., Tseng, Y.-K. and Wang, J.-L. (2006) Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, **62**, 1037–1043.

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

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

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis.* Oxford: Clarendon Press.

Lagakos, S.W. and Hoth, D.F. (1992) Surrogate markers in AIDS: Where are we? Where are we going? *Annals of Internal Medicine*, **116**, 599–601.

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.

Lange, K. (2004) *Optimization*. New York: Springer.

Lesko, L.J. and Atkinson, A.J. (2001) Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annual Review of Pharmacological Toxicology*, **41**, 347–366.

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.

Lin, H., Turnbull, B., McCulloch, C., and Slate, E. (2002) Laten class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, **97**, 53–65.

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 Schluchter, M.D. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497–512.

Liu, L.C. and Hedeker, D. (2006) A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, **62**, 261–268.

MacCallum, R., Kim, C., Malarkey, W., and Kiecolt-Glaser, J. (1997) Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, **32**, 215–253.

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/CRC.

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., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A., and Buyse, M. (2008) The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *Journal of Statistical Planning and Inference*, **138**, 432–449.

Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A., and Buyse, M. (2009) A unified framework for the evaluation of surrogate endpoints in clinical trials. *Statistical Methods in Medical Research*, **00**, 000–000.

Molenberghs, G., Geys, H., and Buyse, M. (2001) Evaluation of surrogate end-points in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.

Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

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. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Morrell, C.H. and Brant, L.J. (1991) Modelling hearing thresholds in the elderly. *Statistics in Medicine*, **10**, 1453–1464.

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.

Ochi, Y. and Prentice, R.L. (1984) Likelihood inference in a correlated probit regression model. *Biometrika*, **71**, 531–543.

Olkin, I. and Tate, R.F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–465 (with correction in **36**, 343–344).

Oort, F.J. (2001) Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, **54**, 49–78.

Pearson, J.D., Morrell, C.H., Gordon-Salant, S., Brant, L.J., Metter, E.J., Klein, L.L., and Fozard, J.L. (1995) Gender differences in a longitudinal study of age-associated hearing loss. *Journal of the Acoustical Society of America*, **97**, 1196–1205.

Pharmacological Therapy for Macular Degeneration Study Group (1997) Interferon α-IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology*, **115**, 865–872.

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

Prentice, R.L. and Zhao, L.P. (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825–839.

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. (1982) Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika*, **69**, 331–342.

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

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Proust-Lima, C., Joly, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2009) Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach. *Computational Statistics and Data Analysis*, **53**, 1142–1154.

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

Regan, M.M. and Catalano, P.J. (1999a) Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, **55**, 760–768.

Regan, M.M. and Catalano, P.J. (1999b) Bivariate dose-response modeling and risk estimation in developmental toxicology. *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 217–237.

Regan, M.M. and Catalano, P.J. (2000) Regression models for mixed discrete and continuous outcomes with clustering. *Risk Analysis*, **20**, 363–376.

Regan, M.M. and Catalano, P.J. (2002) *Combined Continuous and Discrete Outcomes*. In: *Topics in Modelling of Clustered Data*, Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (eds.), London: Chapman & Hall.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal*, **44**, 1–15.

Rizopoulos, D., Verbeke, G. and Molenberghs, G. (2009a) Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, to appear (doi: 10.1111/j.1541-0420.2009.01273.x).

Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009b) Fully exponential Laplace approximation for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, **71**, 637–654.

Rizopoulos, D., Verbeke, G. and Molenberghs, G. (2008) Shared parameter models under random effects misspecification. *Biometrika*, **95**, 63–74.

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.

Roy, J. and Lin, X. (2000) Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, **56**, 1047–1054.

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.

Sammel, M.D., Ryan, L.M., and Legler, J.M. (1997) Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B,* **59**, 667–678.

Schafer J.L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, **57**, 19–35.

Schatzkin, A. and Gail, M. (2002) The promise and peril of surrogate end points in cancer research. *Nature Reviews Cancer*, **2**, 19–27.

Schemper, M. and Stare, J. (1996) Explained variation in survival analysis. *Statistics in Medicine*, **15**, 1999–2012.

Self, S. and Pawitan, Y. (1992) Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology: Methodological Issues*, (N.P. Jewell, K. Dietz and V.T. Farewell, Eds.). Boston: Birkhauser.

Shah, A., Laird, N., and Schoenfeld, D. (1997) A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, **92**, 775–779.

Shannon, C. (1948) A mathematical theory of communication, *Bell System Technical Journal*, **27** 379–423 and 623–656.

Shock, N.W., Greullich, R.C., Andres, R., Arenberg, D., Costa, P.T., Lakatta, E.G., and Tobin, J.D. (1984) Normal human aging: The Baltimore Longitudinal Study of Aging. *National Institutes of Health publication 84-2450.*

Shih, J.H. and Louis, T.A. (1995) Inferences on association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384–1399.

Sivo, S.A. (2001) Multiple indicator stationary time series models. *Structural Equation Modeling*, **8**, 599–612.

Song, X., Davidian, M. and Tsiatis, A. (2002) A semiparameteric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, **58**, 742–753.

Tate, R.F. (1954) Correlation between a discrete and a continuous variable. *Annals of Mathematical Statistics*, **25**, 603–607.

Tate, R.F. (1955) The theory of correlation between two continuous variables when one is dichotomized. *Biometrika*, **42**, 205–216.

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

Therneau, T. and Grambsch, P. (2000) *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Thiébaut, R., Jacqmin-Gadda, H., Chêne, G., Leport, C., and Commenges, D. (2002) Bivariate linear mixed models using SAS PROC MIXED. *Computer Methods and Programs in Biomedicine*, **69**, 249–256.

Thum, Y.M. (1997) Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, **22**, 77–108.

Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005) Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, **92**, 587–603.

Tsiatis, A. and Davidian, M. (2001) A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, **88**, 447–458.

Tsiatis, A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.

Tsiatis, A., DeGruttola, V. and Wulfsohn, M. (1995) Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. Journal of the American Statistical Association, **90**, 27–37.

van der Laan, M.J. and Robins, J.M. (2002) *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

Verbeke, G., Lesaffre, E., and Spiessens, B. (2001) The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, **35**, 419–434.

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

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001) Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, **57**, 7–14.

Wang, Y. and Taylor, J. (2001) Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, **96**, 895–905.

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.

Wu, M. and Carroll, R. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.

Wulfsohn, M. and Tsiatis, A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S. (2001) Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, **50**, 375–387.

Yu, M., Law, N., Taylor, J., and Sandler, H. (2004) Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, **14**, 835–832.

Zhao, L.P., Prentice, R.L., and Self, S.G. (1992) Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society B*, **54**, 805–811.