Limburgs Universitair Centrum
Faculteit Wetenschappen

ANALYSIS OF INCOMPLETE LONGITUDINAL QUALITY OF LIFE DATA

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen: Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

DESMOND CURRAN

Promotor:
Prof. Dr. G. MOLENBERGHS

Co-promotor:
Prof. Dr. R. SYLVESTER

June 2000

Acknowledgement

I would like to express my gratitude to Prof. Dr. Geert Molenberghs for all his help and support during the preparation of this thesis. The time spent collaborating and discussing has been extremely instructive.

I would like to thank the statistics team of the EORTC Data Center, Brussels (especially Prof. Richard Sylvester and Dr. Francesco Pignatti) and the scientific staff of the Center of Statistics in the Faculty of Science at the L.U.C (especially Dr. K. van Steen and Herbert Thijs). I would also like to acknowledge Dr. Jocelyn Kramer for her careful review of the various drafts of this thesis.

Several sections of the thesis are based on joint research with: Neil Aaronson, Marisa Bacchi, Diane Fairclough, Peter Fayers, Sophie Fossa, Elizabeth Hahn, David Machin, Shu-Fang Hsu Schmitz, Bo Standaert, Patrick Therasse and Andrea Troxel.

I am also greatly indebted to my parents and my family who never failed to encourage me and to support me whenever they could.

Desmond Curran
June 2000.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Quality of Life Studies

The first publications on quality of quality of life (QL) of cancer patients in clinical trials appeared about two decades ago. Most of the early efforts, in the eighties, involved instrument development with the intention of producing psychometrically sound instruments for assessing QL. In 1986, the European Organization for Research and Treatment of Cancer (EORTC) initiated a research programme to develop an integrated, modular approach for evaluating the QL of patients participating in international clinical trials. This research resulted in the development of a core questionnaire which is referred to as the EORTC QLQ-C30 (Aaronson *et al* 1993). The QLQ-C30 incorporates nine multi-item scales: five functional scales (physical, role, cognitive, emotional and social); three symptom scales (fatigue, pain and nausea/vomiting); and a global health and QL scale. Six single-item scales are also included (dyspnoea, insomnia, appetite loss, constipation, diarrhoea and financial difficulties). The QLQ-C30 has been found to meet the requisite standards of validity (measuring what it is intended to measure), reliability (measuring with sufficient precision) and responsiveness (ability to detect changes) (Aaronson *et al* 1993). Many other questionnaires have also been developed for assessing QL in cancer clinical trials (e.g., Rotterdam Symptom Checklist (RSCL, de Haes *et al* 1990), Functional Assessment of Cancer Therapy (FACT, Cella *et al* 1993)).

Although all of these questionnaires were designed to produce psychometrically sound in-

struments for assessing QL they introduced other statistical dilemmas due to the nature of the data collected. QL instruments tend to be multidimensional and usually consist of a series of items (i.e., questions) with ordinal response categories. The items may be collapsed subsequently into a number of scales or domains, such as: physical, role, emotional and cognitive functioning. QL data tend to be longitudinal with the questionnaire administered at regular intervals during treatment and subsequent follow-up of patients in a trial. QL data differ from clinical data in several ways. In particular, clinical data may often be collected retrospectively, e.g., from the patient's medical charts. However, once a patient has missed a QL assessment the retrospective collection of the data is hampered by the recall abilities of the patient.

1.2 The Impact of Incompleteness

Difficulties with data collection and compliance appear to be the most important barriers to the successful implementation of QL assessments in clinical research. Kiebert *et al* (1998) discussed the various reasons for missing data in EORTC cancer clinical trials. The authors noted that although certain sources of missing data are unavoidable, such as attrition because of death and withdrawal from the study due to progressive disease or treatment-related toxicities, other sources of missing data can be minimized if procedures and infrastructure are in place. In QL research we encounter two main types of missing data: (1) item non-response (missing data in a questionnaire where a response has not been provided for a question); and (2) unit non-response (the whole questionnaire is missing for a patient). This latter category may be further subdivided into three categories: (a) intermittent missing forms, (b) dropout from the study and (c) late entry into the study (Curran *et al* 1998a). In particular, dropout may be problematic in QL studies since it is likely that patients with the poorest QL scores drop out earlier, especially in cancer clinical trials in patients with advanced disease (Hopwood *et al* 1994). Rubin (1976) described three missing data mechanisms: missing completely at random (MCAR: dropout is independent of observed and unobserved scores), missing at random (MAR: dropout is independent of unobserved scores but dependent on observed scores) and missing not at random (MNAR: dropout is dependent on at least one unobserved score). Section 2.5 provides a more formal definition of the various missing data mechanisms.

The effect of missing data is best seen by means of the following hypothetical example (see Figure 1.1). Suppose 20 patients are entered into a longitudinal randomized study

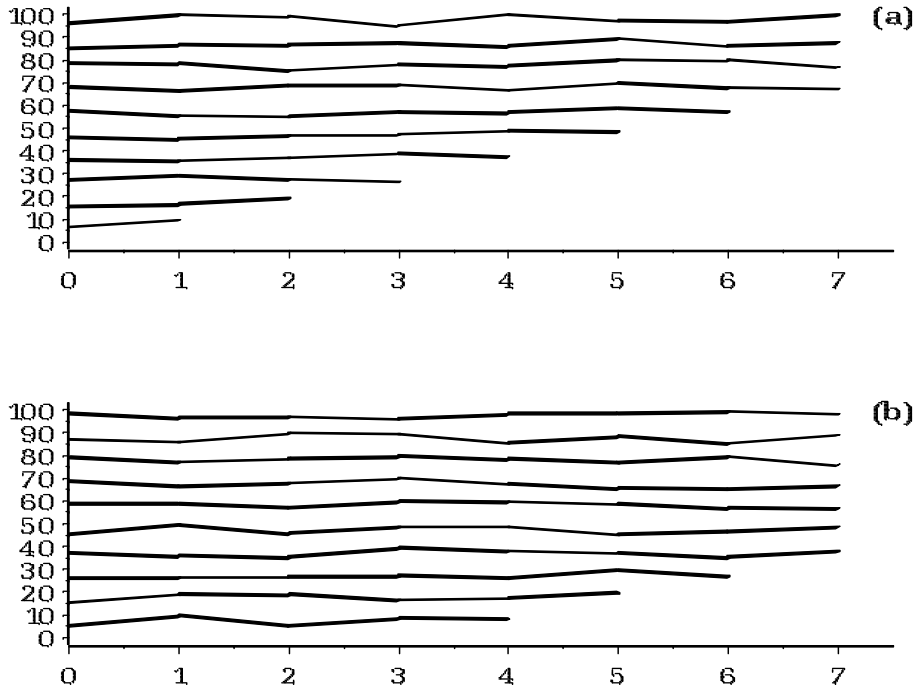

Figure 1.1: *Hypothethical Example I. Plots of individual patient profiles a) Drug A b) Drug B.*

comparing two drugs in patients with advanced cancer (10 patients receive drug A and 10 drug B). Suppose treatment does not influence QL but patients in treatment arm A tend to dropout earlier than patients in treatment arm B suggesting that drug B is more effective than drug A. Now suppose the dropout depends on the previously observed scores with patients with lower scores dropping out earlier in both treatment arms. Under the assumption of MCAR we could base our estimates of the mean score in both arms and the treatment effect on the complete cases. For example, based on a cross-sectional analysis at time point 7 the mean score in treatment arm A is 83.0 compared with 67.7 in treatment arm B suggesting that QL is better in treatment arm A. Thus, the treatment comparison is biased in favour of the inferior drug. However, under the assumption of MAR, the point estimates for the means at time point 7 are 53.9 and 54.0 in arms A and B, respectively. Thus, if the assumption of MAR is true, unbiased estimates may be obtained using appropriate models and corresponding software (e.g., PROC MIXED in SAS). Now suppose, as in Figure 1.2, that QL scores decrease after dropout, indicated by the dotted lines. Due to the nature of dropout these scores are not observed. This phenomenon may occur in clinical trials where

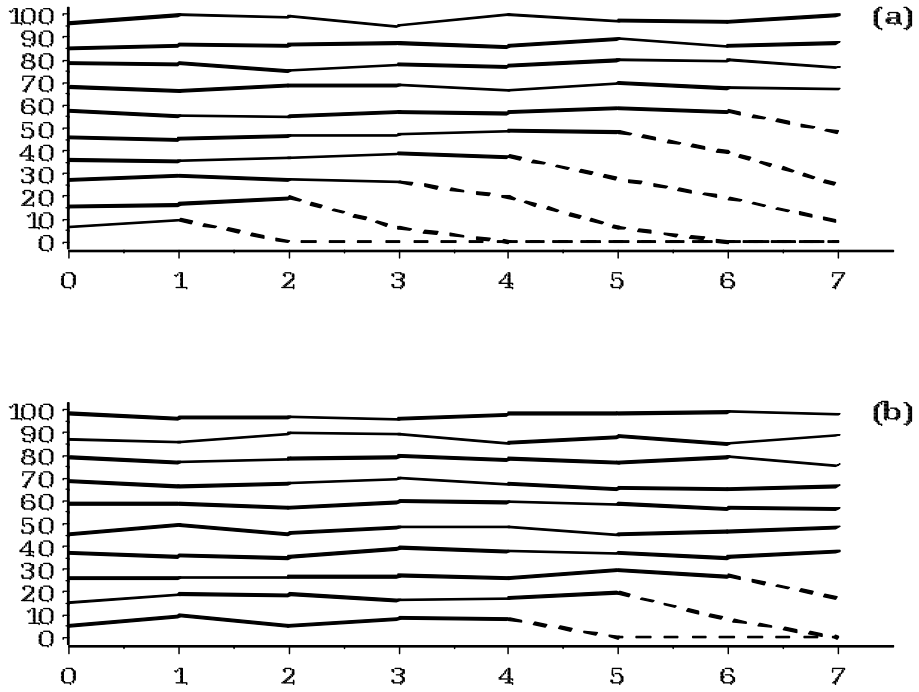

Figure 1.2: *Hypothethical Example II. Plots of individual patient profiles a) Drug A b) Drug B.*

QL scores decrease once the disease has progressed to such a level that patients begin to suffer from disease-related symptoms. The dropout mechanism is dependent on these unobserved scores (i.e., dropout is MNAR). If the scores had been observed the complete-data point estimates at time point 7 would be 41.4 and 49.1 in arms A and B, respectively suggesting that drug B results in a better QL score.

In addition, considering only patients who complete questionnaires using a cross-sectional analysis at time point 7, leads to too optimistic a view of both treatments under study with respect to the QL of patients (Olschewski *et al* 1995). The magnitude of potential bias in standard analyses depends strongly on the reasons why patients do not participate. Possible reasons include: administrative failure to distribute the questionnaire, the patient thought the questionnaire was a violation of privacy or that it was inconvenient (takes too much time), patient refusal or withdrawal, the patient felt too ill, or disease progression. At the time of analysis certain assumptions regarding the missing data have to be made. Such

assumptions can be made with more confidence if the reasons why patients did not complete the forms are known.

Not only do missing data lead to potentially biased results but there can be a severe loss of power if the proportion of missing data is high. In extreme cases this may mean that there is insufficient data to draw any useful conclusions from the study.

1.3 Simple *ad hoc* Methods for Dealing with Incomplete Data

Two simple, common approaches to analysis are (1) to discard subjects with incomplete sequences and (2) simple imputation. The first approach has the advantage of simplicity, although the wide availability of more sophisticated methods of analysis renders this approach inappropriate. It is also an inefficient use of information. It is not difficult to envisage situations where it can be very misleading, and examples of this exist in the literature (Wang-Clow *et al* 1995).

There are several forms of simple imputation. For example, a cross-sectional approach replaces a missing observation by the average of available observations at the same time from other subjects with the same covariates and treatment. A simple longitudinal approach carries the last available measurement from a subject forwards replacing the entire sequence of missing values. A more sophisticated version predicts the next missing value using a regression relationship established from available past data. These methods share the same drawbacks, although not all to the same degree. Under certain dropout mechanisms the process of imputation may recover the actual marginal behaviour required while under other mechanisms it may be wildly misleading, and it is only under the simplest and most ignorable mechanisms that the relationship between imputation procedure and assumption is easily deduced. Little (1994) gives two simple examples where the relationship is clear. A further minor point is that, without further elaboration, the analysis of the completed dataset will underestimate the true variability of the data.

In conclusion, we see that when there are missing values simple methods of analysis do not necessarily imply simple, or even accessible, assumptions and without understanding properly the assumptions being made in an analysis we are not in a position to judge its validity or value. It has been argued that while any particular *ad hoc* analysis may not

represent the true picture behind the data, a collection of such analyses should provide a reasonable envelope within which the truth might lie. This points to the desirability of a sensitivity analysis. However, without a clear formulation of the assumptions being made we are not in a position to interpret such an envelope, and are certainly not justified in assuming that its coverage is in some practical sense inclusive. One way to proceed is to consider a formal framework for the missing value problem, and this leads us to Rubin's classification.

1.4 Modeling Incompleteness

In order to incorporate incompleteness into the modeling process, we need to reflect on the nature of the missing value mechanism and its implications for statistical inference. As described in Section 1.2 Rubin (1976) and Little and Rubin (1987, Ch. 6) make important distinctions between different missing values processes: MCAR, MAR and MNAR. A formal definition of these concepts is given in Chapter 2. If a dropout process is random, then a valid analysis can be obtained through a likelihood-based analysis that ignores the dropout mechanism, provided the parameters describing the measurement process are functionally independent of the parameters describing the dropout process, the so-called parameter distinctness condition. This situation is termed ignorable by Rubin (1976) and Little and Rubin (1987). This leads to considerable simplification in the analysis.

Often, the reasons for dropout are many and varied and it is therefore difficult to justify on a priori grounds the assumption of random dropout. Arguably, in the presence of non-random dropout, a wholly satisfactory analysis of the data is not feasible.

One approach is to estimate from the available data the parameters of a model representing a non-random dropout mechanism. It may be difficult to justify the particular choice of dropout model, and it does not necessarily follow that the data contain information on the parameters of the particular model chosen, but where such information exists the fitted model may provide some insight into the nature of the dropout process and of the sensitivity of the analysis to assumptions about this process. This is the route taken by Diggle and Kenward (1994) in the context of continuous longitudinal data; see also Diggle, Liang and Zeger (1994, Ch. 11). Further approaches are proposed by Laird, Lange, and Stram (1987), Wu and Bailey (1988, 1989), Wu and Carroll (1988), and Greenlees, Reece, and Zieschang (1982). An overview of the different modeling approaches is given by Little (1995).

Also the case of categorical outcomes has received considerable attention. See for example Baker and Laird (1988), Stasny (1986), Baker, Rosenberger, and DerSimonian (1992), Conaway (1992, 1993), Park and Brown (1994), and Molenberghs, Kenward, and Lesaffre (1997).

One feature common to all of the more complex approaches is that they rely on untestable assumptions about the relation between the measurement process and the dropout process. One should therefore avoid missing data as much as possible, and if dropout occurs, information should be collected about the reasons for dropping out. Because different models imply different untestable assumptions that may affect the statistical inferences of interest, it is always advisable to perform a sensitivity analysis.

1.5 Overview

In Chapter 2 terminology is introduced to facilitate discussion of the subject of incomplete longitudinal QL data. The various missing data mechanisms as described by Rubin (1976) are explained. Concepts such as ignorability, separability and bias are also discussed.

Chapter 3 provides a review of the literature. Numerous methods for handling incomplete data are introduced starting with the 'classical' techniques, such as complete case and available case analyses and then progressing to imputation techniques. Both single and multiple imputation methods are described. This is followed by a discussion of likelihood based approaches to analysis of both continuous and categorical data.

Chapter 4 introduces the longitudinal sets of data which will be used throughout the thesis. Chapter 5 focuses upon the issues involved in handling questionnaires which contain one or more missing items and reviews several procedures for dealing with missing items: (1) case deletion (2) simple mean imputation and (3) general imputation methods. Simple mean imputation is the most widely used method of imputation of missing items since it is based on traditional psychometric approaches to scale design and analysis. Examples are provided where this method may not be appropriate and alternative imputation methods may be considered.

Chapter 6 presents various techniques which have appeared in the literature and which may be described globally as summary measures and summary statistics. These techniques are

illustrated using data from an EORTC clinical trial (EORTC 10921) in locally advanced breast cancer. For EORTC trial 10921 it is shown that by choosing different techniques different conclusions may be drawn concerning the QL outcome. This chapter illustrates the limitations of using these procedures, i.e., they are wasteful since they do not use all available information and they may provide biased results since they do not take into account the missing data or the process which creates missing data.

In Chapter 7 two methods of identifying the types of missingness in quality of life (QL) data in cancer clinical trials are explored. The first approach involves collecting information on why the QL questionnaires were not completed. Based on the reasons provided one may be able to distinguish the mechanisms causing missing data. The second approach is to model the missing data mechanism and perform hypothesis testing to determine the missing data processes. Two methods of testing if missing data are missing completely at random (MCAR) are presented and applied to incomplete longitudinal QL data obtained from international multicenter cancer clinical trials. The first method (Ridout 1991) is based on a logistic regression and the second method (Park and Davis 1993) is based on an adaptation of weighted least squares. In one application (advanced breast cancer) missing data was not likely to be MCAR. In the second application (adjuvant breast cancer) the missing data mechanism was and was not likely to be MCAR depending on the scale being studied ('hair loss' and 'anxiety' scales). MCAR and missing at random (MAR) have distinct consequences for data analysis. Therefore it is relevant to distinguish between them. Distinguishing between MAR and missing not at random (MNAR) is not trivial and relies on fundamentally untestable assumptions (Glynn, Laird and Rubin 1986).

Chapters 8 and 9 focus on continuous outcomes. Two alternative approaches to modeling longitudinal data with incomplete measurements have frequently been proposed in the literature, selection models (Diggle and Kenward 1994) and pattern-mixture models (Little 1993, Little 1995 and Hogan and Laird 1997). These modeling frameworks approach the issue of dropout in two distinct ways: in selection models the dropout probability is conditional on the measurement process, whereas in pattern-mixture models the measurement model is conditional on the dropout pattern. Selection models are often used as a sensitivity analysis tool to investigate the (treatment) effect under various assumptions about the dropout mechanism. When fitting selection models assumptions are made which are not fundamentally testable, e.g., the dependence of the dropout process on measurements which have not been obtained. In contrast, in pattern-mixture models the model for the missingness process is usually kept fairly simple, and can reduce to a multinomial distribution, describing the proportion of patients in the different patterns. Also, fitting a selection model may be computationally cumbersome. For pattern-mixture models, the only requirement is

that there are sufficient data in the various patterns to achieve reliable estimates. One then only needs relatively straightforward, non-iterative code to determine marginal quantities such as treatment effect. In Chapter 8 we compare pattern-mixture models with selection models using two datasets: the milk protein content trial described by Diggle and Kenward (1994) and a QL example from an EORTC clinical trial.

The natural parameters of selection models and pattern-mixture models have a different meaning, and transforming a probability model into one of the other frameworks is in general not straightforward, even for normal measurement models. When a selection model is used, as mentioned earlier, one has to make untestable assumptions about the relationship between dropout and missing data (discussion of Diggle and Kenward 1994, Molenberghs, Kenward, and Lesaffre 1997). In pattern-mixture models, it is explicit which parameters cannot be identified. Little (1993) suggests the use of identifying relationships between identifiable and non-identifiable parameters. Thus, even though these identifying relationships are also unverifiable (Little 1995), the advantage of pattern-mixture models is that the verifiable and unverifiable assumptions can easily be separated. In Chapter 9 we present a new strategy for fitting pattern mixture models which leads naturally into the field of sensitivity analysis. We explore the idea of extrapolating incomplete patterns using various identifying restrictions. This idea of extrapolating incomplete patterns was first suggested by Hogan (1999).

QL data not only involves repeated measures but is also usually collected on ordered categorical scales. In the recent statistical literature increasing attention is given to methods that can handle non-continuous outcomes in the presence of missing data. The aim of Chapter 10 is to investigate the effect on statistical conclusions of applying different modeling techniques to QL data generated from an EORTC phase III trial. Chapter 10 focuses on selection models. For information on the use of pattern-mixture models in categorical data we refer to Michiels, Molenberghs, and Lipsitz (1999). In Chapter 10 we first fit a random-effects model, relating a binary longitudinal response (derived from the physical functioning scale of the QLQ-C30) to several covariates. In a second approach, marginal models are fitted, retaining the response variable and the mean structure used before. The fitted marginal models differ with respect to the estimation procedure: generalized estimating equations (GEE), weighted generalized estimating equations (WGEE) and maximum likelihood (ML).

Chapter 2

Terminology

2.1 Introduction

In QL research two major types of missing data may be identified. Firstly, there can be missing items within a form, where a patient may have answered some of the questions but has failed to provide responses to other questions on the same form (Fayers *et al* 1998). We shall describe this situation as 'missing items'. Secondly, patients may fail to complete and return some of the questionnaires that were due during the study period (Curran *et al* 1998a). We will call this 'missing forms'. We will describe each of these situations separately.

2.2 Quality of Life Scales and Items

Most QL instruments consist of a number of questions or items, some or all of which are frequently combined to form scores for a number of scales or subscales. We will focus on the QLQ-C30 and the RSCL (de Haes *et al* 1990) as the basis for our examples. As mentioned in Section 1.1, the QLQ-C30 is a 30-item questionnaire consisting of 5 functional scales (physical, role, cognitive, emotional and social), 3 symptom scales (fatigue, pain, and nausea-and-vomiting), a global health status QL scale, a number of single items assessing additional symptoms commonly reported by cancer patients, and perceived financial impact

Table 2.1: *Emotional Functioning Scale. Adapted from the EORTC QLQ-C30 Scoring Manual (Fayers et al., 1999):In the QLQ-C30, emotional functioning (EF) is assessed by* 4 *items corresponding to questions* 21 *to* 24*, each on a* 4*−point scale.*

	Not at all	A little	Quite a bit	Very much
Did you feel tense?	1	2	3	4
Did you worry?	1	2	3	4
Did you feel irritable?	1	2	3	4
Did you feel depressed?	1	2	3	4

of the disease. Most items are 4-point scales graded 'not at all', 'a little', 'quite a bit' and 'very much'. The first section of the RSCL (de Haes *et al* 1990) comprises 30 items with 4-point scales, covering four domains of physical symptom distress and psychological distress; the second section is an eight-item activity level scale, and finally there is an overall life quality scale. The physical symptom distress scale is further broken down into four sub-scales representing fatigue, gastro-intestinal symptoms, pain related symptoms and chemotherapy related symptoms. In both the QLQ-C30 and the RSCL, as is common with QL instruments, the scales are scored by first summing their constituent items, followed by a simple linear transformation to produce a 0 - 100 standardised score. Such scales are often known as (standardised) Likert Summated Scales (McIver and Carmines 1981). An example of a summated scale from the EORTC QLQ-C30 is provided in Table 2.1.

The method of calculation for such a scale is as follows. Let us call the n questions contributing to a scale Q_i, where in this example n = 4 questions which are Q_{21}, Q_{22}, Q_{23}, Q_{24}. If no items are missing, then the 'Raw Score' is calculated as the average of the items:

$$\text{Raw Score} = \sum Q_i/n = (Q_{21} + Q_{22} + Q_{23} + Q_{24})/4. \tag{2.1}$$

In what follows we shall for brevity refer to standardised summated scales simply as 'scales', whilst the individual items will be called 'items'. Since different scales include items with a different *range* of values, they will have various minimum and maximum values; these items

are on 4-point scales and thus the 'Raw Score' lies between 1 and 4. *Range* is the difference between the maximum and minimum values (here, *range* = 4-1=3)

Therefore, a linear transformation is applied, to standardise the score to 0 to 100; also, this scale is reversed so that high scores indicate a high or healthy level of functioning.

$$\begin{aligned} \text{Standardised Score} &= \{1 - (\text{Raw Score} - 1)/\text{range}\} \times 100 \\ &= \{1 - (\text{Raw Score} - 1)/3\} \times 100. \end{aligned} \tag{2.2}$$

In the presence of missing items within a scale and provided that at least half of the items were completed, the EORTC recommends the following procedure: the scale score is calculated using the completed items which were present for that respondent. Suppose Q_{22} were missing. The 'Raw Score' becomes

$$\text{Raw Score} = \sum Q_i/n = (Q_{21} + Q_{23} + Q_{24})/3. \tag{2.3}$$

Equation 2.2 for transforming the 'Raw Score' to the final Score remains unchanged. Chapter 5 discusses several methods of handling missing items. The more complex situation of missing forms is the main focus of this thesis.

2.3 Missing Forms

In virtually all longitudinal studies the issues of unbalancedness and missing data arise. Some studies are designed such that the number of measurements (e.g., QL forms) per subject is variable or even random. The measurement times themselves can vary across subjects and can be random as well. We term these studies *unbalanced*. In such unbalanced studies it is usually not possible to identify missingness, unless measurement times have been recorded, even for occasions at which no measurement was actually taken. In contrast, in a *balanced* study the number of measurements per subject is fixed and the measurements are usually taken at an approximately common set of occasions. QL data are generally collected at fixed assessment time points and as such can be considered as *balanced*. In this situation, missing observations can be identified without ambiguity. The specific case of *dropout* (i.e., a subject

is completely observed until a certain point in time, after which no more measurements are taken) can be handled in the unbalanced case as well. The treatment of dropout in both balanced and unbalanced cases is very similar. In QL research we distinguish between three types of missing forms:

a. intermittent missing forms,

b. dropout from the study,

c. late entry into the study.

Intermittent missing forms occur when a patient misses an assessment but is later observed. Dropout (or monotone missingness) occurs when a patient, once missing an assessment, is never observed again. Late entry into the study occurs when a complete set of forms is not yet available due to a patient recently being registered in the study, but additional forms are expected in the future.

The main concern when analyzing incomplete QL data is that of bias. Consider the case of a randomized clinical trial where we wish to estimate the overall QL score of patients at one time point on treatment A. Suppose the proportion of patients who respond (return a completed form) is P_r^A and so the proportion of patients who do not respond is P_{nr}^A. We assume that the responders (r) and the non responders (nr) may have different means. Let μ_r^A be the mean score of responders and μ_{nr}^A be the mean score of non-responders in treatment A if they had responded. The mean QL score for all patients in treatment A is

$$\mu^A = P_r^A \mu_r^A + P_{nr}^A \mu_{nr}^A.$$

However, as our sample only contains information on responders, the mean of the observed scores is μ_r^A and so the bias in treatment arm A by using only the responders is

$$\begin{aligned} \text{Bias}_A &= \mu^A - \mu_r^A \\ &= P_r^A \mu_r^A + P_{nr}^A \mu_{nr}^A - \mu_r^A \\ &= P_{nr}^A (\mu_{nr}^A - \mu_r^A). \end{aligned}$$

Thus the bias is proportional to the difference in mean QL score between responders and non-responders and to the proportion of non-responders. This bias is not reduced by increasing

the sample size. In a two arm trial there will be a similar expression for the bias in treatment arm B.

In a clinical trial, the objective is usually to investigate the difference in QL between treatment arms, i.e., $\delta = \mu^A - \mu^B$. However, the observed difference is $\delta_r = \mu_r^A - \mu_r^B$ and so the bias in the treatment (T) difference caused by the missing forms is:

$$\begin{aligned} \text{Bias}_T &= \delta - \delta_r \\ &= (\mu^A - \mu^B) - (\mu_r^A - \mu_r^B) \\ &= \text{Bias}_A - \text{Bias}_B \end{aligned}$$

Therefore, the bias in a treatment comparison is equal to the difference in the bias observed in each treatment arm. A treatment comparison is therefore unbiased if the bias is the same in both treatment arms. Although one may calculate the proportion of responders in both treatment arms, it is generally not possible to calculate the difference in mean QL score between responders and non-responders. Identifying the dropout mechanism and incorporating it in the model may produce less biased results. We will introduce some formal notation and terminology to facilitate and streamline the treatment of the subject of incomplete longitudinal QL data.

2.4 Notation

In this section we build on the standard framework for missing data, which is largely due to Rubin (1976) and Little and Rubin (1987).

Let $i = (1, \ldots, n)$ index patients. In most clinical trials, a fixed number of repeated QL assessments is planned at fixed times: let these times be denoted by $(t_1, \ldots, t_T)$. The QL score for patient i at time j is denoted by Y_{ij}; patient i has T possible measurements $(Y_{i1}, \ldots, Y_{iT})$. Since some of the data are missing, it is useful to assign a series of indicators $(R_{i1}, \ldots, R_{iT})$, where

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \textit{ observed}, \\ 0 & \text{otherwise.} \end{cases}$$

The *missing data indicators* R_{ij} are grouped into a vector $\boldsymbol{R}_i$ which is, of course, of the same

length as $\boldsymbol{Y}_i$.

Let $\boldsymbol{Y}_i^o$ and $\boldsymbol{Y}_i^m$ denote the observed and missing values of $(Y_{i1}, \ldots, Y_{iT})$, respectively. Matrices X_i and Z_i are design matrices for fixed and random effects, respectively. X_i will generally include the times of measurement $(t_1, \ldots, t_T)$ as well as treatment indicators and other fixed covariates such as age, sex, and performance status.

The following terminology is adopted:

Complete data $\boldsymbol{Y}_i$: the scheduled measurements. This is the outcome vector that would have been recorded if there were no missing data.

Missing data process $\boldsymbol{R}_i$. The process generating $\boldsymbol{R}_i$ is referred to as the missing data process.

Full data $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$: the complete data, together with the missing data indicators. Note that, unless all components of $\boldsymbol{R}_i$ equal 1, the full data components are never jointly observed.

Observed data $\boldsymbol{Y}_i^o$.

Missing data $\boldsymbol{Y}_i^m$.

Some confusion might arise between the terms *complete data* introduced here and *complete case analysis*. While the former refers to the (hypothetical) data set that would arise if there were no missing data, 'complete cases' refers to deletion of all subjects for which at least one component is missing.

Note that one observes the measurements $\boldsymbol{Y}_i^o$ together with the dropout indicators $\boldsymbol{R}_i$.

Statistical modeling begins by considering the full data density

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

where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are vectors that parameterize the joint distribution. We will use $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ to describe the measurement and missingness processes, respectively.

In considering the full data density two alternative factorizations can be used to facilitate modeling. Conditioning on $\boldsymbol{R}_i$ results in a pattern-mixture model, while conditioning on $\boldsymbol{Y}_i$ results in a selection model; both are discussed by Little (1995).

Pattern-mixture models are written as

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, Z_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}), \tag{2.4}$$

That is, the model for the responses depends on the particular missingness pattern, and the overall distribution of the longitudinal measurements is a mixture of the conditional distributions. Parameters describing the model for $\boldsymbol{Y}_i$ are estimated in each stratum (determined by $\boldsymbol{R}_i$), and the overall parameters are a weighted average of these estimates, weighted by the proportion of subjects in each stratum. The parameters describing the distribution of the missingness indicators themselves are generally considered a nuisance, and usually a form for $f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi})$ is not specified. This is in some sense an advantage, since model checking for a specific missing data mechanism is difficult. However, a drawback arises because the conditional distributions for $f(\boldsymbol{y}_i | X_i, Z_i, \boldsymbol{\theta})$ are not fully identifiable, with the exception of the pattern where no observations are missing. This problem can be reduced by combining some of the missing data patterns to create a smaller number of strata. For example, patients might be stratified into groups based on the length of their follow-up; an example of this type of pattern-mixture model is given by Fairclough, Peterson and Chang (1998a). When this coarser grouping is used, standard software can often be used to fit the stratified models.

Selection models are written as

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, Z_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, X_i, \boldsymbol{\psi}), \tag{2.5}$$

An underlying distribution is assumed for the complete longitudinal measurements, and the missingness mechanism is modeled as a function of those measurements. This is a joint distribution for the complete data; as with the pattern-mixture model, there may be problems of identifiability. To avoid this, a likelihood for the *observed* data is obtained by starting with the complete data likelihood and integrating out the unobserved responses, according to the specified underlying distribution.

Rubin (1976) defined a taxonomy to describe the various missing data mechanisms using the following terminology based on selection models. The missingness probabilities are denoted by $\pi_{ij} = \mathrm{P}(r_{ij} = 1 | \boldsymbol{y}_i, X_i, \boldsymbol{\psi})$, modeled as a function of covariates X_i, which are related to the missingness probabilities through the parameters $\boldsymbol{\psi}$. Note that in some cases the 'covariates' may also include current or previous values of the response variable $\boldsymbol{Y}_i$. The density $f(r_{ij} | \boldsymbol{y}_i, X_i, \boldsymbol{\psi})$ is that of a Bernoulli random variable with probability π_{ij}.

2.5 Missing Data Mechanisms

Missing data is often described as either 'dropout' or 'intermittent'. The mechanism of missing data also varies: Rubin (1976) and Little and Rubin (1987) make distinctions among different missing value processes (completely random, random, or not at random) as described informally in the introduction.

2.5.1 Missing Completely at Random (MCAR)

An observation is said to be missing completely at random if the missingness probability is independent of all previous, current, and future assessments. Thus the distribution of the missing data mechanism reduces to $f(\boldsymbol{r}_i|\boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|X_i, \boldsymbol{\psi})$. Notice, however, that the missingness mechanism may depend on the values of fixed covariates. In particular, if the covariate matrix includes treatment as a variable then dropout rates may vary by treatment. Covariate dependent dropout is illustrated by Fairclough, Peterson and Chang (1998a) using initial performance status and survival at 6 months for patients with advanced non-small cell lung cancer. Curran *et al* (1998b) provide examples where the missing data were, and were not, likely to be MCAR. In their first example, in an advanced disease setting, the probability of missingness was dependent on the previous QL score; in the second example, in an adjuvant treatment setting, different QL scales yielded different results: more anxiety was observed in patients with incomplete data, but there was no sgnificant difference in burden of hair loss scores between completers and non-completers.

2.5.2 Missing at Random (MAR)

Data are missing at random if the missingness probability does not depend on the missing values $\boldsymbol{Y}_i^m$ but depend on the observed measurements $\boldsymbol{Y}_i^o$, i.e., $f(\boldsymbol{r}_i|\boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, X_i, \boldsymbol{\psi})$. Thus, as with MCAR, when data are MAR the missing data mechanism may be modeled using only the available data and any inference on $\boldsymbol{Y}$ may be based solely on the observed data. An example in which the missing data depend on $\boldsymbol{Y}^o$ is given by Fairclough, Peterson and Chang (1998a). The authors show that the medians of the observed scores on the Perceived Adjustment to Chronic Illness Scale were higher for patients who completed more assessments, suggesting that the probability of a missing assessment was dependent on the

observed values of the outcome.

2.5.3 Missing Not at Random (MNAR)

Finally, when the missingness probability depends on the missing values $\boldsymbol{Y}_i^m$, the process is referred to as *missing not at random*. A MNAR process is also allowed to depend on $\boldsymbol{Y}_i^o$, i.e. $f(\boldsymbol{r}_i|\boldsymbol{y}_i, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, X_i, \boldsymbol{\psi})$. A MNAR missing data mechanism often seems plausible in QL studies: subjects who have worse QL, due to increased toxicity, disease progression, may be more likely to miss assessments. If information concerning the reasons for missingness is collected and included in the missing data model, it may be possible to reduce the mechanism from MNAR to MAR or even MCAR.

2.6 Ignorability

Rubin (1976) addressed the issue of what assumptions are necessary to justify ignoring the missing data mechanism. He established that the extent of ignorability depends on the inferential framework. The full data likelihood contribution for subject i assumes the form

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

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

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

with

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

Under an MAR process, we obtain

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= f(\boldsymbol{y}_i^o|X_i, Z_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, X_i, \boldsymbol{\psi}), \end{aligned} \tag{2.6}$$

i.e., the likelihood factorizes into two components of the same functional form as the general factorization of the complete data. If further $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint in the sense that the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is the product of the individual parameter spaces then inference can be based on the marginal observed data density only. This technical requirement is referred to as the separability condition. In conclusion, when the separability condition is satisfied, *within the likelihood framework*, ignorability is equivalent to the union of MAR and MCAR. Hence, non-ignorability and 'non-randomness' are synonyms in this context. A formal derivation is given in Rubin (1976), where it is also shown that the same requirements hold for Bayesian inference, but that frequentist inference is ignorable only under MCAR. Even when likelihood or Bayesian inference is applied it may be necessary to distinguish between MCAR and MAR depending on the research questions. For example, if $\boldsymbol{Y}$ follows a multivariate Gaussian distribution, then under MCAR the mean structure of $\boldsymbol{Y}$ coincides with the conditional mean structure of $\boldsymbol{Y}$ given no missing data, but this is not so under MAR, except in the generally unrealistic case of uncorrelated $\boldsymbol{Y}_j$. Thus, if the research question involves determining the conditional mean structure of $\boldsymbol{Y}$ given no drop-out, it is necessary to distinguish between MCAR and MAR.

Classical examples of the more stringent condition with frequentist methods are ordinary least squares and the generalized estimating equations approach of Liang and Zeger (1986). These GEE define an unbiased estimator only under MCAR. Robins, Rotnizky, and Zhao (1995) have established that some progress can be made under MAR and even under informative processes. Their method is based on including weights that depend on the missingness probability, proving the point that at least some information on the missingness mechanism should be included and thus that ignorability does not hold.

To determine which methods of statistical analysis will be appropriate, one should initially distinguish the pattern of missing data and identify the mechanism that generates missing data.

Chapter 3

Literature Review

3.1 Introduction

From Chapter 1 it is clear that the issue of incomplete QL data is an important one and needs careful attention. This attention should be focused on two goals: (1) improving compliance and hence reducing the proportion of missing questionnaires and (2) using appropriate statistical techniques to deal with incomplete QL data. However, the QL literature contains very few methodological papers handling the problem of missing data. Also, very few published clinical trial results indicate clearly the extent of missing forms or how they are handled. Zwinderman (1992) proposed a number of solutions to the problem of missing forms and provided some computer programming code in BMDP and SAS for modeling the data using repeated measures analysis of variance, assuming that missing data are missing at random. Zee and Pater (1991) suggested that one method of analyzing incomplete quality of life data is to use a growth curve model, in conjunction with the EM (Expectation-Maximization) algorithm. Beacon and Thompson (1996) illustrated the potential of multi-level modeling in analysing QL data.

A vast amount of work on missing data has been carried out in other research fields such as agriculture, (Diggle and Kenward 1994) clinical trials (Molenberghs and Lessafre 1994, Lessafre, Molenberghs and Dewulf 1996) and survey sampling (Madow, Nissleson and Olkin 1983, Rubin 1987). A lot of this work focused on longitudinal settings. An earlier, very important review was provided by Laird and Ware (1982). Little (1995) provides a systematic

account, also covering the more recent work. Both review articles treat methods for MCAR, MAR, and MNAR.

3.2 Quick Methods

The easiest, but least desirable, approach to a missing data situation is to remove the patients with missing forms from the analysis (Little and Rubin 1987). In QL studies, especially in advanced disease this generally means deleting an unacceptable amount of information. It does however mean that standard complete case methods of analysis can be used (e.g., MANOVA). When forms are MCAR, the reduced data represent a randomly drawn subsample of the original dataset and thus inferences about the values of the population parameters are consistent (Little and Rubin 1987). However, in QL research in cancer clinical trials patients who are in a generally good condition would be expected to have more complete follow-up. Therefore this method has two distinct disadvantages: (1) it reduces the sample size and (2) it may produce biased results if missing data are not MCAR.

An alternative to complete case analysis is to use all available forms. For example, in a clinical trial we may wish to compare two treatments with respect to QL at individual time points. A possible available case analysis would calculate a treatment difference (and standard error) at every time point separately. Wei and Johnson (1985) proposed a test which allows the per time point test statistics to be combined into one overall test statistic. This method still requires the missing data mechanism to be MCAR to yield unbiased estimates. An available case likelihood analysis can be conducted, whereby every subject contributes its available (observed) measurements. For example, such an analysis can be conducted with the SAS procedure MIXED, where subjects with some missing measurements are also included in the analysis.

A widely used method for analysis of data collected serially over time is to reduce the data on each patient to a single summary (Tannock *et al* 1996, Fairclough and Gelber 1996) that reflects some important aspect of the response (e.g., mean, median, min or max). For example, in clinical trials, data on toxicity is usually summarized by taking the worst value recorded during the treatment period. Summary measures are valid only under a MCAR mechanism. Even when data are MCAR, a biased estimate of the treatment effect may be obtained if the number of completed forms is not equivalent in both treatment arms.

Some of these simple methods have been subject to heavy criticism in the scientific literature (Little and Rubin 1987, Curran *et al* 2000a) but are still commonly used in many areas of applied statistical research. For example, many clinical trial reports include a complete case analysis and a last observation carried forward (LOCF) analysis. The limitations of these methods are discussed further in Chapter 6.

3.3 Imputation

Whereas a complete case analysis removes the problem of incomplete sequences by removing them, imputation strategies achieve the same goal by filling in values for the unknown measurements. In survey sampling a large literature has developed on imputing missing items (Madow, Nissleson and Olkin 1983, Pregibon 1977). Methods of imputation include last observation carried forward (LOCF), mean imputation, hot deck imputation, cold deck imputation, and regression imputation (Rubin 1987, Little and Rubin 1987)

3.3.1 Single Imputation

Last observation value carried forward (LOCF) replaces the missing value with the last observation, thereby assuming a constant score over time. Very strong and often unrealistic assumptions have to be made to ensure validity of this method. First, one has to believe that a subject's measurement stays at the same level from the moment of drop out onwards (or during the period they are unobserved in the case of intermittent missingness). In a cancer clinical trial setting, one might believe that the patient's QL scores *decrease* after dropout.

The LOCF depends uniquely on the measurements of the individual for which an imputation has to be generated. As such, we could term it *horizontal* imputation. Next, we will discuss a strategy where the imputed value depends uniquely on the other subjects, i.e., a *vertical* strategy. Simple (or unconditional) mean imputation generally refers to substitution of the mean scores of a group of patients with observed data for the score of patients with unobserved data. A result of simple mean imputation is that the estimate of the variance will be artificially reduced. A more promising form of imputation is to substitute means that are conditioned on other variables or previously observed scores. This method has the advantage that it combines both 'horizontal' and 'vertical' information in creating the

imputations. The method was initially proposed by Buck (1960). He showed that under mild regularity conditions the method is valid for MCAR mechanisms. Little and Rubin (1987) added that the method is valid under certain types of MAR mechanisms. Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant across missingness patterns.

Hot deck imputation refers to selecting at random a score from patients with available data and substituting it for the patient with missing information. The hot deck literally refers to the deck of responses of patients with available data from which we may select a score. Hot deck imputation may involve very elaborate schemes for selecting responses for substitution. For example, one might select a response from only those patients with matching patient covariates (e.g., treatment, sex, age group, performance status, previous QL scores).

A special case of the hot deck procedure is given by 'nearest neighbour hot deck', where a distance measure is defined between patients. For example, a regression analysis is performed to identify which factors are associated with QL. Based on the parameter estimates weights are placed on the values of each covariate to obtain a distance score for each patient. If a patient has a missing QL score then the QL score from the 'nearest' patient with available data is taken. Curran *et al* (1998c) illustrated how both the hot deck and 'nearest neighbour hot deck' procedures could be used in a trial in prostate cancer.

Cold deck imputation refers to replacing a missing value by a constant value from an external source, such as a value from a previous study. A more detailed desciption of these imputation processes may be found elsewhere (Rubin 1987, Little and Rubin 1987).

As with complete case analysis, a major advantage of imputation is that, once the values have been filled in, standard complete data methods of analysis can be used. In contrast, other approaches to missing data require new and specialized computer programs. Of great importance also is that acceptance and understanding of statistical conclusions may be lost if sophisticated mathematical techniques are used for analysis. The choice and the design of an imputation technique allows the user's prior knowledge and experience to be incorporated in the imputation process. Consequently, theoretically sound methods of imputation may be advantageous as the imputation method may be easily applied and at the same time yield easily comprehensible conclusions.

Some problems do exist using single imputation. Dempster and Rubin (1983) write 'The idea of imputation is both seductive and dangerous. It is seductive because it can lull the

user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.' Several pitfalls of imputation techniques in a longitudinal context are discussed in Verbeke and Molenberghs (1997). In addition, by treating the imputed values as actual information, the estimated standard errors will generally be too small. This problem can be overcome using multiple imputation.

3.3.2 Multiple Imputation

Multiple imputation was formally introduced by Rubin (1978). Rubin (1987) provides a comprehensive treatment. Several other sources, such as Rubin and Schenker (1986), Little and Rubin (1987), Tanner and Wong (1987), and Schafer's (1997) book give excellent and easy-to-read descriptions of the technique. Efron (1994) discusses connections between multiple imputation and the bootstrap. An important review, containing an extensive list of references and a large bibliography, is given in Rubin (1996). The idea of multiple imputation (Rubin 1987) is that several values, m say, are imputed instead of just one. As a result m datasets are created. In current imputation practice m is often small (e.g., $m = 5$, Meng 1994). Conducting a multiple imputation analysis requires repeating the same standard complete data analysis several times, e.g., calculating the summary statistic and variance for each of the m imputed datasets. The m separate analyses may then be combined into one inference using the rules given by Rubin (1987). With the rapid development of computer technology this has become relatively straightforward. Thus, multiple imputation retains the ability to perform complete-data analysis as in single imputation whilst at the same time reflecting the uncertainty of the imputed value. Multiple imputation also has the advantage that the accuracy of the standard errors is improved. By imputing several values for a single missing component, this uncertainty is explicitly acknowledged.

Rubin (1987) points to another very useful application of multiple imputation. Rather than merely accounting for *sampling uncertainty*, the method can be used to incorporate *model uncertainty*. Indeed, when a measurement is missing but the researcher has a good idea about the probabilistic measurement and missingness mechanisms, then constructing the appropriate distribution from which to draw imputations is, at least in principle, relatively straightforward. In practice there may be considerable uncertainty about some parts of the joint model. In that case, several mechanisms for drawing imputations might seem equally plausible. They can be combined in a single multiple imputation analysis. As such, multiple

imputation can be used as a tool for sensitivity analysis. Multiple imputation is described in more detail in Chapter 9.

3.4 Likelihood Based Methods

Longitudinal analysis of repeated measurements with incomplete data is a relatively new research area. In 1976 Rubin's revolutionary paper introduced a taxonomy to describe the various processes which generate missing data. In the subsequent twenty years this methodology has flourished and many mathematical and statistical techniques have been developed in the pursuit of correct and efficient methods of handling incomplete data. These methods vary widely in their ease of implementation, their robustness to modeling assumptions, and their ability to handle different kinds of missingness patterns and missingness mechanisms. Recently a few books have emerged covering these developments (Diggle, Liang and Zeger 1994, Verbeke and Molenberghs 1997). Most of the methods focus on multivariate normal data with less attention given to categorical outcomes.

3.4.1 Continuous Outcomes

As mentioned in Chapter 2, the extent of ignorability depends on the inferential framework. For likelihood or Bayesian inference when missing data are MCAR or MAR, it follows that an ignorable analysis is valid (Rubin 1976). This means that the likelihood contribution of a given subject is proportional to the density (or probability mass) associated with its set of observed measurements. Such a feature is particularly appealing since it avoids explicit modeling of the non-response mechanism, as well as imputing values for the missing measurements. The only requirement is that the actual implementation used to maximize the observed data (log-)likelihood is capable of handling observed sets of repeated measures of varying length. When assessments are missing, there are usually different sample sizes at different assessment times. Some software packages do not allow measurement sequences of unequal length; notable exceptions include SAS PROC MIXED, BMDP-5V, MLWiN and SPlus which allow mixed models with missing data to be fitted. Intermittent missing data patterns are allowed. The mixed model contains fixed and random effects and usually takes the following form:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\alpha} + Z_i\boldsymbol{a}_i + \boldsymbol{w}_i(t) + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

where X_i and Z_i are design matrices for fixed and random effects, respectively, $\boldsymbol{\alpha}$ are fixed effects and $\boldsymbol{a}_i$ are random-effects parameters with $\boldsymbol{a}_i \sim N(0, D)$. Further, $\boldsymbol{w}_i$ are realizations of a Gaussian stochastic process and $\boldsymbol{\varepsilon}_i$ is a vector of normally distributed error terms. Various covariance structures for the association of repeated measures can be assumed; autoregressive, compound symmetry, or unstructured forms are common choices.

Missing data in the context of the linear mixed model is extensively treated in Verbeke and Molenberghs (1997). Several extensions to standard mixed models have been proposed in the context of longitudinal measurements in clinical trials. Zee (1998) proposed growth curve models where the parameters relating to the polynomial in time are allowed to differ according to the various health states experienced by the patient (e.g., on treatment, off treatment, post-relapse, etc.).

While less attention has been devoted to MNAR, there has been a growing literature on the subject, in particular when missingness is confined to dropout. For Gaussian data, the landmark paper of Diggle and Kenward (1994) as well as its discussion deserves attention. The procedure consists of specifying a linear mixed model for the measurements, together with a logistic regression to describe the dropout process. It has been implemented in the suite of SPlus functions termed OSWALD (Smith, Robertson and Diggle 1996). Random effects based approaches are discussed in Wu and Bailey (1989) and Wu and Carroll (1988).

Schluchter (1992) proposed a joint mixed effects model for the longitudinal assessments and the time to dropout. Suppose the time to dropout, or censoring, is denoted by T_i. The joint model allows the T_i (or a function of the T_i, in this case the log) to be correlated with the random effects $\boldsymbol{a}_i$. The model is as follows:

$$\begin{bmatrix} \boldsymbol{a}_i \\ \log(T_i) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu_t \end{bmatrix}, \begin{bmatrix} B & \sigma_{bt} \\ \sigma'_{bt} & \tau^2 \end{bmatrix} \right).$$

For example, patients with steeper rates of decline in measurements over time (as measured by the random effects $\boldsymbol{a}_i$) may be more likely to fail early. This model allows MNAR data in the sense that the time of dropout is allowed to depend, through the covariance parameter σ_{bt}, on the rate of change in the underlying measurements. Intermittent missing data are assumed to be MCAR.

3.4.2 Categorical Outcomes

Less research is devoted to repeated categorical data (binary or ordinal data). Standard log-linear models are particularly easy to fit but are less useful since they are not reproducible or upward compatible (Liang, Zeger and Qadish 1992). A model for a vector of repeated measures Y is said to be reproducible if the corresponding model for a subvector of Y is described by a subset of the parameter vector of which the elements retain their meaning. This is not true for the log-linear model since, for example, the main effects are interpreted as conditional logits. Thus, passing to a subvector changes the subset on which one conditions and hence changes the meaning of the parameters. Therefore, so-called marginal models are promising in this respect.

Generalized estimating equations (GEE) (Liang and Zeger 1986, Zeger, Liang and Albert 1988) appear to be of interest. The method is attractive because, rather than having to make full distributional assumptions, the researcher can suffice with specifying the marginal expectation of the repeated measures (as in a cross-sectional study, using a generalized linear model). The correlation structure is accounted for by working assumptions. This procedure yields consistent and asymptotically normal point estimates. When the so-called robust or sandwich estimator is used for the asymptotic covariance matrix, valid standard errors are also obtained. However, this method is frequentist in nature. This implies that the technique can only be used with incomplete data if the missingness process is MCAR. Proposals have been made to overcome this problem by appropriately weighting the terms in the estimating equations (Robins, Rotnitzky and Zhao 1995). This adaptation of GEE is valid under MAR and even with MNAR.

Molenberghs, Kenward and Lesaffre (1997) developed a method for ordinal longitudinal data. They coupled a model for repeated categorical data with a logistic regression for the dropout process and maximized the resulting likelihood by means of the EM algorithm. The multivariate Dale model (Lesaffre and Molenberghs 1994), also called the multivariate odds ratio model, was used to model the repeated measurements. These authors show that their model can be used with MCAR and MAR processes. An extension to the informative case is given in Molenberghs, Kenward, and Lesaffre (1997). These authors used the EM algorithm to estimate the model parameters, which include the measurement parameters (marginal logits, odds ratios to describe the association, and logistic regression parameters to describe the dropout probability, given both observed and missing measurements). This method was compared to ordinary and weighted estimating equations (Robins, Rotnitzky and Zhao 1995) in Fitzmaurice, Laird and Lipsitz (1994). Other marginal models for categorical data are

given in Lang and Agresti (1994) and Glonek and McCullagh (1995).

3.4.3 Remarks

Although, much has been achieved in the last few decades there is still considerable potential for research in missing data methods in QL settings. Many methods have been developed in the general statistical literature but few of these techniques have been applied to QL examples. This thesis endeavours to collate and contrast existing methods of handling missing data for both continuous and categorical outcomes. We investigate the assumptions that are required to ensure validity of results. In addition, we explore further the use of pattern mixture models as an alternative to selection models in the continuous data setting.

Chapter 4

Datasets

This chapter introduces the longitudinal sets of data which will be used throughout this thesis. EORTC trial 10921, a study of locally advanced breast cancer is presented in Section 4.1. SIAK (Schweizerisches Institut fuer Angewandte Krebsforschung) study 20/90, comparing the effectiveness and toxicity of 4-OH-Androstenedione with Megestrol acetate as second line hormonal treatment in advanced breast cancer patients is described in Section 4.2. Section 4.3 is devoted to IBCSG study VI-14 which was designed to investigate adjuvant chemotherapy and/or endocrine treatment in patients with operable breast cancer. The Milk Protein Content Trial, a study frequently used in analysis of longitudinal data with dropout, is presented in Section 4.4. Sections 4.5 and 4.6 introduce two prostate studies from the EORTC Genito-Urinary Tract Cancer Co-operative Group.

4.1 EORTC 10921: Locally Advanced Breast Cancer Trial

EORTC trial 10921 (Therasse *et al* 1998) is an international, intergroup dose-intensified study. This randomized phase III study was designed to compare six four-weekly cycles of CEF (cyclophosphamide, epirubicin and 5FU) versus six two-weekly cycles of dose-intensified EC (epirubicin and cyclophosphamide) + G-CSF (growth colony stimulating factor) in patients with locally advanced/inflammatory breast cancer. The expected duration of standard

treatment was 24 weeks compared with 12 weeks in the intensified treatment arm. Between June, 1993 and April, 1996, 448 patients were entered in the trial with 224 patients being randomized into the CEF arm and 224 into the EC+G-CSF arm. The main endpoint of the trial was progression-free-survival. To date, duration of survival and progression-free survival are not significantly different between the two treatment arms.

QL was considered to be a mandatory part of the protocol. The QL assessment consisted of 2 generic questionnaires, the EuroQoL (The EuroQoL Group 1990) and the RAND MOS (Ware and Sherbourne 1992), a cancer-specific questionnaire, the EORTC QLQ-C30 (Aaronson *et al* 1993), and a study-specific breast module. The global health status/QL scale from the EORTC QLQ-C30 was specified as the QL domain on which the main analysis would be performed. This scale was constructed using the scoring procedures for the EORTC Core Quality of Life Questionnaire EORTC QLQ-C30 version 1.0 (Fayers *et al* 1999), i.e., the scale score was calculated by averaging items within the scale and transforming the average score linearly to a 0 to 100 scale, with higher scores representing a better global health status/QL.

The planned schedule of assessment in both treatment arms was as follows: at randomization, every month for the first three months, every three months for the first year, at 18 months and every eight months thereafter until disease progression. The current analysis concentrates on the QL assessments during the first year. A window (time frame) for acceptance of questionnaires was defined for each assessment. To allow for some delay in the schedule of treatment (and hence QL assessment) more time was allowed after than before the scheduled assessment point (e.g., at week 4 questionnaires were accepted if they were completed within one week before and up to 2 weeks after the scheduled assessment). This reduces the possibility of patients being omitted from the analysis because of delayed chemotherapy cycles.

Of the 448 patients included in the trial 11 patients were considered ineligible (8 due to disease stage, 1 due to prior treatment for breast cancer and 2 due to previous or concurrent malignancy). For 2 patients eligibility was not verifiable due to inadequate source documentation. Seventeen patients from two institutions were excluded from the QL study, as officially translated EORTC QLQ-C30 questionnaires were not available for these languages during the study. A further 15 patients were excluded because they were judged unfit to complete the QL questionnaires (9 in the CEF arm and 6 in the EC+G-CSF arm). A total of 403 patients were eligible for the QL analysis, 199 in the CEF arm and 204 in the EC+G-CSF arm.

The compliance with the QL assessments during the first 12 months of the study is presented

Table 4.1: *EORTC Trial 10921. Compliance with QL assessment by treatment arm.*

CEF							
Months	0	1	2	3	6	9	12
Expected	199	199	199	195	179	167	155
Received	169	157	148	156	124	114	103
%	85	79	74	80	69	68	66

EC+G-CSF							
Months	0	1	2	3	6	9	12
Expected	204	204	202	199	194	185	170
Received	173	169	158	141	133	127	104
%	85	83	78	71	69	69	61

in Table 4.1. Some patients completed more than one valid QL questionnaire within a given time window. For these patients the questionnaire which was completed first during this period was retained. The main reason for patients going off study was progression of disease. Table 4.2 presents the patterns of completed questionnaires. Ninety-three patients completed QL questionnaires at all seven assessment time points. Monotone dropout patterns (i.e., a complete series of questionnaires before dropout) were observed in 189 cases (this includes the latter 93 patients). Intermittent missing questionnaires was also a problem with 115 patients having exactly 1 missing questionnaire and the remaining 94 patients having more than 1 missing questionnaire in a series before dropout. Five patients did not complete any questionnaires during this period.

4.2 SIAK 20/90: Postmenopausal Advanced Breast Cancer Study

One hundred and seventy-seven postmenopausal advanced breast cancer patients were accrued into a multicenter randomized phase III trial (SIAK 20/90), which aimed at comparing

Table 4.2: *EORTC Trial 10921. Patterns of missing data.*

Months								Months							
0	1	2	3	6	9	12	Frequency	0	1	2	3	6	9	12	Frequency
+	+	+	+	+	+	+	93	+	+				+		2
+	+	+	+	+	+		27	+	+			+	+		2
+	+	+	+	+			21	+	+		+			+	2
+	+	+	+				16	+		+					2
+	+	+					14	+		+		+			2
+	+						8	+		+	+		+		2
+							10	+		+	+	+			2
+	+	+	+	+		+	14		+	+	+		+	+	2
+	+	+	+		+	+	11		+		+				2
+		+	+	+	+	+	11		+		+	+	+		2
+	+		+				11			+			+	+	2
+	+		+	+	+		8			+	+				2
+	+	+		+	+	+	7					+		+	2
	+	+	+	+	+	+	7		+	+		+	+	+	1
+	+	+		+			7	+	+		+	+		+	1
+	+		+	+	+	+	6	+	+			+			1
+	+	+	+		+		6	+		+				+	1
+	+		+	+			6	+		+		+	+		1
	+	+					5	+				+			1
+	+	+	+			+	4	+			+				1
+			+	+	+	+	4	+			+	+			1
+		+	+				4		+	+				+	1
+			+		+	+	4		+	+		+			1
	+	+	+		+		4		+	+		+		+	1
+	+	+			+	+	3		+	+	+				1
+	+		+		+	+	3		+	+	+	+		+	1
+		+	+	+	+		3		+						1
+	+				+	+	3		+				+		1
+		+			+	+	3		+				+	+	1
	+	+	+	+			3		+			+	+		1
+	+			+	+	+	2		+			+	+	+	1
		+	+	+	+	+	2			+					1
+	+	+		+	+		2			+				+	1
+		+	+		+	+	2			+		+	+	+	1
	+	+	+	+	+		2			+	+		+	+	1
	+		+	+	+	+	2			+	+	+			1
+		+		+	+	+	2			+	+	+		+	1
+	+	+		+		+	2					+	+	+	1
+	+	+			+		2				+	+			1
+	+					+	2								

the effectiveness and toxicity of 4-OH-Androstenedione (arm A, 91 patients) vs Megestrol acetate (arm B, 86 patients) as second line hormonal treatment. QL data were collected to evaluate secondary endpoints such as impact of treatment on QL and QL as a prognostic factor for time to treatment failure. Patients were treated continuously until treatment failure, i.e., disease progression, unacceptable toxicity, death or patient refusal. The clinical visits were scheduled at week 2, months 1, 2, 3, then every 2 months and at treatment failure. QL assessments were collected during clinical visits at randomization (baseline), months 1, 3, 5, 7, 9 and 11. Thus, patients who did not have a premature treatment failure (i.e., before month 11) should have completed 7 QL assessments.

QL was measured by 7 Linear Analogue Self-Assessment (LASA) scales, ranging from 0 to 100, for physical well-being, mood, fatigue, appetite disturbance, hot flushes, dizziness and perceived adjustment to chronic illness (PACIS). Reasons for the missingness of QL questionnaires were documented if available and classified as 'administrative problems', 'patient refusal', 'language problems' and 'others' (e.g., physician refusal, no clinical visit). For this thesis, only the PACIS scale was considered. For ease of interpretation the original scores were reversed (100 - original score), so that higher scores represent better QL.

The reasons for missing values for each patient sometimes differed from assessment to assessment. For example, consider a patient who completed assessments at time points 1, 2 and 4. The missing value at time point 3 was intermittently missing and caused by administrative problems, while the missing values at time points 5, 6 and 7 were due to missing PACIS score within the received QL questionnaire (i.e., item non-response), and to dropout caused by patient refusal and premature treatment failure, respectively.

The dominant type of missingness in this example was dropout. Table 4.3 presents the number of dropouts at each time point and the cumulative dropout rates. About half of the patients dropped out before month 5, which was consistent with the median times to treatment failure of 120 days in arm A and 111 days in arm B. Most of the dropouts were caused by premature treatment failure; however, 20% of patients (17% in arm A and 23% in arm B) dropped out for reasons other than treatment failure. For the latter subgroup Table 4.4 lists the number of dropout values (before treatment failure) per patient and the reasons for missingness. Intermittent missing values were relatively infrequent. The number of intermittent missing values per patient varied between 0 and 4. Most cases involved missing at baseline. Table 4.5 lists the number of intermittent missing values per patient and the reasons for missingness.

Eight patients (2 in arm A and 6 in arm B) did not give any data on their QL, i.e., were

Table 4.3: *SIAK Trial 20/90. Number of patients with dropout-missing values and cumulative dropout rates.*

Dropout time	Arm A (91 pts)			Arm B (86 pts)		
	No. of pts dropped out due to		Cumulative dropout rate (%)	No. of pts dropped out due to		Cumulative dropout rate (%)
	Treatment failure	Other reasons		Treatment failure	Other reasons	
Baseline	0	2	2.2	0	6	7.0
Month 1	7	4	14.3	14	2	25.6
Month 3	28	1	46.2	21	1	51.2
Month 5	9	1	57.1	10	1	64.0
Month 7	6	0	63.7	8	2	75.6
Month 9	10	1	75.8	4	1	81.4
Month 11	3	4	83.5	3	5	90.7
Total	63	13		60	18	

Table 4.4: *SIAK Trial 20/90. Number of before-treatment-failure dropout-missing values per patient and the causes of their missingness among patients who dropped out for reasons other than premature treatment failure.*

No. of dropout missing values per patient	Arm A (13 pts)	Arm B (18 pts)	Cause of missingness	Arm A (18 values)	Arm B (29 values)
1	8	12	Administrative	7	11
2	5	2	Patient refusal	6	7
3	0	3	Language problems	2	0
4	0	1	Other	1	9
			Missing PACIS	2	2

Table 4.5: *SIAK Trial 20/90. Number of intermittent-missing values per patient and the reasons for missingness.*

No. of intermittent missing values per patient	Arm A (15 pts)	Arm B (16 pts)	Cause of missingness	Arm A (20 values)	Arm B (23 values)
1	10	12	Administrative	9	11
2	5	2	Patient refusal	3	2
3	0	1	Others	3	6
4	0	1	Missing PACIS	5	4

not compliant (see Table 4.3), and thus were excluded from the analysis. The reasons for the missingness of their QL questionnaires were 'administrative problems' ($n = 4$), 'patient refusal' ($n = 2$), 'language problems' ($n = 1$) and 'others' ($n = 1$).

4.3 IBCSG VI-14: Operable Breast Cancer Study

Two hundred and nineteen patients randomized into International Breast Cancer Study Group (IBCSG) studies VI-14 (Sabbioni *et al* 1996) and also participating in an ancillary study of immunological and psychosocial evaluation were observed for a minimum of 6 months (during adjuvant chemotherapy and/or endocrine treatment for operable breast cancer). Clinical factors (age, menopausal and nodal status), sociodemographic factors (level of education, language) and assigned treatment were investigated. Immunological and QL assessments were planned at day 1 (baseline), months 3, 6, 12 and 24. At the time of the clinical visit, patients were given self-administered questionnaires and interviews. The QL questionnaries included LASA (IBCSG) and a series of ordered categorical scales. Some scales were single- and others multi-item. We prospectively selected for this example: 'anxiety' (summary score as an average of 5 items each with 6 ordered response categories ranging from 0=no to 5=very much), and 'burden related to hair loss' (1 item with 6 ordered categories, the same as for anxiety). When an item belonging to a multi-item scale was missing, the total scale value was defined as missing. The reasons for missing QL questionnaires were 'prospectively' collected and defined as: 'local organization problems', 'patient refusal', 'language problems', 'health related problems', 'relapse/death' and 'others'.

One-hundred and seventy patients had node negative (78%) breast cancer, 140 (64%) were

pre-menopausal and 163 (74%) were younger than 60 years. Five patients (2%) did not participate in the psychosocial-QL part of the investigation because of refusal (3), local administrative (1) or language problems (1).

The analysis was based on 642 questionnaires covering the QL assessment in the first 6 months. There were few missing QL questionnaires (18/642, 3%; 4 local problems, 5 refusals, 2 health related problems, 6 relapses or deaths and 1 for other reasons). In this highly compliant group, the small amount of missing data was a mixture of intermittent-missing questionnaires and dropouts (due to relapse or death and refusal).

4.4 The Milk Protein Content Dataset

The ' Milk Protein Content Dataset' is used frequently as a sample dataset for longitudinal data with dropout. Diggle (1990) and Diggle and Kenward (1994) analyzed the data after taking it from Verbyla and Cullis (1990) who in turn had discovered the data at a workshop at Adelaide University in 1989. The data consist of assayed protein content of milk samples taken weekly during 19 weeks from 79 Australian cows. The cows entered the experiment after calving and were randomly allocated to one of three diets: barley, mixed barley-lupins and lupins alone, with 25, 27 and 27 animals in the three groups, respectively. All cows remained on study during the first fourteen weeks, whereafter the sample reduced to 59, 50, 46, 46, and 41, respectively, due to dropout. This means that dropout was as high as 48% by the end of the study. Table 4.6 shows the number of cows per arm and per dropout pattern.

4.5 EORTC 30893: Poor Prognosis Prostate Cancer Trial

EORTC trial 30893 was designed as a prospective multicenter randomized phase III study comparing orchidectomy and orchidectomy plus mitomycin C (15 mg/m^2 intravenously every six weeks until progression) in patients with poor prognosis metastatic prostate cancer. The main endpoint of the trial was survival.

Table 4.6: *Milk Protein Content Trial. Number of cows per arm and per dropout pattern.*

	Diet		
Dropout week	Barley	Mixed	Lupins
Week 15	6	7	7
Week 16	2	3	4
Week 17	2	1	1
Week 18			
Week 19	2	2	1
Completers	13	14	14
Total	25	27	27

A shortened version of the EORTC QLQ-C30 supplemented by disease and treatment specific items was used to assess QL. Scale scores were constructed using the standard procedures recommended by the EORTC Quality of Life Study Group (Fayers *et al* 1999), i.e., scores were calculated by averaging items within scales and transforming average scores linearly to a 0 to 100 scale, with higher scores representing a higher level of functioning or a higher level of symptoms. In Chapter 8 Sections 8.2 and 8.5 we focus on the global health/QL scale as the primary QL outcome whereas in Chapter 10 the physical functioning scale is the primary scale of interest. Further details on the clinical analysis and the QL analysis are described elsewhere (de Reijke *et al* 1999, Fossa *et al* 1999). The planned schedule of assessment in both treatment arms was as follows: at randomization, every six weeks during the first nine months, every three months thereafter until progression of disease and at the time of disease progression.

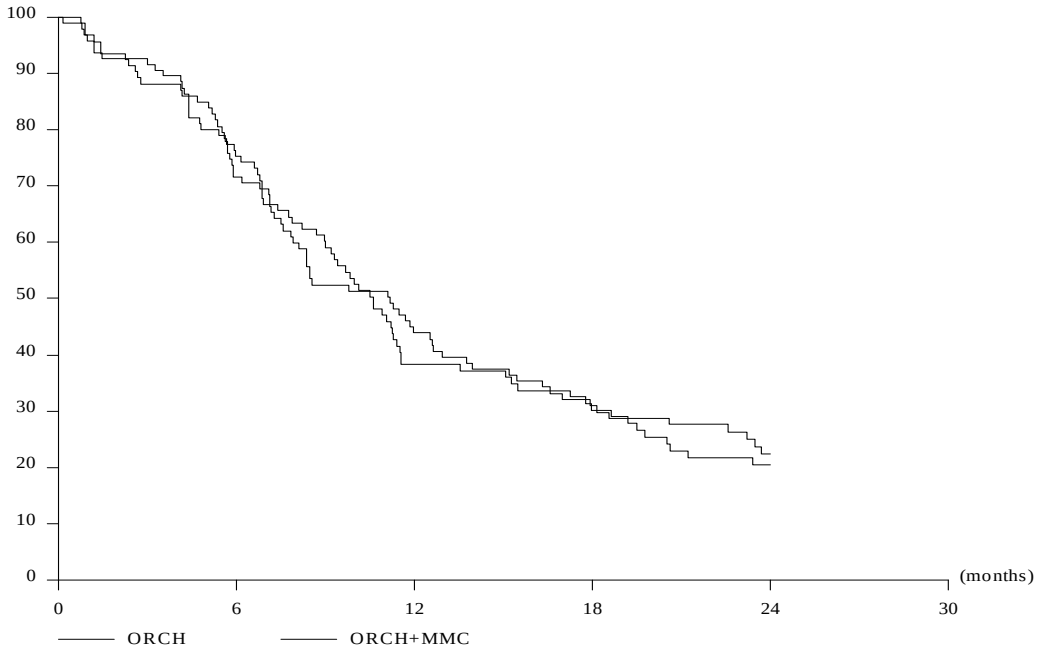

Figure 4.1: *EORTC Trial 30893. Progression free survival by treatment arm.*

Orchidectomy								
Schedule	0	6	12	18	24	30	36	12m
Expected	93	89	84	83	75	61	56	38
Received	51	67	64	59	51	40	43	38
%	55	75	76	71	68	66	77	100

Orchidectomy + mitomycin C								
Schedule	0	6	12	18	24	30	36	12m
Expected	96	91	88	83	74	66	59	40
Received	61	69	72	63	47	46	36	33
%	64	76	82	76	64	70	61	83

Between February 1990 and May 1995, 189 patients were entered into EORTC trial 30893 (93 patients were randomized into the orchidectomy alone (Orch) treatment arm and 96

into the orchidectomy + mitomycin C (Orch+MMC) treatment arm.) Figure 4.1 presents a Kaplan-Meier plot of progression-free survival and a table of QL assessment compliance. The median duration of progression free survival was 10.1 and 11.5 months in the Orch and the Orch+MMC arms, respectively. The main reason for patients going off-study was progression or death. As may be seen in Figure 4.1, the attrition of patients is substantial in both treatment arms. The compliance rate is lower at baseline than at later points in the study. This is explained in part by the fact that although baseline QL should have been completed before orchidectomy, for 40 (21%) patients orchidectomy was performed prior to randomization and hence these patients did not complete a questionnaire before randomization.

4.6 EORTC 30903: Hormone-Resistant Prostate Cancer Trial

EORTC trial 30903 was designed as a prospective multicenter randomized phase III study comparing flutamide versus prednisone in hormone resistant metastatic prostate cancer patients. The main endpoint of the trial was survival. Flutamide and prednisone were administered daily until progression after which patients were treated according to the investigators discretion. Progression was defined as either: an increase in pain score by ≥ 1 category; an increase in daily analgesic dose by $\geq 25\%$; any need to give additional anti-pain treatment, e.g., radiotherapy; deterioration of WHO performance status by ≥ 1 category. Quality of life should have been evaluated at randomization, 3 and 6 weeks later, and at subsequent six weekly intervals. Because it was not clearly defined in the protocol most institutions did not perform QL assessments after progression. The EORTC QLQ-C30 was used to assess QL. In this report we focus on the Global health status/QL scale of the EORTC QLQ-C30. The scale scores were constructed using the standard procedures recommended by the EORTC Quality of Life Study Group (Fayers *et al* 1999), i.e., scores were calculated by averaging items within scales and transforming average scores linearly to a 0 to 100 scale, with higher scores representing a better QL.

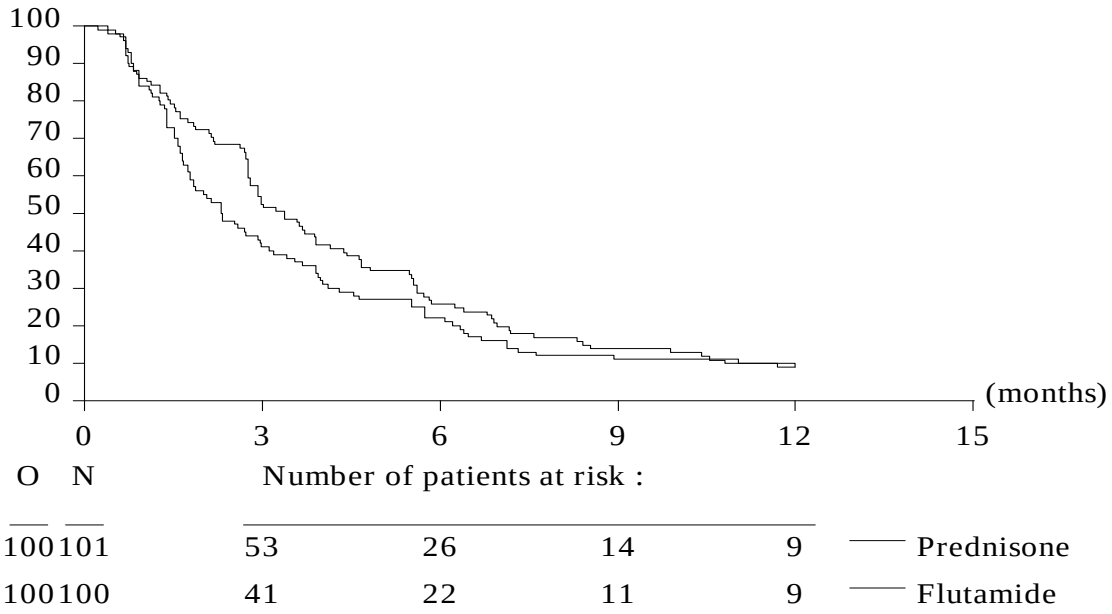

Figure 4.2: *EORTC Trial 30903. Progression free survival by treatment arm.*

Prednisone						
Schedule	0	3	6	12	18	24
Expected	101	97	83	65	42	34
Received	92	77	57	41	29	25
%	91	79	69	63	69	73

Flutamide						
Schedule	0	3	6	12	18	24
Expected	100	97	78	44	30	27
Received	89	73	60	32	21	19
%	89	75	77	73	70	70

Between January 1992 and March 1998, 201 patients were entered into EORTC trial 30903 (101 patients were randomized into the prednisone treatment arm and 100 into the flutamide

treatment arm.) Figure 4.2 presents a Kaplan-Meier plot of progression-free survival and a table of QL assessment compliance. The median duration of progression free survival was 3.4 and 2.3 months in the prednisone and flutamide arms, respectively. The main reason for patients going off-study was progression or death. As may be seen in Figure 4.2, the attrition of patients is substantial in both treatment arms. For this reason, only the assessments up until 24 weeks were used in this analysis. Table 4.7 shows the various missingness patterns. Twenty patients completed QL questionnaires at all five assessment time points. Monotone dropout patterns (i.e., a complete series of questionnaires before dropout) were observed in 104 cases. Intermittent missing questionnaires was also a problem with 46 patients having exactly 1 missing questionnaire and the remaining 11 patients having more than 1 missing questionnaire in a series before dropout.

As the main objective was to investigate (a) differences between treatment groups during treatment and (b) change from baseline it was decided to analyze change scores from baseline i.e., observed scores − baseline scores. This transformation resulted in response scores which were approximately normally distributed and more continuous in nature than the original QL score.

4.7 Remarks

The datasets described above exhibit the main characteristics of QL datasets, e.g. QL is assessed longitudinally with measurements missing both intermittently and due to dropout of patients from the study. These characteristics will be explored further in the ensuing chapters. The milk dataset was chosen as an ideal example to illustrate some of the limitations of selection models in dropout matters, see Chapters 8 and 9.

Table 4.7: *EORTC Trial 30903. Patterns of missing data.*

Weeks							
3	6	12	18	24	>24	N	Percent
+	+	+	+	+	+	16	9.9
+	+	+	+	+		4	2.5
+	+	+	+		+	3	1.9
+	+	+	+			8	5.0
+	+	+		+	+	4	2.5
+	+	+		+		4	2.5
+	+	+				20	12.4
+	+		+	+	+	6	3.7
+	+		+	+		1	0.6
+	+		+			5	3.1
+	+					28	17.4
+		+	+	+	+	2	1.2
+		+	+	+		2	1.2
+		+	+			2	1.2
+		+		+	+	1	0.6
+		+				3	1.9
+			+	+	+	3	1.9
+			+	+		1	0.6
+			+			1	0.6
+						28	17.4
	+	+	+	+	+	2	1.2
	+	+	+	+		3	1.9
	+	+	+			2	1.2
	+	+		+	+	1	0.6
	+	+				1	0.6
	+		+	+	+	1	0.6
	+		+		+	1	0.6
	+					5	3.1
		+		+		1	0.6
		+				1	0.6
			+	+	+	1	0.6

Chapter 5

Missing Items

5.1 Introduction

The problem of missing items was described briefly in Chapter 2. A search of the literature reveals very few papers handling the problem of missing items, e.g. the extent of missing items or how missing values were treated in the analyses. Morris and Coyle (1994) recommend estimating summary scale scores using the mean of the other observed scale items (simple mean imputation, described in Section 2.2). As with missing forms, when items are missing the central statistical issues are bias and power:

1. For example, it might be the case that more ill patients, or patients with more problems, are less willing or less able to complete the questionnaires satisfactorily, or that patients with no problems are less convinced about the need to return comprehensive information. Thus any analyses which ignore the presence of missing data may result in biased conclusions about the changing QL of patients.

2. Even with a small proportion of missing values for each item, the cumulative effect can result in a substantial proportion of patients having one or more missing items during the follow-up period. Analyses based solely upon those patients for whom complete data are available may have severe loss of power because cumulative exclusion of patients results in too few patients remaining in the final analyses.

This chapter considers the implications of and possible solutions for missing items, and proposes some specific solutions for imputing values of missing items. Some widely advocated and commonly adopted procedures are shown to be inappropriate under certain conditions, and alternative methods are suggested for these situations.

5.2 Extent of the Problem

Medical Research Council (MRC) and European Organization for Research and Treatment of Cancer (EORTC) experience in a variety of trials suggests that for most items between 0.5% and 2% of values will be missing from returned QLQ-C30 forms, and similar figures apply to most of the items on the RSCL. Thus, overall, the problem of missing items might be regarded as unimportant. However, there are two important considerations. Firstly, as already stated, since each questionnaire contains about 30 questions, a 1% missing rate would, if it occurred at random, imply that about a quarter of patients ($1\text{-}.99^{30} = 26\%$) could have a missing item on their initial QL assessments whilst even a 0.5% rate could result in 14% missing. Furthermore, at each subsequent assessment there will be additional missing data and thus many patients are likely to have some degree of missing data. Thus any method of analysis which excludes patients who have missing values may result in a seriously reduced data set. However, analysis of the patterns of missingness reveal that it does not occur randomly. Patients who omit answers to one question are more likely to omit answers to other questions, and often there is a pattern of non-response to several consecutive questions, despite the successive questions usually being unrelated to each other. This leads to far fewer forms containing missing values than would be expected by chance alone. Review of 7000 forms in 6 MRC trials indicates that 92% of forms contained complete information regarding 29 out of the 30 questions in the first section of the RSCL, although one question (as explained below) presented particular problems. The proportion of forms with missing data varied considerably from study to study, from 4% to 14%.

Secondly, some items may present particular problems. In particular, the RSCL question "(to what extent have you been bothered by) Decreased sexual interest" produces far more serious problems with patient compliance and for that reason was excluded when estimating the overall missing item rate of 1%. This and similar questions addressing sexuality issues on the QLQ-C30 supplementary modules frequently present high rates of missing values. In all MRC trials, females were far more likely to avoid this question than males - for example, in a trial of palliative radiotherapy for advanced bladder cancer (BA09), 35% of females had

missing data but only 16% of males. In the RSCL the only question apart from 'decreased sexual interest' which showed a significant gender difference was 'loss of hair' (p=0.005), with twice as many females (1.5%) to males (0.8%) avoiding this question. Several questions manifested age effects, with older patients being more likely to leave questions unanswered. Dividing the data about the median age of 65, these included 'shortness of breath' with 1.2% of older patients, versus 0.3% of younger patients (p=0.001); 'sore mouth', 1.4% v. 0.4% (p=0.001) ; 'shivering', 1.1% v. 0.4% (p=0.001); 'acid indigestion', 1.2% v. 0.6% (p=0.003); 'feeling tense', 1.0% v. 0.4% (p=0.007); 'burning eyes', 1.1% v. 0.5% (p=0.008); and most items on the activity scale. These associations with gender and age suggest that at least for some items the missingness is covariate-dependent. Missingness may even be MNAR, with patients experiencing problems being less likely to admit to them.

5.3 Reasons for Missing Data

It is important to consider the reasons underlying the occurrence of missing data, since this may indicate which methods of analysis are plausible and realistic. Some items may be MCAR while others are MAR or MNAR. Unfortunately there are several potential reasons for missing items. The following are likely to be the principal causes in QL studies.

Patient Forgot

Patients may forget to complete, or may overlook, a few questions. Many newer protocols instruct staff to check forms for completeness, and to ask the patient to fill in any missing items.

Patient Felt Too Ill / Too Distressed

Patients may wish to avoid some questions which are embarrassing or cause distress. For example, questions relating to sexual activity, performance and ability are often avoided. However, there are also issues about interpretation and applicability of such questions to those who are very elderly or living alone. This type of missingness is unlikely to be MCAR and may even be MNAR.

Questions Not Understood or 'Not Applicable'

Some questions may be badly worded, and the patient may not know how to respond. For example, 'do you have trouble going up stairs?' might be left blank by patients that (a) live in ground floor apartments, or (b) have such severe trouble climbing stairs that they no longer attempt to do so. In (a) this may be considered as 'not applicable' and equivalent to MCAR, whilst (b) may be considered to be 'not applicable' but is definitely not MCAR. Few QL instruments have provision for patients to indicate that an item has been left blank because it is not applicable, and even fewer provide space for the reasons to be given. Without such information it is difficult to identify if missing data is MCAR.

5.4 Estimation of Scale Scores

When individual items are missing there are problems in calculating values for the summated scales. For a scale that is based upon a number of items, of which one or more is missing, there are in general three main methods that may be adopted.

5.4.1 Treat the Score for the Scale as Missing

If any of the constituent items are missing, the scale-score for that patient is excluded or treated as missing for all statistical analyses (often referred to as listwise deletion or complete case analysis). This method is the simplest and most naive approach to the analysis but results in overall loss of data.

5.4.2 Simple Mean Imputation

The scale score can be estimated from the mean of those items which are available. This is a widely adopted approach which is very simple to implement (Fayers *et al* 1999, de Haes *et al* 1990, Ware *et al* 1993). There are two mathematically equivalent ways of describing this method. Suppose two items are missing from a five-item scale. The mean of the three known results may be calculated, and this mean is then used to replace the two missing values. The

scale score can then be estimated. This is equivalent to calculating the scale score using only those three items for which we do have known values. In its most common form, application of this rule is usually restricted to cases where the respondent has completed at least half of the items in the scale. Section 2.2 describes how to implement simple mean imputation for the emotional functioning scale of the QLQ-C30.

5.4.3 General Imputation Methods

The objective of imputation is to replace the missing data by estimated values which preserve the relationships between items and which reflect as far as possible the most likely true value. If properly carried out, imputation should reduce the bias that can arise by ignoring non-response. By filling in the gaps in the data, it also restores balance to the data and permits simpler analyses. Hence imputation is an attractive procedure - provided one can be sure that the conditions are appropriate and that unintended bias is not being introduced. As was mentioned in Chapter 3 a variety of techniques have been proposed, some of which are mathematically and computationally quite difficult to apply. Perhaps as a consequence, general imputation methods do not appear to have been widely used with QL instruments.

5.5 Psychometric Theory

Conventional psychometric theory holds that scales should ideally be unidimensional; that is, a scale (or a subscale) should measure a single underlying construct. This theory states that any observed score X is the sum of two components, a true score T and an error ϵ:

$$X = T + \epsilon$$

where ϵ is randomly distributed with mean zero and variance σ^2_ϵ. The distribution of ϵ is assumed to be independent of T. The assumption of independence between T and ϵ implies that

$$\sigma^2_X = \sigma^2_T + \sigma^2_\epsilon$$

Let two tests have observed scores $X_1 = T_1 + \epsilon_1$ and $X_2 = T_2 + \epsilon_2$ that satisfy the assumptions of the theory. Then the tests are said to be 'parallel' (Nunnally and Bernstein 1994) if $T_1 = T_2$ and $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2}$. The logical justification for adding items lies in the intuitively

appealing notion that for any pair of parallel items, their sum is more reliable than any single item. Consider a series of k parallel items Y_i for which $Y_i = T + \epsilon_i$, and let

$$X = \Sigma_i Y_i = kT + k\Sigma_i \epsilon_i$$

which leads to

$$\sigma_X^2 = k^2\sigma_T^2 + k\sigma_\epsilon^2 \tag{5.1}$$

As seen from (5.1), for a sum of items, the weight of the variance of the true score is the square of the weight of the variance of the error. Therefore, increasing the number of items decreases the weight of the error variance compared with the true score variance, thus providing a more reliable estimate of the true score.

In practice, all items contributing to a scale should show reasonably strong correlation with the overall scale, and only weak correlations with other scales. This additionally implies that items should be highly correlated with other items in the same scale. These assumptions arise because many areas of psychometric research attempt to measure postulated 'latent constructs' such as intelligence, educational attainment, or personality. For these latent constructs the aim is to devise test items which fulfil the requirements of 'parallel tests'. Under this concept, each item is expected to be an equally good measure of an underlying latent construct for the scale. This principle underpins the use of summated or 'Likert' scales, and justifies using a simple sum of the individual items as a summary scale score. The scale score provides an estimate of the supposed latent construct. The parallel items in a psychometric questionnaire will show correlation with each other by virtue of the fact that they are all measuring the same underlying latent construct or subject's ability. If they are good measures of the latent construct, the correlation will appear to be high. This correlation should be entirely a consequence of the latent construct, and not influenced by other factors. Therefore, if one conditions upon the value of the latent construct, the test items should be independent of each other and of all other factors (Nunnally and Bernstein 1994). This concept of parallel test items also underpins the measures of reliability, such as Cronbach's alpha:

$$\alpha_X = \frac{k}{(k-1)}\left(1 - \frac{\Sigma_i \sigma_{Y_i}^2}{\Sigma_i \sigma_{Y_i}^2 + 2\Sigma_{i<j}\sigma_{Y_iY_j}^2}\right) \tag{5.2}$$

Cronbach's alpha is proportional to the part of the variance of X (the sum) which comprises the covariance of each pair of items. Under traditional test theory, all items should be parallel tests and should be highly correlated, leading to high values of alpha reliability. Under the

conditions of parallel tests simple mean imputation will be adequate. This follows because if the tests are truly parallel, each item is by definition an equally good estimator of the scale score. Therefore the score for a scale can be estimated using the available non-missing items, and the main consequence of missing items will be a lack of precision. If there are high correlations between the items, as manifested by a high Cronbach's alpha reliability measure, the imputed value will provide a good and precise estimate of the scale score. Even when the items are not strictly parallel tests, the use of simple mean imputation may still be a reasonable procedure. However, in some situations this form of imputation may be unsuitable for use with QL scales, and examples of these are described below. In such cases simple mean imputation may be inefficient or may lead to biased estimates of the score values.

5.6 Statistical Considerations

In most clinical trials the number of missing items will be only a small percentage of the total data. However, when this percentage becomes large simple mean imputation will tend to result in underestimation of the variance or standard deviation. This arises because all missing values are being estimated as being equal to the mean, whilst in reality the true data values would have been scattered around the mean value and are subject to variability. Hence, under simple mean imputation, there will be a tendency to estimate confidence intervals and percentiles which are too narrow. When large amounts of data are missing it may be appropriate to consider adding random variability or 'noise' to the imputed values, so as to ensure the standard deviation remains at the expected level.

More serious implications arise when the missing data does not occur completely at random; this may lead to biased comparisons which compromise the treatment comparisons of a trial. Unfortunately it is usually difficult to distinguish values which are missing completely at random from those which are not. The main indications of the need to use more sophisticated methods involve testing the missing data mechanism, and also checking whether there are grounds for suspecting that the data may not be missing completely at random.

5.6.1 Treat the Score for the Scale as Missing

This method is the simplest and most naive approach to the analysis. Thus careful consideration should be given before using it. When data are missing completely at random treating the score for the scale as missing results in a reduced data set which represents a randomly drawn sub-sample of the full data set. Hence inferences about the values of QL for the trial patients can be considered reasonable. Provided the missingness rate is low (e.g., <2%) this approach may be considered since the effect on overall results will probably be negligible. However, if the missingness rate is higher then it becomes important to identify the reasons for missingness and the missing data mechanism. This may be done either by collecting the reasons for missingness or by testing the missing data mechanism as described in Chapter 7. However, testing the missing data mechanism may be problematic if the proportion of missingness is small since the power to detect a difference between responders and non-responders may be low. As such, although the null hypothesis of MCAR is not rejected it may not be possible to completely rule out a MAR or even MNAR process.

If the reasons for missingness or the model for the missingness suggest that the missing data is not MCAR or is covariate-dependent then treating the scale score as missing may lead to serious bias in particular when values are MNAR. An example is the decreased sexual interest item on the RSCL, as discussed above. Since one plausible assumption is that patients experiencing problems are likely to be more reticent concerning this question, missing items may occur more frequently when there are sexual problems. Thus 'missing' might frequently imply 'very much' problem. If this assumption is correct, simply excluding the scores for these patients with missing values could result in misleading and biased conclusions about the prevalence and severity of problems.

One warning sign of a potential problems is the rate of missing data. MRC and EORTC experience suggests that 0.5% or 2% are common random missing rates, and therefore if an item has an appreciably higher missing value rate there may be some consistent reason which should be explored. One example where higher percentages of values may be missing is activities of daily living scales which include an item about 'able to go to work' - this is frequently missing amongst elderly patients, suggesting that they may ignore the preliminary covering sentence 'we do not want to know whether you actually do these, only whether you are able to at the moment'. In this case simple mean imputation is suspect and alternative methods should be considered.

5.6.2 Simple Mean Imputation

The following checks should be made before using simple mean imputation. Most of these checks fall into two broad categories. Firstly, is the probability of missingness dependent on either covariates or observed QL scores? Secondly, for patients in whom the item is not missing, does the item in question behave differently from other items in the same subscale or is it correlated with items external to its own scale?

Patients With Missing Items Should be Similar to Other Patients

Patients with missing items may be different from those with data available. For example, if older male patients tend to omit responses to questions about social functioning, it might be doubtful whether simple mean imputation remains unbiased. As mentioned in the previous section it is useful to test if missingness is dependent on either demographic or clinical covariates or observed QL scores from both the item of interest and other items. In addition, the correlations of 'decreased social functioning' with sex and age may hint at MNAR. Unfortunately it is not possible to distinguish MNAR on this basis alone. In particular, when the response is missing it may well be indicative of poor social functioning. If the missingness is covariate-dependent or dependent on observed QL scores, the estimated score should reflect this in some way.

Item-Means and Item-Variances Within a Scale Should be Similar

If items within a scale do not have the same means then the use of mean imputation may not be justied. We will provide an example where the means and variances within a scale are different using pre-treatment data taken from MRC colorectal trial CR04.

EXAMPLE

This study used the subscale for chemotherapy-related symptoms, which is formed by summing the 5 items shown in Table 5.1. From the table we see that 43% of patients experienced some problems with heartburn / belching (item s22), but at most 15% of patients had prob-

Table 5.1: *MRC Trial CR04. RSCL Chemotherapy-related symptoms subscale, from pre-treatment data.*

Item	Not at all 1 %	A little 2 %	Moderately 3 %	Very much 4 %	Mean*
22:Heartburn / belching	57	27	10	6	22
24:Tingling hands / feet	85	12	1	2	7
26:Pain in mouth when swallowing	95	2	2	1	3
27:Loss of hair	90	8	1	1	4
28:Burning eyes	93	5	2	0	3

* The mean scores have been standardised to lie between 0 and 100, as in Equation 2.2.

lems with each of the other items. The mean score for s22 is 22, which is very different from that of the other items. If there are missing values for s22, it would be a mistake to use the average of the other non-missing items, since this would consistently tend to underestimate the problems that we know are experienced by nearly half of all patients. Equally, if one of s24, s26, s27 or s28 is missing, we should not base the estimate upon calculations incorporating s22 since that would tend to overestimate the score for those patients. Similarly, the variance of s22 is different from the other items. A 'proper' imputation procedure would take into account the entire distribution of scores in this example.

Sometimes the treatment or disease may be expected to result in a high frequency of particular symptoms (such as heartburn/belching), in which case this data may also have been collected on a toxicity report form; it may then be possible to impute missing values from these forms. Examination of correlations and cross tabulations of variables may also be useful in suggesting alternative imputation rules.

It should be noted that the between-item correlation may be high despite items having different mean values. Cronbach's alpha may also be high. Hence high correlations and alphas are not sufficient to justify the use of mean imputation; they are necessary, but not sufficient.

Table 5.2: *Physical Functioning Scale. Adapted from the EORTC QLQ-C30 (version 2.0).*

		No	Yes
1	Do you have any trouble doing strenous activities, like carrying a heavy shopping bag or a suitcase?	1	2
2	Do you have any trouble taking a long walk?	1	2
3	Do you have any trouble taking a short walk?	1	2
4	Do you have to stay in a bed or a chair for most of the day?	1	2
5	Do you need help with eating, dressing, washing yourself or using the toilet?	1	2

The Scale Should Not be Ordered, Hierarchical or a Guttman Scale

Some scales are 'hierarchical' and have an implicit ordering of responses. An example of this is the QLQ-C30 scale for physical functioning, which is based upon questions 1 to 5 (Table 5.2). If a patient replies 'Yes' to Question 3, about trouble taking a short walk, it would not be sensible to base a missing value for Question 2 on the average of the answered items; clearly those who have difficulty with short walks would have even greater problems with a long walk. In this case the structure of the questionnaire may imply that the replies to some questions will restrict the range of plausible answers to other questions. Ordered scales are often called Guttman scales, although technically this description only applies to a particular type of hierarchical scale.

Hierarchical or ordered scales are different in nature from the more usual psychometric scales. The items are not 'parallel tests' with similar expected mean values, but are chosen such that different items correspond to different levels of functioning. Mean scores for one level cannot be used as estimates of missing values for another level. These scales also violate the assumption that all items contributing to the scale should have high correlations with all other items in the scale (Nunnally and Bernstein 1994). Finally, hierarchical scales contain additional information which should be considered when estimating the scale score. Simple mean imputation is not suitable for such scales. If we consider the 'long walk' and 'short

walk' questions, two simple cases are straightforward: (a) Assume the answer to 'long walk' is missing. If the patient has trouble taking short walks, then it would seem reasonable to assume that long walks would cause difficulty too; we therefore impute a value of 'yes' for trouble taking long walks. (b) Assume 'short walk' is missing. If the patient has no difficulty with long walks, we may assume there is unlikely to be difficulty with short walks; we therefore impute a value of 'no' trouble for short walks. One simple extension to the within-patient imputation in this situation is to consider across patients: if the individual in question had no difficulty with short walks, a missing value for long walks could be estimated by calculating the proportion of patients who could take short walks and were also able to manage long walks and impute a value from a Bernouli distribution.

Items Within a Scale Should be Strongly Correlated

The fundamental rationale for simple mean imputation is that a missing item is best estimated by using the values of the other items within the same scale, and that other items and factors may be ignored. This is only sensible if there are reasonably high positive correlations between the items. For a valid and homogeneous scale it is expected that all the component items will be fairly highly correlated with each other. Thus psychometric theory commonly advocates within-scale items should be strongly correlated (but not too strongly correlated, or else there is redundancy). Section 5.5 indicated that Cronbach's alpha is closely related to between-item correlation. As such, a high alpha coefficient is an indication that all items are highly correlated. An alpha less than 0.35 indicates that mean imputation is unreliable, whilst an alpha greater than 0.85 supports its use.

However, correlations may be weak for items within some QL scales, especially with more heterogeneous symptom scales. Therefore it is important to check the magnitude of correlations before using imputation rules. As an example where it would be dangerous to use mean imputation, we consider the QLQ-C30. A combined data set of 178 patients from MRC studies TE17 (adjuvant chemotherapy for high risk stage 1 teratoma) and LU16 showed a correlation of 0.40 between items Q20 and Q25 of the EORTC QLQ-C30 cognitive functioning scale which contains two items: Q20 - 'Have you had difficulty concentrating on things like reading a newspaper or watching television?', and Q25 - 'Have you had difficulty remembering things?' Thus it is dubious whether Q20 can be imputed from Q25 and vice versa.

Furthermore, to justify using simple mean imputation in preference over other methods, not

only should items within a scale be strongly correlated with each other, but in addition correlations with items in other scales, and with external factors, should be low relative to the within-scale correlations; otherwise more efficient methods of imputation would be preferable. An example of the influence of external factors and variables might be the association of performance status with physical functioning; if a patient is known to have a very poor performance status, it would be wise to take that into account if the value for 'able to take long walks' is missing. If the within-scale correlations are weak or if there are strong correlations with other items or factors, simple mean imputation becomes increasingly suspect.

Item 'Not Applicable'

It is unclear how to estimate scale scores when some constituent items are missing through not being applicable. When patients return missing for 'Do you have any trouble doing strenuous activities like carrying a heavy shopping bag or a suitcase' because they never try to perform such activities. The decision how best to allow for non-applicable items will depend partly upon the scientific question being posed. In the example cited, it might be arguable that 'not applicable' represents major problems in terms of QL implications. Therefore, it is important to provide a facility for patients to respond 'not applicable' in situations when they wish deliberately to leave a response blank. Too many questionnaires fail to distinguish between 'missing' and 'not applicable' responses.

5.6.3 General Imputation Procedures

One simple form of imputation is 'last value carried forward' which takes the previously completed value for that patient. However, this assumes that the patient score remains constant over time. As mentioned in Section 3.3 this assumption is very strong and is not likely to be valid in cancer clinical trials in advanced disease. Other forms of imputation which allow shifts in the patient score over time may be considered more appropriate. Conditional mean imputation is an attractive alternative.

Conditional Mean Imputation

Conditional mean imputation allows one to substitute means that are conditioned on other variables or previously observed scores, and therefore also allows for shifts in the general population over time. The method was initially proposed by Buck (1960). Let us describe it first for a single multivariate normal sample. The first step is to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ from the complete cases. This step builds on the assumption that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a subject with missing components, the regression of the missing components ($\boldsymbol{Y}_i^m$) on the observed ones ($\boldsymbol{Y}_i^o$) is

$$\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o \sim N(\boldsymbol{\mu}^m + \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{Y}_i^o - \boldsymbol{\mu}_i^o), \boldsymbol{\Sigma}^{mm} - \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om}). \tag{5.3}$$

Superscripts o and m refer to 'observed' and 'missing' components respectively. The second step calculates the conditional mean from this regression and substitutes it for the missing values. In this way, 'vertical' information (estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) is combined with 'horizontal' information ($\boldsymbol{Y}_i^o$) thus allowing both within and between subject data to be taken into account. In general it yields consistent point estimates when the missingness mechanism is MCAR and it is also valid under certain types or MAR mechanisms (Little and Rubin 1987). Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant across missingness patterns.

Example

EORTC trial 10850 (Curran *et al* 1998d) was designed as a phase III trial to compare mastectomy versus breast conserving surgery in operable breast cancer patients. The QL questionnaire contained a scale assessing body image of which one question focused on being self conscious when seen nude in front of a husband/partner. There was a 13% non-response rate, with non-responders tending to be older.

The Cronbach's alpha reliability coefficient was 0.79 for the body image scale. All item-scale correlations (corrected for overlap) exceeded the 0.40 criterion for item-convergent validity for both multi-item scales. Scaling successes were observed in all cases; that is,

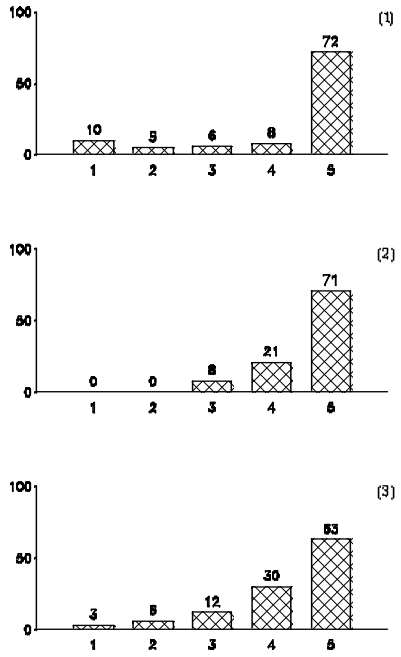

Figure 5.1: *EORTC Trial 10801. Distribution of scores for 'nude' question: (1) observed scores, (2) imputed simple mean scores, (3) imputed conditional mean scores.*

all items correlated significantly higher with their own scale (corrected for overlap) than with other scales. All item means and variances were similar suggesting that simple mean imputation may be appropriate. Figure 5.1 displays the observed score and imputed scores using both simple mean imputation and conditional mean imputation. The precise wording of the question was 'I feel self conscious about being seen nude by my husband/partner' with possible response categories: (1) all of time (2) most of the time (3) some of the time (4) little of the time (5) none of the time. Conditional mean imputation was performed taking the observed scores for other items in the body image scale and covariates age and randomized treatment into account. Although imputed scores correlated highly (Pearson's Rho=0.635), the imputed conditional mean scores indicated a shift towards more problems than was seen by the scores obtained using simple mean imputation. Although differences were small there is an indication that taking covariates into account may be useful, particularly if the proportion of missing items is high. In this study QL was assessed cross-sectionally, preventing us from incorporating within patient repeated scores in the imputation procedure.

Off course, conditional mean scores assumes that the scores for the item come from an approximately normal conditional distribution. This may not be true for all items. However, this method is easy to implement, takes into account the previous score and also allows for a shift in the distribution of scores.

Categorical Data Imputation

An improvment to the methods described above is to use methods specifically designed for categorical data. Schaffer developed a set of algorithms for imputing incomplete categorical data using saturated multinomial models in conjunction with the EM algorithm. Although the algorithms are conceptually simple the notation required to describe them is somewhat intricate. The main advantage of the EM algorithm is that the general theory (Dempster, Laird and Rubin 1977) assures that each iteration increases the likelihood. The EM algorithm consists of an *expectation step* (E step) and a *maximization step* (M step). Given the current value of the parameter vector, the E step computes the expected value of the complete data log-likelihood, given the observed data and the current parameters, which is called the *objective function.* Next, the M step determines the parameter vector maximizing the objective function. One then iterates between the E and M steps until convergence.

Curran *et al* (1998c) used the 'nearest neighbour hot deck' to impute missing QL scores taking clinical factors into account. The authors showed how a 'distance' score could be calculated for each patient using the parameter estimates from a logistic regression.

Comments

The imputation methods reviewed here are not the only ones. Little and Rubin (1987) and Rubin (1987) mention several others. Almost all imputation techniques suffer from the following limitations:

1. The performance of imputation techniques is unreliable. Situations where they do work are difficult to distinguish from situations where they prove misleading.
2. Imputation often requires ad hoc adjustments to yield satisfactory point estimates.
3. The methods fail to provide simple correct precision estimators.

Section 9.5 describes the procedure of multiple imputation in detail. By imputing several values for a single missing component, the uncertainty due to missing data is explicitly acknowledged.

5.7 Remarks

Missing items frequently arise in QL studies, and are a nuisance in that they can affect computation of subscales. There is also the question as to whether values are missing completely at random. Simple mean imputation is one of the most widely practiced methods of allowing for missing values. However, there are many conditions which should be satisfied for this. Several checks have been proposed, which should be applied before deciding how to handle missing items. With few exceptions prescriptive values for these checks are not provided, since any value is subjective and also suitable values may vary according to the nature of the data set.

Given the number of assumptions underlying mean imputation, one might be tempted to assume that it is rarely of value. Nevertheless, provided the scales are unidimensional and constructed in accordance with standard psychometric theory, most of the conditions should in theory be well satisfied. In some cases it may seem likely that the scales are reasonably homogeneous, and also that items may be MCAR. Unfortunately, QL scales are more frequently heterogeneous than most psychometric scales, including as they do items relating to disease symptoms and treatment side effects. Thus it is especially important to consider whether the attributes of well-behaved psychometric scales are realised.

Fortunately, apart from particular questions such as those about sexuality, the proportion of missing items is usually small. Rarely do more than one or two percent of patients omit any particular item. Nonetheless, procedures should be instituted whereby the forms are checked for completeness as soon as they are received. It should be emphasised to patients that they should complete all questions if at all possible. However, missing items are always likely to occur, and the issues have to be addressed. It is disappointing to note that reports of assessments of QL in clinical trials frequently ignore the problems of biases arising from non-random patterns of missing data. Presumably one reason this issue may be ignored is that it is not clear how to analyse such data! However, at the very least a sensitivity analysis should be conducted by examining the potential impact of different levels of bias upon the inferences drawn from the observed data.

Chapter 6

Summary Measures and Summary Statistics

6.1 Introduction

Quality of life assessment has rapidly become an integral part of clinical research resulting in many studies yielding vast quantities of data. However, the question as to what is the best way to analyze and present the results has not been sufficiently addressed. Researchers have sought a practical solution to the conflict of complexity of QL datasets and the desire to simplify presentation of results. Nevertheless, controversies surrounding quality of life analyses have remained, mainly due to the fact that a standard questionnaire consists of numerous categorical scales, assessed at frequent time points during the study, and also because patients may drop out of the study at various times.

Summary measures have been widely accepted as useful methods for reporting results from longitudinal studies (Matthews 1993, Fairclough 1997). In essence a summary measure collapses the complete set of measurements of an individual into a single number. A summary measure should be chosen to reflect some important aspect of the repeated measurements. For example, in cancer clinical trials, data on toxicity is often summarized by taking the worst value recorded for each patient during the entire treatment period.

Within clinical trials, QL is usually reported at repeated time-points before, during and after treatment. Summary measures may be useful for simplifying the repeated structure of the data. A few such measures that could be considered are the mean, median, and the minimum or maximum score recorded for each individual patient. The summary measures for all patients are then analyzed using an appropriate univariate method.

Tannock *et al* (1996) in a clinical trial in prostate cancer patients, used two summary measures in analyzing the QL data. Each patient's score for each QL domain was summarized using the median and the best score. These were subsequently converted to median and best change scores by subtracting the patient's baseline score. Differences in the summary scores between the two treatment groups were assessed with the Wilcoxon rank-sum test.

A distinction should be made between summary measures and summary statistics. A summary measure reduces the measurements for one individual to one single number whereas a summary statistic reduces the measurements of a group of individuals to one number. Similarly, a summary statistic may be a summary of the group differences in QL between two treatment strategies. For example, several authors have compared treatments with respect to QL at individual time points (e.g., using a t-test or a Wilcoxon test). Seymour *et al* (1996) in a study of colorectal cancer patients, presented summary statistics at each time point for each treatment group and used exact χ^2 tests to compare the QL scores in the two treatment groups. Although the sample size may vary at each time point, this method makes use of all available data. However, there are problems with using this method (see Section 6.3.1 of this Chapter).

Using a practical example from an EORTC clinical trial (see Section 4.1), we investigated a number of methods of analysis that have been presented in the literature. A number of summary measures and summary statistics are discussed and it is shown how the choice of method of analysis influences the study results. Examples are provided to show where it may be useful to include summary statistics and measures to reflect an important aspect of the study. The advantages and disadvantages of each method and the basic assumptions that are required when using these methods are discussed.

In 1996 the steering committee for EORTC trial 10921 drew up an analysis plan for the study. It was decided that the primary analysis would be based on an Area Under the Curve (AUC) analysis. This was mainly due to the expectation that the intensified treatment would initially result in a reduced QL but patients would recover more rapidly due to the shorter duration of treatment whereas patients in the standard arm would experience side-effects of treatment over a longer period of time. Therefore it was assumed that the most appropriate

method of balancing the short-term side effects of intensified treatment (EC+GCSF) with the extended side-effects of the standard treatment (CEF) was to perform an AUC analysis. In this chapter we respect the original analysis plan while examining other methods of analysis which have been presented in the literature.

6.2 Summary Measures

Several types of summary measures may be employed in the analysis of QL data. These can be categorized as follows: (1) simple summary measures, e.g., minimum, maximum, median or mean QL score for a patient; (2) time to occurrence of event where, for example, the time to observation of the first minimum or maximum score is taken; (3) area under the curve where both time and QL are summarized into one single number for each individual.

6.2.1 Simple Summary Measures

In cancer clinical trials where a new experimental chemotherapy is being investigated one may wish to investigate if the experimental treatment is less toxic than the standard treatment while achieving equivalent efficacy results. In such trials where there is an interest in reducing toxicity and maintaining an acceptable QL, the worst symptom score may be of particular importance as a summary measure.

In contrast, in clinical trials involving patients with advanced disease (e.g., symptomatic disease), the treatment provided may be palliative in nature, i.e., directed at symptom relief. In such trials, where the primary objective may be to reduce the patient's suffering and thus improve the quality of remaining life, a useful summary measure could be the best score with respect to symptom relief or the best QL score.

In EORTC trial 10921, a trial designed to show superiority, a plot of the individual patient scores (not shown) indicated a good deal of variation in within-patient scores in both treatment arms. Therefore, it was thought that the mean or median score over all sequences would provide useful insight into the patients' QL. The mean score also gives an indication of the frequency of episodes or intensity of problem over time. For example, two treatment groups may have comparable best or worst scores but one treatment group may have consis-

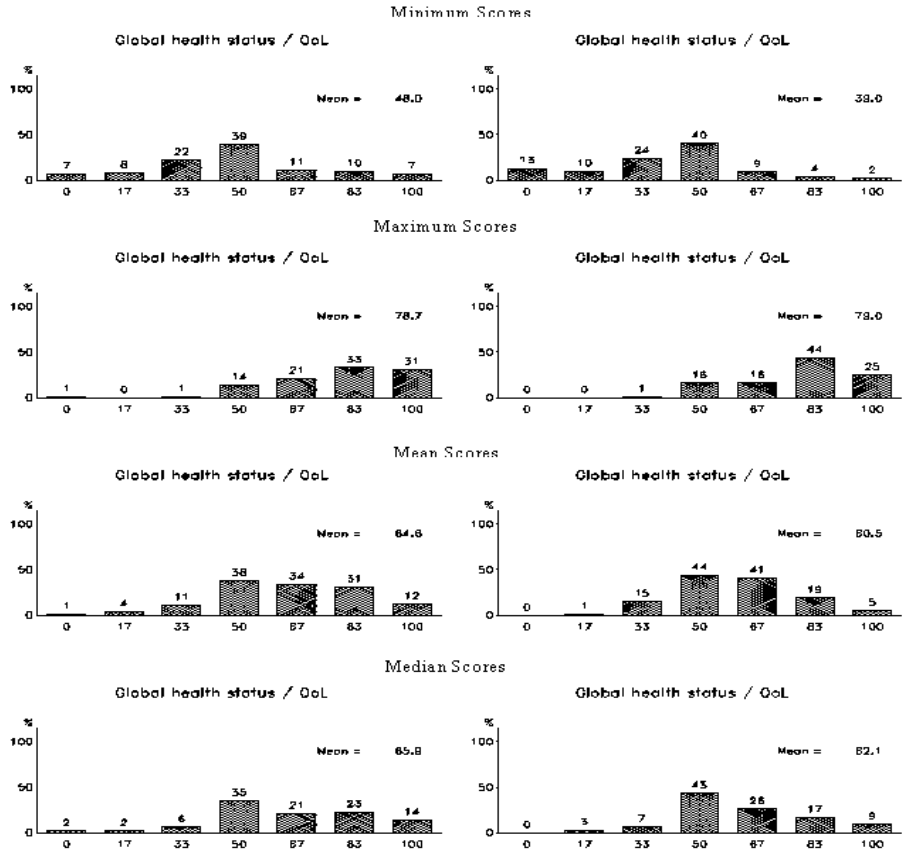

Figure 6.1: *EORTC Trial 10921. Summary measures for the global health status/QL score. Note: for presentation purposes the scores have been grouped into equally spaced intervals with midpoints 0, 17, 33, 50, 67, 83, 100. The X axis represents the % of patients with scores in each interval.*

tently lower scores. Since the mean is an average of all observed scores it would reflect this phenomenon. On the other hand, when two means are being compared it is also important to investigate the spread of scores (i.e., variance). If two treatments provide similar means but the variance is significantly larger in one treatment group, from a conservative point of view one might prefer the treatment which provides more consistent results as it minimizes worst case scenarios.

Figure 6.1 presents these summary measures for patients in trial 10921 during the first year. The Wilcoxon rank-sum test was used to compare QL scores in the two treatment groups. A significant group difference was observed in terms of minimum ($p < 0.001$), mean ($p = 0.016$)

and median QL scores ($p = 0.041$) during the first year in favor of the CEF treatment arm. Note that the differences in mean QL score and median QL score, although statistically significant, were relatively small. The null hypothesis of no treatment difference could not be rejected for the maximum summary measure. These results would suggest that there is a dip in the QL score during the first year in the EC+G-CSF arm.

6.2.2 Time to Occurrence of a Summary Measure

Time to event analyses are frequently used in cancer clinical trials. In most cases the event is death or disease progression and the associated time periods are referred to as duration of survival and time to disease progression, respectively. Some trials have also investigated time to a certain increase in a tumor marker such as prostate specific antigen (PSA) in prostate cancer. In QL research it may also be useful to use this approach to investigate the time at which QL is at its worst or at its best, or when a certain decline in QL scores is observed.

In the analysis plan of trial 10921 it was hypothesized that QL would initially deteriorate in both treatment arms due to treatment toxicity, but that it would increase thereafter due to relief of symptoms related to the tumor. One might expect that this increase would occur more rapidly in the intensified treatment arm due to the shorter duration of treatment and the fact that treatment included G-CSF. Thus an interesting question was to investigate at what point in time, during the first year, patients reported their maximum QL score. Towards this end, the maximum QL score during the first year was obtained for each patient. An event was defined as a maximum QL score greater than the patient's baseline QL score. If the patient's maximum score was not greater than that at baseline or if the patient dropped out before observing a maximum score greater than that at baseline the patient was censored at the time of the last available assessment during the first year. The time to event was defined as the time to the first maximum QL score. In Figure 6.2, maximum QL scores tended to be observed earlier in the CEF arm. However, at months 6 and 9, there was a greater tendency for maximums to be reached in the EC+G-CSF arm. No significant difference was observed between the two treatment groups ($p = 0.507$). Approximately, 50% of patients in both treatment arms observed a maximum score greater than baseline during the first year.

In the above approach, due to the original categorical nature of the EORTC global health status/QL scale, a score greater than baseline can be interpreted as an improvement of at least 8 points on a 0-100 scale. A similar approach would be to define a specific minimum

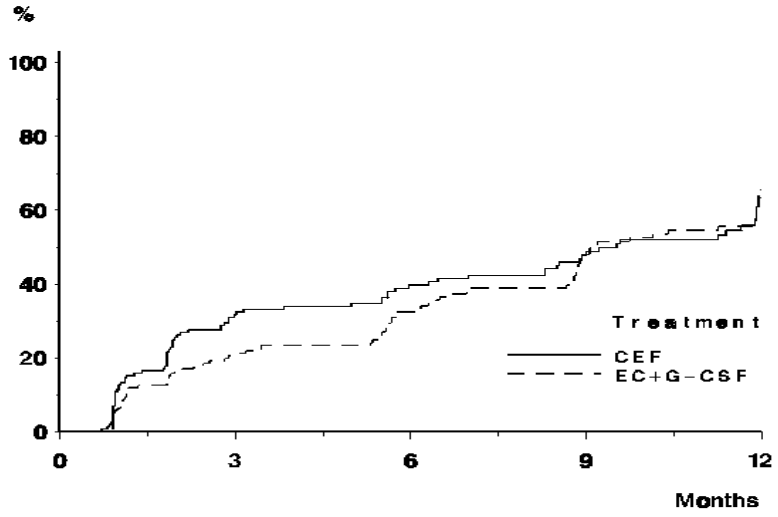

Figure 6.2: *EORTC Trial 10921. Time to maximum QL score.*

level of improvement in QL (e.g., a change of 20 points or 0.5 of a standard deviation) and to define the time to event as the time to reaching such a minimal improvement.

6.2.3 Area Under the Curve

The area under the curve is calculated by summing areas under the graph between each pair of consecutive observations. Thus, the AUC is a weighted average of the QL scores at each individual time point weighted by the time between observations. Using the trapezium rule (hence assuming linear change over time between assessment points) the area under the curve for a patient i is calculated as

$$AUC_i = \frac{1}{2} \sum_{j=0}^{n-1} (t_{j+1} - t_j)(y_{ij} + y_{i,j+1}), \tag{6.1}$$

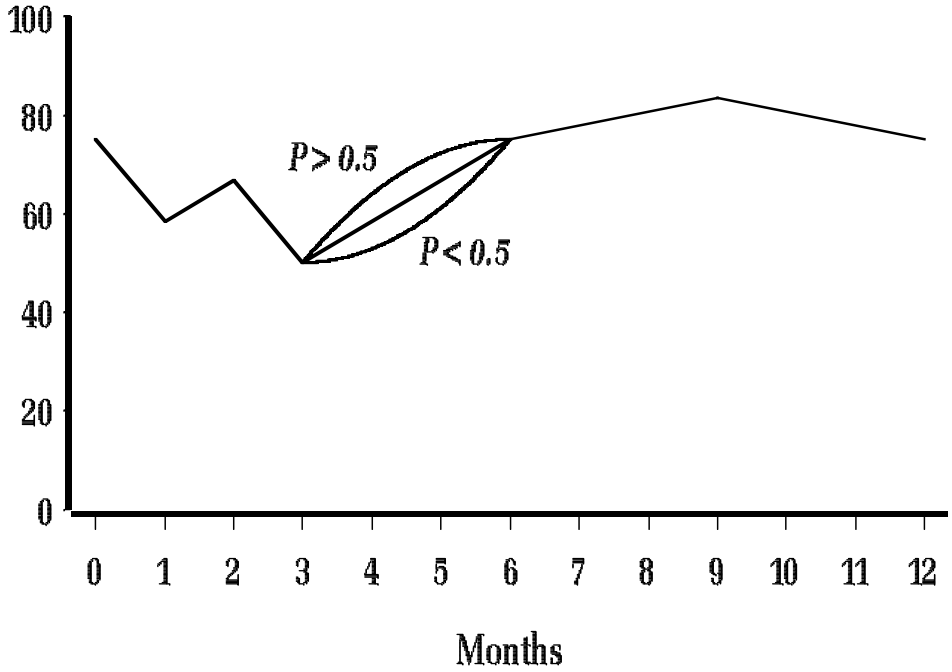

Figure 6.3: *EORTC Trial 10921. Global health status/QL score during the first year for an individual patient.*

where y_{ij} represents the individual's scale score at time $t_j \in \{0, 1, 2, 3, 6, 9, 12 \text{ months}\}$. It is a weighted average of the QL scores at each individual time point, weighted by the spacing of the assessments.

In QL analysis, time may be considered as discrete or continuous, i.e., the time of assessment may be taken as the planned time of assessment (e.g., time points 0, 1, 2 months which is discrete) or as the actual observed assessment time points (e.g., -1, 29, 61 days which is continuous). In trial 10921 we chose to treat time as a continuous variable. The AUC was compared between the two treatment groups using the Wilcoxon rank sum test. No significant difference was observed between the two groups ($p = 0.882$).

One of the advantages of the AUC method is that a sensitivity analysis may easily be performed to investigate the change in QL between two consecutive assessments. For example, in trial 10921 QL was assessed at months 1, 2 and 3 and then at months 6, 9 and 12. Of

particular importance was the question 'Did patients recover rapidly after treatment with EC+G-CSF?' Generally, the AUC is calculated assuming a linear change in QL scores between consecutive assessment time points. However, the formula for calculating the AUC may be changed to allow the rate of change between assessments to occur non-linearly (see also Figure 6.3) as follows.

$$AUC_i = \sum_{j=0}^{n-1} (t_{j+1} - t_j)(\min(y_{i,j}, (y_{i,j+1}) + \rho(\mid (y_{i,j} - y_{i,j+1}) \mid)), \tag{6.2}$$

For the special case where $\rho = 1/2$, Equations (6.1) and (6.2) are equivalent. A sensitivity analysis may be performed to investigate if changing the parameter ρ modifies the conclusions with respect to treatment effect. An alternative sensitivity analysis would be to investigate the effect of dropout, analogous in principle to the approach taken by Hollen *et al* (1997) who imputed a score of 0 on the day of death.

Note, the AUC method should give approximately the same results as taking the mean as a summary measure if the time between assessments is equal. When using the mean each score is given equal weight whereas the AUC method weights the scores according to the time between assessments. We used the trapezium rule to calculate the AUC and compared the resulting summary measures using the Wilcoxon rank sum test. No significant difference was observed between the two treatment arms. Recall that when the mean was used as a summary measure there was a significant difference between the two treatment groups ($p = 0.016$). When using the AUC, the area under the curve between baseline and month 3, which contains four assessments, is given the same weight as the AUC between months 6 and 9, which contains two assessments. Thus, the area under the curve provides a more balanced estimate of the overall QL during the first year than does the mean.

6.2.4 Limitations of Summary Measures

When using summary measures, as with any type of statistical analysis, care has to be taken that bias is not being introduced. For example, where the worst score for a symptom is taken as a summary measure, the results may be biased if patients with a high level of a symptom are unable to complete a self assessment questionnaire and are thus not able to report their worst level of symptoms. This would lead to a biased estimate of the level of symptoms in each treatment arm. Similarly, this may result in a biased estimate of the relative effect of one treatment versus the other.

When using summary measures, bias may be introduced into the comparison if the follow-up periods are not similar in the two treatment arms. Additionally, the rates of completing questionnaires should be high and equivalent across treatment arms. In trial 10921 the progression-free survival and QL compliance were similar in the two treatment arms. Summary measures may not be appropriate in studies where dropout of patients is high as they ignore the problem of incomplete data.

In 'time to occurrence of event' analyses there may be some difficulty in defining an event and defining censoring. In the above analysis, it was assumed that the majority of patients would observe the event of interest during the first year and thus censoring would have less impact on the treatment comparison. However, if the timing of censored observations is different between the two groups there may be problems with interpretation of results.

Generally, summary measures such as the minimum and maximum score are very sensitive to outliers (i.e., extreme observations). When analyzing categorical data this may not be a problem. However, for continuous data one might consider using more robust estimators (e.g., mean or median).

6.3 Summary Statistics

Since QL measurements are typically obtained via repeated assessments over time, it is generally assumed that QL data should be analyzed as such, taking the repeated measurements into account. However, this is often hampered by the structure of the data; i.e., QL data are usually measured on ordered categorical response scales and a proportion of questionnaires will be missing both intermittently and due to dropout of patients from the study. Statistical techniques for longitudinal, ordered categorical, incomplete data are limited. This has led QL researchers to perform separate analyses at each assessment time point. This method of analysis is usually referred to as cross-sectional analysis or available case analysis as it uses all available data at each assessment time point.

6.3.1 Cross-sectional Analysis

If the distribution of QL scores is approximately normally distributed, it may be appropriate to perform simple t-tests within each cross-sectional analysis. Often non-parametric tests, such as Wilcoxon or Mann-Whitney tests, may be more appropriate, in that many QL questionnaires yield skewed distributions with notable ceiling or floor effects (i.e., proportion of patients with either none or severe problems). If there is a large difference in the mean QL score between the two treatment groups one might also expect that the variance may be quite different in both treatment groups, particularly with skewed distributions, suggesting that if a standard t-test is to be performed the variance in each group should first be investigated.

In trial 10921 we compared the two treatments with respect to global health status/QL score at each time point using a Wilcoxon test. The results are presented graphically in Figure 6.4. The plot indicates that, compared to the standard regimen, the intensified regimen had a significant negative impact on QL during the first three months. At month 6 the QL score returned to pre-treatment levels in the intensified arm, while patients in the standard arm tended to have a poorer QL score. No significant differences were observed between the two groups at months 9 and 12.

The main disadvantage of this method is that different sets of patients contribute at different time points depending on the pattern of missing data. Thus, this method yields problems of comparability across time points. Additionally, it does not control for any potential biases in the treatment comparisons which may occur due to dropout of patients. The procedure of examining differences between groups of patients at each assessment time point also leads to inflated Type I and II errors due to multiple testing. An adequate adjustment of the significance level of each test (Pocock, Geller and Tsiatis 1987, Hochberg 1988), or a combination of individual test statistics into a global statistic (Wei and Johnson 1985) is then essential. If many statistical tests are being performed, it is possible to use a more restrictive significance value such as $P<0.01$ or 0.001, thereby reducing the risk of Type I errors. In the next section, we will discuss the Wei-Johnson (1995) procedure which allows the per time point test statistics to be combined into one overall test statistic.

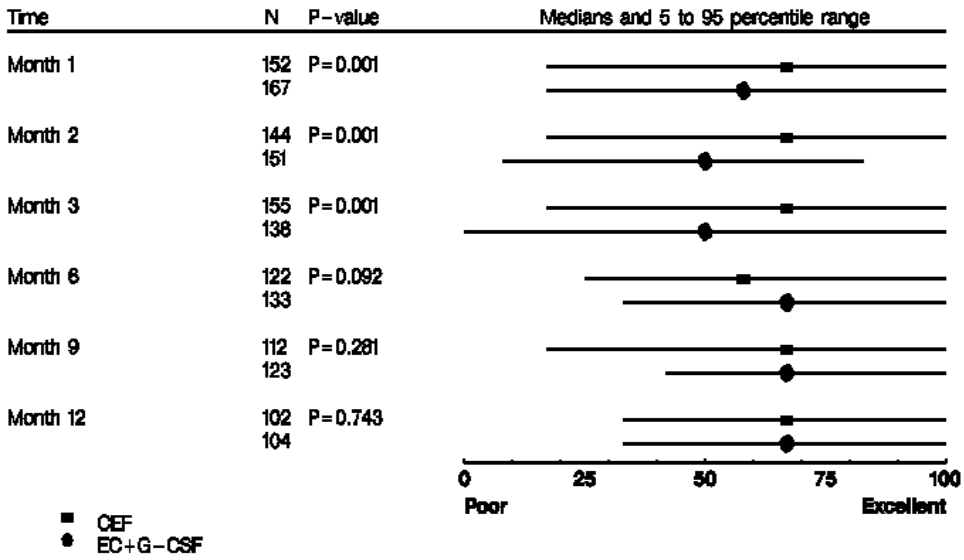

Figure 6.4: *EORTC Trial 10921. Cross sectional analysis of global health status/QL score during the first year.*

6.3.2 Wei-Johnson

Overall tests of significance are generally preferable to comparisons per time point. Overall tests allow general statements about effects, are statistically more powerful and provide a safeguard against multiple comparisons. When overall tests yield statistically significant results, they can be followed by exploratory comparisons per time point. Wei and Johnson (1985) proposed a test, which allows cross-sectional tests to be combined in an overall treatment comparison. They illustrated how cross-sectional Wilcoxon tests, t-tests or tests for 2x2 tables could be combined. In trial 10921 cross-sectional analysis to compare the QL scores between the two treatments at each of the timepoints were performed using the Mann-Whitney test. These were then combined using the Wei-Johnson method as follows. Let U_j be the Mann-Whitney test statistic obtained for each cross-sectional analysis j, then the Wei-Johnson statistic is defined as:

$$WJ = \frac{\boldsymbol{w}'\boldsymbol{U}}{(\boldsymbol{w}'V\boldsymbol{w})^{1/2}},$$

where V is the covariance matrix of $\boldsymbol{U}= (U_1, \cdots, U_j)$ and w_j are weights representing the weight given to each cross-sectional analysis j. They are usually chosen to yield a test which maximizes certain local asymptotic powers. No particular parametric model of dependence is imposed on the repeated measurements of each individual. Although it is mentioned by Wei and Johnson (1985) that an MCAR process is assumed, a simulation study performed by Chirwa (1996) suggested that the Wei-Johnson test was fairly powerful, irrespective of the underlying missingness mechanism.

The estimate of the Wei-Johnson test statistic, when taking equal weights for each time point (i.e., $w_j = 1$ for all j), is WJ=3.404 ($p < 0.001$). Defining the weights relative to the time period between measurements for the six assessments at 1, 2, 3, 6, 9, and 12, respectively yields $w = (0.5, 0.5, 0.5, 1.5, 1.5, 1.5)$. Note the average of all the weights (w_j^s) should be 1 (since the first 3 assessments are taken 1 month apart they receive a weight of 0.5 and the last 3, taken 3 months apart, receive a weight 1.5). In this setting, the Wei-Johnson test statistic is $WJ = 1.812$ ($p < 0.070$). Various alternative techniques for estimating weights could be employed including estimating weights based on the number of patients contributing to the analysis at each time point.

Note, the Wei-Johnson procedure has all the disadvantages of the cross-sectional analysis presented in Section 6.3.1 except that it produces an overall test and therefore reduces the number of statistical comparisons. Although the Wei-Johnson procedure allows one to analyse the data in a longitudinal fashion, it does not tell us anything about the correlation structure. For example, one might expect that QL measurements taken close together are strongly correlated while measurements taken at larger intervals are less correlated. While cross-sectional studies are useful in describing between-patient variability, explaining within-patient variability necessitates the study of the repeated assessments over time.

6.4 Categorical Data

Although all of the above methods were illustrated using the global health status/QL scale, most of the methods (perhaps with the exception of the AUC method) can be applied to the other scales which have fewer potential categories such as the single item scales of the EORTC QLQ-C30. For single item scales with binary or ordered categorical response scales another summary measure could be the frequency of events. This could be particularly useful for scales assessing symptoms or toxicity. For example, in the QLQ-C30 one might

be interested in the number of times patients responded that they had 'Quite a bit' or 'Very much' trouble sleeping during the treatment period.

When performing cross-sectional analyses with ordinal response categories a useful approach is to compare proportions of patients with a certain category. For example, instead of comparing the distribution of insomnia scores between the two treatment groups, one could calculate the proportion of patients in each group who report having 'Quite a bit' or 'Very much' trouble sleeping. A proportion is one summary statistic that is easy to describe and facilitates understanding of results. Additionally, two proportions can easily be compared using a chi-square test. Cumulative proportions are of particular relevance when dealing with ordinal data and may be analyzed using odds ratios. The Wei-Johnson procedure may also be used to combine tests of proportions at several time points into one overall test.

6.5 Remarks

The main differences between the two treatments groups occurred during the first 3 months where QL scores are significantly lower in the EC+G-CSF arm (illustrated by the cross-sectional analysis and minimum scores). The scores are borderline significantly different at 6 months in favor of the EC+G-CSF probably due to the fact that patients were still receiving treatment in the CEF arm whereas patients in the EC+G-CSF arm had recovered from treatment toxicities. At later time points there are no significant differences between the two groups. As the summary measures: minimum, maximum, mean and median attach equal weight to each assessment they ignore the time between assessments. The Wei-Johnson method allows one to explore how associating different weights to the various assessments yields different conclusions. The AUC method yields an overall score for each patient, which is calculated as a cumulative weight of all QL scores weighted according to the time between assessments. Thus, the difference between methods can be explained in part by the weighting of time, e.g., short term differences versus overall differences.

Due to the complicated nature of QL datasets, care should be taken when choosing a method for analysis since there is a risk of drawing incorrect conclusions if inappropriate statistics are used. In this Chapter, we illustrated that the conclusions drawn may differ depending on the type of analysis performed. By definition summary measures and statistics do not use all the data collected and as such they may be considered wasteful, e.g. they do not take into account how patients' scores change over time. Due to this, study conclusions should not

be based solely on one summary measure or summary statistic, but should be supported by additional analyses in the form of a sensitivity analysis. Although, performing a sensitivity analysis using summary measures and summary statistics provides more insight and a better understanding of the data it may not lead to a complete picture of events. In general, summary measures and summary statistics assume that there is no bias in the treatment comparisons due to intermittent missing data or due to patients dropping out of the study. This may not be the case if patients do not complete the questionnaire because they are unfit to do so or if they dropout of the study due to progressive disease or worsening clinical condition. In practice it is usually impossible to conclude definitively whether dropout causes a bias in the treatment comparison or not, since the required information is not available. It is therefore important to prospectively collect the reasons for missing QL assessments. Chapters 2 and 7 discuss how the bias due to dropout may be investigated and how it may affect the study results. Curran *et al* (1998a) concluded that it is important to identify whether there is differential dropout in the two groups. This is likely to occur when the clinical outcomes (e.g., time to progression or progression-free survival) differ between the two treatment groups.

Recently, much attention is being given to longitudinal (repeated measures) analyses with possibly intermittent missing data and missing data due to dropout of individuals from a study (Diggle and Kenward 1994, Diggle, Liang and Zeger 1994, Molenberghs, Kenward and Lesaffre 1997). Chapter 8 illustrates how graphical techniques may be used to explore the longitudinal structure of the repeated measurements followed by statistical modeling of the longitudinal measurements using selection models and pattern-mixture models. Unlike summary measures and summary statistics longitudinal data analysis approaches make more use of the available data.

Chapter 7

Identifying the Types of Missingness in QL Data

7.1 Introduction

In the previous chapter we noted that summary measures and summary statistics do not take into account the dropout process. Using longitudinal data techniques as described later in Chapters 8, 9 and 10 the need for exploring the dropout process becomes more explicit. Identifying the missing data mechanisms can be viewed from two complementary perspectives: (1) collecting information on why the QL questionnaires were not completed and (2) hypothesis testing of the missing data processes. The first approach is a pragmatic one and is based on prospectively collecting as much information as possible to determine the reasons why questionnaires are missing. From these it may be possible to decide if ignoring the missing questionnaires will bias the analysis. The second approach is based on modeling the missing data mechanism to test if the data are MCAR, MAR or MNAR. Two methods from the literature for testing MCAR are presented with application to QL data from clinical trials. The first method is based on fitting a logistic regression whereas the second method is based on an adaptation of weighted least squares (WLS). Testing MNAR is also discussed. An argument is provided to illustrate that it may not be possible to test this hypothesis.

7.2 Why Has the QL Questionnaire Not Been Completed?

One advantage of studying QL as an integral part of a clinical trial is that additional clinical information is collected at each visit. In the past, information related to the patients' survival status, disease status, symptoms and toxicity was useful in determining retrospectively why further QL data had not been obtained. However, this information was useful for explaining only a portion of the missing questionnaires. Therefore, more recently researchers have prospectively included questions on the clinical case report forms (CRF's, e.g., treatment and follow-up forms) in an attempt to capture more information on why questionnaires were missing. These questions generally have the following format: Has the patient filled in the current quality of life questionnaires, 0=no, 1=yes. If no, please state the main reason

1 = patient felt too ill

2 = clinician or nurse felt the patient was too ill

3 = patient felt it was inconvenient, takes too much time

4 = patient felt it was a violation of privacy

5 = patient did not understand the actual language / illiterate

6 = administrative failure to distribute the questionnaire to the patient

7 = other, please specify

The survival and disease status of the patient are also collected on the CRF's. Initially, attempts to distinguish between treatment toxicity and disease related symptoms were made. However, as a nurse or a data manager usually administers the QL questionnaire it may be difficult for him or her to distinguish between the two.

7.3 Hypothesis Testing for MCAR

Recall from Chapter 2, when likelihood and Bayesian inference is used and when only the measurement model parameters are of interest then the distinction between MCAR and MAR

is of minor concern. Although there are still a few issues related to the estimation of standard errors (Molenberghs and Kenward 1997). In addition, even when likelihood and Bayesian inference is applied it may be neccessary to distinguish between MCAR and MAR depending on the research questions. For example, if $\boldsymbol{Y}_i$ follows a multivariate Gaussian distribution, then under MCAR the mean structure of $\boldsymbol{Y}_i$ coincides with the conditional mean structure of $\boldsymbol{Y}_i$ given no dropout, but this is not so under MAR, except in the generally unrealistic case of uncorrelated $\boldsymbol{Y}_{ij}$. Thus, if the research question involves determining the conditional mean structure of $\boldsymbol{Y}_i$ given no dropout it is necessary to distinguish between MCAR and MAR. Moreover, frequentist techniques, such as generalized estimating equations (Liang and Zeger 1986) are only valid under MCAR. It is then crucial to discriminate between MCAR and MAR.

In the literature a number of methods have been described for testing the hypothesis of MCAR (Little 1988, Diggle 1989, Ridout 1991, Park and Davis 1993, Lipsitz, Laird and Harrington 1994, Heitjan and Basu 1996). Two methods are presented and discussed in this chapter. The first method proposed by Ridout (1991) is based on a logistic regression, whereas the second method proposed by Park and Davis (1993) is based on an adaptation of weighted least squares (WLS). Some applications are provided in the context of incomplete longitudinal QL data obtained from international multicenter cancer clinical trials.

7.3.1 Ridout Method

In 1991 Ridout proposed a method for testing completely random dropout using a logistic regression (Cox 1970). This method assumes a monotone pattern of missing data, i.e., that the baseline assessment is available for all patients and at subsequent assessments a proportion of patients drop out and never complete the questionnaire again (see Figure 7.1). Thus intermittent-missing questionnaires are not taken into account in this method.

For each time point T_k, identify the subset of patients S_k (collection of patients i such that $i = 1, \cdots, n_k$) for whom an assessment is available at that time point and identify the subset of patients s_k for whom it is the final assessment before they drop out of the study. The subset of patients s_k consists of the patients i such that $n_{k+1} < i \leq n_k$ (see Figure 7.1). Testing for completely random dropout involves testing the assumption that the scores from the s_k patients are a random sample of the scores from the S_k patients. The pool of $S'_k s$ constitute the sample for the regression. The response variable is dropout or not at time k.

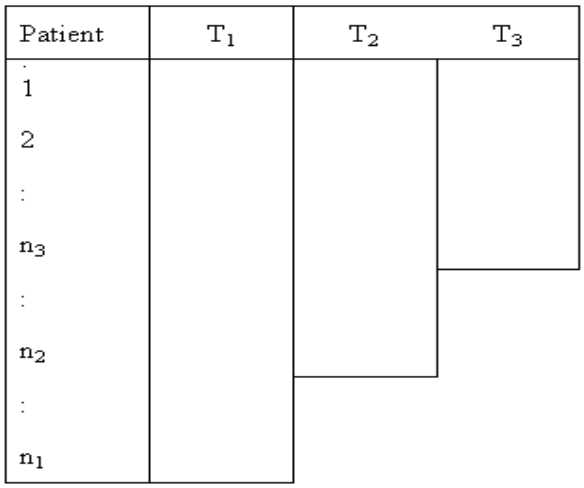

Figure 7.1: *Hypothetical Example. A monotone pattern of missing data.*

The logistic regression model is given by:

$$\text{logit}[\text{pr}(\text{dropout})] = \alpha + (X, \boldsymbol{Y})\beta,$$

where α is the intercept, β is a vector of parameters, X is an array consisting of covariates such as treatment and time of assessment and $\boldsymbol{Y}$ is an array of observed QL scores. Note: for MCAR the dropout mechanism may depend on the values of fixed covariates. In particular, if the covariate matrix includes time and/or treatment as a variable then the model allows the dropout rates to vary over time and/or treatments. This is usually referred to as 'covariate dependent dropout' (Little 1995). A logistic regression may be performed using standard statistical software such as the LOGISTIC procedure in SAS.

Example

We will illustrate testing for missingness with the logistic model using QL data collected in a postmenopausal advanced breast cancer trial conducted by the Swiss group. See Section 4.2 for details on the dataset.

Let k denote the possible times for dropping out where $k = 1, \cdots, 6$ corresponding to months 1, 3, 5, 7, 9 and 11, respectively. The assessments of PACIS were used in 3 different ways as an explanatory variable (expressed as Y in the logistic equation below): Model 1 – the last assessment at time $k-1$, Y_{k-1}; Model 2 – the difference between baseline and the last assessment, $(Y_{k-1} - Y_0)$; Model 3 – the last 2 assessments at times $k-1$ and $k-2$, expressed as $(Y_{k-1} + Y_{k-2})$ and $(Y_{k-1} - Y_{k-2})$ Note that for models 2 and 3 only those patients who had not dropped out by month 1 were included in the analysis, thus $k = 2, \cdots, 6$.

Three other factors, i.e., treatment arm (Trt), cause of dropout (CD, for the first dropout-missing value), and dropout time (Time, for the first dropout-missing value), were suspected to have an influence on the missing mechanism and thus also included in the full regression model. The full model was

$$\begin{aligned} \text{logit}(\Pr(\text{dropout})) &= \alpha + \alpha_{Trt}X_{Trt} + \alpha_{CD}X_{CD} + \alpha_{Time}X_{Time} + \beta_Y \boldsymbol{Y} \\ &\quad + \beta_{Trt}(X_{Trt} \times \boldsymbol{Y}) + \beta_{CD}(X_{CD} \times \boldsymbol{Y}) + \beta_{Time}(X_{Time} \times \boldsymbol{Y}), \end{aligned}$$

where α is the overall mean, X are the dummy variables for Trt (1 variable), CD (1 variable), and Time (5 variables for Model 1, 4 variables for Models 2 and 3), and β are the corresponding coefficients of Y and the interaction terms.

The results of the regression analysis are presented in Table 7.1. Comparing line 2 with line 1 of Model 1 by likelihood ratio test indicates that the interaction terms were not important. Lines 3 and 4 showed that treatment arm and the cause of dropout were not important either. However line 5 indicated the significance of dropout time. Comparing line 2 with 8 and line 4 with 6 suggested covariate Y_{k-1} should not be ignored; thus, the missing mechanism of dropouts was very probably not completely at random. The result of Model 2 was a little different. Comparison of lines 2, 3, 4 and 5 with line 1 showed that the interaction terms, cause of dropout, treatment and dropout time were all not significant. Nevertheless, comparing line 2 with 8 and line 5 with 7 also suggested that covariate $(Y_{k-1} - Y_0)$ should not be ignored and the missing mechanism of dropouts was probably not completely at random. Model 3 with $(Y_{k-1} + Y_{k-2})$ and $(Y_{k-1} - Y_{k-2})$ showed results similar to Model 2.

There was no significant difference in clinical effectiveness in terms of response rate and time to treatment failure between the two arms. From the above results one can see that the difference in probability of dropping out for PACIS between arms was not significant. Although 20% of the patients dropped out for other reasons before treatment failure, this was not a significant factor for the missing mechanism of dropouts, which might be explained as being confounded with other factors, e.g., dropout time. As seen in Table 4.3, the increase in

Table 7.1: *SIAK Trial 20/90. Results of logistic regression analysis.*

	Null hypothesis
1	Full model
2	$\beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
3	$\alpha_{CD} = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
4	$\alpha_{CD} = \alpha_{Trt} = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
5	$\alpha_{CD} = \alpha_{Trt} = \alpha_{Time}* = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
6	$\alpha_{CD} = \alpha_{Trt} = by = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
7	$\alpha_{CD} = \alpha_{Trt} = \alpha_{Time}* = by = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0
8	$\beta_y = \beta_{Trt} = \beta_{CD} = \beta_{Time}$*=0

	Model 1 (N=573)		Model 2 (N=399)		Model 3 (N=399)	
	df	-2 log L	df	-2 log L	df**	-2 log L
1	557	606.374	385	458.824	378	452.560
2	564	608.310	391	468.175	390	463.051
3	565	609.675	392	472.425	391	465.989
4	566	612.155	393	473.030	392	467.256
5	571	636.744	397	479.761	396	473.269
6	567	624.598	-	-	-	-
7	-	-	398	484.554	398	484.554
8	565	621.581	392	474.509	392	474.509

* Corresponding to 5 dummy variables in Model 1
and 4 dummy variables in Model 2.
** Because of 2 Y variables in Model 3 the number of β
parameters are doubled when compared with Model 2.

cumulative dropout rate between months 1 and 3 was much larger than in other consecutive time intervals; therefore, it was not a surprise to see the significance of dropout time in Model 1. All 3 models suggested the dropout mechanism was probably not completely at random.

Table 7.2: *IBCSG Study VI-14. Number of patients by response profiles (Anxiety scale).*

Response category (baseline, 3rd month, 6th month): N= No anxiety, Y = anxiety, M =missing																		
N	N	N	N	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	M	M	M	M
N	N	Y	Y	N	N	Y	Y	Y	M	M	Y	M	M	M	Y	Y	N	N
N	Y	N	Y	N	Y	N	Y	M	Y	M	M	Y	N	M	Y	N	N	M
73	10	9	10	16	7	8	57	4	4	3	1	1	3	1	4	1	1	1

Table 7.3: *IBCSG Study VI-14. Proportion with 'Anxiety'.*

	Baseline	3rd month	6th month
Incomplete data (n=24)	0.647	0.833	0.643
Complete data (n=190)	0.463	0.442	0.442

7.3.2 Park and Davis Method

The weighted least square (WLS) methods proposed by Grizzle, Starmer and Koch (GSK) in 1969 have been further developed for the analysis of incomplete longitudinal categorical data (Stanish, Gillings and Koch 1978, Woolson and Clarke 1984). These methods assume that the missingness mechanism is completely at random (MCAR). When the response variable is categorical (with few response categories), the number of measurement times is small and the sample size is relatively large within each category of the cross-classification of response and time, a general linear models approach based on WLS can be used to produce Wald statistics for testing hypotheses. Park and Davis (PD) (1993) proposed a simple test of the missing data mechanism in incomplete repeated categorical data in the framework of the GSK method.

The test is an extension of the test of Little (1988) and uses a test criterion given in general form by Wald. The method is briefly summarized as follows: consider a single response variable that has c response categories (including the category for missing or unknown response) and n subjects with the response variable measured at t time points. Each of the subjects has a response profile belonging to one of $c^t - 1$ possible categories (examples: YYN is a

response profile corresponding to response 'yes', 'yes' and 'no' at 3 time points for a binary variable; YNM is a response profile corresponding to response 'yes', 'no' and 'missing' at 3 time points). We may define H strata according to the missing data patterns, and for the h^{th} pattern the regression models are defined as:

$$E[F(p_h)] = X_h\beta_h$$

with p_h being the vector of sample proportions, $F(p_h)$ a vector of u_h functions of p_h, X_h is a u_h model matrix and β_h is a n $\times$ 1 vector of unknown parameters. The model allows different estimators of β_h for $h = 1, \cdots, H$.

The missing data mechanism can be examined by testing the homogeneity of model parameters using the Wald statistic for $H_0 : \beta_1 = \cdots = \beta_H$. If H_0 holds, then the distribution of F_h does not depend on h and the missing data process may be considered MCAR. If H_0 does not hold, then the distribution of F_h is likely to depend on the missing data patterns and the missing data process is probably not MCAR. In addition to testing the missing data mechanism, the model allows other linear hypotheses to be tested for β.

The PD method requires at least moderately large samples for each stratum so that the estimates of coefficients are approximately normally distributed. When there are many strata and/or the strata sample sizes are considerably different, it is better to use a 2-strata approach (i.e., 'complete' data at all time points versus 'incomplete' with at least one missing observation). In settings where there is a dominant missing data pattern or a monotone missing data pattern this may be a reasonable approach. Its advantages are: it is useful for specific missing data patterns, it is flexible and easy to apply and reduces the number of parameters to be estimated. PD methodology can be implemented using standard statistical software such as the CATMOD procedure in SAS (Davis 1992). An application of this approach in longitudinal incomplete categorical QL data obtained from an international multicenter clinical trial is given below.

Example

We will use data from a study in operable breast cancer to demonstrate how the PD method may be used in a QL setting. See Section 4.3 for details of the dataset. Due to the rather small amount of missing data, we categorized the 2 QL variables as binary outcome (anxious, not anxious) plus a third category for missing and used a 2 strata approach: 'complete' vs 'incomplete'. We were mainly interested in formally testing if the parameter estimates

for the 'complete' were significantly different from the 'incomplete' data. We modeled the marginal probability of 'anxiety' (defined as a scale score >1) or 'burden related to hair loss' (defined as a scale score >1) at each time point. The vector of response functions was $p_i = (p_{0c}, p_{3c}, p_{6c}, p_{0i}, p_{3i}, p_{6i})$ where the subscript 'c' is for 'complete' data and 'i' for 'incomplete' data and the numbers represent the time (months). When the data are incomplete, the components of p_i must be calculated as ratios of sums of the multinomial proportions corresponding to the response profiles. In PROC CATMOD, these operations are specified as a series of linear, logarithmic and exponential transformations of the elements of the vector p_i^* of multinomial proportions. In general, a composite link function is used:

$$p_i = \exp(A_2 \log(A_1 p_i^*)).$$

The matrix A_1 has ct rows and as many columns as there are observed response profiles and A_2 is a $(c-1)t \times ct$ matrix.

In both examples, we fitted a saturated model with separate intercepts and linear and quadratic time effects for the 2 strata of 'complete' and 'incomplete'. A 3 degrees of freedom contrast was used to test whether the parameter estimates for the 'complete' differ significantly from those for the 'incomplete'. We did not explore any further model reduction.

The following were the results for the 'anxiety' scale (214 patients). Overall, the proportion with 'anxiety' was 0.48 at baseline (207 patients) and was rather stable for the subsequent 2 assessments (0.47 at 3^{rd} and 0.46 at 6^{th} month respectively). Anxiety scale compliance was not significantly associated with clinical or sociodemographic factors. Twenty-four of the two hundred and fourteen (11%) patients had at least one missing value in the 3 QL assessments and defined the stratum of incomplete data, the other 190 had complete data. Since the outcome measure had three possible values (no, yes or missing) and was assessed at 3 time points, there were $ct - 1 = 27 - 1 = 26$ possible response patterns (8 for 'complete' and 18 for 'incomplete'). We observed 11 missing response patterns (and 5 distinct missing patterns, see Table 7.2) with missing at the 3^{rd} month being the most frequent. The observed proportions with 'anxiety' at each time point are presented in Table 7.3. At each time point the proportion with 'anxiety' was higher in patients with 'incomplete' data, in particular at the 3^{rd} month. The saturated model results showed that there was no significant time effect and the Wald statistic indicated a highly significant difference between 'complete' and 'incomplete' cases (χ^2 13.09, 3 df, $p = 0.0044$).

We considered also the 'subjective burden related to hair-loss' scale (214 patients). Overall, the proportion with'burden' was 0.09 at baseline (202 patients) and increased to 0.57 and

Table 7.4: *IBCSG Study VI-14. Number of patients by response profiles (Burden related to hair loss).*

Response category (baseline, 3rd month, 6th month): N= No burden, Y = burden, M =missing																		
N	N	N	N	Y	Y	Y	Y	Y	Y	N	N	N	N	N	M	M	M	M
N	N	Y	Y	N	N	Y	Y	Y	M	Y	N	M	M	M	Y	Y	N	N
N	Y	N	Y	N	Y	N	Y	M	M	M	M	Y	N	M	Y	N	Y	N
62	15	23	71	4	1	4	7	1	1	3	1	2	4	3	6	2	1	3

Table 7.5: *IBCSG Study VI-14. Proportion with 'Burden related to hair loss'.*

	Baseline	3rd month	6th month
Incomplete data (n=27)	0.133	0.706	0.500
Complete data (n=187)	0.086	0.562	0.503

0.50 in the subsequent 2 assessments. Scale compliance was not significantly associated with clinical or sociodemographic factors. Twenty seven of the two-hundred and fourteen (13%) patients had at least one missing value and defined the stratum of 'incomplete' (11 distinct missing response patterns, see Table 7.4). Missing at baseline was the most frequent. The observed proportions with 'burden' at each time point are presented in Table 7.5. At each time point, except at 3^{rd} month, the proportion with 'burden' was similar in patients with 'complete' or 'incomplete' data. As before, we were interested in modeling the probability of 'burden' at different time points in the 2 strata. The saturated model results showed that there was a significant time effect and the Wald statistic indicated that the parameters for 'complete' and 'incomplete' cases were not significantly different (χ^2 1.73, 3 df, $p = 0.63$).

7.4 Hypothesis Testing for MNAR

Molenberghs, Goetghebeur and Lipsitz (1997) demonstrated that testing the assumptions of MAR or alternatively for MNAR is not trivial. The authors suggest that testing will almost

always rest on strong assumptions which are often untestable. Glynn, Laird and Rubin (1986) developed an argument to illustrate these issues. An illustration of this argument for QL data is given below.

Suppose in a clinical trial QL is assessed at two time points, e.g., pre-treatment Y_1 and post-treatment Y_2. Assume Y_1 is always observed and Y_2 is either observed ($t = 2$) or missing ($t = 1$). Let us further simplify the notation by suppressing dependence on parameters and additionally adopting the following conventions:

$$\begin{aligned} g(t|y_1, y_2) &= f(t|y_1, y_2), \\ p(t) &= f(t), \\ f_t(y_1, y_2) &= f(y_1, y_2|t). \end{aligned}$$

Equating the selection model and pattern-mixture model factorizations yields:

$$\begin{aligned} f(y_1, y_2)g(d = 2|y_1, y_2) &= f_2(y_1, y_2)p(t = 2), \\ f(y_1, y_2)g(d = 1|y_1, y_2) &= f_1(y_1, y_2)p(t = 1). \end{aligned}$$

Since we have only two patterns, this simplifies further to

$$\begin{aligned} f(y_1, y_2)g(y_1, y_2) &= f_2(y_1, y_2)p, \\ f(y_1, y_2)[1 - g(y_1, y_2)] &= f_1(y_1, y_2)[1 - p], \end{aligned}$$

of which the ratio yields:

$$f_1(y_1, y_2) = \frac{1 - g(y_1, y_2)}{g(y_1, y_2)} \frac{p}{1 - p} f_2(y_1, y_2).$$

All selection model factors are identified, as are the pattern-mixture quantities on the right hand side. However, the left hand side is not entirely identifiable. We can further separate the identifable from the non-identifiable quantities:

$$f_1(y_2|y_1) = f_2(y_2|y_1) \frac{1 - g(y_1, y_2)}{g(y_1, y_2)} \frac{p}{1 - p} \frac{f_2(y_1)}{f_1(y_1)}. \tag{7.1}$$

In other words, the conditional distribution of the second measurement given the first one, *in the incomplete first pattern*, about which there is no information in the data, is identified by equating it to its counterpart from the complete pattern, modulated via the ratio of the "prior" and "posterior" odds for dropout ($p/(1-p)$ and $g(y_1, y_2)/(1-g(y_1, y_2))$, respectively), and via the ratio of the densities for the first measurement.

Table 7.6: *EORTC Trial 08925. Cross tabulation of QL scores by dropout pattern.*

R=2			R=1	
	Y_2			Y_2
Y_1	No	Yes	Y_1	Missing
No	12	7	No	3
Yes	8	19	Yes	5

Table 7.7: *EORTC Trial 08925. Predicted counts for the MAR and MNAR models respectively.*

MAR			MNAR		
	Y_2			Y_2	
Y_1	No	Yes	Y_1	No	Yes
No	1.89	1.11	No	1.53	1.47
Yes	1.48	3.52	Yes	1.02	3.98

Thus, while an identified selection model is seemingly less arbitrary than a pattern-mixture model, it incorporates *implicit* restrictions. Indeed, precisely these are used in (7.1) to identify the component for which there is no information.

This clearly illustrates the need for sensitivity analysis. Due to the different nature of the selection and pattern-mixture models, specific forms for each of the two contexts will be presented in Chapters 8 and 9, respectively. In Chapter 8 we will describe a general strategy for fitting pattern-mixture models. Chapter 9 is devoted to a formal juxtaposition of several strategies for pattern-mixture modeling.

Let us consider the following example: the EORTC QLQ-C30 includes the question 'Are you limited in any way in doing either your work or doing household jobs?' with possible response categories 'no' and 'yes'. Suppose the QLQ-C30 was assessed at baseline $\boldsymbol{Y}_1$ and post-treatment $\boldsymbol{Y}_2$. In total, 54 patients had a baseline QL questionnaires in EORTC study 08925. Forty-six patients had assessments at both baseline and at the first assessment during treatment while the remaining 8 patients completed only the first questionnaires. The results

are presented in Table 7.6.

Factorizing the joint distribution of $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ and R and fitting a logistic model for the missing data mechanism yields

$$f(\boldsymbol{Y_1}, \boldsymbol{Y_2}, R = 1) = f(\boldsymbol{Y_1}, \boldsymbol{Y_2}) \frac{e^{(\beta_0 + \beta_1 \boldsymbol{Y_1} + \beta_2 \boldsymbol{Y_2})}}{1 + e^{(\beta_0 + \beta_1 \boldsymbol{Y_1} + \beta_2 \boldsymbol{Y_2})}}$$

We consider 2 particular cases: $\beta_2 = 0$ (MAR) and $\beta_1 = 0$ (MNAR). Such restrictions are necessary to ensure a unique solution. Both models are saturated in the sense that the predicted counts coincide with the observed data. The predicted counts for both the MAR and MNAR models are given in Table 7.7. Although differences are small in this case, there is no formal way to discriminate between the two models in terms of observed data.

Little (1995) suggests that underidentifiability is a serious problem with non-ignorable missing data models. There may be a problem in estimating the parameters of the missing data mechanism simultaneously with the parameters of the complete data model. Molenberghs, Goetghebeur and Lipsitz (1997) provided examples where models provided almost similar fits to the observed data, but yielded completely different predictions for the unobserved data.

7.5 Remarks

Two approaches of identifying the types of missing data in QL research have been discussed: (1) collecting information on why the QL questionnaires were not completed and (2) hypothesis testing of the missing data process. Both have their intrinsic difficulties. For example, in the first approach it may be difficult to collect information on the clinical CRF's on why questionnaires are missing. Often clinical CRF's are completed retrospectively by retrieving data from the patient's medical chart. If the information is not recorded in the patient's chart then it may be irretrievable. In addition, the person responsible for completing the clinical CRF's is seldom responsible for administering the QL questionnaire. Some cancer research organisations include a cover sheet with the QL questionnaire which includes questions requesting the reasons for missing questionnaires. However, if the patient does not

complete the QL questionnaire due to inadequate administrative procedures in the hospital then it is likely that the cover sheet will not be completed either.

Two methods for testing between MCAR and MAR have been implemented. The first method, based on a logistic regression analysis, was applied in the setting of postmenopausal advanced breast cancer where the efficacy of two second line hormonal treatments were compared. The focus was on the perceived adjustment to chronic illness scale (PACIS). In this advanced disease setting where the focus was on a health related scale, the hypothesis of MCAR was rejected, i.e., the probability of dropout was dependent on the previous QL score. The second method, based on an adaptation of WLS, was applied in an adjuvant setting in patients with operable breast cancer. The focus was on anxiety and burden related to hair loss. In this setting, more anxiety was observed in patients with incomplete data and thus missingness was related to the anxiety score, i.e., data were not likely to be MCAR. However, for the burden of hair loss scale no significant difference was observed between completers and non-completers indicating that the hypothesis of MCAR could not be rejected. Thus, one would conclude that the missing data mechanism may depend on the scale under investigation. It is important to check the assumptions about missingness on scale level, since on a questionnaire level there could be more than one coexisting missing data process. The missing data process may also vary between disease settings. For example, it may be more likely that missing data are MCAR or MAR in adjuvant settings whereas in advanced diseases the missing data may often be MAR or MNAR.

Heitjan and Basu (1996) investigated the consequences of mis-specifying the missing data mechanism. They demonstrated that MCAR and MAR have distinct consequences for data analysis. Distinguishing between the missing data mechanisms is necessary to determine which types of analysis are appropriate. For example, in the situation where data are not MCAR, analyses such as complete case analyses may be biased. In addition, graphical presentations of summary statistics (e.g., means or proportions) of available cases over time may be misleading since scores at later time periods may be seriously biased. In the logistic regression example the change score between the two previous assessments was predictive of dropout indicating that patients with a decreasing score were more likely to dropout. Thus, in this example imputation methods such as last value carried forward described in Section 3.3.1 are not suitable.

Both the logistic regression and PD approaches are based on modeling the missing data mechanism and may be sensitive to the model specification. They are both easy to apply in practice with existing software. The logistic regression is more suitable if the primary objective of the analysis is to investigate the dropout mechanism (e.g., dropout or not as

in the first example). The PD approach allows one to test if the 'completers' are different from the 'non-completers'. The PD approach requires at least moderately large samples for each stratum, and thus a two strata approach, which in fact also requires moderate to large sample sizes, was used.

In the PD approach the patient setting was represented by a 'healthy' patient population scheduled to attend regular appointments for adjuvant chemotherapy administration or visits during adjuvant endocrine therapy. The patient group size and the observed high compliance may not provide the ideal setting to test hypotheses about the missing data process. In addition, since power may be low, it is important to remember that accepting H_0 of homogeneity of stratum-specific parameters does not imply its correctness.

An alternative 2-step WLS approach has been studied by Lipsitz, Laird and Harrington (1994). The first step (estimates of multinomial probabilities) uses maximum likelihood. The second step (noniterative WLS) is the same as the PD. It has the advantage of not being model dependent, but, it must be carried out separately in each covariate stratum. They also provide a test for the null hypothesis of MCAR versus the alternative of MAR. Stratification according to missing data patterns has been considered also by Dawson (1994) (continuous outcome, stratification of summary statistic tests). He found that stratification of the analysis tends to result in an increase of power and improves the robustness to violations of missing data.

Analysis of longitudinal data is even more complex when data may be missing for several reasons. In the two examples provided above it was shown that in some situations QL data were unlikely to be MCAR. This is not surprising as in cancer clinical trials, especially in advanced disease, one would expect patients with a poorer health related QL to complete fewer QL questionnaires because they are too ill or because they drop out of the study early. In some cases it may be possible to determine the QL scores of a random sample of patients by using alternative modes of administration such as telephone interview or by obtaining proxy scores from members of the patients family.

Molenberghs, Goetghebeur and Lipsitz (1997) showed that sensitivity to model specification may be a serious problem. When fitting a model certain assumptions have to be made about the relationship of the missing data process and the unobserved data. Since these assumptions are fundamentally untestable it is prudent to calculate estimates on a variety of models, rather than relying exclusively on one model, especially when the amount of missingness is considerable.

Chapter 8

Continuous Longitudinal Data

8.1 Introduction

Building on the methodology for incomplete data developed in Chapter 2, the current chapter presents two examples of analyzing longitudinal continuous measurements with incomplete data. In Chapter 6 we studied longitudinal data using both summary measures and summary statistics. However, we concluded that these methods were wasteful and did not account for dropout adequately. In this Chapter we focus on the linear mixed model. We illustrate its use when considering two alternative factorizations of the complete data $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$. As introduced in Chapters 2 and 3, conditioning on $\boldsymbol{R}_i$ results in a pattern-mixture model, while conditioning on $\boldsymbol{Y}_i$ results in a selection model; both are discussed by Little (1995).

In Section 8.2, we present a selected set of plots to underpin the model building. We distinguish between two modes of display: (1) individual profiles and (2) averaged over (sub)populations. Both ways are used to present three fundamental aspects of the longitudinal structure: (1) the average evolution; (2) the variance function, (3) the correlation structure. Each of those will be discussed in turn. In addition, the variogram will be discussed.

In Section 8.3, we describe the linear mixed model and show how it may be extended for use in either a selection or pattern-mixture modeling framework. In Section 8.4 we introduce the milk dataset, a commonly used dataset in the statistical literature. We show how both

pattern-mixture and selection modeling frameworks can be used to analyze the data and we illustrate how an informal sensitivity analysis on incomplete longitudinal data may be conducted. In Section 8.5 we show how the definition of QL fits in with the analysis of longitudinal data. We illustrate how additional information may be obtained using longitudinal modeling techniques.

8.2 Graphical Exploration

8.2.1 Introduction

With longitudinal data it is useful to explore the data using graphical techniques before advancing to model fitting (Diggle, Liang, and Zeger 1994). These may include both graphical exploration of individual responses and means plotted against time, and an exploration of the variance-covariance structure using residual plots, scatter plots and the variogram (Diggle, Liang and Zeger 1994).

8.2.2 Example

As an illustration of the use of graphical exploration of longitudinal continuous measurements, we used the data reported in Section 4.5.

Individual profiles and means
An obvious plot to consider when exploring longitudinal data is a plot of the response variable against time. Although the QLQ-C30 Global health status score is on a 0 to 100 scale, it is derived from two seven point categorical response items. As a result, in a graphical exploration of individual profiles patients with identical profiles will be superimposed which may result in a misleading plot. To overcome this problem one could generate random numbers r_i from a uniform distribution (e.g., -3 to 3 points) for each individual patient (jittering). Let Y_{ij} represent the QL score for patient i ($i = 1, \cdots, N$) at time point j ($j = 1, \cdots, T$) . The generated random numbers r_i are then added to the QL scores for each individual patient i, i.e., $G_{ij} = Y_{ij} + r_i$ for all time points j. Even with the attributes of the jittered score, in clinical trials that include many patients the plot may become too

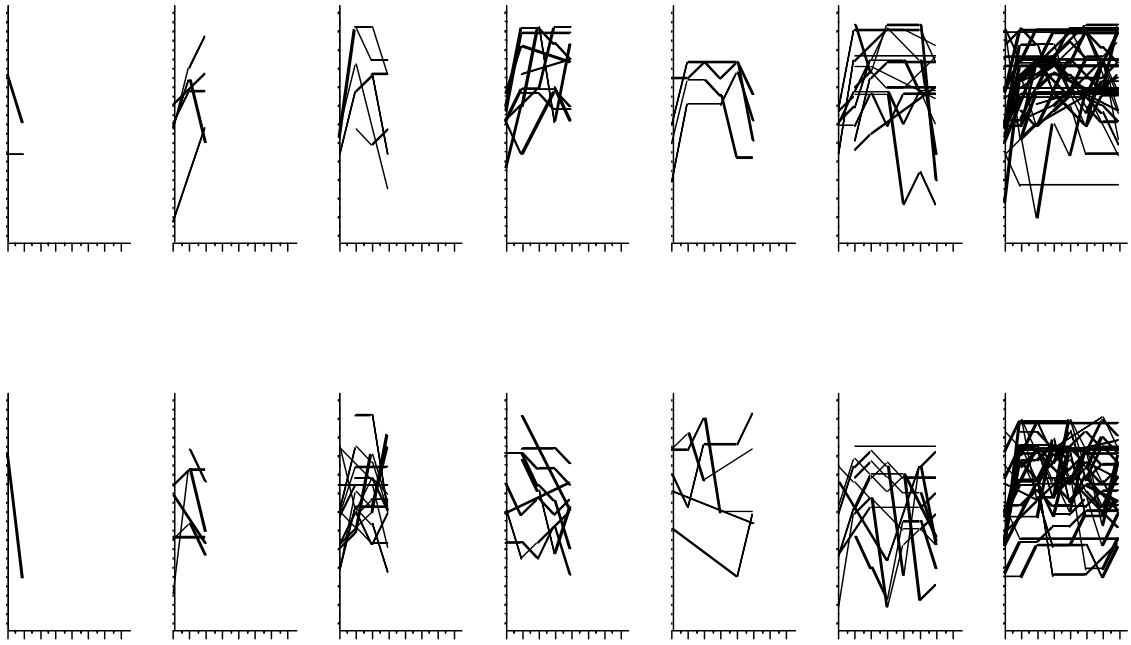

Figure 8.1: *EORTC Trial 30893. Individual profiles by dropout pattern and treatment: Top: Orchidectomy, Bottom: Orchidectomy + mitomycin C.*

cluttered. One solution is to divide patients into subgroups. In Figure 8.1 the modified scores G_{ij} are presented over time j according to treatment group and dropout pattern. Note there is considerable variation in between-patient and within-patient scores. Figure 8.2 presents the mean profiles by dropout time and treatment group. For the majority of subgroups, except for those including patients who drop out at week 6, there is an increase in mean QL scores between baseline and week 6. This is in line with clinical experience which suggests that cancer-related symptoms such as fatigue are often alleviated within a few days after orchidectomy leading to a better QL score. The QL scores in the orchidectomy alone arm appear to increase initially for all subgroups and only decrease before patients dropout of the study. The QL scores in the orchidectomy + MMC arm do not appear to increase to the same extent as those for the orchidectomy alone arm which is probably due to the toxicity observed with MMC.

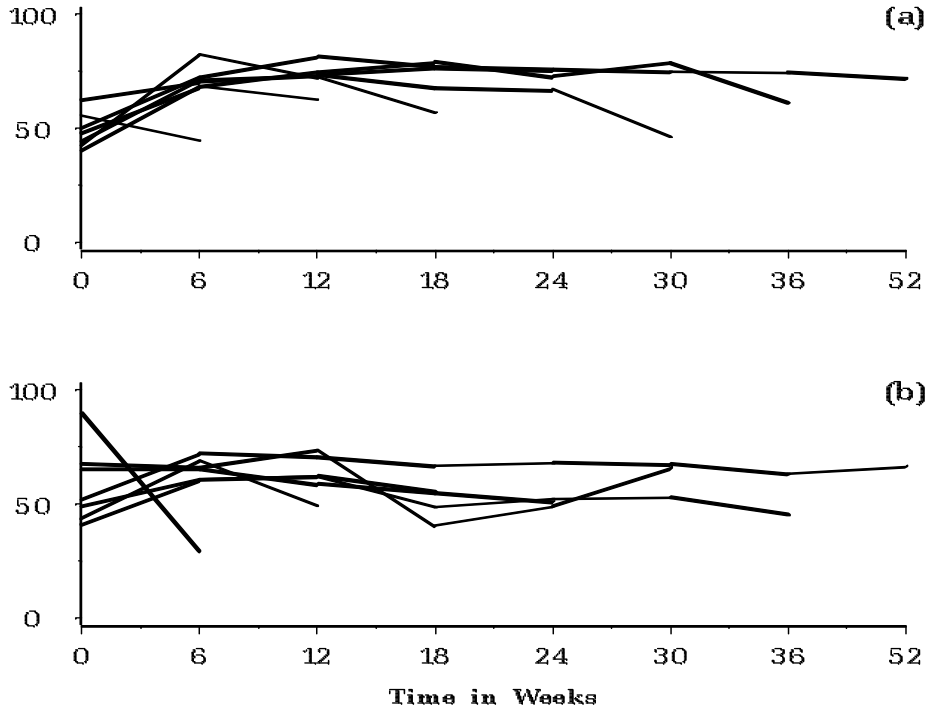

Figure 8.2: *EORTC Trial 30893. Mean profiles by dropout pattern and treatment a) orchidectomy b) orchidectomy + mitomycin C.*

Variance-covariance structure

The variance-covariance structure was investigated using several methods. Initially, an 8-dimensional scatter plot matrix of the data was generated as shown in Figure 8.3. The diagonal elements display the distribution of QL scores at each assessment time point. For presentation purposes the scores were divided into categories (<10, 10-30, 30-50, 50-70, 70-90, >90). The histograms confirm the earlier finding of an initial improvement in QL scores at week 6. Thereafter, the distributions are similar, allowing for the decreasing number of patients contributing to plots at later time points. The scatter plots (off diagonal) of assessments taken closer together (e.g., near the diagonal) appear to exhibit larger correlations than those taken further apart, suggesting perhaps an autoregressive covariance structure. Standardized residuals were obtained using ordinary least squares estimates. Similar plots as for Figures 8.1 and 8.3 were generated using the standardized residuals (data not shown).

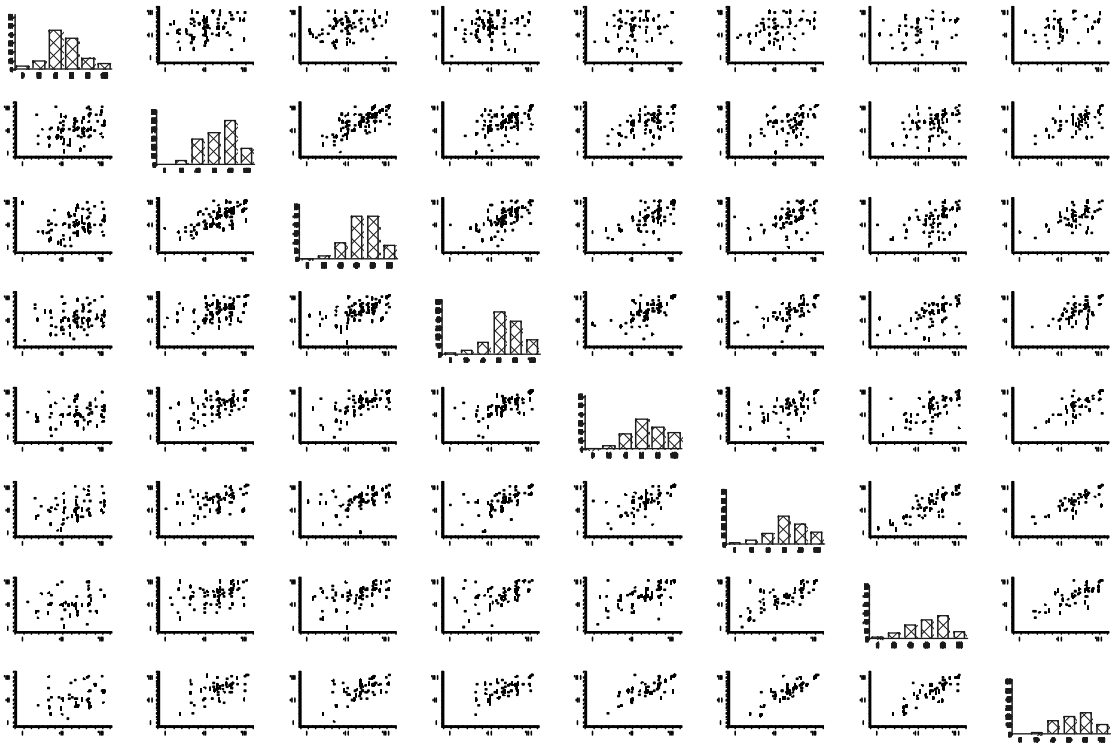

Figure 8.3: *EORTC Trial 30893. Eight dimensional scatter plot matrix.*

Variogram

Diggle (1994) and Diggle, Liang and Zeger (1994) promote the so-called semi-variogram to picture the variance components: measurement error, serial correlation, and random effects. It is easily estimated even with irregular observation times (but in such cases may require some smoothing). Given a stationary mean-zero stochastic process $Y(t)$, where t denotes time, with constant variance, the variogram is defined as

$$V(u) = \frac{1}{2} E\left\{[Y(t) - Y(t-u)]^2\right\}.$$

A specific form is discussed in Appendix A.1.

We constructed the sample variogram for our setting in Figure 8.4. The intercept of the variogram provides a rough estimate of the measurement error. This shows that it represents about 40% of the total process variance (indicated by the full horizontal line). The variogram increases over time indicating a decrease in correlation as time increases. This decay in serial

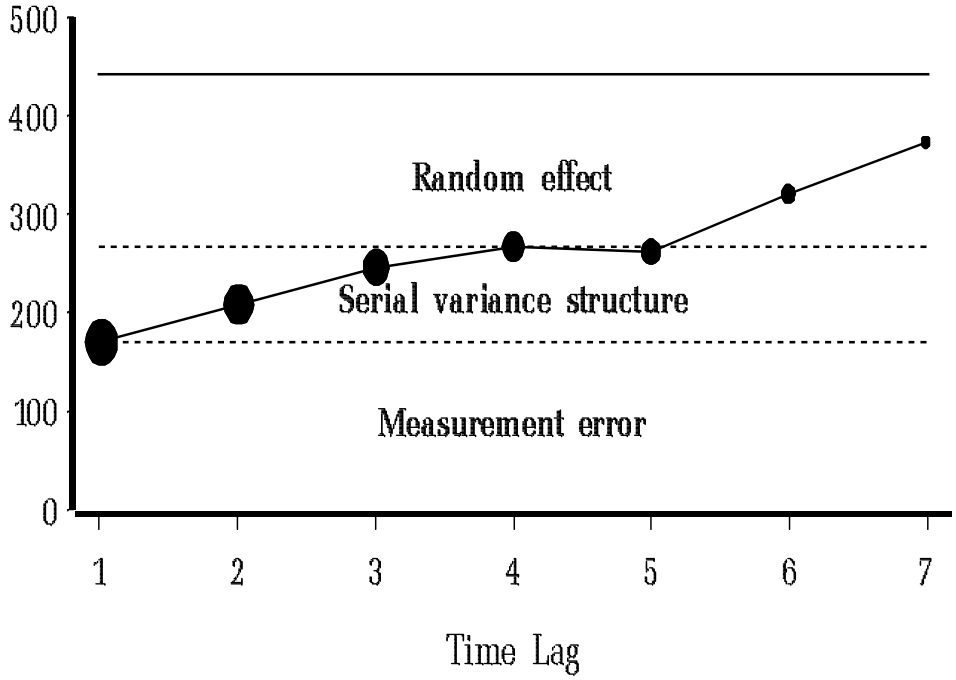

Figure 8.4: *EORTC Trial 30893. Variogram.*

correlation was confirmed in the scatterplot matrix: moving farther from the diagonal (for observations that are further apart in time), the degree of correlation appears to decrease. The variogram appears to level off at lag 4 and increases again at lags 6 and 7. The latter increase is unreliable due to the small numbers of observations (illustrated by the size of the dot). The difference between the estimate of the variogram at lag 4 and the total variance (i.e., the process variance), provides an estimate of the subject-level component of the variability (i.e., the random effect). In summary, the variogram suggests a random effect, a serial decay in correlation and measurement error. However, since the variogram was built on just 8 repeated measures with few observations at lags 5, 6 and 7 the aforementioned components are difficult to assess. Nevertheless, the construction of the variogram is an important tool to formulate an initial model, and in particular the variance components therein.

8.3 Linear Mixed Model

8.3.1 Introduction

A continuous outcome, or an appropriate transformation of it, is often regarded as drawn from an approximately normal distribution. For such outcomes, the linear mixed model (Laird and Ware 1982) is well developed. Let Y_{ij}, grouped into vector $\boldsymbol{Y}_i$, represent the QL score for patient i $(i = 1, \cdots, N)$ at time point j $(j = 1, \cdots, T)$. The mixed-effects model can be written as

$$\boldsymbol{Y}_i = X_i\boldsymbol{\alpha} + Z_i\boldsymbol{a}_i + \boldsymbol{w}_i(t) + \boldsymbol{\varepsilon}_i, \tag{8.1}$$

where X_i and Z_i are design matrices for fixed and random effects, respectively, $\boldsymbol{\alpha}$ are fixed effects and $\boldsymbol{a}_i$ are random-effects parameters with $\boldsymbol{a}_i \sim N(0, D)$. Further, $\boldsymbol{w}_i$ are realizations of a Gaussian stochastic process and $\boldsymbol{\varepsilon}_i$ represents measurement error. The variance is

$$\text{Var}(\boldsymbol{Y}_i) = Z_i D Z_i^{'} + \sigma^2 H_i + \tau^2 I,$$

where σ^2 refers to the variance of the serially correlated process, $H_i = (h_{jk}) = (\rho(t_j, t_k))$ to the associated correlation matrix, τ^2 pertains to the measurement error variability and finally I is a $T \times T$ identity matrix.

Let $D_i = d$ identify dropout time, where $D_i = T + 1$ if the sequence of measurements is complete.

8.3.2 Selection Model

As discussed in Section 2.4 selection models arise when the joint likelihood of the measurement process and the dropout process is factorized as:

$$f(\boldsymbol{y_i}, D_i \mid X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i} \mid X_i, \boldsymbol{\theta}) f(D_i \mid \boldsymbol{y_i}, X_i, \boldsymbol{\psi})$$

where the first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the measurements. The linear mixed-effects model is used to model the responses, as in (8.1), together with a logistic regression to describe the dropout process. Let $g_j(y_{ij}, h_{ij})$ represent the conditional probability of

dropout at time j given the measurement at time j and the history of the measurement process $h_{ij} = (Y_{i1}, Y_{i2}, \cdots, Y_{ij-1})$ up until time $j-1$. Modeling the dropout mechanism may be simplified by allowing dropout to depend on the current measurement and immediately preceding measurement only with corresponding regression coefficients ψ_1 and ψ_2. This leads to the logistic expression

$$\text{logit}(g_j(y_{ij}, h_{ij})) = \psi_0 + X_i\boldsymbol{\psi}_c + \psi_1 y_{ij} + \psi_2 y_{ij-1}$$

where ψ_0 represents the intercept and $\boldsymbol{\psi}_c$ is a vector of parameters for covariates X_i. A likelihood ratio test is used to test the hypothesis of $\psi_1 = 0$ (i.e., MAR) and similarly to test the hypothesis of $\psi_1 = \psi_2 = 0$ (i.e., MCAR). Informative dropout models which combine a linear mixed model for the measurements, together with a logistic regression were fitted in OSWALD (a suite of macros written in SPlus) (Smith, Robertson and Diggle 1996) and a study specific program written in GAUSS, a modified version of a program written by Verbeke and Molenberghs (1997) and Thijs, Molenberghs and Verbeke (1999).

8.3.3 Pattern-Mixture Model

In contrast with selection models, pattern-mixture models result when the joint distribution of $\boldsymbol{Y_i}$ and D_i is factorized as

$$f(\boldsymbol{y_i}, D_i \mid X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{D_i} \mid D_i, X_i, \boldsymbol{\theta}) f(D_i \mid X_i, \boldsymbol{\psi}).$$

Thus, the model for the responses is written as a conditional model, depending on the particular missingness pattern. The marginal distribution of the longitudinal measurements is a mixture of the conditional distributions, given the pattern of missingness. Parameters describing the model for $\boldsymbol{Y_i}$ are estimated in each stratum (determined by D_i), and the overall parameters are obtained using a weighted average of these estimates, weighted by the proportion of subjects in each stratum. The parameters describing the distribution of the missingness indicators themselves are generally considered a nuisance, and in fact a model for

$$f(D_i \mid X_i, \boldsymbol{\psi})$$

can be relatively simple. Indeed, since the measurement process is not involved, a model is often based merely on the multinomial probabilities of occurrence of the various dropout patterns. The parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missing process respectively.

Since the measurement model is dependent on dropout, the equation given in (8.1) is modified to reflect this:

$$\begin{cases} \boldsymbol{Y_i} = X_i\boldsymbol{\alpha}(d_i) + Z_i\boldsymbol{a}_i + \boldsymbol{w}_i(t) + \boldsymbol{\varepsilon}_i, \\ \qquad \boldsymbol{a}_i \sim N(0, D(d_i)), \\ \qquad \boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i(d_i)). \end{cases} \tag{8.2}$$

Thus, a priori, the fixed effects as well as the covariance parameters are allowed to vary unconstrained according to the dropout pattern. In models where a particular effect, such as treatment effect or treatment-by-baseline interaction, is pattern dependent an additional calculation is required to obtain the marginal effect. Let us illustrate this for the marginal treatment effect. Let γ_d represent the parameters for the treatment effect in pattern d $(d = 1, \cdots, P)$, and let π_d denote the proportion of patients in each of the P patterns. Then the estimate of the marginal treatment effect β is given by:

$$\beta = \sum_{d=1}^{P} (\gamma_d \pi_d). \tag{8.3}$$

The variance is obtained using the delta method. Precisely, the matrix of derivatives of β is given by:

$$\begin{aligned} A &= \frac{\partial \beta}{\partial(\gamma_1, \gamma_2, \cdots, \gamma_P, \pi_1, \pi_2, \cdots, \pi_P)} \\ &= (\pi_1,\ \pi_2,\ \cdots, \pi_P,\ \gamma_1,\ \gamma_2,\ \cdots, \gamma_P). \end{aligned}$$

An asymptotic variance expression of the treatment parameter is

$$Var(\beta) = AVA^T, \tag{8.4}$$

where

$$V = \left(\begin{array}{c|c} Var(\boldsymbol{\gamma}) & 0 \\ \hline 0 & Var(\boldsymbol{\pi}) \end{array}\right).$$

The estimate of the variance-covariance matrix of the estimates $\hat{\gamma}_d$ is obtained from standard statistical software (e.g., the SAS procedure MIXED). Since the proportions of patients in each dropout pattern form a multinomial distribution the covariance matrix can be estimated as follows: $\text{Var}(\boldsymbol{\pi}) = [\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']/n$ where $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_P)'$. A Wald statistic for the hypothesis of no treatment effect $\beta = 0$ is then given by $\beta[AVA']^{-1}\beta'$, which follows approximately a χ^2_1 null distribution.

8.4 The Milk Protein Content Trial

8.4.1 Introduction

The primary objective of the milk protein experiment was to describe the effects of diet on the mean response profile of milk protein content over time. Previous analyses of the same data are reported by Diggle (1990, Chapter 5), Verbyla and Cullis (1990), Diggle, Liang, and Zeger (1994), and Diggle and Kenward (1994), under different assumptions and with different modeling of the dropout process. Diggle (1990) assumed random dropout whereas Diggle and Kenward (1994) concluded that dropout was non-random, based on their selection model. It has already been noted in Chapter 7 that appropriate care should be taken with non-random selection models for their reliance on unverifiable assumptions.

In addition to the usual problems with this type of models, serious doubts have been raised about even the appropriateness of the "dropout" concept in this study. Cullis (1994) warned that the conclusions inferred from the statistical model are very unlikely since usually there is no relation between dropout and a relatively low level of milk protein content. In the discussion of the Diggle and Kenward (1994) paper one is informed by Cullis that Valentine, who originally conducted the experiment, had previously revealed the real reasons for dropout. The explanation elucidates that the experiment terminated when feed availability declined in the paddock in which animals were grazing. Thus, this would imply that a non-random dropout mechanism is very implausible. A non-random dropout mechanism would wrongly relate dropout to response while on the contrary dropout depends on food availability only. Thus, there are actually no dropouts but rather five cohorts representing the different starting times. Together with Cullis (1994) we conclude that especially with incomplete data a statistical analysis should not proceed without a thorough discussion with the experimenters.

The complex and somewhat vague history of the dataset probably is the main cause for so many conflicting issues related to the analysis of the milk data. At the same time, it becomes a perfect candidate for sensitivity analysis. Modeling will be based upon the linear mixed effects model with serial correlation (8.1), introduced in Section 8.3. In Section 8.4.2 we examine the validity of the conclusions made in Diggle and Kenward (1994) by incorporating subject matter information into the method of analysis. As dropout was due to design the method of analysis should reflect this. We will investigate two approaches. The first approach involves restructuring the dataset and then analyzing the resulting dataset using a

selection modeling framework, whereas the second method involves fitting pattern-mixture models taking the missingness pattern into account. Both analyses consider the sequences as *unbalanced in length* rather than as a formal instance of dropout.

8.4.2 Informal Sensitivity Analysis

Since there has been some confusion about the actual design employed we cannot avoid making subjective assumptions such as the following: several matched paddocks are randomly assigned to one of three diets: barley, lupins or a mixture of the two. The experiment starts as the first cow experiences calving. At the end of the first five weeks, all 79 cows have entered their randomly assigned, randomly cultivated paddock. By week 19, all paddocks appear to approach the point of exhausting their food availability (in a synchronous fashion) and the experiment is terminated for all animals simultaneously.

All previous analyses assumed a fixed date for entry into the trial and the crucial issue then becomes how the dropout process should be handled and analyzed. However, it seems intuitive that since entry into the study was at random time points (i.e., after calving) and since the experiment was terminated at a fixed time point, that this time point should be the reference for all other time points. It is therefore also appealing to reverse the time axis and to analyze the data backwards, starting from time of dropout. Under the aforementioned assumptions we have found a partial solution to the problem of potentially non-random dropout since dropout has been replaced by ragged entry. Note however that a crucial simplification arises: since entry into the trial depends solely on calving and gestation it can be thought of as totally independent of the unobserved responses.

A problem with the alignment lies in the fact that virtually all cows showed a very steep decrease in milk protein content immediately after calving, lasting until the third week into the experiment. This behavior could be due to a special hormone regulation of milk composition following calving which lasts only for a few weeks. Such a process is likely to be totally independent of diet and, probably, can also be observed in the absence of food, to the expense of the animal's natural reserves. Since entry is now ragged, the process is spread and influences mean response level during the first eight weeks. Of course one might construct an appropriate model for the first three weeks with a separate model specification, in analogy to the one used in Diggle and Kenward (1994). Instead, we prefer to ignore the first three weeks, analogous in spirit to the approach taken in Verbyla and Cullis (1990). Hence, we have time series of length 16, with some observations missing at the beginning. Figure 8.5 displays

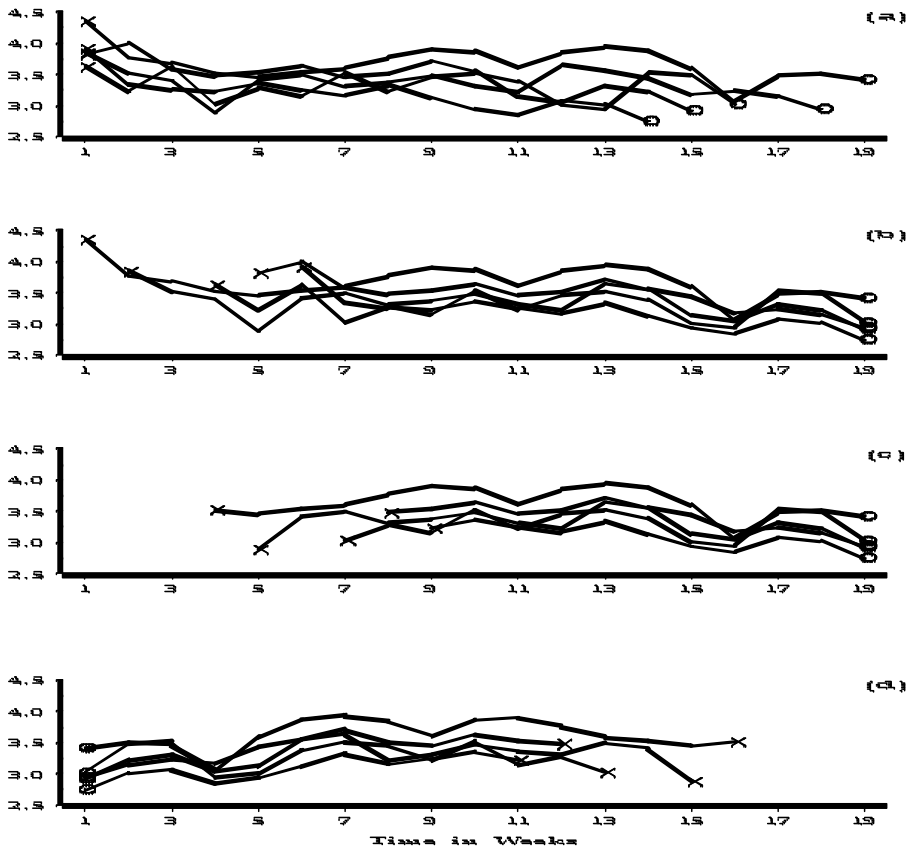

Figure 8.5: *Milk Protein Content Trial. Data manipulations on 5 selected cows. (a) Raw profiles; (b) Right aligned profiles; (c) Deletion of the first three observations; (d) Profiles with time reversal.*

the data manipulations for 5 selected cows. In Figure 8.5a the raw profiles are shown. In Figure 8.5b the plots are right aligned. Figure 8.5c illustrates the protein content levels for the 5 cows with the first three observations deleted and Figure 8.5d presents these profiles when time is reversed. In order to explore the patterns after transformation, we plotted the newly obtained mean profiles. Figures 8.6a and 8.6b display the mean profiles before and after the transformation, respectively. Notice that the mean profiles have become parallel in Figure 8.6b. To address the issue of correlation we shall compare the two variograms (see also Section 8.2). The two pictures shown in Figure 8.7 are very similar although slight differences

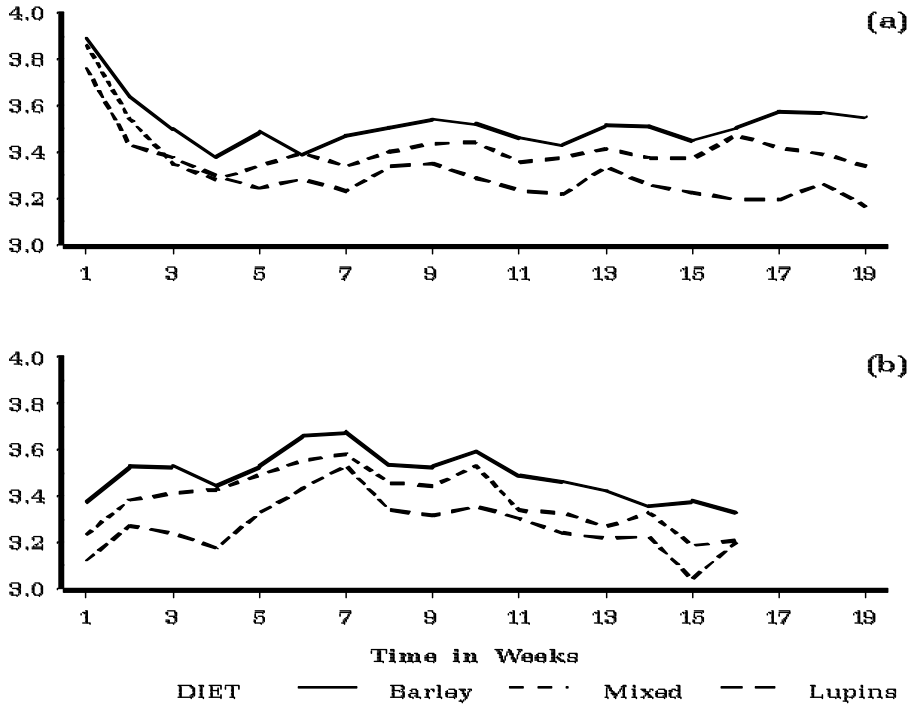

Figure 8.6: *Milk Protein Content Trial. Mean response profiles on the original data and after aligning and reverting.*

can be noted in the estimated process variance which is slightly lower after transformation. Complete decay of serial correlation appears to happen between time lags 9 and 10 in both variograms. There is virtually no evidence of random effects as the serial correlation levels off towards the process variance.

Table 8.1 presents the maximum likelihood estimates for the Diggle and Kenward MCAR model and the corresponding estimates after aligning and reverting. Analogous in principle to Diggle and Kenward (1994), time was taken as linear in the model. We notice that the mean parameters and standard errors remain similar. Although they are, strictly speaking, not directly comparable, the random intercept, already suspected to be negligible after studying the variogram, can be excluded after looking at the random intercept estimate which is very close to zero. Following Diggle and Kenward (1994), no attempt to model an increase towards the end of the experiment is made. Indeed the newly obtained profiles would rather suggest a decrease. As in the Diggle and Kenward (1994) paper we observed that the mean

Table 8.1: *Milk Protein Content Trial. Maximum likelihood estimates (standard errors) of random and non-random dropout models, fitted to the milk protein contents data. Dropout starts from week 15 onwards.*

	DK (CRD)	After aligning and reverting (CRD)
Mean parameters		
Intercept	3.56 (0.04)	3.45 (0.06)
Lupins	-0.21 (0.05)	-0.21 (0.08)
Mixed	-0.10 (0.05)	-0.12 (0.08)
Variance parameters		
Variance random intercept	1.424e-008	1.167e-008
Serial variance	0.094	0.135
Measurement error variance	0.020	0.015
Serial process parameter	0.211	0.097

protein response profile for the barley diet was consistently higher than for the mixed diet and similarly the mean protein response profile for the mixed diet was consistently higher than for the lupins diet. We reject the null hypothesis of no diet effect (F=1.98 on 32 d.f., P=0.001).

The analysis using aligned and reverted data shows little difference if compared to the original analysis by Diggle and Kenward (1994). It would be interesting to know what mechanisms determined the systematic increase and decrease observed for the three parallel profiles illustrated in Figures 8.6b and 8.8. It is difficult to envisage that the parallelism of the profiles and their systematic peaks and troughs shown in Figures 2b and 4 are due entirely to chance. Indeed, many of the previous analyses debated the influence on variability of factors common to the paddocks cultivated with the three different diets (e.g. meteorological factors) that had not been reported by the experimenter. These factors may account for a large amount of variability in the data. Hence, the data exploration performed in this analysis may prove to be a useful tool in gaining insight into the response process. For example, we notice that after transformation, the inexplicable trend towards an increase in milk protein content, as the paddocks approach exhaustion has, in fact, vanished or even reverted to a possible decrease. This was also confirmed in the stratified analysis where the protein level content

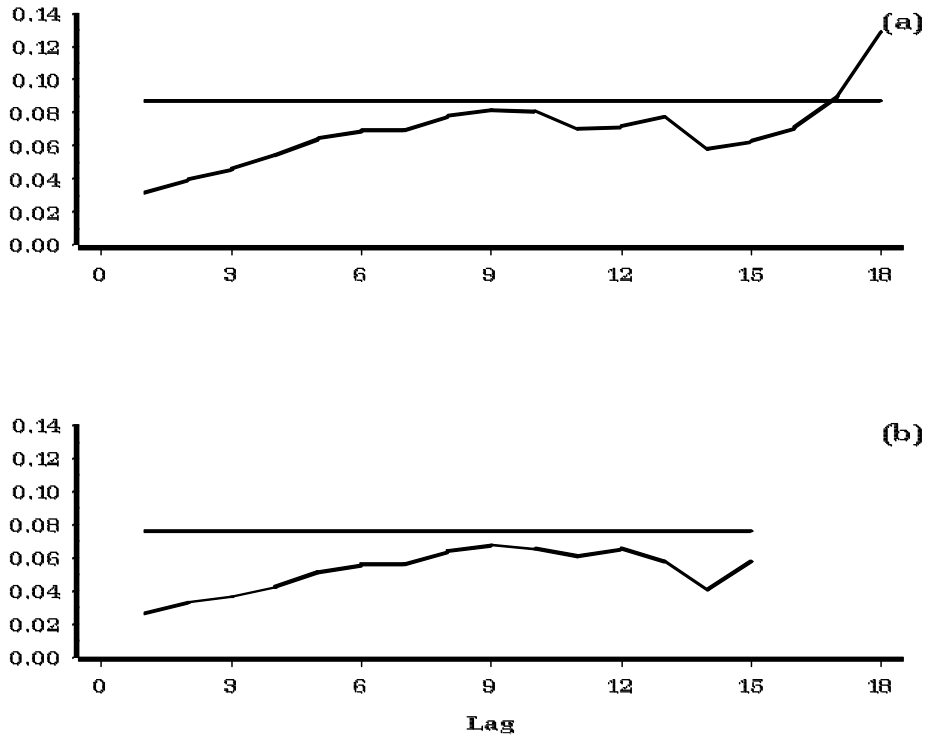

Figure 8.7: *Milk Protein Content Trial. Variogram for the original data and after aligning and reverting.*

tended to decrease prior to termination of the experiment (see Figure 8.8).

An alternative method of analysis is based on the premise that the protein content levels form distinct homogenous subgroups of cows based on their dropout pattern. This leads very naturally to pattern-mixture models. Parameters in (8.1) are now made to depend on pattern, as in (8.2). In its general form, the fixed effects as well as the covariance parameters are allowed to vary unconstrained according to the dropout pattern. Alternatively, simplifications can be sought. For example, diet effect can vary linearly with pattern or can be pattern-independent. In the latter case, this effect becomes marginal. When the diet effect is pattern dependent, an extra calculation is necessary to obtain the marginal diet effect. Precisely, the marginal effect can be computed as in (8.3) while the delta method variance expression is given by (8.4).

Denoting the parameter for diet effect $\ell = 1, 2$ (difference with the barley group) in pattern

$t = 1, 2, 3$ by $\beta_{\ell t}$ and letting π_t be the proportion of cows in pattern t, then the matrix A assumes the form

$$A = \frac{\partial(\beta_1, \beta_2)}{\partial(\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \pi_1, \pi_2, \pi_3)} \tag{8.5}$$

$$= \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & 0 & 0 & 0 & \beta_{11} & \beta_{12} & \beta_{13} \\ 0 & 0 & 0 & \pi_1 & \pi_2 & \pi_3 & \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}. \tag{8.6}$$

Note that the simple multinomial model for the dropout probabilities could be extended when additional information concerning the dropout mechanism is available. For example, if covariates are known or believed to influence dropout, the simple multinomial model can be replaced by logistic regression or time-to-event methods (Hogan and Laird 1997).

Table 4.6 presents the dropout pattern by time in each of the three diet groups. As few dropouts occured in weeks 16, 17 and 19 these three dropout patterns were collapsed into a single pattern. Thus three patterns remain with 20, 18 and 41 cows, respectively. The corresponding pattern probabilities are

$$\widehat{\boldsymbol{\pi}} = (0.253160.227850.51899)', \tag{8.7}$$

with asymptotic covariance matrix

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\pi}}) = \begin{pmatrix} 0.00239 & -0.00073 & -0.00166 \\ 0.00073 & 0.00223 & -0.00150 \\ 0.00166 & -0.00150 & 0.00316 \end{pmatrix}. \tag{8.8}$$

These figures, apart from giving an indication of the relative importance of the various patterns, will be needed to calculate marginal effects (such as marginal treatment effect) from pattern-mixture model parameters.

The model fitting results are presented in Table 8.2. The most complex model for the mean structure assumes a separate mean for each diet by time by dropout pattern combination. As the variogram indicated no random effects the covariance matrix was taken as first order autoregressive with a residual variance term $\sigma_{jk} = \sigma^2 \rho^{|j-k|}$. Also the variance-covariance parameters are allowed to vary according to the dropout pattern. This model is equivalent to including time and diet as covariates in the model and stratifying for dropout pattern and provides a starting point for model simplification through backward selection. The protein

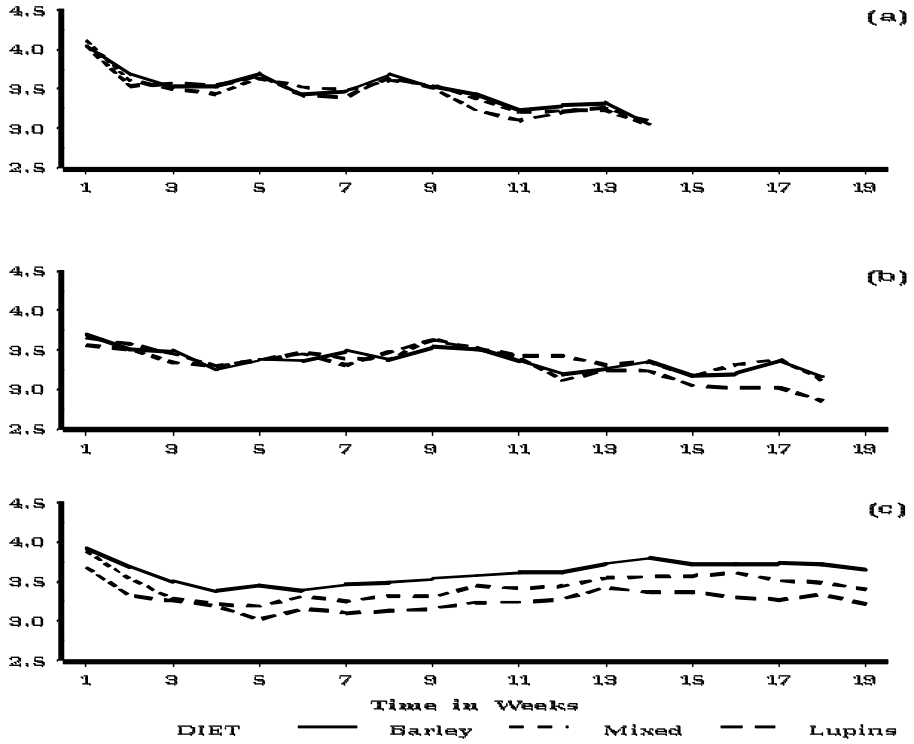

Figure 8.8: *Milk Protein Content Trial. Mean response level per diet and per dropout pattern.*

content levels over time are presented by pattern and diet in Figure 8.8. Note that the protein content profiles appear to vary considerably according to missingness pattern and time. Additionally, Diggle and Kenward (1994) suggested an increase in protein content level towards the end of the experiment. This observation is not consistent for the three plots in Figure 8.8. In fact, there is a tendency for a decrease in all diet by pattern subgroups prior to dropout.

To simplify the covariance structure presented in Model 1, Model 2 assumes the residual covariance parameter is equal in the three patterns. The likelihood ratio test indicates that Model 2 compares favorably with Model 1, suggesting a common residual variance (measurement error component) parameter (2 for the three groups; see Table 8.2 for details). However, comparing Model 3 with Model 2 we reject a common variance-covariance structure in the three groups.

Next, we investigate the mean structure. In Model 4, the three-way interaction between

Table 8.2: *Milk Protein Content Trial. Model fit summary for pattern-mixture models.*

	Mean	Covar
1	Full interaction	AR1(d), meas(t)
2	Full interaction	AR1(d), meas
3	Full interaction	AR(1), meas
4	Two-way interactions	AR1(d), meas
5	diet, time, pattern, diet×time, diet×pattern	AR1(d), meas
6	diet, time, pattern, diet×time, time×pattern	AR1(d), meas
7	diet, time, pattern, diet×pattern, time×pattern	AR1(d), meas
8	time, pattern, time×pattern	AR1(d), meas
9	time, diet(time)	AR(1), meas
10	time, diet	AR(1), meas

	par	-2ℓ	Ref	G^2	df	p
1	162	-474.93				
2	160	-470.49	1	4.44	2	0.109
3	156	-428.26	2	42.23	4	<0.001
4	100	-439.96	2	30.53	60	0.999
5	70	-202.40	4	237.56	30	<0.001
6	96	-430.55	4	9.41	4	0.052
7	64	-405.04	4	35.92	36	0.520
8	58	-378.22	7	26.82	6	<0.001
			6	52.33	38	0.061
9	60					
10	24					

Covar:	Covariance model	df:	Degrees of freedom
Par:	Number of parameters	P:	P-value
-2ℓ:	-2 times log-likelihood	AR(1):	Autoregresive order 1
Ref:	Reference model	d:	By dropout pattern
G^2:	Likelihood ratio test statistic		

pattern, time, and diet is removed. This simplified model is acceptable when contrasted to Model 2, based on $p = 0.999$. Models 5, 6 and 7 are fitted to investigate the pairwise interaction terms. Comparing Models 5 and 4 suggests a strong interaction between dropout pattern and time. Model 6 results in a borderline decrease in goodness of fit ($p = 0.052$). From Table 8.2 we observe that Model 7 is a plausible simplification of Model 4. Moreover, there is an apparent lack of fit for Model 8, which only includes one interaction term, time

and pattern, when compared to Model 7. In conclusion, among the models presented, Model 7 is the preferred one to summarize the data as it is the simplest model consistent with the data. However, Model 6 should be given some attention as well. In analogy to Diggle and Kenward (1994), we attempted to include time as a separate linear factor for the first three weeks and the subsequent 16 weeks. These models did not improve the fit (results not shown).

The objective of the experiment was to assess the influence on diet on protein content level. With selection models, the corresponding null hypothesis of no effect can be tested using, for example, the standard F tests on two numerator degrees of freedom as provided by the SAS procedure MIXED or similar software. In the pattern-mixture framework, such a standard test can be used only if the treatment effects do not interact with pattern. Otherwise, the marginal treatment (diet) effect has to be determined as in (8.3) and the delta method can be used to test the hypothesis of no effect. In Model 6 the diet effect is independent of pattern while the reverse holds for Model 7. Reparameterizing Model 6 by including the diet effect and diet and time interaction as one effect in the model provides us with an appropriate F test for the three diet profiles. The F test rejects the null hypothesis of no diet effect ($F = 1.57$ on 38 degrees of freedom, $p = 0.015$). In the corresponding selection model, Model 9, we remove all the terms from Model 6 which include pattern. In that case, the F test is *not* significant ($F = 1.26$ on 38 degrees of freedom, $p = 0.133$). The difference in the tests may be explained by the variance parameters which were larger in the selection model in the absence of stratification for pattern, thereby effectively diluting the strength of the difference. Additionally, the standard errors for the estimates of the fixed effects were slightly smaller in the pattern-mixture model. This is not surprising as in the model fitting we found that the means and variance parameters were dependent on pattern. Thus, stratifying for pattern results in more homogenous subgroups of cows reducing the variance within each group and subsequently providing more precise estimates for the diet effect.

Using Model 7, we test the global null hypothesis of no diet effect in any of the patterns. This analysis can be seen as a stratified analysis where a diet effect is estimated separately within each pattern. This model results in a significant F test for the diet effect ($F = 6.05$, on 6 degrees of freedom, $p < 0.001$). Alternatively, we can consider the pooled estimate for the diet effect, provided by equation (8.3), and calculate the test statistic using the delta method. This test also indicates a significant diet effect ($F = 17.82$ on 2 degrees of freedom, $p < 0.001$) as does the corresponding selection model, Model 10 ($F = 8.51$ on 2 degrees of freedom, $p < 0.001$).

Figure 8.9 presents the diet by time parameter estimates for selection Model 10, for the

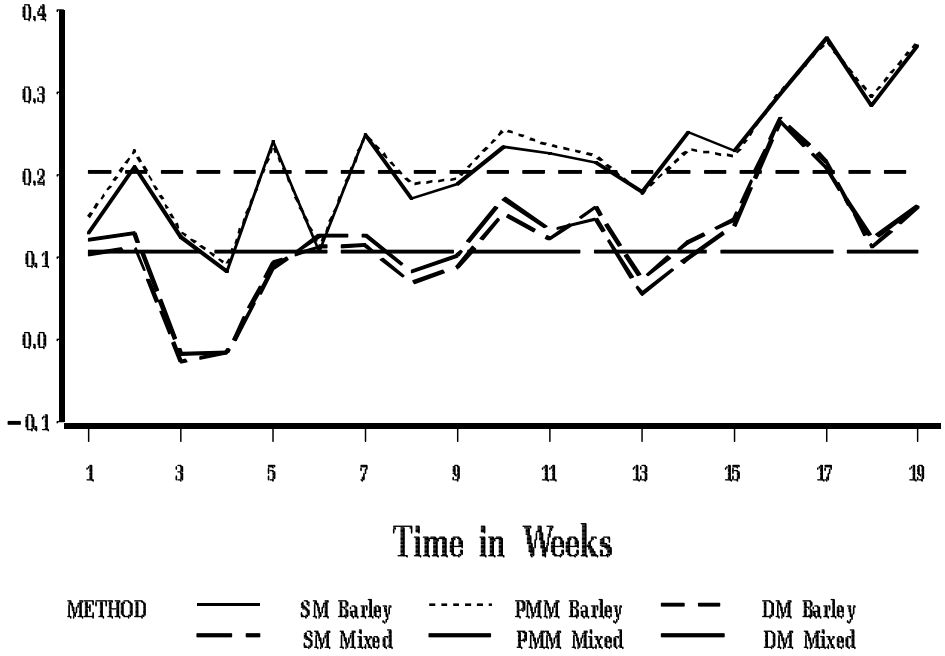

Figure 8.9: *Milk Protein Content Trial. Diet effect over time for the selection model (SM), the corresponding pattern-mixture model (PMM), and the estimate obtained after weighting the PMM contributions using the delta method (DM).*

corresponding pattern-mixture Model 7 and the weighted average estimates used in the delta method. The estimates for the selection model and the pattern-mixture model appear to differ only slightly. Since the model building within both families is done separately, this is a very reassuring sensitivity analysis outcome.

In conclusion, including pattern in the model improves the model fit significantly. In particular, the time by pattern and diet by pattern interactions are maintained in Model 7, which is considered to be the most parsimonius model consistent with the data (Figure 8.8). In addition, the covariance parameters are also dependent on the missingness pattern. Dividing cows into more homogenous groups based on their missingness patterns reduces the unexplained variation in the data and subsequently provides more precise parameter estimates.

The analyses discussed here provide an alternative to those obtained by Diggle and Kenward

(1994) but generally do not contradict them. Rather, they convey the message that the use of sensitivity analysis should become standard practice when dropout occurs. We strongly stress the importance of careful data verification to be undertaken prior to any statistical analysis. To this end we might add that the effect of an erroneous initial description of a dataset should not be underestimated as it can lead to subsequent mis-modeling of the data, thus adding confusion to an already complex undertaking of analyzing longitudinally measured observations.

Our analysis of the correlation structure appears to agree with the general conclusions retained in the Diggle and Kenward (1994) analysis. Particularly, it is interesting to notice the absence of random effects. We do not completely share the surprise expressed by Diggle and Kenward (1994) since it should be noted that the study animals are highly selected through centuries of cow eugenics and race selection. Had we dealt with wild animals, the role played by random effects would most likely have been much more substantial. To explain the absence of random effects we may assume that there were additional eligibility criteria for the trial (e.g., a specific breed of cow), which made random effects even more unlikely.

Analyzing a dataset using various approaches to answer a particular question is seen as a simple and informal way of sensitivity analysis, as is supplementing the main analysis with additional ones to gain extra insight. Each method used requires certain assumptions about the measurement process and the dropout process. In particular, pattern-mixture models and selection models approach the issue of dropout in different ways. It may also be useful to investigate the fundamental assumptions concerning the design of the experiment since dropout may be design driven. This example, and in particular the absence of genuine dropout, illustrates once more that care has to be taken when analyzing longitudinal outcomes with a non-rectangular structure.

8.5 EORTC Trial 30893

8.5.1 Introduction

The definition of QL which is universally accepted, is that QL is: (1) multidimensional (i.e., it comprises elements of a patient's emotional, social and physical well being); (2) it is a process (i.e., subject to change over a patient's lifetime); and (3) it is subjective (relies

primarily on the patient's own judgement). These three points all need to be taken into account in the design and the analyses of QL data in clinical trials.

Although QL is considered to be multidimensional, frequently one QL scale or domain is taken as the primary QL endpoint and is analysed as such. The other scales are then considered secondary in nature and analysed in an exploratory fashion. For example, when the EORTC QLQ-C30 is included in a clinical trial the global health/QL scale may be taken as the primary QL endpoint to answer the question 'Is there a difference in QL between two or more treatment groups?' For the remaining scales an exploratory, hypothesis-generating analysis may then be performed where the objective is to explore the data rather than to draw definitive conclusions from the results. Alternatively, if several scales are thought to be equally important, it is possible to base conclusions on multiple comparisons, using a more restrictive significance level (e.g., $P<0.01$ or 0.001) in order to reduce the risk of Type I errors (Hochberg 1988).

The second point in the definition of QL illustrates that it is a process and consequently it is subject to change over time, depending on the effects of random occurrences or planned interventions in an individual's life. To understand QL it is necessary to understand how it changes over time. One might consider that QL measurements that are taken close together may be strongly correlated while measurements that are taken at larger intervals may be less correlated. Statistically, we refer to this notion as *serial correlation* (Diggle, Liang and Zeger 1994). While cross-sectional studies (see Section 6.3.1) are useful in describing between-patient variability, explaining within-patient variability necessitates the study of repeated assessments over time.

The last point alluded to the patients' self assessment of QL. Since some patients may report consistently high scores and others consistently low scores, it is useful to capture this subject-specific effect in the statistical model. In a broad part of the literature, this concept is referred to as *random effects* (Diggle, Liang and Zeger 1994).

In this section we illustrate how the issues raised above may be taken into account in the analysis of QL datasets with missing data. This is done using as an example EORTC trial 30893 (see Sections 4.5 and 8.2.2) fitting both pattern-mixture and selection models. Choosing an appropriate model in both families and comparison of the results can be viewed as a sensitivity analysis.

8.5.2 Pattern-Mixture Model

As few dropouts occurred at weeks 6 and 12, these patients were grouped into one dropout pattern (see Figure 8.1). Similarly, patients who drop out at either week 24 or week 30 were collapsed into a single pattern. Thus five patterns remain with 19, 22, 28, 25 and 72 patients in the five patterns, respectively. Of course, the actual measurement times were preserved even in the collapsed patterns. Several baseline clinical variables were considered as covariates in the model. These included demographic variables: age, WHO performance status (see Appendix A.2), presence of chronic disease and pain assessed by the clinician; and disease descriptive variables: T stage, N stage and tumor grade (see Appendix A.3). In a univariate analysis WHO performance status and presence of chronic disease were significantly correlated with QL while there was a (non-significant) trend for an association between QL and age. As age is a known prognostic factor in patients with prostate cancer which could also influence the underlying dropout process, it was decided to keep it in the model. All other variables were not significantly correlated with QL scores. The model fitting results are presented in Table 8.3. The most complex model (model I) for the means structure includes the main effects of the clinical variables and assumes a separate mean for each treatment-by-time-by-dropout pattern combination. As suggested by the variogram the covariance matrix was taken as autoregressive order 1 with $\sigma_{jk} = \sigma^2\rho^{|j-k|}$ with a residual covariance term ϕ^2 and a random-effects component ν^2. The variance-covariance parameters are allowed to vary according to the dropout pattern. This model is equivalent to including time and treatment as covariates in the model statement and stratifying for dropout pattern. It provides a starting point for model simplification through backward selection. To simplify the covariance structure presented in Model I, in Model II the random-effect component is omitted. The likelihood ratio test indicates that Model II compares favourably with Model I. An explanation for this may be that the random effect is partially explained by the baseline covariates included in the model (age, WHO performance status and presence of chronic disease). Comparing Model III and IV with Model II indicates that the covariance structure cannot be simplified further.

In Model V we removed the 3-way interaction term between pattern, time and treatment. This model yields a significant likelihood ratio test statistic (p=0.003) when compared with Model II, rejecting the hypotheses of no 3-way interaction. To reduce the mean structure further we examined the graphical presentations in order to generate hypotheses. As mentioned earlier in Section 8.2.2, the QL scores appeared to increase immediately after orchidectomy and remained stable thereafter until the assessment prior to dropout. Thus we generated two indicator variables to achieve this: T_0 (0: Time= 0; 1: Time> 0) and T_X (1: Time= $D-1$;

Table 8.3: *EORTC Trial 30893. Model fit summary for pattern-mixture models.*

	Mean	Covariance model
I	BCV + pat×time×trt	AR(1)d+ν^2+$\phi^2 d$
II	BCV + pat×time×trt	AR(1)d+$\phi^2 d$
III	BCV + pat×time×trt	AR(1)d+ϕ^2
IV	BCV + pat×time×trt	AR(1)+$\phi^2 d$
V	BCV + trt time trt×pat trt×time pat×time	AR(1)d+$\phi^2 d$
VI	BCV + trt T_0 T_X trt×pat trt×T_0 pat×T_0×trt trt×T_X pat×T_X×trt	AR(1)d+$\phi^2 d$
VII	BCV trt T_0 T_X trt×pat trt×T_0 pat×T_0×trt trt×T_X pat×T_X	AR(1)d+$\phi^2 d$
VIII	BCV trt T_0 T_X trt×pat trt×T_0 pat×T_0 trt×T_X pat×T_X	AR(1)d+$\phi^2 d$
IX	BCV trt T_0 T_X trt×pat trt×T_0 pat×T_0×trt trt×T_X	AR(1)d+$\phi^2 d$
X	BCV trt T_0 T_X trt×pat trt×T_0 pat×T_0×trt pat×T_X	AR(1)d+$\phi^2 d$

	par	-2ℓ	Ref	G^2	df	p
I	75	7000.73				
II	74	7000.37	I	0.36	1	0.549
III	70	7020.52	II	20.15	4	<0.001
IV	66	7029.52	II	29.15	8	<0.001
V	54	7041.59	II	41.22	20	0.003
VI	48	7028.05	II	27.68	26	0.374
VII	44	7029.69	VI	1.64	4	0.802
VIII	40	7051.74	VII	22.05	4	<0.001
IX	40	7043.35	VII	13.66	4	0.008
X	43	7032.09	VII	2.40	1	0.121

Par:	Number of parameters	P:	P-value
-2ℓ:	-2 times log-likelihood	AR(1):	Autoregresive order 1
Ref:	Reference model	d:	By dropout pattern
G^2:	Likelihood ratio test statistic	BCV:	Baseline clinical variables
df:	Degrees of freedom	pat:	Pattern, trt: Treatment

0: otherwise, where D represents time of dropout). All possible treatment-by-pattern-by-time indicator combinations were included in Model VI. From Table 8.3 we observe that model VI is indeed a plausible simplification of model II. To further simplify the model we fitted models VII to X. These model fits suggested that the three way interaction between T_X, pattern and treatment and the 2-way interaction T_X-by-treatment could be omitted. However a three way interaction between T_0, pattern and treatment could not be rejected.

In conclusion, among the models presented, model X is the preferred one to summarize the data as it is the simplest model consistent with the data.

As there is an interaction between treatment effect and pattern, the delta method (see Equation 8.3) was used to test the marginal treatment effect. There appeared to be an imbalance in baseline QL score between the two groups. For this reason we performed two analyses, one adjusting for the baseline difference and the other ignoring the imbalance at baseline. We reparametrized model X by including the two way interaction terms pattern-by-T_0 and pattern-by-treatment in the model, thus obtaining directly interpretable estimates for the treatment effect for each pattern (i.e., pattern-by-treatment) and the 3-way interaction terms T_0-by-pattern-by-treatment. Note, since T_0 is coded 0 at baseline and 1 thereafter the interaction term T_0-by-pattern-by-treatment provides us with treatment effects by pattern adjusted for differences at baseline. For both the unadjusted and adjusted analyses, we obtained the pooled estimate for the treatment effect using equation (8.3), and calculated a test statistic using the delta method.

For the unadjusted analysis, the marginal treatment effect β was given by (8.3). Thus

$$\begin{aligned}\beta &= \gamma_1\phi_1 + \gamma_2\phi_2 + \gamma_3\phi_3 + \gamma_4\phi_4 + \gamma_5\phi_5 \\ &= -6.64(19/166) - 3.50(22/166) - 5.02(28/166) - 15.28(25/166) - 5.81(72/166) \\ &= -6.89\end{aligned}$$

The asymptotic estimate of the variance of the treatment effect is

$$\widehat{Var}(\hat{\beta}) = A \left(\begin{array}{c|c} Var(\gamma_{id})_{id} & 0 \\ \hline 0 & Var(\phi_d)_d \end{array} \right) A^T = 5.51$$

A Wald test is $\chi^2_1 = 8.61$ on 1 d.f which gives $p = 0.003$.

However, adjusting for the baseline imbalance resulted in an estimated treatment effect $\hat{\beta} = -5.936$ ($\widehat{\text{Var}}(\hat{\beta}) = 13.57$, $F = 2.60$, on 1 d.f., with $p = 0.107$).

8.5.3 Selection Model

The final model as selected in Section 8.5.2 was used as a starting point for the measurement model, excluding the pattern terms. Additionally, the term T_X was omitted as this term is

pattern dependent. Including this term would result in fitting a pseudo-likelihood model, i.e., $f(Y/D)f(D/Y)$. The model proposed by Diggle and Kenward (1994) and fitted in Oswald was primarily developed for monotone dropout. Therefore, due to poor compliance at baseline and our primary interest of examining the dropout mechanism, we omitted the baseline assessment from the analysis.

As is often the case with fitting (non-random) selection models, attaining convergence was not straightforward and we think it is useful to report on the problems encountered. A clear indication of convergence when fitting selection models is that the parameters for fixed effects for MAR and MCAR are identical. Initially the baseline covariates WHO performance status, age and presence of chronic disease were included in the measurement model. Increasing the number of iterations substantially did not result in convergence. The model was fitted using a program written in GAUSS to allow more flexibility in terms of exploration of the fitted variance-covariance matrix, the hessian matrix and residuals. After convergence of the MNAR model the inverse of the hessian matrix was not positive definite. An examination of the eigenvalues indicated a negative value for age indicating a potential saddle point. Plotting the profile likelihood indicated that the profile likelihood was relatively flat close to 0 which may have resulted in difficulties in estimation of the hessian matrix. As the maximum likelihood estimate of the age parameter was close to 0, we removed age from the model. Refitting the model in Oswald yielded consistent results, but still not fully convergent, between the MCAR and MAR models. Further simplification of the measurement model was not justifiable.

The variance parameters were included based on the results of the variogram. In contrast to the pattern-mixture model, the selection model indicated that the random intercept contributed significantly to the model (P<0.001). The measurement error appeared to be larger in the MNAR model than for the MCAR and MAR models whereas the serial variance parameter estimate was smaller in the MNAR model. For the dropout model we used a forward selection procedure initially fitting an MCAR model including covariates age, WHO performance status, presence of chronic disease and time. The model fits suggested that only time be kept in the model. Table 8.5.3 presents the final model fits for the three models: MCAR, MAR, and MNAR. In the MCAR model the parameters ψ_1 and ψ_2 are set to 0 indicating no dependence on the current measurement and immediately preceding measurement. Comparing the MAR model with the MCAR model using the likelihood ratio test we reject the null hypothesis of a MCAR dropout process (likelihood ratio test statistic of 21.98 on 1 d.f., $p < 0.001$). In addition, a comparison between the MNAR and MAR models yields a likelihood ratio test statistic of 0.76 (P=0.383). The treatment effect for the MAR model, -6.06

Table 8.4: *EORTC Trial 30893. Model fit summary for selection models.*

Parameter	MCAR	MAR	MNAR
Intercept	79.49 4.54	79.38 4.45	83.82
Presence of chronic disease	-4.56 2.79	-4.45 2.73	-6.79
WHO performance status	-4.75 2.30	-4.76 2.25	-5.56
Treatment	-2.96 1.65	-3.03 1.63	-3.80
Time	-0.82 0.28	-0.83 0.28	-0.43
Time by treatment	-0.43 0.28	-0.41 0.28	-0.44
Random intercept (v^2)	121.31	115.37	140.37
Measurement error (σ^2)	213.71	218.30	169.00
Serial variance (τ^2)	91.72	93.36	106.35
serial correlation exp (ϕ)	0.27	0.26	0.41
ψ_0	-3.80	-2.20	-4.06
ψ_{time}	0.47	0.45	0.53
ψ_1	0	0	0.06
ψ_2	0	-0.02	-0.06
Deviance	9396.24	9374.26	9373.50
Likelihood ratio test statistic		21.98	0.76

(-3.03 $\times$ 2: treatment is coded 1, -1) is significant($p = 0.011$) and similar in magnitude to the estimate obtained by the MNAR model and the pattern-mixture model not adjusted for baseline difference. The dropout parameters in the MNAR model suggest that the dropout increases when the prevailing QL scores are low ($(\psi_1 + \psi_2)/2 = 0.0$) and when there is an increase in QL score after dropout ($(\psi_1 - \psi_2)/2 = 0.12$). From a QL researchers point of view this is not logical since one might expect in an advanced cancer clinical trial that QL scores would continue to decrease after dropout. The final model, most consistent with the data, suggested a linear change in QL over time as shown in Figure 8.10 which presents both the observed and predicted profiles for patient 16. As the measurement model specified a linear, approximately horizontal, change over time it is not surprising that the expected value at the time of dropout indicates an increase from the previous value as illustrated by the dropout parameter estimates. This result suggests over-specification in the model. It is impossible to fit the decrease before dropout as presented in Figure 8.2 using a selection model as this decrease is spread over the various assessment time points dependent on the dropout time. Accordingly, its effect on the measurement parameter estimates in a selection

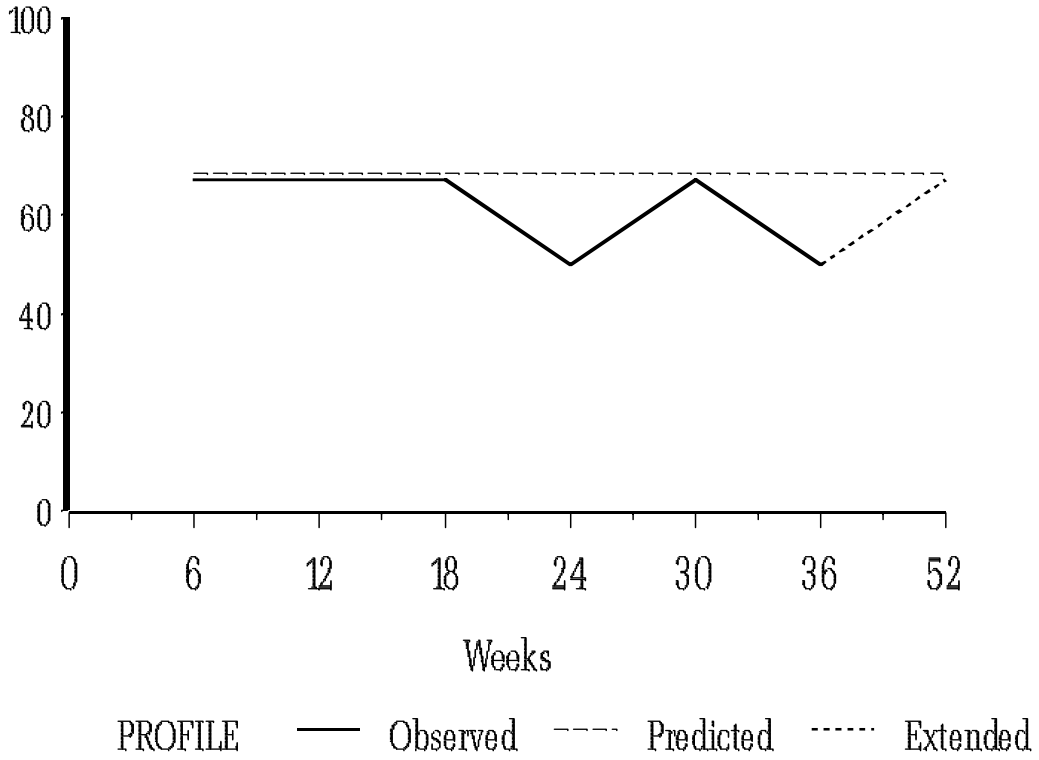

Figure 8.10: *EORTC Trial 30893. A selected patient profile and its fitted values using a selection model.*

model is diluted.

8.5.4 Remarks

The purpose of this section was to present the analysis of a QL study taking into account the definition of QL and the inherent structure of the dataset caused by attrition of patients. Section 8.2 presented the initial data exploration using graphical techniques. This step was revealing for a number of reasons. First, the mean structure indicated that the QL scores varied according to dropout pattern and treatment group, suggesting that a pattern-mixture model could be appropriate. Second, the mean structure suggested an initial increase after randomization followed by a stable period thereafter until the assessment prior to dropout. This information was employed to generate hypotheses during model fitting which resulted

in a simplified mean structure for the measurements.

The covariance structure suggested by the variogram and the scatter plots led us quickly to the appropriate covariance structure. Although the variogram suggested a random effect, the pattern-mixture model fit did not confirm this. Since the variogram is geared towards the selection model, this finding is not surprising. It may be explained by the inclusion of dropout pattern as a covariate and also by including pattern dependent variance components explaining a proportion of the variation which may otherwise have been ascribed to a subject-level component. Many authors have analysed QL data using cross-sectional methods ignoring the longitudinal structure of the data. However, the AR(1) covariance structure highlights the correlation between measurements taken at separate time points, particularly when assessments are made close together. Neglecting to analyze the data in a longitudinal fashion would therefore be wasteful and result in a loss of information.

The pattern-mixture model indicated that the measurement process was dependent on dropout pattern and treatment.The delta method was used to obtain an appropriate test statistic for the treatment effect. As there was an imbalance in QL scores between the two treatment groups at baseline we performed two analyses. The first analysis ignored the baseline difference and resulted in a significant treatment effect in favor of the orchidectomy alone arm. However, in the second analysis, which took into account the baseline imbalance, a significant treatment difference could not be detected. Since randomization between the two treatment arms was performed using the minimization technique stratifying for potentially important factors such as institution and WHO performance status, any imbalance in QL scores may have been entirely due to chance. Therefore, adjusting for a random occurrence may result in a biased analysis. On the other hand, although the imbalance may have occurred due to chance, it may be interesting to observe the impact of adjusting for the imbalance. The adjusted analysis may be seen as a sensitivity analysis.

An estimate of the marginal effect was obtained using a weighted average of the parameter estimates, weighted by the proportion of subjects in each dropout pattern. In this study the dropout rates were similar in both treatment groups. However, if the dropout rates vary between the two treatment arms then it may be advisable to weight the estimates according to the proportion of patients in each treatment by pattern subgroup to reduce the potential bias in the estimation of the marginal treatment effect.

This analysis is a sensitivity analysis in the sense that the data were analyzed under different assumptions about the measurement and dropout process. Although selection models and pattern-mixture models are considered to be probabilistically equivalent, they shed different

light in the context of a real data analysis. For example, in pattern-mixture models the overall distribution of the longitudinal measurements is a mixture of the conditional distributions, given the pattern of missingness. Therefore, the overall distribution may not necessarily be multivariate normal. Additionally, a pattern-mixture analysis has the flavour of an analysis stratified by dropout pattern, yielding homogenous groups resulting in smaller standard errors within each pattern leading to more precise estimates of the overall treatment effect. In addition it allows one to correct for the potentially confounding effect of dropout pattern.

The use of selection models in QL settings appears to be intuitive because the dropout process is thought of as being dependent on the measurement process. In contrast, the interpretation of pattern-mixture models is not so obvious since it implies that the QL scores for an individual are dependent on the time that patient will drop out (i.e., in the future). However, for pattern-mixture models, the assumptions which are made during model fitting are clearer, e.g., in pattern-mixture models patients only contribute to parameter estimates prior to dropout, whereas in selection models all patients contribute to the model at all time points whether they have dropped out or not. Additionally, in selection models the likelihood for both the dropout model and measurement models are maximized simultaneously resulting in a maximized joint likelihood during which some assumptions are made which are not fundamentally testable, e.g., the dependence of the dropout process on measurements which have not been obtained. In contrast, in pattern-mixture models the missingness process is usually fairly simple, and can reduce to a multinomial distribution, describing the proportion of the different patterns. Also, fitting a selection model may be computationally cumbersome. For pattern-mixture models, the only requirement is that there are sufficient data in the various patterns to achieve reliable estimates. One then only needs fairly straightforward non-iterative code to determine marginal quantities.

Model building using pattern-mixture models does not allow one to test if the dropout process is MNAR, although Molenberghs *et al* (1998) derived identifying restrictions which can be used in a pattern-mixture context to correspond to MAR (see Section 9.2). With selection models, the dropout probability is estimated conditional on the measurements, allowing the dropout process to be tested. Several authors have suggested that caution be exercised when fitting MNAR models (Glynn, Laird and Rubin 1986, Molenberghs *et al* 1998) as assumptions are made in model fitting which are untestable, namely regarding the relationship between non-response and the missing measurements. However, Verbeke and Molenberghs (1997) argue that restricting the model building exercise to a MAR mechanism is equally dangerous since the MAR assumption is itself fundamentally untestable. The selection model suggested that the missing data were not MNAR. The treatment effects under both MAR and MNAR were similar, supporting the earlier conclusions regarding the

unadjusted treatment comparisons. In the analyses of QL data with dropout it is advisable to fit both families of models to examine the extent of agreement in results.

Chapter 9

Sensitivity Analysis for Pattern-Mixture Models

9.1 Introduction

Chapter 8 illustrated the use of pattern-mixture models in QL settings. As demonstrated they provide an alternative formulation for the common selection model factorization. In Chapter 7 we observed that pattern-mixture models are underidentified, which is clearly seen by means of the Glynn, Laird, and Rubin (1986) 'paradox' (Section 7.4). Consequently, Little (1993, 1994, 1995) suggested the use of so-called identifying restrictions to overcome this under-identification: inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. Little (1993) shows how these constraints can be used to identify all the parameters in the model and so obtain estimates for these and the marginal probabilities. For example, *complete case missing value* (CCMV) restrictions (Little 1993) essentially equate conditional distributions beyond time t, i.e., those unidentifiable from this dropout group, with the same conditional distributions from the completers. All in all, while some authors perceive this under-identification as a drawback, we believe it is an asset since it forces one to reflect on the assumptions made. On the other hand, neither of the two examples in Chapter 8 (the milk dataset in Section 8.4 and the QL study in Section 8.5) made use of identifying restrictions. For the milk protein trial it was suggested that the experiment terminated when feed availability

declined in the paddock in which animals were grazing. If this is true then we are not interested in what happened after the experiment terminated and we do not need to extrapolate our results further. For the QL study we may wish to answer two scientific questions: (1) what is the difference with respect to QL between the two treatment groups while patients remain on-study (progression-free)? or (2) what is the difference with respect to QL between the two treatment groups while patients remain alive? To answer the latter question we need to extrapolate our results. This may be done using identifying restrictions. Thus, we have two strategies to build a full data model in the pattern-mixture context: identifying restrictions, and the inclusion of pattern as a covariate as shown in Chapter 8. We will show in this chapter how using different identifying restrictions can serve as a starting point for sensitivity analysis.

While identifying restrictions impose a careful reflection on the unidentified part of the distribution, including pattern as a covariate is more implicit about the assumptions made to identify the full distribution. In this respect, identifying restrictions are open to some of the criticism of selection models. The identifying restrictions strategy is harder to implement, unless in fairly simple settings, such as for a single normal sample or for contingency tables (Little 1993, 1994). This chapter provides a tool to conduct such a strategy in realistic longitudinal settings.

Section 9.2 describes the relationship between MAR and the pattern-mixture framework and Section 9.3 discusses the use of sensitivity analysis when fitting pattern-mixture models. The identifying restriction strategy is described in detail in Section 9.4. Multiple imputation, a tool used in the identifying restrictions strategy, is reviewed in Section 9.5. Application to the milk dataset and a QL dataset are discussed in Sections 9.6 and 9.7 respectively. Some remarks and suggestions for alternative routes of sensitivity are offered in Section 9.8.

9.2 Pattern-Mixture Models and MAR

The missing data taxonomy is usually presented in the selection modeling framework rather than in the pattern-mixture context. Here we show that pattern-mixture models can be classified similarly, and further that the intermediate MAR category is connected to particular kinds of restrictions on the parameters of a pattern-mixture model in the case of monotone missingness.

Assume a complete measurement sequence is of length n. Recall that the classical taxonomy considers the structure of $f(d|\boldsymbol{y})$. The missing data are MAR if a subject's missingness mechanism depends on its observed outcomes only, $f(d = t + 1|y_1, \ldots, y_n) = f(d = t + 1|y_1, \ldots, y_t)$, for $t = 1, \ldots, n$.

We will now show how pattern-mixture models can be classified using exactly the same taxonomy as is used for selection models. Furthermore, we establish a link between this classification and the identifying restrictions proposed in Little (1993). Clearly, selection models and pattern-mixture models coincide under the MCAR assumption. Next, we show that MAR can be expressed in a pattern-mixture framework through restrictions, related to the *complete case missing value* (CCMV) restrictions (Little 1993), which we will call *available case missing value* (ACMV) restrictions. Little's CCMV restrictions set a conditional density of unobserved components given a particular set of observed components equal to the corresponding conditional density in the subgroup of completers. ACMV restrictions equate this conditional density to the one calculated from the subgroup of all patterns for which all required components have been observed.

In our setting of longitudinal data with dropouts, CCMV can be defined formally as the condition that for each $t \geq 2$ and for $j < t$:

$$f(y_t|y_1, \ldots, y_{t-1}, d = j + 1) = f(y_t|y_1, \ldots, y_{t-1}, d = n + 1),$$

whereas ACMV is the condition that for all $t \geq 2$ and $j < t$:

$$f(y_t|y_1, \ldots, y_{t-1}, d = j + 1) = f(y_t|y_1, \ldots, y_{t-1}, d > t). \tag{9.1}$$

If there are only 2 time points ($n = 2$), then ACMV and CCMV coincide. With these definitions, Molenberghs *et al* (1998) have shown that, for longitudinal data with dropouts, MAR $\Longleftrightarrow$ ACMV.

An interesting aside of this theorem is that, since MAR corresponds to a set of (untestable) restrictions (ACMV) in the pattern-mixture framework, MAR itself is also untestable. Precisely, *given* MAR, standard (observed data) methods can be used but the assumption of MAR itself cannot be tested. This fact is often overlooked in the selection framework.

The restrictions discussed here will be incorporated in strategies to fit pattern-mixture models in Sections 9.6 and 9.7.

9.3 Pattern-Mixture Models and Sensitivity Analysis

Pattern-mixture models have gained renewed interest in recent years (Little 1993, 1994, Hogan and Laird 1997). Several authors have contrasted selection models and pattern-mixture models. This is done either (1) to answer the same scientific question, such as marginal treatment effect or time evolution, based on these two rather different modeling strategies, or (2) to gain additional insight by supplementing the selection model results with those from a pattern-mixture approach. Examples can be found in Verbeke, Lesaffre, and Spiessens (1998), Curran, Pignatti, and Molenberghs (2000b), and Michiels *et al* (1999) for continuous outcomes. The categorical outcome case has been treated in Molenberghs, Michiels, and Lipsitz (1999), and Michiels, Molenberghs, and Lipsitz (1999). Further references include Ekholm and Skinner (1998), Molenberghs, Michiels, and Kenward (1998), Little and Wang (1996), Hedeker and Gibbons (1997), and McArdle and Hamagani (1992).

Sensitivity analysis for pattern-mixture models can be conceived in many different ways. Crucial aspects are whether pattern-mixture and selection modeling are to be contrasted with one another as presented in Chapter 8 or whether the pattern-mixture modeling is the central focus of interest. It is natural to conduct sensitivity analysis *within* the pattern-mixture family. The key area where sensitivity analysis should be focused is on the unidentified components of the model and the way(s) in which this is handled.

Little (1993, 1994) advocated the use of identifiying restrictions and presented a number of examples. We will outline a general framework for identifying restrictions in Section 9.4, with CCMV (introduced by Little 1993), ACMV, and neighboring case missing value restrictions (NCMV) as important special cases. Recall that ACMV is the natural counterpart of MAR in the PMM framework. This provides a way to compare ignorable selection models with their counterpart in the pattern-mixture setting. Michiels, Molenberghs, Lipsitz (1999) took up this idea in the context of binary outcomes, with a marginal global odds ratio model to describe the measurement process (Molenberghs and Lesaffre 1994).

9.4 Identifying Restriction Strategies

We restrict attention to monotone patterns. In general, let us assume we have $t = 1, \ldots, T$ dropout patterns where the dropout indicator is $d = t + 1$. For pattern t, the complete data density is given by

$$f_t(y_1, \ldots, y_T) = f_t(y_1, \ldots, y_t) f_t(y_{t+1}, \ldots, y_T | y_1, \ldots, y_t). \tag{9.2}$$

The first factor is clearly identified from the observed data, while the second factor is not. It is assumed that the first factor is known or, more realistically, modeled using the observed data. Then identifying restrictions are applied in order to identify the second component.

While, in principle, completely arbitrary restrictions can be used by means of any valid density function over the appropriate support, strategies which relate back to the observed data deserve privileged interest. One can base identification on all patterns for which a given component, y_s say, is identified. A general expression for this is

$$f_t(y_s|y_1,\dots y_{s-1}) = \sum_{j=s}^{T} \omega_{sj} f_j(y_s|y_1,\dots y_{s-1}), \quad s = t+1,\dots,T. \tag{9.3}$$

We will use $\boldsymbol{\omega}_s$ as shorthand for the set of ω_{sj}'s used. Every $\boldsymbol{\omega}_s$ which sums to one provides a valid identification scheme.

Let us incorporate (9.3) into (9.2):

$$f_t(y_1,\dots,y_T) = f_t(y_1,\dots,y_t) \prod_{s=0}^{T-t-1} \left[\sum_{j=T-s}^{T} \omega_{T-s,j} f_j(y_{T-s}|y_1,\dots,y_{T-s-1}) \right]. \tag{9.4}$$

Expression (9.4) clearly shows which information is used to complement the observed data density in pattern t in order to establish the complete data density.

Let us consider three special but important cases. Little (1993) proposes CCMV which uses the following identification:

$$f_t(y_s|y_1,\dots y_{s-1}) = f_T(y_s|y_1,\dots y_{s-1}), \quad s = t+1,\dots,T. \tag{9.5}$$

In other words, information which is unavailable is always borrowed from the completers. This strategy can be defended in cases where most of the subjects are complete and only small proportions are assigned to the various dropout patterns. Also, extension of this approach to non-monotone patterns is particularly easy.

Alternatively, the nearest identified pattern can be used:

$$f_t(y_s|y_1,\dots y_{s-1}) = f_s(y_s|y_1,\dots y_{s-1}), \qquad s = t+1,\dots,T. \tag{9.6}$$

We will refer to these restrictions as *neighboring case missing values* or NCMV.

The third special case of (9.3) will be ACMV of which the definition is presented in (9.1). Thus, ACMV is reserved for the counterpart of MAR in the PMM context. Let us derive the corresponding $\boldsymbol{\omega}_s$ vectors. Expression (9.3) can be restated as

$$f_t(y_s|y_1,\ldots,y_{s-1}) = f_{(\geq s)}(y_s|y_1,\ldots,y_{s-1}), \tag{9.7}$$

for $s = t+1,\ldots,T$. Here, $f_{(\geq s)}(.|.) \equiv f(.|.,d>s)$, with d an indicator for time of dropout, which is one more than the length of the observed sequence. Now, we can transform (9.7) as follows:

$$\begin{aligned} & f_t(y_s|y_1,\ldots,y_{s-1}) \\ &= f_{(\geq s)}(y_s|y_1,\ldots,y_{s-1}) \\ &= \frac{\sum_{j=s}^{T}\alpha_j f_j(y_1,\ldots,y_s)}{\sum_{j=s}^{T}\alpha_j f_j(y_1,\ldots,y_{s-1})} \qquad (9.8) \\ &= \sum_{j=s}^{T}\frac{\alpha_j f_j(y_1,\ldots,y_{s-1})}{\sum_{j=s}^{T}\alpha_j f_j(y_1,\ldots,y_{s-1})} f_j(y_s|y_1,\ldots,y_{s-1}). \qquad (9.9) \end{aligned}$$

Next, comparing (9.9) to (9.3) yields:

$$\omega_{sj} = \frac{\alpha_j f_j(y_1,\ldots,y_{s-1})}{\sum_{\ell=s}^{T}\alpha_\ell f_\ell(y_1,\ldots,y_{s-1})}. \tag{9.10}$$

We have now derived two equivalent explicit expressions of (9.1). Expression (9.8) is the conditional density of a mixture, whereas (9.3) with (9.10) is a mixture of conditional densities. Clearly, $\boldsymbol{\omega}$ defined by (9.10) consists of components which are nonnegative and sum to one. In other words, a valid density function is defined.

Restrictions (9.3), with the CCMV, NCMV, and ACMV forms as special cases, can be incorporated in a comprehensive strategy to fit pattern-mixture models.

9.4.1 Strategy Outline

We will briefly sketch the strategy. Several points which require further specification will be discussed in subsequent sections.

1. Fit a model to the pattern-specific identifiable densities: $f_t(y_1,\ldots,y_t)$. This results in a parameter estimate, $\hat{\boldsymbol{\gamma}}_t$.

2. Select an identification method of choice.

3. Using this identification method, determine the conditional distributions of the unobserved outcomes, given the observed ones:

$$f_t(y_{t+1}, \ldots, y_T | y_1, \ldots, y_t). \tag{9.11}$$

4. Using the methodology outlined in Section 9.5, draw multiple imputations for the unobserved components, given the observed outcomes and the correct pattern-specific density (9.11).

5. Analyze the multiply-imputed sets of data using the method of choice. This can be another pattern-mixture model, but also a selection model or any other desired model.

6. Inferences can be conducted in the way described in Sections 9.5.1 and 9.5.2.

Special Case: 3 Measurements

In this case, there are only three patterns and identification (9.4) takes the following form:

$$\begin{aligned} f_3(y_1, y_2, y_3) &= f_3(y_1, y_2, y_3), & (9.12)\\ f_2(y_1, y_2, y_3) &= f_2(y_1, y_2) f_3(y_3 | y_1, y_2), & (9.13)\\ f_1(y_1, y_2, y_3) &= f_1(y_1) \left[\omega f_2(y_2|y_1) + (1-\omega) f_3(y_2|y_1)\right] \\ &\quad \times f_3(y_3|y_1, y_2). & (9.14) \end{aligned}$$

Since $f_3(y_1, y_2, y_3)$ is completely identifiable from the data, and for $f_2(y_1, y_2, y_3)$ there is only one possible identification, given (9.3), the only place where a choice has to be made is for pattern 1. Setting $\omega = 1$ corresponds to NCMV, while $\omega = 0$ implies CCMV. Using (9.10) in this particular case, ACMV corresponds to

$$\omega = \frac{\alpha_2 f_2(y_1)}{\alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}. \tag{9.15}$$

The conditional density $f_1(y_2|y_1)$ in (9.14) can be rewritten as

$$f_1(y_2|y_1) = \frac{\alpha_2 f_2(y_1, y_2) + \alpha_3 f_3(y_1, y_2)}{\alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}.$$

9.4.2 Drawing from the Conditional Densities

In the previous section, we have seen how general identifying restrictions (9.3), with CCMV, NCMV, and ACMV as special cases, lead to the conditional densities for the unobserved components, given the observed ones. This came down to deriving expressions for $\boldsymbol{\omega}$, such as in (9.10) for ACMV. This endeavor corresponds to items 2 and 3 of the strategy outline (9.4.1). In order to carry out item 4, we need to draw imputations from these conditional densities.

Let us proceed by studying the special case of three measurements first. To this end, we consider identification scheme (9.12)–(9.14) and we start off by avoiding the specification of a parametric form for these densities. The following steps are required:

1. Estimate the parameters of the identifiable densities: $f_3(y_1, y_2, y_3)$, $f_2(y_1, y_2)$, and $f_1(y_1)$. Then, for each of the m imputations, we have to execute the following steps.
2. Draw from the parameter vectors as in the first step on page 130. It will be assumed that in all densities from which we draw, this parameter vector is used.
3. **For pattern 2.** Given an observation in this pattern, with observed values (y_1, y_2), calculate the conditional density $f_3(y_3|y_1, y_2)$ and draw from it.
4. **For pattern 1.** We now have to distinguish three substeps.
   (a) Given y_1, and the proportions α_2 and α_3 of observations in the second and third patterns, respectively, determine ω. Every ω in the unit interval is valid. Special cases are:
   - For NCMV, $\omega = 1$.
   - For CCMV, $\omega = 0$.
   - For ACMV, ω is calculated from (9.15). Note that, given y_1, this is a constant.

   Generate a random uniform variate, U say. (Note that, strictly speaking, this draw is unnecessary for the boundary NCMV and CCMV cases.)

   (b) If $U \leq \omega$, calculate $f_2(y_2|y_1)$ and draw from it. Otherwise, do the same based on $f_3(y_2|y_1)$.

   (c) Given the observed y_1 and given y_2 which has just been drawn, calculate the conditional density $f_3(y_3|y_1, y_2)$ and draw from it.

All steps but the first one have to be repeated M times, to obtain the same number of imputed datasets. Inference then proceeds as outlined in Sections 9.5.1 and 9.5.2.

When the observed densities are estimated using linear mixed models, $f_3(y_1, y_2, y_3)$, $f_2(y_1, y_2)$, and $f_1(y_1)$ produce fixed-effect and variance parameters. Let us group all of them in $\boldsymbol{\gamma}$ and assume a draw is made from their distribution, $\boldsymbol{\gamma}^*$ say. To this end, their precision estimates need to be computed. These are easily obtained in most standard software packages, such as SAS.

Let us illustrate this procedure for (9.13). Let us assume that the ith subject has only two measurements, and hence belongs to the second pattern. Let its design matrices be X_i and Z_i for the fixed effects and random effects, respectively. Its mean and variance for the *third* pattern are:

$$\begin{aligned} \boldsymbol{\mu}_i(3) &= X_i\boldsymbol{\beta}^*(3), & (9.16)\\ V_i(3) &= Z_iD^*(3)Z_i' + \Sigma_i(3), & (9.17) \end{aligned}$$

where (3) indicates that the parameters are specific to the third pattern.

Now based on (9.16)–(9.17), and the observed values $y_i = (y_{i1}, y_{i2})'$, the parameters for the conditional density follow immediately:

$$\begin{aligned} \boldsymbol{\mu}_{i,2|1}(3) &= \boldsymbol{\mu}_{i,2}(3) + V_{i,21}(3)[V_{i,11}(3)]^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,2}(3)),\\ V_{i,2|1}(3) &= V_{i,22}(3) - V_{i,21}(3)[V_{i,11}(3)]^{-1}V_{i,12}(3), \end{aligned}$$

where a subscript 1 indicates the first two components and a subscript 2 refers to the third component. Draws from every other conditional density are entirely similar.

In several cases, the conditional density is a mixture of normal densities. Then, drawing from (9.3) consists of two steps:

- Draw a random uniform variate U to determine which of the $n - s + 1$ components one is going to draw from. Specifically, the kth component is chosen if

$$\sum_{j=s}^{k-1} \omega_{sj} \leq U < \sum_{j=s}^{k} \omega_{sj},$$

  where $k = s, \ldots, n$. Note that if $k = 1$, the left hand sum is set equal to zero.

- Draw from the kth component.

All of these steps have been combined in a SAS macro.

9.5 Multiple Imputation

In Section 9.4 multiple imputation was used as a tool in developing identifying restrictions strategies. The concept of multiple imputation refers to replacing each missing value with more than one imputed value. The goal is to combine the simplicity of imputation strategies, with unbiasedness in both point estimates and measures of precision. In Section 3.3.1 we noted that some simple imputation procedures may yield inconsistent point estimates as soon as the missingness mechanism surpasses MCAR. This could be overcome to a large extent with conditional mean imputation, but the problem of underestimating the variability of the estimators is common to all methods since they all treat imputed values as observed values. By imputing several values for a single missing component, this uncertainty is explicitly acknowledged.

Rubin (1987) points to another very useful application of multiple imputation. Rather than merely accounting for *sampling uncertainty*, the method can be used to incorporate *model uncertainty*. Indeed, when a measurement is missing but the researcher has a good idea about the probabilistic measurement and missingness mechanisms, then constructing the appropriate distribution to draw imputations from is, at least in principle, relatively straightforward. In practice there may be considerable uncertainty about some parts of the joint model. In that case, several mechanisms for drawing imputations might seem equally plausible. They can be combined in a single multiple imputation analysis. As such, multiple imputation can be used as a tool for sensitivity analysis.

Suppose we have a sample of N, i.i.d. $n \times 1$ random vectors $\boldsymbol{Y}_i$. Our interest lies in estimating some parameter vector $\boldsymbol{\theta}$ of the distribution of $\boldsymbol{Y}_i$. Assume notation is as in Chapter 2.4. Multiple imputation fills in $\boldsymbol{Y}^m$ using the observed data $\boldsymbol{Y}^o$, several times, and then the completed data are used to estimate $\boldsymbol{\theta}$.

As discussed by Rubin and Schenker (1986), the theoretical justification for multiple imputation is most easily understood using Bayesian concepts, but a likelihood-based treatment of the subject is equally possible. If we knew the joint distribution of $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ with

parameter vector $\boldsymbol{\gamma}$ say, then we could impute $\boldsymbol{Y}_i^m$ by drawing a value of $\boldsymbol{Y}_i^m$ from the conditional distribution

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\gamma}). \tag{9.18}$$

Note that we explicitly distinguish the parameter of scientific interest $\boldsymbol{\theta}$ from the parameter $\boldsymbol{\gamma}$ in (9.18). Since $\boldsymbol{\gamma}$ is unknown, we must estimate it from the data, say $\hat{\boldsymbol{\gamma}}$, and presumably use

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \hat{\boldsymbol{\gamma}}) \tag{9.19}$$

to impute the missing data. In Bayesian terms, $\boldsymbol{\gamma}$ in (9.18) is a random variable of which the distribution is a function of the data. In particular, we first obtain the distribution of $\boldsymbol{\gamma}$ from the data, depending on $\hat{\boldsymbol{\gamma}}$. The construction of model (9.18) is referred to by Rubin (1987) as the *Modeling Task*.

After formulating the distribution of $\boldsymbol{\gamma}$, the imputation algorithm is:

1. Draw $\boldsymbol{\gamma}^*$ from the distribution of $\boldsymbol{\gamma}$.
2. Draw $\boldsymbol{Y}_i^{m*}$ from $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\gamma}^*)$.
3. Using the completed data, $(\boldsymbol{Y}^o, \boldsymbol{Y}^{m*})$, and the method of choice (i.e., maximum likelihood, restricted maximum likelihood, method of moments, partial likelihood), estimate the parameter of interest $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{Y}) = \hat{\boldsymbol{\theta}}(\boldsymbol{Y}^o, \boldsymbol{Y}^{m*})$ and its variance (called *within*-imputation variance) $U = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$.
4. Independently repeat steps 1–3, M times. The M datasets give rise to $\hat{\boldsymbol{\theta}}^{(m)}$ and $U^{(m)}$, for $m = 1, ..., M$.

Steps 1 and 2 are referred to as the *Imputation Task*. Step 3 is the *Estimation Task*. Of course, one wants to combine the M inferences into a single one. Parameter and precision estimation and hypothesis testing will be discussed next.

9.5.1 Parameter Estimation

The M within-imputation estimates for $\boldsymbol{\theta}$ are pooled to give the multiple imputation estimate:

$$\hat{\boldsymbol{\theta}}^* = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\theta}}^{(m)}.$$

Suppose that complete-data inference about $\boldsymbol{\theta}$ would be made by $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \sim N(\mathbf{0}, U)$. Then one can make normal-based inferences for $\boldsymbol{\theta}$ based upon

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^*) \sim N(\mathbf{0}, V), \tag{9.20}$$

where

$$V = \hat{W} + \left(\frac{M+1}{M}\right)\hat{B}, \tag{9.21}$$

$$\hat{W} = \frac{\sum_{m=1}^{M} U^{(m)}}{M} \tag{9.22}$$

is the average within-imputation variance, and

$$\hat{B} = \frac{\sum_{m=1}^{M}(\hat{\boldsymbol{\theta}}^{(m)} - \hat{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}}^{(m)} - \hat{\boldsymbol{\theta}}^*)'}{M-1} \tag{9.23}$$

is the between-imputation variance (Rubin 1987). Rubin and Schenker (1986) report that a small number of imputations ($M = 2, 3$) already yields a major improvement over single imputation. Upon noting that the factor $(M + 1)/M$ approaches 1 for large M, (9.21) is approximately the sum of the within and the between imputations variability.

Multiple imputation is most useful in situations where $\boldsymbol{\gamma}$ is an easily estimated set of parameters characterizing the distribution of $\boldsymbol{Y}_i$, while $\boldsymbol{\theta}$ is complicated to estimate in the presence of missing data, and/or when obtaining a correct estimate for the variance is non-trivial with incomplete data.

9.5.2 Hypothesis Testing

Testing hypotheses could be based on the asymptotic normality results (9.20) and (9.21). However, the rationale for using asymptotic results and hence χ^2 reference distributions is not

just a function of the sample size, N, but also of the number of imputations, M. Therefore, Li, Raghunathan, and Rubin (1991) propose the use of an F reference distribution. Precisely, to test the hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, they advocate the following method to calculate p values:

$$p = P(F_{k,w} > F), \tag{9.24}$$

where k is the length of the parameter vector, $F_{k,w}$ is an F random variable with k numerator and w denominator degrees of freedom, and

$$F = \frac{(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)' W^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)}{k(1+r)}, \tag{9.25}$$

$$w = 4 + (\tau - 4)\left[1 + \frac{(1 - 2\tau^{-1})}{r}\right]^2,$$

$$r = \frac{1}{k}\left(1 + \frac{1}{M}\right)\mathrm{tr}(BW^{-1}), \tag{9.26}$$

$$\tau = k(M-1).$$

It is interesting to note that, when $M \to \infty$, the reference distribution of F approaches an $F_{k,\infty} = \chi^2/k$ distribution, in line with intuition. Good operational characteristics of this procedure are reported in Li, Raghunathan, and Rubin (1991).

Clearly, procedure (9.24) can be used as well when not the full vector $\boldsymbol{\theta}$, but one component, a subvector, or a set of linear contrasts, is the subject of hypothesis testing. When a subvector is of interest (a single component being a special case), the corresponding submatrices of B and W need to be used in (9.25) and (9.26). For a set of linear contrasts $L\boldsymbol{\theta}$, one should use the appropriately transformed covariance matrices: $\tilde{W} = LWL'$, $\tilde{B} = LBL'$, and $\tilde{V} = LVL'$.

9.6 Analysis of the Milk Data

In order to illustrate the methodology described in this chapter, we will apply it to the milk data. As was described in Sections 4.4 and 8.4, 79 cows were included in the study. Three dropout patterns were defined with 20, 18 and 41 cows, respectively.

We will apply each of the identifying restriction strategies presented in Section 9.4, to these data. First, starting models will be fitted (Section 9.6.1). Second, it will be illustrated

how hypothesis testing can be performed, given the pattern-mixture parameter estimates and their estimated covariance matrix (Section 9.6.2). Third, model simplification will be discussed and applied (Section 9.6.3).

9.6.1 Fitting a Model

In order to apply the identifying restriction strategy, one needs to fit a model to the observed data first. We opted for a complex model for the mean structure, while keeping the variance-covariance structure relatively simple. The mean structure was defined as a full interaction model, i.e., time by diet and the variance-covariance structure was defined as first-order autoregressive with a residual variance term as suggested in Section 8.4. A model was fitted separately within each pattern thus providing parameters specific to each pattern. Of course, not all parameters are estimable. For example, in the first pattern the time effects at 15-19 weeks are unidentified. This initial model provides a basis for identifying restriction models. Using the methodology detailed in Section 9.4.1, a SAS macro, was written to conduct the multiple imputation, fitting of imputed datasets, and combination of the results into a single inference.

The initial multiple imputation results are presented graphically: Figure 9.1 presents the mean response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. For patterns 1 and 2 there is some variability in the estimated profiles across the three restrictions towards the end of the study, although this may be in part due to random variation. Since the data in pattern 3 are complete, there is of course no difference between the profiles obtained with each of the identifying restriction techniques. Notice that in pattern 3 the profiles tend to increase over time from week five onwards for all patterns, in particular for Barley, and then level off at the end of the study. In contrast, the profiles in pattern 1 tend to decrease. Notice also that using the CCMV identifying restriction results in an increase in the Barley diet group for pattern 1 which is consistent with imputation taking information from the complete cases in pattern 3. Restrictions using ACMV and NCMV tend to provide lower mean responses than for CCMV. This is shown more clearly in Figure 9.2. Recall that pattern 1 includes dropouts at week 15 only, pattern 2 includes dropouts occuring in weeks 16, 17 and 19 and pattern 3 includes completers. Thus, NCMV restrictions use information from pattern 2 to impute missing values for both patterns 1 and 2 up until week 18, ACMV randomly chooses between patterns 2 and 3 depending on (9.10), whereas CCMV always takes information from the complete cases in pattern 3. Note at week 19 it is only possible to impute data using the conditional distribution obtained

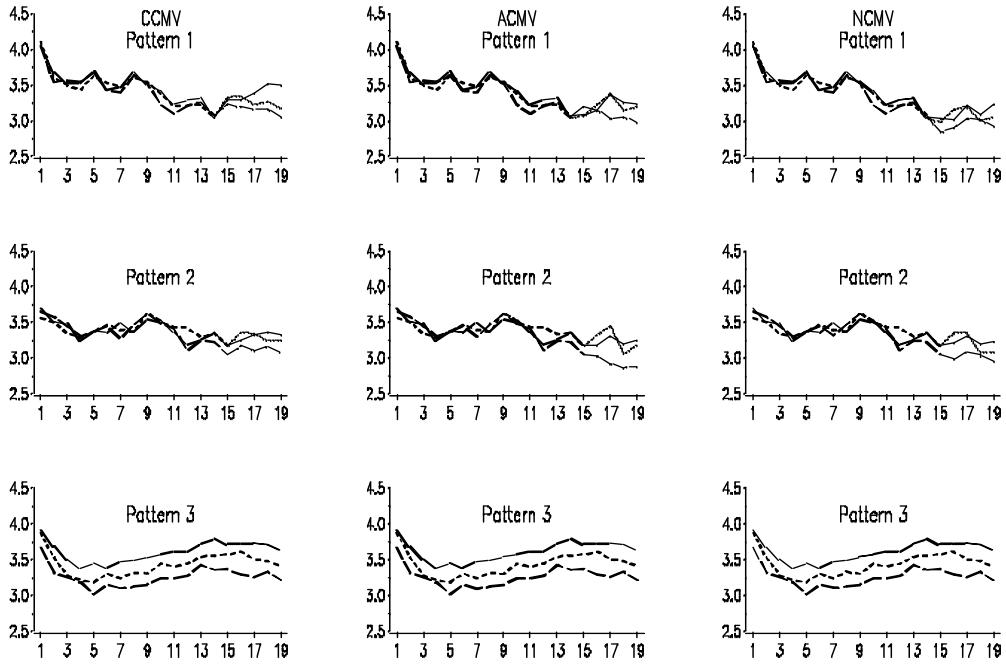

Figure 9.1: *Milk Protein Content Trial. Mean response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. Full line: Barley, Broken Line: Mixed, Dotted line: Lupins*

from the complete cases in pattern 3.

In general, CCMV extrapolates rather towards a rise whereas NCMV and ACMV seem to predict lower mean responses. This conclusion needs to be considered carefully. Note, the results obtained by Diggle and Kenward (1994) suggested that there was highly significant evidence indicating a MNAR dropout process and suggesting that the probability of dropout increased when either the prevailing level of protein content was low or when the increment between the last and current protein content was high. This would suggest an increase in protein content level after dropout which is consistent with the results obtained from the CCMV identifying restrictions. However Curran, Pignatti and Molenberghs (2000b) suggested that this result was contradictory and might be taken as an indicator that there was need for reflection on the model for the dropout mechanism.

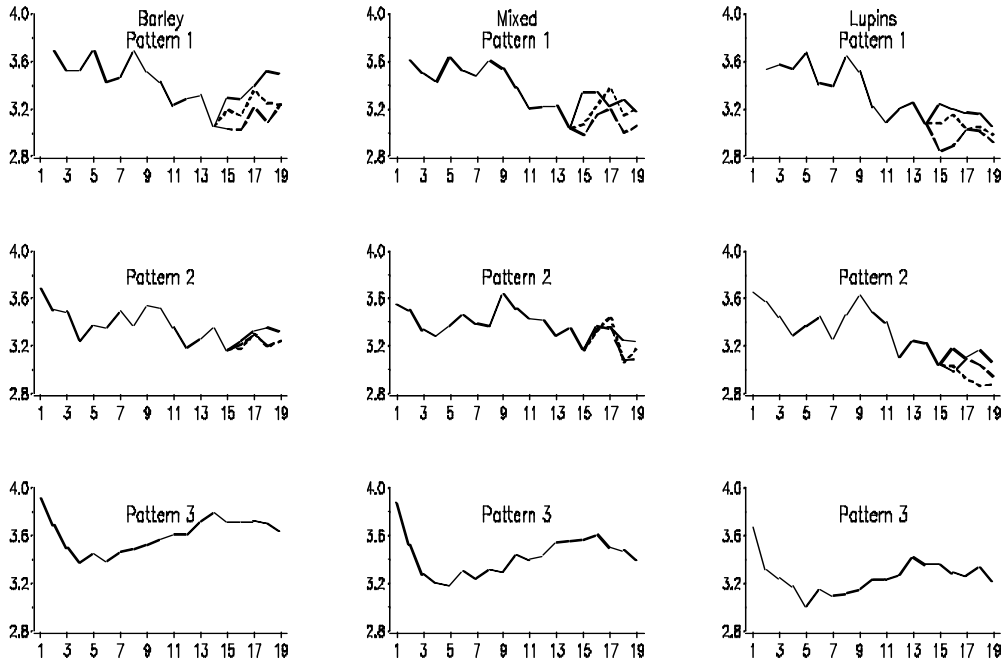

Figure 9.2: *Milk Protein Content Trial. Mean response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. Full line: CCMV, Dotted line: ACMV, Broken line: NCMV*

The ACMV and NCMV predictions look more plausible since the mean scores show some declining profiles. Should one want to explore the effect of assumptions beyond the range of (9.3), one can allow ω_s to include components outside of the unit interval. In that situation, one has to ensure that the resulting density is still non-negative over its entire support. Finally, completely different restrictions can be envisaged as well.

9.6.2 Hypothesis Testing

For ease of exposition, let us assume we are interested in a single effect such as treatment effect. In the particular case of the milk data, this translates into the time*diet interaction parameter. For simplicity and for future use, we will generically refer to the parameter

of interest as *treatment effect*. In the simplest case of a single parameter for the effect of interest, the corresponding selection model would contain exactly this single treatment effect parameter, turning the hypothesis testing task into a very straightforward one. If there were several treatment effect parameters, such as in a three-armed trial such as the milk study or in an analysis where interactions between treatment and other effects are included, standard hypothesis testing theory could be applied.

Some pattern-mixture models will have a treatment effect parameter specific to each pattern. This was the case in Sections 8.4 and 8.5. Let us note in passing that this does not need be the case. We will show in Section 9.6.3 that treatment effect is reduced to two single parameters independent of pattern. In such cases, the assessment of treatment effect is no more difficult than in a corresponding selection model. Therefore, this section will focus on the situation where there are pattern-dependent treatment effects.

It is useful to point out a strong analogy with post-hoc stratification, where pattern plays the role of a stratifying variable. A selection model corresponds to a pooled analysis, where data from all patterns (strata) are pooled, without correction for the "confounding effect" stemming from heterogeneity across dropout patterns. A pattern-mixture model on the other hand does correct for pattern and hence, in a sense, for the confounding effect arising from pattern. If treatment effect does not interact with pattern, such as in Section 9.6.3, then a simple, so-called *corrected*, treatment effect estimate is obtained. Finally, if treatment effect interacts with pattern there is heterogeneity of treatment effect across patterns (cf. heterogeneity of the relative risks in epidemiological studies).

One can calculate the same quantity as would be obtained in the corresponding selection model. Then, the *marginal* treatment effect is calculated, based on the pattern-specific treatment effects and the weighting probabilities, perhaps irrespective of whether the treatment effects are homogenous across patterns or not. This was done using equation 8.3 in Chapter 8.

Precisely, let $\beta_{\ell t}$ represent the treatment-effect parameter estimates $\ell = 1, \ldots, g$ (assuming there are $g + 1$ groups) in pattern $t = 1, \ldots, n$ and let π_t be the proportion of patients in pattern t. Then, the estimates of the marginal treatment effects β_ℓ are:

$$\beta_\ell = \sum_{t=1}^{n} \beta_{\ell t} \pi_t, \qquad \ell = 1, \ldots, g. \tag{9.27}$$

The variance is obtained using the delta method. Precisely, it assumes the form

$$\mathrm{Var}(\beta_1, \ldots, \beta_g) = AVA', \tag{9.28}$$

where

$$V = \left(\begin{array}{c|c} \text{Var}(\beta_{\ell t}) & 0 \\ \hline 0 & \text{Var}(\pi_t) \end{array} \right) \tag{9.29}$$

and

$$A = \frac{\partial(\beta_1, \ldots, \beta_g)}{\partial(\beta_{11}, \ldots, \beta_{ng}, \pi_1, \ldots, \pi_n)}. \tag{9.30}$$

The estimate of the variance-covariance matrix of the $\widehat{\beta}_{\ell t}$ is obtained from statistical software (e.g., the 'covb' option in the MODEL statement of the SAS procedure MIXED). The multinomial quantities are obtained from the pattern-specific sample sizes. In the case of the milk data, these quantities are presented in (8.7) and (8.8). A Wald test statistic for the null hypothesis $H_0 : \beta_1 = \cdots = \beta_g = 0$ is then given by

$$\boldsymbol{\beta}_0'(AVA')^{-1}\boldsymbol{\beta}_0, \tag{9.31}$$

where $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_g)'$.

We will now apply both testing approaches to the milk dataset. The CCMV case will be discussed in detail. The two other restriction types are entirely similar.

There are six treatment effects, one for each pattern by diet effect ($k = 2$) as in Equation (8.5) of Section 8.4. Hence, multiple imputation produces 5 vectors of 6 treatment effects which are averaged to produce a single treatment effect vector. In addition, the within, between, and total covariance matrices are calculated:

$$\boldsymbol{\beta}_{CC} = (0.1413, 0.0692, 0.3506, 0.0523, 0.1765, 0.0555)', \tag{9.32}$$

$$W_{CC} = \begin{pmatrix} 0.0109 & -1E-18 & 2E-19 & 0.0051 & -1E-18 & 2E-19 \\ -1E-18 & 0.0071 & -5E-19 & -1E-18 & 0.0036 & -7E-19 \\ 2E-19 & -5E-19 & 0.0037 & 6E-20 & -5E-19 & 0.0018 \\ 0.0051 & -1E-18 & 6E-20 & 0.0101 & -1E-18 & 1E-19 \\ -1E-18 & 0.0036 & -5E-19 & -1E-18 & 0.0071 & -7E-19 \\ 2E-19 & -7E-19 & 0.0018 & 1E-19 & -7E-19 & 0.0036 \end{pmatrix}, \tag{9.33}$$

$$B_{CC} = \begin{pmatrix} 0.0070 & 0.0018 & 0.0001 & 0.0069 & 0.0020 & 0.0001 \\ 0.0018 & 0.0014 & 9E-6 & 0.0012 & 0.0001 & 1E-5 \\ 0.0001 & 9E-6 & 8E-7 & 0.0001 & 2E-5 & 5E-7 \\ 0.0069 & 0.0012 & 0.0001 & 0.0076 & 0.0020 & 4E-5 \\ 0.0020 & 0.0001 & 2E-5 & 0.0020 & 0.0010 & 8E-6 \\ 0.0001 & 1E-5 & 5E-7 & 4E-5 & 8E-6 & 5E-7 \end{pmatrix}, \tag{9.34}$$

Table 9.1: *Milk Protein Content Trial. Tests of treatment effect for CCMV, NCMV, and ACMV restrictions.*

Parameter	CCMV	NCMV	ACMV
Stratified analysis:			
k	6	6	6
τ	24	24	24
denominator d.f. w	360.7	332.0	166.3
r	0.284	0.301	0.496
F statistic	4.64	4.48	3.79
p value	< 0.001	< 0.001	0.001
Marginal Analysis:			
Marginal effects (s.e.)	0.233(0.003)	0.233(0.004)	0.226(0.004)
	0.117(0.003)	0.134(0.003)	0.121(0.004)
k	2	2	2
τ	8	8	8
denominator d.f. w	49.3	45.3	24.1
r	0.317	0.339	0.604
F statistic	8.48	8.46	6.72
p value	0.001	0.001	0.005

and

$$T_{CC} = \begin{pmatrix} 0.0194 & 0.0022 & 0.0001 & 0.0133 & 0.0024 & 0.0001 \\ 0.0022 & 0.0088 & 1E-5 & 0.0015 & 0.0036 & 2E-5 \\ 0.0001 & 1E-5 & 0.0037 & 0.0001 & 3E-5 & 0.0018 \\ 0.0133 & 0.0015 & 0.0001 & 0.0192 & 0.0024 & 0.0001 \\ 0.0024 & 0.0036 & 3E-5 & 0.0024 & 0.0083 & 9E-6 \\ 0.0001 & 2E-5 & 0.0018 & 0.0001 & 9E-6 & 0.0036 \end{pmatrix}. \tag{9.35}$$

In the stratified case, we want to test the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Using (9.32)–(9.34), we can apply the multiple imputation results described in Section 9.5.2.

Note that, even though the analysis is done per pattern, the between and total matrices have non-zero off-diagonal elements. This is because imputation is done based on information from *other* patterns, hence introducing inter-pattern dependence. Results are presented in Table 9.1. All results are significant, in line with earlier evidence from Section 8.4 .

For the marginal parameter, the situation is more complicated here than in Section 8.4. Indeed, the theory of Section 9.5.2 assumes inference is geared towards the original vector, or linear contrasts thereof. Formula (8.3) displays a non-linear transformation of the parameter vector and therefore needs further development. First, consider $\boldsymbol{\pi}$ to be part of the parameter vector. Since there is no missingness involved in this part, it contributes to the within matrix, but not to the between matrix. Then, using (9.28), the approximate within matrix for the marginal treatment effect is

$$W_0 = A'WA + \boldsymbol{\beta}'\text{Var}(\boldsymbol{\pi})\boldsymbol{\beta},$$

with, for the between matrix, simply

$$B_0 = A'BA.$$

The latter formula consists of one term only, since there is no between-variance for $\boldsymbol{\pi}$. Note A is provided in Equation (8.5).

The results are presented in the second panel of Table 9.1. All three p values are similar and all agree on the significance of the treatment effect. The reason for the small differences observed in significance is to be found in the way the treatment effect is extrapolated beyond the period of observation. Indeed, the highest p value is obtained for NCMV restriction and, from Figure 9.1, we see that the differences between diet groups is less pronounced for this restriction method.

9.6.3 Model Reduction

Standard model building guidelines for the linear mixed-effects model can be used without any problem in a selection model context, but the pattern-mixture case is more complicated. Of course, the same general principles can be applied, taking into account the intertwining between the mean or fixed-effects structure and the components of variability.

In addition to these principles, one has to reflect on the special status of *pattern* in a pattern-mixture model. Broadly, we can distinguish between two cases as presented in Chapter 8 where pattern was included as a covariate or using identifying restrictions as presented in this chapter. In fact, the identifying restriction strategy leaves the method of analysis to be used after imputation unspecified, as mentioned in the strategy outline (Section 9.4.1). In our analysis, we have chosen to conduct a per-pattern global analysis, using pattern as

Table 9.2: *Milk Protein Content Trial. F-tests for multiple imputation estimates for CCMV, NCMV, and ACMV restrictions.*

Effect	F	K	DDF	P
CCMV				
3-way interaction	0.360	72	3946.6	0.999
diet×time	0.907	36	1584.7	0.626
diet×pattern	1.626	4	201.5	0.169
time×pattern	5.318	36	3143.0	<0.001
diet effect	9.881	2	52.0	<0.001
ACMV				
3-way interaction	0.316	72	2140.3	1.000
diet×time	0.762	36	938.6	0.844
diet×pattern	1.427	4	193.0	0.227
time×pattern	4.745	36	2394.3	<0.001
diet effect	8.880	2	34.5	0.001
NCMV				
3-way interaction	0.256	72	1742.2	1.000
diet×time	0.624	36	564.2	0.959
diet×pattern	1.482	4	117.0	0.212
time×pattern	4.421	36	971.7	<0.001
diet effect	5.536	2	15.9	0.015

a covariate, but it is possible to conduct a per-pattern analysis or even to use selection modeling. The only requirement is that the *proper* nature of the imputation is preserved (Rubin 1987).

As mentioned in Section 9.6.1 the most complex model for the means structure in conjunction with an autogressive covariance structure and a residual covariance component was fitted in each pattern. The estimates from the model fitting were then used to extrapolate the incomplete patterns. Although in model simplification we could initially begin with a complex model for the covariance structure we decided to use the covariance structure as suggested in Section 8.4, i.e., a separate autoregressive structure per pattern and a residual component common to all patterns. It is useful to note that the parameter vector may be quite large in pattern-mixture models especially when all parameters are allowed to depend on pattern. Indeed, Hogan and Laird (1997) noted that in order to estimate the large number of parameters in general pattern-mixture models, one has to make the awkward requirement

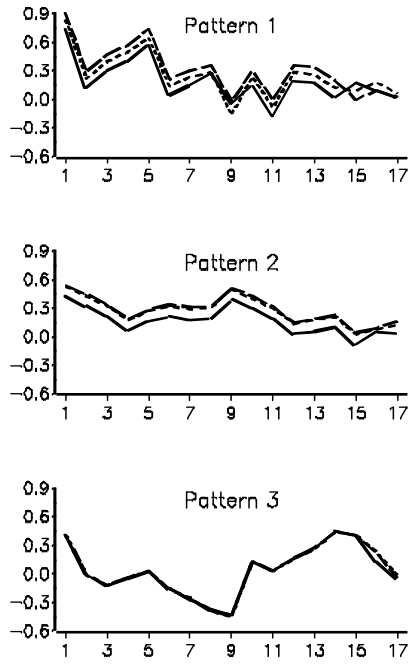

Figure 9.3: *Milk Protein Content Trial. Parameter estimates for the time effects for all three patterns using the identifying restrictions CCMV, ACMV and NCMV. Full line: CCMV, Dotted line: ACMV, Broken line: NCMV*

that each dropout pattern occurs sufficiently often. In the milk dataset we grouped cows who dropout out at weeks 16, 17 and 19 to reduce problems of estimation.

Our initial model fitted to the imputed data contained all 3-way interactions between diet, time and pattern. Model reduction was performed using the procedures defined in Section 9.5.2. The details concerning model simplification for all identifying restrictions are presented in Table 9.2. From a sensitivity analysis viewpoint it is comforting that model simplification under all 3 identifying restictions led to the same final model. This is probably not surprising since the proportion of dropout was small and restricted towards the end of the study. In addition there were only three dropout patterns with most observations in the complete cases. However, some differences were observed with respect to the F statistics and the denominator degrees of freedom. In contrast, with model fitting results in Section 8.4 the interaction between diet and pattern is excluded. This is mainly due to the fact that a

different parameterization was used in Model IV and VI of Section 8.4, i.e., the main effect for diet was not included in the model while interaction terms were allowed. As can be seen from Table 9.1 where the same parameterization was used consistent results are obtained. In the models described in Table 9.2 all main effects, two-way interaction terms and the 3-way interaction were included initially thus allowing straightforward interpretation of tests. For completeness, as the diet effect was not dependent on pattern in the final model we tested the diet effect on 2 degrees of freedom. Similar results were obtained to those shown in Table 9.1. The NCMV procedure demonstrated the least significant p value. This again illustrates that the NCMV restriction results in extrapolations which are consistent with the observed data for patterns 1 and 2 as shown in Figure 9.1.

The final parameter estimates for the time effects for all three patterns by identifying restrictions are presented in Figure 9.3. The remaining parameter estimates are presented in Table 9.3. Small differences in precision may be observed with the NCMV restrictions resulting in less precise estimates.

9.7 Analysis of QL Data

In Section 9.6 we applied the methodology developed for identifying restrictions to the milk data. In this section we will show how this approach may be used in the analysis of longitudinal QL data. QL data in longitudinal studies may be missing for a variety of reasons including progression of disease, treatment toxicity or patient refusal. However, most methods of analysis focus on a single dropout mechanism and do not take into account multiple reasons for dropout or patterns of missing data. In addition, these methods are based on strong assumptions which are not fundamentally testable because of the missing data. Using the identifying restriction strategy described in Section 9.4 to impute missing data, thus extrapolating incomplete patterns, we can incorporate both the reasons for missingness and the patterns of missingness into the imputation process. Employing several identifying restrictions allows us to perform a sensitivity analysis thus adressing the uncertainty caused by dropout. These concepts will be illustrated using an example from an EORTC trial.

EORTC trial 30903 was designed as a prospective multicenter randomized phase III study comparing flutamide versus prednisone in hormone resistant metastatic prostate cancer. Quality of life should have been evaluated at randomization, 3 and 6 weeks later, and at subsequent six weekly intervals. For more information on the data see Section 4.6. As illus-

Table 9.3: *Milk Protein Content Trial. Multiple imputation estimates and standard errors for CCMV, NCMV, and ACMV restrictions.*

EFFECT	ESTIMATE	St. Dev.	DDF	P
CCMV				
INTERCEPT	3.09	0.110	14.2	<0.001
Diet 1	0.23	0.054	57.2	<0.001
Diet 2	0.12	0.053	52.1	0.029
Pattern	0.20	0.105	28.1	0.068
Pattern 1	0.02	0.120	24.7	0.850
Variance Pattern 3	0.06	0.008	2759.6	<0.001
AR(1) Pattern 3	0.87	0.030	326.0	<0.001
Variance Pattern 1	0.05	0.013	321.3	<0.001
AR(1) Pattern 1	0.94	0.029	13.8	<0.001
Variance Pattern 2	0.07	0.014	102.7	<0.001
AR(1) Pattern 2	0.80	0.057	66.1	<0.001
Residual	0.02	0.002	31.1	<0.001
ACMV				
INTERCEPT	2.98	0.123	10.8	<0.001
Diet 1	0.23	0.061	22.8	0.001
Diet 2	0.13	0.053	59.7	0.014
Pattern 3	0.31	0.118	16.8	0.018
Pattern 1	0.03	0.104	184.2	0.739
Variance Pattern 3	0.05	0.009	2948.7	<0.001
AR(1) Pattern 3	0.88	0.026	2798.1	<0.001
Variance Pattern 1	0.06	0.016	25.8	0.002
AR(1) Pattern 1	0.92	0.035	19.4	<0.001
Variance Pattern 2	0.07	0.014	57.2	<0.001
AR(1) Pattern 2	0.80	0.050	2521.2	<0.001
Residual	0.03	0.002	304.6	<0.001
NCMV				
INTERCEPT	2.98	0.111	14.2	<0.001
Diet 1	0.22	0.076	9.9	0.017
Diet 2	0.12	0.065	15.1	0.095
Pattern 3	0.32	0.098	57.8	0.002
Pattern 1	-0.02	0.142	12.6	0.875
Variance Pattern 3	0.06	0.009	191.6	<0.001
AR(1) Pattern 3	0.88	0.029	104.9	<0.001
Variance Pattern 1	0.05	0.014	46.7	0.001
AR(1) Pattern 1	0.92	0.036	17.4	<0.001
Variance Pattern 2	0.07	0.016	24.2	<0.001
AR(1) Pattern 2	0.80	0.052	285.6	<0.001
Residual	0.02	0.003	15.6	<0.001

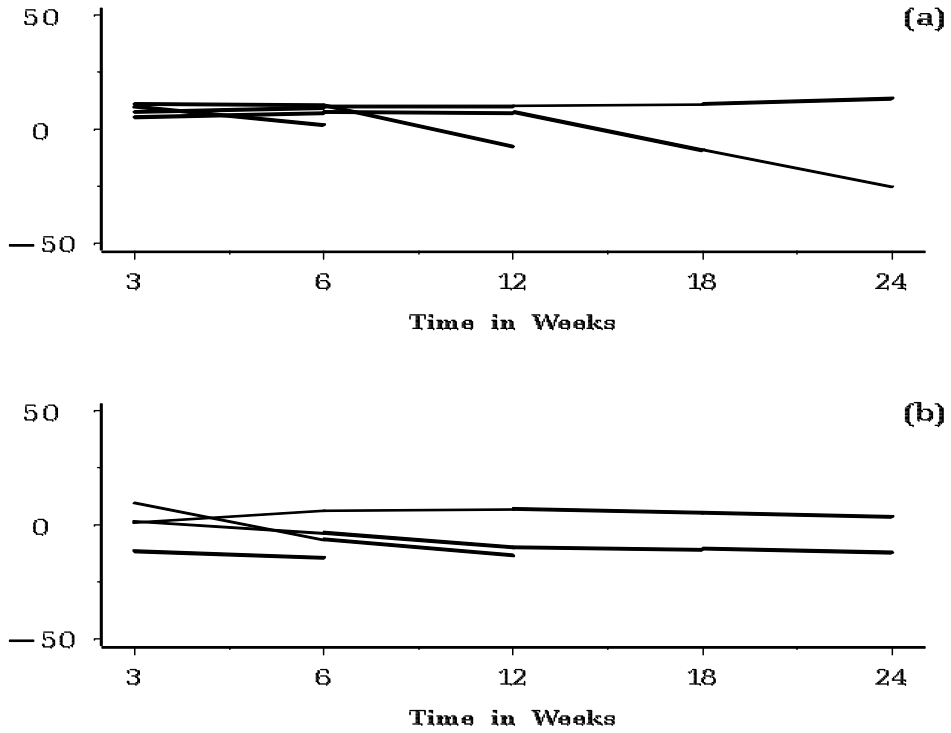

Figure 9.4: *EORTC Trial 30903. Mean profiles by dropout pattern and treatment a) prednisone b) flutamide.*

trated in Section 8.2 it is useful to explore longitudinal QL data using graphical techniques before advancing to model fitting. Initially we plotted the mean responses against time by treatment group and dropout pattern (see Figure 9.4). The dropout patterns were defined based on the dropout times, i.e., 3, 6, 12, 18, 24 and >24 weeks. However, as only 16 and 18 patients dropped out at week 24 and 30, respectively, patterns 4 and 5 were collapsed into 1 pattern. From Figure 9.4 it appears that the scores in the prednisone arm increase from baseline to 3 weeks, but tend to decrease just before dropout suggesting that dropout is not completely at random. In contrast the mean scores in the flutamide arm show very little change during the treatment period. The variance-covariance structure was investigated using several methods. A 5-dimensional scatter plot matrix of the data was generated as shown in Figure 9.5. The diagonal elements display the distribution of QL scores at each assessment time point. For presentation purposes the scores were divided into categories using midpoints: $-100, -75, -50, -25$, 0, 25, 50, 75, 100. The scatter plots (off diagonal) of assessments taken closer together (e.g., near the diagonal) appear to exhibit larger cor-

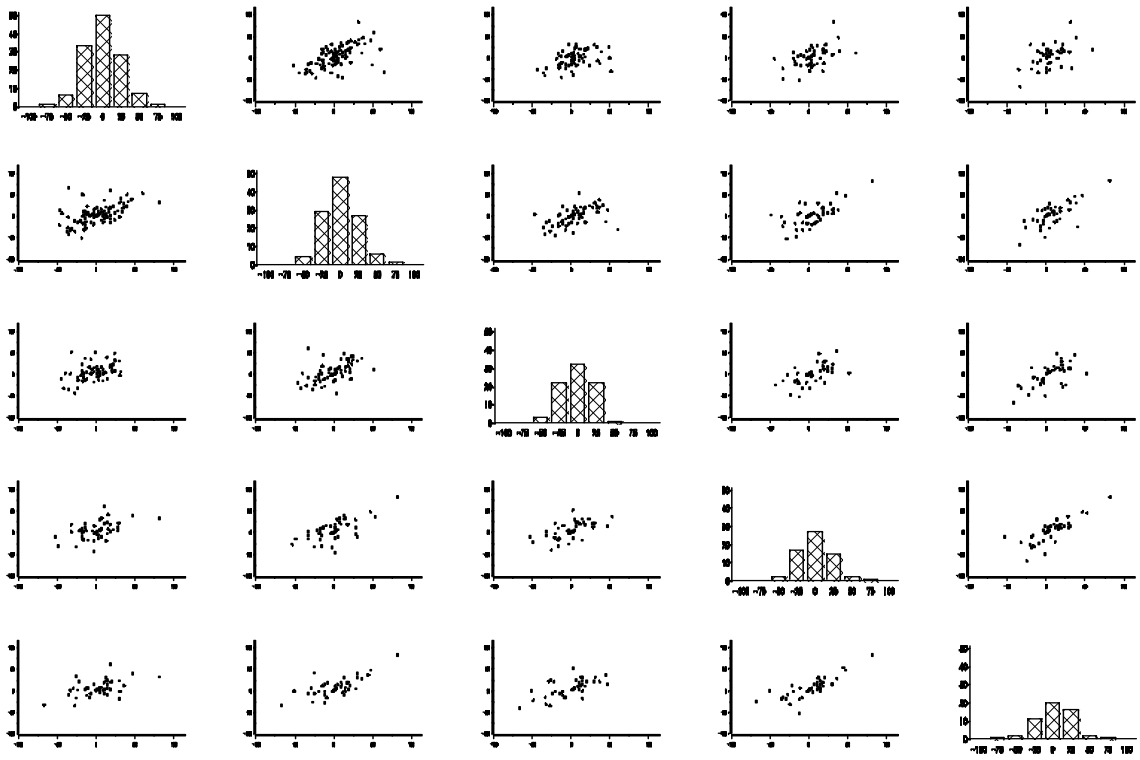

Figure 9.5: *EORTC Trial 30903. Scatterplot of change scores from baseline.*

relations than those taken further apart, suggesting an autoregressive covariance structure. As change scores were of interest no attempt was made to estimate the variogram. Thus the hypothesized variance structure was considered to be autoregressive with measurement error.

9.7.1 Pattern Mixture Model Fitted to Observed Data

We will fit a pattern-mixture model to the observed data and use standard model simplification techniques as presented in Chapter 8 to obtain a final model. This will be compared with the final models obtained using the identifying restrictions in Section 9.7.3.

Several baseline clinical variables were considered as covariates in the model. These included demographic variables: age, WHO performance status and pain assessed by the clinician.

The model fitting results are presented in Table 9.4. The most complex model (model I) for the means structure includes the main effects of the clinical variables and assumes a separate mean for each treatment-by-time-by-dropout pattern combination. As suggested by the scatterplot the covariance matrix was taken as autoregressive order 1 ($\sigma_{jk} = \sigma^2\rho^{|j-k|}$) with a residual covariance term (ϕ^2). A random intercept was not included as the response variable of interest was change score from baseline. The variance-covariance parameters are allowed to vary according to the dropout pattern. This model is equivalent to including time and treatment as covariates in the model statement and stratifying for dropout pattern. It provides a starting point for model simplification through backward selection. In models II and III we attempted to simplify the variance-covariance structure. Comparing models I and II indicated that indeed the residual covariance term could be considered equal between patterns. However, comparing model III with model II indicates that the AR1 term is significantly different between patterns.

In model IV we removed the 3-way interaction term between pattern, time and treatment. This model yields a non-significant likelihood ratio test statistic (p=0.086) when compared with model II suggesting that the means structure could be simplified further. To reduce the mean structure further we fitted models V to VII. In conclusion, among the models presented, model VI is preferred as it is the simplest model consistent with the data.

9.7.2 Fitting Models to the Imputed Data

The models used for the imputation process included baseline covariates and an interaction between time and treatment for the means model while the variance-covariance was defined as unstructured. As was described in Section 9.6.1 a separate model was fitted within each pattern. The resulting parameter estimates and their estimated asymptotic covariance matrices were used to extrapolate the patterns as described in Section 9.4.2. The multiple imputation, fitting of imputed datasets, and combination of the results into a single inference was performed using the SAS macro.

Although the macro automatically imputed scores for patients after death, for those patients who died before week 24, we subsequently deleted these imputed values. Some authors recommend imputation of values after death as they reduce bias caused by death (Hollen et al 1997). Imputation of scores after death is a controversial issue. We prefer not to impute values after death for the following reasons:

Table 9.4: *EORTC Trial 30903. Model fit summary for pattern-mixture models.*

	Mean	Covariance model
1	Full interaction	AR1(d), meas(d)
2	Full interaction	AR1(d), meas
3	Full interaction	AR(1), meas
4	Two-way interactions	AR1(d), meas
5	BCV, trt, time, pattern, trt∗pattern, time∗pattern	AR1(d), meas
6	BCV, trt, time, pattern, time∗pattern	AR1(d), meas
7	BCV, trt, time, pattern	AR1(d), meas

	par	-2ℓ	Ref	G^2	df	p
1	61	3851.54				
2	56	3855.74	1	4.2	5	0.521
3	46	3913.28	2	57.54	10	0.001
4	46	3872.24	2	16.5	10	0.086
5	42	3873.82	4	1.58	4	0.812
6	37	3880.32	5	6.5	5	0.261
7	27	3920.87	6	40.55	10	0.001

Par:	Number of parameters	P:	P-value
-2ℓ:	-2 times log-likelihood	AR1:	Autoregresive order 1
Ref:	Reference model	(d):	By dropout pattern
G^2:	Likelihood ratio test statistic	BCV:	Baseline clinical variables
df:	Degrees of freedom	trt:	Treatment

1. It is difficult to justify any imputed value after death

2. Various studies using utility measures have shown that on average patients are only willing to give up small amounts of survival time in exchange for much improved QL (Rosendahl et al 1999). If survival times are similar between treatment groups then the bias in parameter estimation due to death will be small. If survival times are significantly different between two groups then QL is generally considered to be a secondary issue. In addition, studies which demonstrate a significant survival benefit may also show an improvement in QL due to reduction in symptoms related to disease.

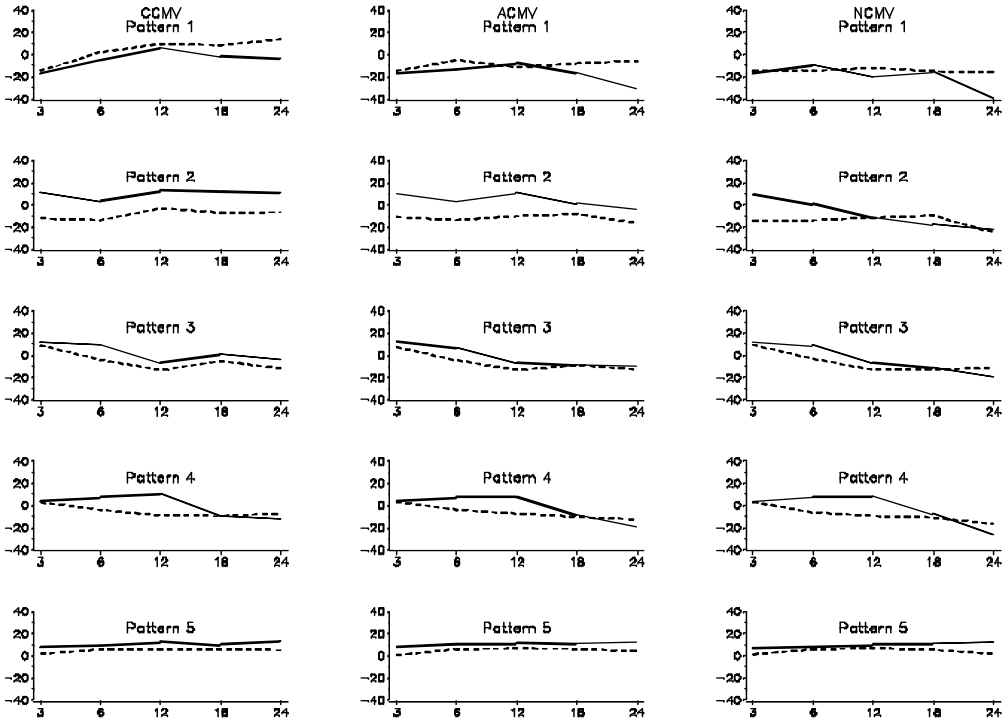

Figure 9.6: *EORTC Trial 30903. Mean response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. Full line: Prednisone, Dotted line: Flutamide*

In Section 4.6 it was shown that there was a considerable number of intermittent missing values, i.e., 46 patients had exactly 1 missing questionnaire and 11 patients had more than 1 missing questionnaire. Intermittent missing values were imputed using information from a patient's own missing data pattern. This is preferable to using information from other patterns since the patterns contain homogenous groups of patients as illustrated in Section 9.7.1 where it was demonstrated that the means model and the covariance structure were pattern dependent. On the other hand linear mixed models in the longitudinal setting treat intermittent missing values as MAR. Therefore, in theory it is not necessary to impute intermittent missing values if one considers MAR to be a valid assumption. However, if the reasons for missingness are available then this information can be included in the imputation process thus resulting in less biased estimates.

The initial multiple imputation results are presented graphically. Figure 9.6 shows the mean

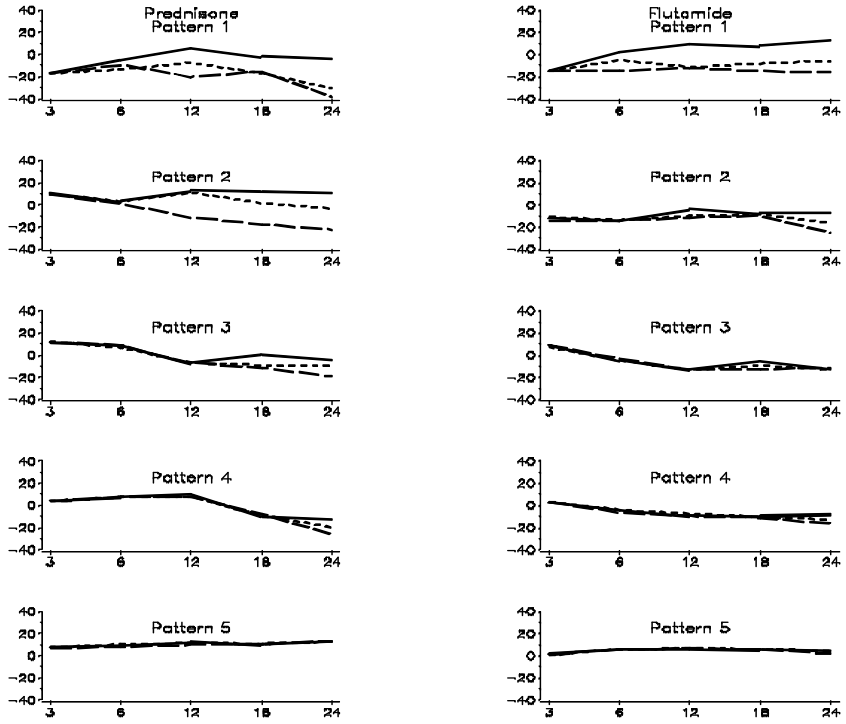

Figure 9.7: *EORTC Trial 30903. Mean response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. Full line: CCMV, Dotted line: ACMV, Broken line: NCMV*

response profiles for the multiply imputed datasets using the identifying restrictions CCMV, ACMV and NCMV. For patterns 1 to 4 there is some variability in the estimated profiles across the three restrictions. Using the CCMV identifying restriction results in an increase in both treatment arms in pattern 1, whereas using ACMV and NCMV restrictions results in the scores in the flutamide arm remaining approximately the same over time while in the prednisone arm they tend to decrease towards the fourth and fifth assessment. In general ACMV and NCMV tend to provide lower mean responses than for CCMV.

Figure 9.7 displays an alternative way of presenting the mean scores. Roughly speaking, CCMV extrapolates rather towards a rise whereas NCMV seems to predict more of a decline. Further, ACMV predominantly indicates a steady state. This conclusion needs to be considered carefully. Since these patients drop out mainly because they progress, a rise in QL seems unlikely. Hence, it is possible that the dropout mechanism is not CCMV, since this strategy

always refers to the 'best' group, in the sense that it groups patients who stay longer in the study and hence have on average a better prognosis. ACMV, which compromises between all strategies may be more realistic, but NCMV may be even better since information is borrowed from the nearest pattern, which is then based on the nearest patients in terms of dropout time and perhaps prognosis and quality of life evolution. However, recall that the identification is done sequentially, and hence even under NCMV, the parameter estimates for pattern 1 are identified borrowing from the remaining patterns chronologically.

9.7.3 Model Reduction

Using the knowledge gained from Section 9.7.1 we decided to define the covariance structure allowing a separate autoregressive structure per pattern and a residual component common to all patterns. The means model was initially defined using the baseline covariates and an interaction between time, treatment and pattern. Table 9.5 displays the F-tests obtained during model simplification. All 3 identifying strategies resulted in the same final model (except for ACMV) as was found in Section 9.7.1. The time by pattern interaction was not significant in the ACMV model. As stated before, the ACMV restriction extrapolated patterns predominantly indicating a steady state thus reducing the interaction between time and pattern.

Although all the restrictions did not yield the same final model we prefer to show the parameter estimates based on the model including the time by pattern interaction. The final parameter estimates for the time effects for this model are presented in Figure 9.8. The remaining parameter estimates are presented in Table 9.6.

In the 'imputation' step we included only baseline covariates and an interaction between time and treatment term. The model used for imputing the missing values will generally differ from the model used in the analysis. The primary objective in the 'imputation' model is to incorporate enough information in the model to ensure unbiased estimates of the missing values. For example, time dependent covariates such as performance status, disease status, weight loss, cumulative dose and treatment toxicity may be included in the 'imputation' model. Of course, these factors should not be included in the 'analysis' model as they are factors which are influenced by treatment and may confound the treatment comparisons. As was mentioned in Section 9.7 information concerning the reasons for dropout may be useful to identify homogenous groups of patients who dropout for the same reason. In 1995, the EORTC began collecting reasons for missing QL questionnaires in all phase III cancer

Table 9.5: *EORTC Trial 30903. F-tests for multiple imputation estimates for CCMV, NCMV, and ACMV restrictions.*

Effect	F	K	DDF	P
CCMV				
3-way interaction	0.760	16	406.6	0.731
trt1*pat	0.886	4	438.8	0.472
trt1*time	0.598	4	50.7	0.666
time*pat	2.590	16	266.2	0.001
ACMV				
3-way interaction	0.671	16	513.8	0.824
trt1*pat	0.712	4	904.7	0.584
trt1*time	1.241	4	305.5	0.293
time*pat	1.591	16	275.4	0.070
NCMV				
3-way interaction	1.461	16	601.1	0.109
trt1*pat	0.486	4	771.5	0.746
trt1*time	0.578	4	56.8	0.680
time*pat	2.589	16	599.8	0.001

clinical trials which include a QL component. The usefulness of this extra information has not been investigated as these studies have yet to mature.

9.8 Remarks

In this chapter, we have illustrated three distinct strategies to fit pattern-mixture models. In this way, we have brought together several existing practices. Little (1993, 1994a) proposed identifying restrictions, which we formalized here using the connection with MAR (Section 9.2) and multiple imputation (Section 9.5).

By contrasting these strategies on a single set of data, one obtains a range of conclusions

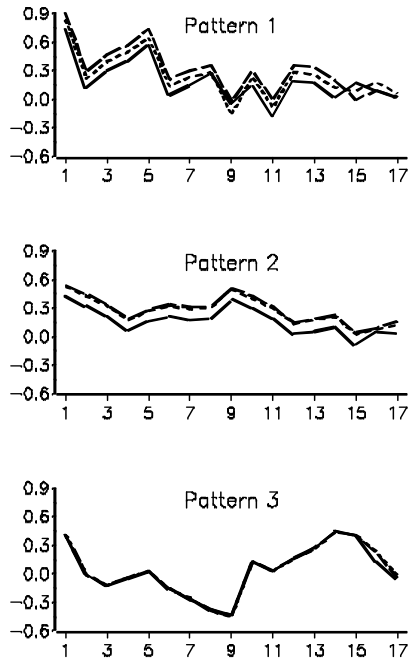

Figure 9.8: *EORTC Trial 30903. Parameter estimates for the time effects for all three patterns using the identifying restrictions CCMV, ACMV and NCMV. Full line: CCMV, Dotted line: ACMV, Broken line: NCMV*

rather than a single one, which provides insight into the sensitivity to the assumptions made. Especially with identifying restrictions, one has to be very explicit about the assumptions and moreover this approach offers the possibility to consider several forms of restrictions. Special attention should go to the ACMV restrictions, since they are the MAR counterpart within the pattern-mixture context.

In addition, a comparison between the selection and pattern-mixture modeling approaches is useful to obtain additional insight into the data and/or to assess sensitivity. This has been done, informally, in Chapter 8 using both the milk data set and a QL example. While these methods are computationally simple, it is important to note that there is a price to pay. Indeed, simplified models, qualified as "assumption rich" by Sheiner, Beale and Dunne (1997), are also making untestable assumptions, just as in the selection model case. Indeed, using the fitted profiles to predict the evolution, within a pattern, past the time of dropout

Table 9.6: *EORTC Trial 30903. Multiple imputation estimates for CCMV, NCMV, and ACMV restrictions.*

	CCMV			ACMV			NCMV		
EFFECT	EST	S.D.	P	EST	S.D.	P	EST	S.D.	P
INTERCEPT	15.27	13.95	0.274	14.46	13.78	0.294	3.55	12.88	0.783
MAGE	-0.64	2.98	0.830	-1.73	3.05	0.570	-4.35	2.77	0.118
WHO 1	4.19	4.31	0.331	4.79	4.29	0.264	4.88	4.16	0.245
WHO 2	7.42	5.28	0.161	6.42	5.71	0.264	2.83	5.90	0.637
WHO 3	9.88	8.17	0.227	8.90	7.97	0.264	8.77	7.95	0.275
Pain 4	-10.86	13.10	0.408	-10.56	12.65	0.404	3.62	11.82	0.760
Pain 1	-17.48	13.16	0.185	-17.22	12.79	0.179	-2.24	11.04	0.839
Pain 2	-10.86	12.83	0.397	-9.70	13.11	0.460	0.93	11.61	0.936
Pain 0	-20.18	13.70	0.141	-17.66	14.36	0.221	-3.42	12.79	0.789
Trt	7.62	2.98	0.011	6.60	2.97	0.027	4.75	3.02	0.123
Pattern 1	-8.21	8.75	0.349	-29.79	12.52	0.032	-36.21	9.62	<0.001
Pattern 4	-20.35	5.29	<0.001	-25.81	5.63	<0.001	-29.45	5.27	<0.001
Pattern 3	-18.71	7.28	0.014	-20.73	7.51	0.008	-26.75	7.57	0.001
Pattern 2	-9.15	5.77	0.114	-18.82	7.69	0.024	-30.81	7.98	0.001
Variance Pattern 1	599.82	230.92	0.016	643.66	253.14	0.024	444.35	155.39	0.015
AR(1) Pattern 1	0.90	0.14	0.001	0.82	0.14	<0.001	0.24	0.39	0.567
Variance Pattern 4	354.59	107.21	0.002	362.92	116.86	0.007	309.94	88.72	0.001
AR(1) Pattern 4	0.88	0.17	0.004	0.81	0.21	0.013	0.83	0.08	<0.001
Variance Pattern 3	322.84	134.41	0.031	393.74	136.94	0.015	312.72	78.28	<0.001
AR(1) Pattern 3	0.84	0.24	0.021	0.64	0.24	0.036	0.51	0.14	<0.001
Variance Pattern 2	280.15	137.43	0.077	380.93	134.42	0.018	373.58	83.16	<0.001
AR(1) Pattern 2	0.83	0.21	0.010	0.69	0.19	0.008	0.39	0.17	0.041
Variance Pattern 5	311.52	109.58	0.014	344.94	90.62	0.001	333.43	82.21	<0.001
AR(1) Pattern 5	0.94	0.15	0.003	0.88	0.18	0.007	0.95	0.04	<0.001
Residual	128.23	79.59	0.178	90.35	90.62	0.373	119.34	24.45	<0.001

is based on extrapolation. Still, the need for assumptions and their implications are more obvious. It is not possible for example to assume an unstructured time trend in incomplete patterns, except if one restricts attention to the time range from onset until dropout. In contrast, assuming a linear time trend allows estimation in all patterns containing at least two measurements. However, it is less obvious what the precise nature of the dropout mechanism is, whereas in the identifying restrictions setting the assumptions are clear from the start.

The identifying restrictions strategy provides further opportunity for sensitivity analysis, beyond what has been presented here. Indeed, since CCMV and NCMV are extremes, it is very natural to consider the idea of *ranges* in the allowable space of $\boldsymbol{\omega}_s$. Clearly, any $\boldsymbol{\omega}_s$ which consists of non-negative elements that sum to one is allowable, but also the idea of extrapolation could be useful, where negative components are allowed, given they provide

valid conditional densities.

Chapter 10

Longitudinal Categorical Data

10.1 Introduction

Generally QL is assessed using self-report questionnaires containing items (questions) with ordinal or binary response categories. Some of these items are subsequently collapsed into subscales which are also discrete in nature. In the literature these scales are frequently analyzed using the assumption of normality (possibly after transformation) or alternatively using non-parametric methods in cross-sectional analysis ignoring the longitudinal characteristics of the data. In this Chapter we analyze the physical functioning scale (PF) of the QLQ-C30 (version 1.0), which is a linear combination of five binary response items transformed to a 0 to 100 scale, with higher scores representing a higher level of functioning. The data were obtained from EORTC trial 30893, which was designed as a prospective multicenter randomized phase III study comparing orchidectomy and orchidectomy plus mitomycin C (15 mg/m^2 intravenously every six weeks until progression) in patients with poor prognosis metastatic prostate cancer. For more information on the dataset see Sections 4.5 and 8.2.

Most methods based on generalized linear models methodology (i) are useful for both discrete and continuous outcomes, (ii) do not require a constant number of repeated measurements per experimental unit, (iii) allow for differing measurement times across subjects and flexible covariate structures (discrete or continuous, time-independent or -dependent), and (iv) can accommodate missing data (MCAR).

Generalized linear models for longitudinal data can be categorized into three families. Firstly, the generalized linear model can be expressed as a marginal model where the marginal expectation $\mu_{it} = E(y_{it})$ ($i = 1, \cdots, N$ refers to an experimental unit and $t = 1, \cdots, n_i$ refers to a measurement time) is directly modelled in terms of covariates of interest, the marginal expectation being the average response over the subpopulation that shares a common value of the covariate vector. Associations among repeated observations are modelled separately. Secondly, it can be expressed as a random-effects model. Here the outcomes are modelled conditional on an unobserved (latent) random effect or a set of random effects. Subject-specific random effects are assumed to account for all the within-subject correlation that is present in the data. The individual-specific effects are used to explicitly model the heterogeneity among individuals. Thirdly, we mention conditional models in which an outcome is modelled conditional on the other outcomes or at least a set of other outcomes. For instance in transition (Markov) models, the conditional expectation of a current response, given past responses, is assumed to follow a generalized linear model.

In the linear model case, a marginal interpretation can be given to regression coefficients arising from each of the three approaches. However, whenever a nonlinear link function is imposed (e.g., in the case of binary outcome variables), the three approaches give different interpretations for the regression coefficients. More specifically, marginal models are most appropriate for making inferences about population averages. They are often applied in a clinical trial setting, since there the focus is generally on assessing average differences between treatment arms. Whereas marginal models follow a so-called population-averaged approach, random-effects models adopt a subject-specific approach. In the latter situation, regression coefficients have interpretations in terms of the influence of covariates on both an individual's response and the average response of the population. In transition models, different assumptions about time-dependence generally imply different interpretations of the regression coefficients.

As outlined in Chapter 2 models may be further classified into selection and pattern-mixture models. In this chapter we focus on selection models. For additional information on pattern-mixture models see Chapters 8 and 9. In this chapter the emphasis is on marginal models for a binary response. Given the initial goals of the clinical trial, this type of model is the most reasonable one. In particular, the expectation of the binary response at time t is related to a time trend and a set of covariates by the known linear logistic link function. Various methods for estimating the parameters of these (marginal) models are examined: likelihood based or using alternatives to likelihood theory. Within the likelihood framework, we propose a model which parameterizes the association in terms of marginal odds ratios (Molenberghs and Lesaffre 1994). Alternatively, we estimate the parameters of proposed

marginal models by using the generalized estimating equation approach (GEE) and by using weighted generalized estimating equations (WGEE). For technical details of model definitions and estimation procedures see Section 10.2.

10.2 Model Formulation and Estimation Procedures

Univariate generalized linear models have three components: (i) a random component identifying the response variable $\boldsymbol{Y} = (Y_1, ..., Y_n)$ in assuming a specific probability distribution for $\boldsymbol{Y}$, e.g., normal or binomial, (ii) a systematic component specifying the explanatory variables used as predictors in the model, e.g., age, treatment and time, and (iii) a link function which describes the functional relationship between the systematic component and the expected value of the random component. For some specific functions $a(.), b(.)$, and $c(.)$ (McCullagh and Nelder 1989), we assume that the independent random variables $Y_1, ..., Y_n$ arise from the distribution

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}.$$

This implies that if ϕ is known, the random component is determined by an exponential family model with canonical parameter θ.

Means and variances of $\boldsymbol{Y}$ are found using properties of the score function

$$U = \frac{\partial}{\partial\theta}\{l(\theta, \phi; y)\} = \frac{y - b'(\theta)}{a(\phi)}, \tag{10.1}$$

with $l(\theta, \phi; y)$ denoting the log-likelihood function as a function of θ and ϕ. It is easily shown that $E(Y) = b'(\theta)$ and $\text{Var}(Y) = b''(\theta)a(\phi)$.

In situations where it is not possible to construct a likelihood function (e.g., because the specific random mechanism by which the data are generated is unknown or the mean-variance relationship is different from the one implied by the model), quasi-likelihood can be used as a method for statistical inference. Whereas the random component of a generalized linear model assumes a specific distribution for the response Y_i, quasi-likelihood assumes only a form for the functional relationship between the mean and the variance.
More specifically, we assume that Y_i $(i = 1, ..., n)$ has mean $\mu_i(\boldsymbol{\beta})$, with $\boldsymbol{\beta} = (\beta_1, ...\beta_p)'$ the parameters of interest. Moreover, we set $\text{Var}(\boldsymbol{Y}) = \phi V(\mu)$, where $V(.)$ is a known function,

$V(\mu) = \text{diag}\{V(\mu_1), ..., V(\mu_n)\}$, and ϕ a possibly unknown scale parameter. Note that ϕ is assumed to be constant for all individuals and does not depend on $\boldsymbol{\beta}$, and that $\text{Var}(Y_i)$ only depends on μ_i: $\text{Var}(Y_i) = \phi V(\mu_i)$.
Taking

$$U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}, \tag{10.2}$$

it follows that

$$E(U_i) = 0, \tag{10.3}$$

$$\text{Var}(U_i) = \frac{1}{\phi V(\mu_i)}, \tag{10.4}$$

and

$$E\left(\frac{\partial U_i}{\partial \mu_i}\right) = -\text{Var}(U_i), \tag{10.5}$$

properties that also hold for log-likelihood derivatives (see Equation (10.1)). The quasi-likelihood for μ_i based on the data y_i is defined by

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y - t}{\phi V(t)}\, dt,$$

and behaves like a log-likelihood function for μ_i (provided the integral exists).

Extending the theory to the longitudinal setting, a generalized linear model needs to account for correlations among the multiple observations for an individual. The description of the model minimally requires specification of (i) a linear component $\boldsymbol{\eta_i} = X_i\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ a p-vector of (usually) unknown parameters, and (ii) a monotonic differential (vector) link function g (e.g., logit functions) describing how the expected value of $\boldsymbol{Y_i}$, denoted by $\boldsymbol{\mu_i}$, is related to the linear predictor $\boldsymbol{\eta_i} = g(\boldsymbol{\mu_i})$. The variance of $\boldsymbol{Y_i}$ is given by

$$\text{Var}(\boldsymbol{Y_i}) = \frac{\phi \boldsymbol{V}(\boldsymbol{\mu_i})}{w_i}, \tag{10.6}$$

where the dispersion parameter ϕ is a (possibly unknown) constant, w_i is a known weight for each observation, and $\boldsymbol{V}(.)$ is a known variance function. The correlation structure among the different time points is accounted for through a separate parameter vector $\boldsymbol{\alpha}$ in

$$\text{Corr}(\boldsymbol{Y_i}) = \boldsymbol{R}(\boldsymbol{\alpha}).$$

Note that the response vectors $\boldsymbol{Y_i}$ are independent for $i = 1, \cdots, N$, and that information retrieved from $\text{Var}(\boldsymbol{Y_i})$ and $\text{Corr}(\boldsymbol{Y_i})$ can be merged into one matrix $\boldsymbol{V_i(\alpha, \mu_i)}$.

A random-effects model is a generalized linear model accounting for both fixed effects $\boldsymbol{\beta}$ and random-effects parameters $\boldsymbol{b_i}$ as in

$$\boldsymbol{\eta_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} \quad \text{and} \quad \boldsymbol{\mu_i} = E(\boldsymbol{Y_i} \mid \boldsymbol{b_i}).$$

Note that now

$$\text{Var}(\boldsymbol{Y_i} \mid \boldsymbol{b_i}) = \frac{\phi \boldsymbol{V}(\boldsymbol{\mu_i})}{w_i} \quad \text{and} \quad \text{Corr}(\boldsymbol{Y_i} \mid \boldsymbol{b_i}) = \boldsymbol{R}(\boldsymbol{\alpha})$$

and that the parameter vector $\boldsymbol{b_i}$ satisfies the moment assumptions

$$E(\boldsymbol{b_i}) = 0 \quad \text{and} \quad \text{Cov}(\boldsymbol{b_i}) = G,$$

with G being a general covariance matrix. This model is fitted in Section 10.5.1.

The score equation for a generalized linear model in the longitudinal setting, is given by

$$U(\boldsymbol{\mu}) = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}} \right)' \text{Var}^{-1}(\boldsymbol{V_i(\alpha, \mu_i)})(\boldsymbol{Y_i} - \boldsymbol{\mu_i}) = \boldsymbol{0}. \tag{10.7}$$

In the full likelihood case, the variance-covariance structure of $\boldsymbol{Y_i}$ is completely determined. In practice however, $\text{Var}(\boldsymbol{Y_i})$ is often a function of $\boldsymbol{\mu_i}$ (via Equation (10.6)), and therefore unknown in advance. Hence Liang and Zeger (1986) proposed the GEE approach where besides the variance description as given in Equation (10.6) with $w_i = 1$, the form of a so-called working correlation matrix R needs to be specified. This working correlation matrix may depend on a vector of unknown parameters $\boldsymbol{\alpha}$, which is the same for all subjects. In fact, $\boldsymbol{\alpha}$ is called a working correlation matrix because with non-normal responses the actual correlation among a subject's outcomes may depend on the mean values. Although the matrix R can differ from subject to subject, we commonly use a working correlation matrix that approximates the average dependence among repeated observations over subjects.

Non-Likelihood Based Estimation The generalized estimating equations (GEE) approach is one of the most popular approaches to the analysis of correlated binary data (Liang and Zeger 1986, Zeger and Liang 1986, Zeger, Liang and Albert 1988). It is an extension of quasi-likelihood to longitudinal data analysis. The method is semi-parametric in that the estimating equations are derived without full specification of the joint distribution of a subject's observations. Only the likelihood for the (univariate) marginal distributions and a

working covariance matrix for the vector of repeated measurements from each subject need to be specified. The GEE's are solved by iterating between quasi-likelihood methods for estimating $\boldsymbol{\beta}$ and a robust (i.e., empirically based) method for estimating α, using the sandwich estimator, as a function of $\boldsymbol{\beta}$. It yields consistent estimates for $\boldsymbol{\beta}$ and the corresponding variances, even with misspecification of the structure of the covariance matrix. The efficiency loss relative to maximum likelihood methods is often minimal.

Zhao and Prentice (1990) and Liang, Zeger and Qaqish (1992) extended the GEE method (GEE2) while simultaneously estimating regression parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\alpha}$. In practice, this requires modeling the third and fourth moments of y_{ij}, instead of just modeling the mean and variance as in the previous case (also referred to as GEE1). Lipsitz and Kim (1994) extended Liang and Zeger's method to models for the correlation between repeated nominal and ordinal categorical responses; in particular, when the repeated responses are binary, their methods reduce to Liang and Zeger's method.

Since GEE's applied to incomplete data lead to valid estimates only if the missingness process is MCAR (this is often unrealistic for QL data), extensions of the method were investigated in the literature. In the weighted GEE approach (WGEE) proposed by Robins, Rotnitzky and Zhao (1995) an individual's contribution to the usual GEE is re-weighted by the inverse estimated probability of drop-out at the time of attrition through w_i in Equation (10.6). It leads to consistent estimates even with MAR. Troxel, Lipsitz and Troyen (1996) proposed weighted estimating equations (using the GEE structure of Liang and Zeger 1986) for data with non-ignorable non-response. Another class of weighted estimating equations was introduced by Robins, Rotnitzky and Zhao (1995): inverse probability of censoring weighted estimators in semi-parametric regression models.

Several marginal models are fitted, implementing weighted generalized estimating equations (Section 10.5.2) and generalized estimating equations (Section 10.5.3). Since the covariance matrix $\boldsymbol{V_i(\alpha, \mu_i)}$ is usually not fully known (note that variances follow from specifications for the mean structure), we can only make a plausible guess, leading to a so-called *working correlation matrix*. Based on this plausible guess for $\boldsymbol{V_i(\alpha, \mu_i)}$, the estimating equations (10.7) are solved, using a multivariate version of the iteratively weighted least squares algorithm. Note that for the weighted generalized estimating equations, specification of the weights w_i has the effect of weighting the contributions of the likelihood function by their value. The weights are obtained from a logistic regression for dropout.

Likelihood-Based Estimation
Within a likelihood-based analysis, the important question is to distinguish between missing

at random (MAR) and missing not at random (MNAR). Standard likelihood-based methods that ignore a MNAR dropout mechanism (such as the MIXED procedure in SAS developed for continuous outcomes) are subject to bias. Even when questions involving the dropout pattern are in place, for example if we condition on not having dropped out, then the dropout mechanism is important and even a shift between MCAR and MAR will change the conclusions. For multivariate categorical responses, likelihood-based regression models can be grouped into random-effects models, marginal models and conditional models.

Random effects based approaches are discussed in Wu and Carroll (1988) and Wu and Bailey (1989). Wu, Hunsberger and Zucker (1994) found that the 'conditional linear model' estimators of Wu and Bailey compared favorably in simulations with ML estimation under random dropout and also with nonparametric rank procedures described by e.g., Wei and Lachin (1984). Pulkstenis, Ten Have and Landis (1998) presented a selection model for binary data where response and dropout are independent conditional on the random effect.

Molenberghs and Lesaffre (1994) developed a full likelihood method for the analysis of ordinal categorical responses, allowing time-varying and subject-specific covariates: the multivariate Dale model, which is a marginal model. The model is based on an extension of the two-dimensional Plackett distribution (Plackett 1965) and on the bivariate global cross-ratio model described by Dale (1986) and McCullagh and Nelder (1989). The latter generalized linear model incorporates the multivariate Dale model (Molenberghs and Lesaffre 1994) in the case of the logit link for the marginal mean functions. McCullagh and Nelder (1989) expressed the link function in terms of joint probabilities $X\boldsymbol{\beta} = \boldsymbol{\eta} = C\ln(A\boldsymbol{\mu})$, with X a design matrix, $\boldsymbol{\mu}$ the vector of joint probabilities, A a matrix consisting zeros and ones, so that $A\boldsymbol{\mu}$ contains the marginal probabilities of all orders: the probabilities of each outcome separately, the probabilities for the cross-classification of all pairs of outcomes, for all triples, etc. Contrasts of log-probabilities are equated to a vector of linear predictors $\boldsymbol{\eta}$ using the contrast matrix C (of which elements are either 0, 1, or -1). The multivariate Dale model specifies the joint distribution by combining (proportional odds) logistic models for each outcome separately, with pointwise and higher order global odds ratios to describe pairwise and higher order associations (Molenberghs and Lesaffre 1994). Consequently patients can drop out at random without biasing the parameter estimates. The resulting likelihood is maximized by means of the EM algorithm.

In Section 10.5.4 we adopt this full maximum likelihood approach while fitting a multivariate Dale model. Further useful references which illustrate the method are Kenward, Lesaffre and Molenberghs (1994) (dealing with missing cases at random) and Molenberghs, Kenward and Lesaffre (1997) (covering non-random dropout).

Fitzmaurice, Laird and Lipsitz (1994) described a likelihood-based method for analysing balanced but incomplete longitudinal binary responses that are assumed to be missing at random. Following the approach outlined in Zhao and Prentice (1990), they focused on marginal models in which the marginal expectation of the response variable is related to a set of covariates. The association between binary responses is modeled in terms of conditional log odds-ratios (a hybrid Marginal-Conditional Model). The maximum likelihood estimates are obtained via an EM algorithm.

10.3 Exploratory Analysis

We now turn to the data introduced in Sections 4.5 and 8.2. In this Chapter, we choose to collapse the possible categories of PF into a binary outcome (PF ≤ 60 versus PF > 60). In Section 8.5 where the global health status/QL score of the QLQ-C30 was the response of interest, which has a minimum of 13 potential categories, the assumption of normality was made. However, for the PF score the necessity of acknowledging the categorical nature of the response is more pertinent and applying normal theory is less appropriate since the PF score may have at most 6 levels and the distribution of scores is notably more skewed. In order to get an idea of a plausible mean and correlation structure between repeated measurements, we initially treat PF as if continuous. We apply standard analysis techniques described in Section 8.2 for formulating the common mean structure in all models and for proposing (a) plausible covariance structure(s).

For our purposes, it is felt appropriate to focus on QL assessments during the first year only, as the majority of patients had dropped out before this time point. Briefly, an exploratory data analysis (residual profile plots, variogram function, scatterplot matrix,$\cdots$) was followed by a continuous longitudinal analysis to reduce an initial (possibly over-elaborated) structure. Evidence was found to explain part of the total variability by a subject-level component (a random intercept), a decaying serial correlation and so-called measurement error.

Based on model fit results obtained by temporarily treating PF as if continuous, besides effects for time and treatment, we included in all further models a time-treatment interaction effect as well as effects for the dichotomized baseline characteristics age (1: age < 68, 0: else), WHO performance status (0: WHO PS 0 or 1, 1: WHO PS 2, see Appendix A.2) and chronic disease status (1: yes, 0: no). In addition to their association with the PF score, the selected baseline characteristics are also known prognostic factors for patients

with metastatic prostate cancer. Note that since they influence survival time, they might affect underlying dropout processes.

In the remainder, time will be treated as categorical. The original 8 assessment time points of interest are collapsed into the following 4 time categories: baseline, 6-12 weeks, 18-24-30 weeks, 36-52 weeks. Combining the various time points was done for several reasons: reduction in the number of parameters, thus improving the estimation process and reducing the number of tests to be performed; limiting problems occurring due to sparse data.

10.4 Evidence Against MCAR

Chapter 2 introduced the various missing data mechanisms and noted that for a frequentist-likelihood approach such as the GEE approach of Liang and Zeger (1986) unbiased estimators are obtained only under MCAR. A WGEE approach or a likelihood-based method is valid under MAR as well. Hence, since the validity of a method depends on the inferential framework, we first investigated the type of dropout in our data set. Using the logistic regression model described in Chapter 7 we tested MCAR against MAR taking baseline covariates in to account:

$$\text{logit}(\text{pr}(\text{dropout at time } t+1)_i) = \alpha + (X_i, \boldsymbol{Y_i})\boldsymbol{\beta},$$

where α is the intercept, $\boldsymbol{\beta}$ is a vector of parameters, X_i is a matrix consisting of covariates for patient i such as treatment and time of assessment and $\boldsymbol{Y_i}$ is a vector of observed physical functioning scores.

Table 10.1 displays results of various logistic regression models for the probability of dropout given time (as continuous class variable), a treatment indicator and the PF score. The first fitted model includes 15 dummy variables representing the 16 possible time/treatment combinations and the physical functioning score PF. Models 2, 3, 7 and 8 are submodels of Model 1. Note that from Model 4 onwards time is treated as continuous. Comparisons of Models 1 and 2, 2 and 3, 1 and 3, 7 and 8, suggest that the dropout rate does not depend on treatment. Comparing Models 2 and 7, 3 and 8, provides strong evidence that the dropout rate varies with time. The validity of the proportional odds assumption for dropout over time is supported by comparing Models 3 and 6.

We further investigate, in Model 10, if the probability of dropout depends on the change from the current to the previous observed PF score and/or the sum of the two most recently

Table 10.1: *EORTC Trial 30893. Logistic regression results for testing the dropout mechanism.*

	Model	-2 ln L	df
1	Timeclass | Treatment + Y_t	609.455	822
2	Timeclass + Treatment + Y_t	613.462	829
3	Timeclass + Y_t	613.462	830
4	Time | Treatment + Y_t	623.918	834
5	Time + Treatment + Y_t	623.999	835
6	Time + Y_t	623.999	836
7	Treatment + Y_t	678.364	836
8	Y_t	678.445	837
9	Time + Y_t	441.212	579
			(= 582-3)
10	Time + $(Y_t - Y_{t\text{-}1}) + (Y_t + Y_{t\text{-}1})$	435.168	578
			(= 582-4)

Comparisons	G^2	df	p-value
1-2	4.007	7	0.779
2-3	0.000	1	>0.999
1-3	4.007	8	0.856
7-8	0.081	1	0.776
2-7	64.902	7	<0.001
3-8	64.983	7	<0.001
3-6	10.537	6	0.104
9-10	6.044	1	0.014

observed scores. The estimated Model 10 is

$$\begin{aligned}\text{logit}(\Pr(\text{dropout at time } t+1)) &= -1.788 + 0.050\,t - 0.026(Y_t - Y_{t\text{-}1}) \\ &\quad - 0.009(Y_t + Y_{t\text{-}1}),\end{aligned} \tag{10.8}$$

where both the effect for $(Y_t - Y_{t\text{-}1})$ and $(Y_t + Y_{t\text{-}1})$ are highly significant (p-values of 0.0001 and 0.0002, respectively). Equation (10.8) indicates that the probability of dropout increases with time, with a decrease of PF $(Y_t - Y_{t\text{-}1})$, and with a low overall score $(Y_t + Y_{t\text{-}1})$. This

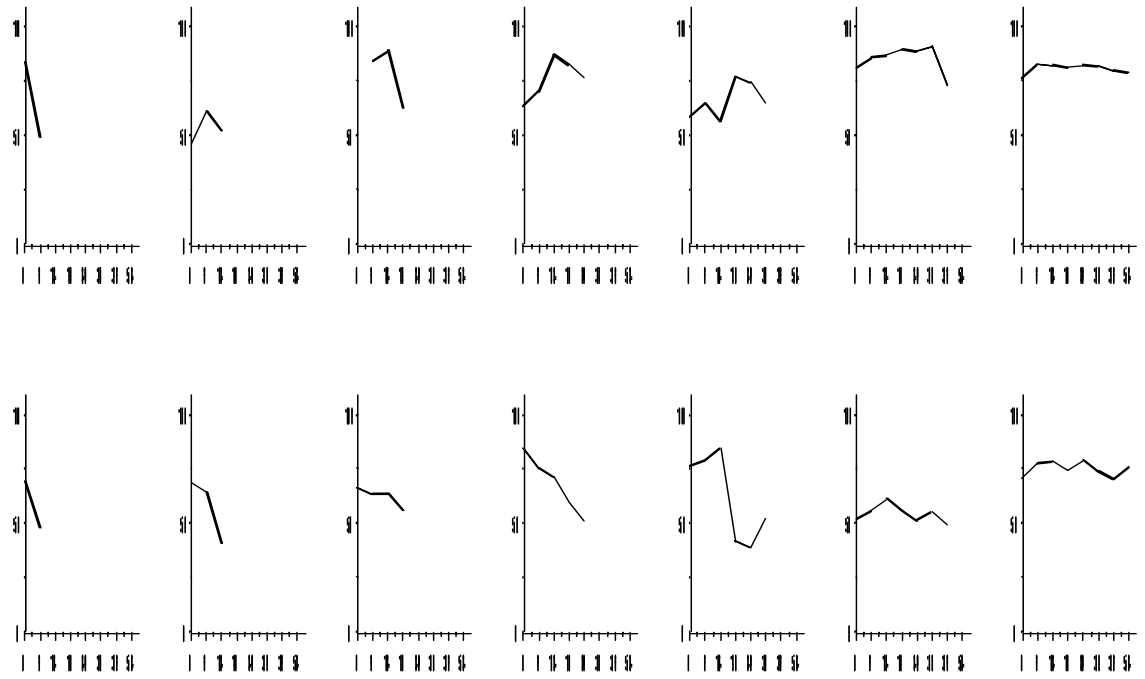

Figure 10.1: *EORTC Trial 30893. Average profile lines per treatment and time.*

is fully in line with graphical observations (Figure 10.1). In fact, the probability of dropping out, increases by a factor 0.597 (95% Wald confidence limits: [0.470,0.758] for every increase of 20 points in the change $(Y_t - Y_{t-1})$. Similar calculations with respect to $(Y_t + Y_{t-1})$ give an estimated factor of 0.830 with confidence limits [0.751,0.916]. Comparing this with Model 9, the likelihood ratio test suggests an improved model fit by accounting for the last but one observed PF score also (chi-square of 6.044, p-value = 0.014). Note that the number of observations in Model 9 was reduced, in contrast with Model 6, to ensure comparability with Model 10.

Hence, we strongly reject the hypothesis of MCAR. The latter observation is very important in the framework of model building and selection. Indeed, as mentioned in Diggle and Kenward (1994), if dropouts are not MCAR, we need to think carefully what are the relevant inferences. Modeling the dropout process also in this context allows the researcher to answer a richer class of inferential questions. The assessment of treatment differences among completers is such a question.

10.5 Different Approaches to Model Longitudinal QL Data

10.5.1 Random-Effects Models

Restricting attention to binary outcomes, we fit a random-effects model to our data using the GLIMMIX macro in SAS. Software such as MLWiN, Egret and MIXOR would also have served the same purpose. Based on the strong evidence to allow for a serial correlation (Section 10.3), and our interest in random effects, the covariance structure is specified by a random component and a serial correlation structure of the first order autoregressive type AR(1). The binary response (1: physical functioning score PF $>$ 60, 0: otherwise) is related to the covariates mentioned before, using the logit link function. We additionally investigate the effect of covariance structures with either a serial component and measurement error or only taking the serial component into account. Resulting parameter estimates and standard errors are listed in Table 10.2.

It is important to realize that all estimates listed in Table 10.2 are defined on the logit function scale. Hence, the intercept estimate (0.7642) in Model 3 is interpreted as the log odds of having a better PF score for patients younger than 68 in the orchidectomy arm with a relatively good physical performance status at baseline and no associated chronic disease. For fixed effects estimates as for WHO PS, the estimate (-0.6628) in Model 3 is interpreted as the decrease with respect to the previously mentioned log odds for patients with an associated worse performance status at baseline. Note, the parameter estimates are in line with clinical experience suggesting that old age, poor WHO PS and presence of chronic disease negatively influence PF. Additionally, PF increases after orchidectomy reflecting the known effect with respect to symptom relief.

Compared to Models 1 and 2, Model 3 leads to the most precise estimates for treatment, age, WHO PS and chronic disease. Parameters involving time levels of T are estimated with the highest precision in Model 2. All parameter estimates in Model 1 have larger associated standard errors compared to Models 2 and 3. Note the effect of having fewer observations at 36 or 52 weeks on the precision of the estimates involving these time points. All three models can pick up a significant effect for baseline chronic disease status. For example, the odds of having a physical functioning score $>$ 60 is reduced by a factor exp(-0.8832) $\approx$ 0.41 in Model 1 for patients with an associated chronic disease at baseline compared to those

Table 10.2: *EORTC Trial 30893. Glimmix Estimates.*

	Model 1 AR(1) + random intercept				Model 2 AR(1) + measurement error			
Effect	Estimate	s.e.	t	p-value	Estimate	Std. Error	t	p-value
INTERCEPT	1.0757	0.4085	2.6333	0.0092	0.8112	0.3166	2.5622	0.0113
Trt	-0.4823	0.4838	-0.9969	0.3192	-0.4613	0.3720	-1.2401	0.2167
Age	-0.4745	0.3555	-1.3347	0.1825	-0.2889	0.2637	-1.0956	0.2747
WHO PS	-0.8272	0.5275	-1.5682	0.1173	-0.6678	0.3972	-1.6813	0.0946
Chronic disease	-0.8832	0.3642	-2.4250	0.0156	-0.7011	0.2689	-2.6073	0.0100
T(36-52)	0.0253	0.3936	0.0643	0.9487	0.1325	0.3288	0.4030	0.6871
T(18-30)	0.6391	0.3459	1.8476	0.0651	0.4336	0.2833	1.5305	0.1263
T(6-12)	0.6493	0.3148	2.0626	0.0396	0.4252	0.2512	1.6927	0.0910
Trt*T(36-52)	-0.8141	0.5609	-1.4514	0.1471	-0.5982	0.4683	-1.2774	0.2019
Trt*T(18-30)	-1.3642	0.4756	-2.8684	0.0043	-0.8280	0.3878	-2.1351	0.0331
Trt*T(6-12)	-0.7301	0.4336	-1.6838	0.0927	-0.4397	0.3426	-1.2834	0.1997
Diff at baseline	0.4823	0.4838	1.0000	0.3192	0.4613	0.3720	1.2401	0.2154
Trt Diff 6-12Wks	1.2123	0.4083	2.9691	0.0031	0.9010	0.3096	2.9102	0.0037
Trt Diff 18-30Wks	1.8465	0.4115	4.4872	0.0001	1.2893	0.3171	4.0659	0.0001
Trt Diff 36-52Wks	1.2964	0.4947	2.6206	0.0090	1.0594	0.3812	2.7791	0.0056

	Model 3 AR(1)			
Parameters	Estimate	Std. Error	t	p-value
INTERCEPT	0.7642	0.3036	2.5171	0.0128
Trt	-0.4265	0.3651	-1.1682	0.2443
Age	-0.2682	0.2377	-1.1283	0.2608
WHO PS	-0.6628	0.3597	-1.8426	0.0672
Chronic disease	-0.6298	0.2435	-2.5864	0.0106
T(36-52)	0.1649	0.3571	0.4618	0.6443
T(18-30)	0.4687	0.3120	1.5022	0.1336
T(6-12)	0.4322	0.2572	1.6804	0.0934
Trt*T(36-52)	-0.5484	0.5058	-1.0842	0.2787
Trt*T(18-30)	-0.8741	0.4259	-2.0524	0.0405
Trt*T(6-12)	-0.4361	0.3523	-1.2379	0.2162
Diff at baseline	0.4265	0.3651	1.1682	0.2431
Trt Diff 6-12Wks	0.8626	0.3048	2.8301	0.0048
Trt Diff 18-30Wks	1.3007	0.3098	4.1985	0.0001
Trt Diff 36-52Wks	0.9749	0.3830	2.5454	0.0111

Trt=Treatment; T(36-52)=Time referring to 36 and 52 weeks; Trt*T(36-52)= effect of treatment at T(36-52); Diff at baseline= treatment difference at baseline; Trt Diff 6-12Wks= treatment difference at T(6-12) (similar definitions for T(18-30) and T(6-12), Trt*T(18-30) and Trt*T(6-12), Trt Diff 18-30Wks, Trt Diff 36-52Wks).

without. The decrease in WHO PS is borderline significant, with a p-value of 0.067. The decreasing p-values associated with WHO PS, from Model 1 over 2 to Model 3, seem to indicate that the random variability at baseline and measurement error are taken over by WHO PS in Model 3. This is not surprising, since the baseline physical functioning score and WHO PS are known to be strongly correlated.

Using a contrast statement significant treatment differences at 6-12 weeks, at 18-30 weeks and at 36-52 weeks are detected for all models. The mean baseline physical functioning score in the orchidectomy treatment arm is 71.44 compared to 64.02 in the orchidectomy plus Mitomycin C arm. This apparent difference is not significant in each of the three models. Performing analyses using change scores from baseline, is not appropriate for these data, since only 113 patients provided a PF score at baseline. However, we did take into account the differences at baseline in comparing treatment differences at a specific time point with baseline treatment differences (see also Table 10.1). For these data, both adjusted and unadjusted analysis for the baseline difference are to be recommended and may be seen as a type of sensitivity analysis. Note that considering treatment differences at a specific point in time implies drawing conclusions within a cross-sectional analysis only. In all models, treatment differences at 18-30 weeks appeared to be significantly different from baseline differences.

The fact that substantive conclusions remain the same over the three models suggests that a reasonable level of fit is reached by all models. We note that these results were similar to those from the linear mixed model with PF as continuous response (not shown).

10.5.2 Weighted Generalized Estimating Equations

As in section 10.4 we propose a logistic model for the dropout such that

$$\text{logit}(\text{pr}(\text{dropout})_i) = \alpha + X_i^T \beta,$$

where X_i^T is a matrix of covariates, initially including PF as binary, time as continuous, treatment, age, chronic disease, WHO PS as well as a time-treatment interaction term. Using 5% as a significance level, we can subsequently omit WHO PS, the interaction term, age and chronic disease status. We choose to keep treatment in the model (although the relatively high p-value of 0.826 strongly suggests no relationship between treatment and the probability of dropout). Subjects at a specific point in time are then inversely weighted by their estimated probability of being observed (w_i in Equation 10.6). Those patients who are unlikely to be observed in the sample are given increased weight in order to compensate for the other subjects with low observation probability who are in fact not observed. The parameter estimates obtained under AR(1), unstructured (UN) and exchangeable (EXCH) assumptions, are given in Table 10.3.

Parameter estimates seem to be less comparable over the three correlation structures com-

Table 10.3: *EORTC Trial 30893. Weighted GEE Estimates.*

		Empirical- based			Model -based		
Parameter	Estimate	Std Err	Z	Pr> $\|Z\|$	Std Err	Z	Pr> $\|Z\|$
WGEE - *exchangeable*							
INTERCEPT	0.6164	0.3084	1.9990	0.0456	0.3346	1.8423	0.0654
Trt at baseline	-0.4234	0.3790	-1.1170	0.2640	0.4038	-1.0490	0.2944
Age	-0.2714	0.2881	-0.9421	0.3461	0.2670	-1.0160	0.3094
WHO PS	-0.7026	0.4414	-1.5920	0.1114	0.4067	-1.7270	0.0841
Chronic disease	-0.7593	0.2836	-2.6770	0.0074	0.2733	-2.7790	0.0055
T(36-52)	0.0227	0.3521	0.0643	0.9487	0.3215	0.0705	0.9438
T(18-30)	0.6221	0.3693	1.6848	0.0920	0.3187	1.9518	0.0510
T(6-12)	0.5532	0.3187	1.7358	0.0826	0.3117	1.7747	0.0759
Trt*T(36-52)	-0.4499	0.4758	-0.9457	0.3443	0.4548	-0.9893	0.3225
Trt*T(18-30)	-0.9373	0.4591	-2.0420	0.0412	0.4321	-2.1690	0.0301
Trt*T(6-12)	-0.3806	0.4022	-0.9462	0.3440	0.4271	-0.8910	0.3729
WGEE - *AR(1)*							
INTERCEPT	0.7016	0.3136	2.2374	0.0253	0.3380	2.0757	0.0379
Trt at baseline	-0.5571	0.3803	-1.4650	0.1429	0.4124	-1.3510	0.1768
Age	-0.2274	0.2904	-0.7829	0.4337	0.2558	-0.8888	0.3741
WHO PS	-0.7793	0.4521	-1.7240	0.0847	0.3912	-1.9920	0.0463
Chronic disease	-0.7618	0.2869	-2.6550	0.0079	0.2623	-2.9040	0.0037
T(36-52)	0.0241	0.3642	0.0662	0.9472	0.3780	0.0638	0.9491
T(18-30)	0.5441	0.3630	1.4988	0.1339	0.3645	1.4928	0.1355
T(6-12)	0.4708	0.3171	1.4849	0.1376	0.3109	1.5142	0.1300
Trt*T(36-52)	-0.4345	0.4856	-0.8948	0.3709	0.5337	-0.8143	0.4155
Trt*T(18-30)	-0.8161	0.4545	-1.7960	0.0726	0.4941	-1.6520	0.0986
Trt*T(6-12)	-0.2344	0.4077	-0.5749	0.5653	0.4252	-0.5512	0.5815
WGEE - *Unstructured*							
INTERCEPT	0.4140	0.3152	1.3131	0.1891	0.3380	1.2248	0.2207
Trt at baseline	-0.2693	0.3836	-0.7020	0.4827	0.4080	-0.6601	0.5092
Age	-0.3190	0.2920	-1.0930	0.2745	0.2820	-1.1310	0.2579
WHO PS	-0.7514	0.4478	-1.6780	0.0933	0.4400	-1.7080	0.0877
Chronic disease	-0.7070	0.2860	-2.4720	0.0134	0.2897	-2.4410	0.0147
T(36-52)	-0.0232	0.3432	-0.0676	0.9461	0.2933	-0.0790	0.9370
T(18-30)	0.6093	0.3689	1.6518	0.0986	0.3357	1.8152	0.0695
T(6-12)	0.6328	0.3163	2.0008	0.0454	0.3126	2.0243	0.0429
Trt*T(36-52)	-0.5375	0.4820	-1.1150	0.2647	0.4111	-1.3080	0.1910
Trt*T(18-30)	-0.9896	0.4650	-2.1280	0.0333	0.4607	-2.1480	0.0317
Trt*T(6-12)	-0.5071	0.4059	-1.2490	0.2115	0.4306	-1.1780	0.2389

Note:
INTERCEPT: Orchidectomy effect at baseline
Trt at baseline: Adjuvant chemotherapy effect at baseline

pared to the associated empirical-based standard errors. Note for instance the estimated intercept term and treatment effect parameter. Moreover, the parameter associated with T(36-52) appears to change direction in affecting the probability of having a high (> 60) PF score when changing UN (-0.023, s.e. 0.343) into AR(1) (0.024, s.e. 0.364) or EXCH (0.023, s.e. 0.352). This however should not be overemphasized since this effect is non-significant (see standard errors) and is probably due to random variation only.

No straightforward choice between either correlation structures can be made solely based on comparisons between model- and empirical based variance-covariance matrices (not shown). However, at the end of Section 10.5.1, we concluded that a reasonable level of fit was reached by all models and that the exclusion of subject-specific variability (the random intercept) seemed to lead to a more significant WHO PS effect. Note that under EXCH in the WGEE approach, the WHO PS effect has an associated p-value of 0.112, compared to 0.085 under AR(1) and 0.093 under UN. In addition, the fixed effects parameter estimates appear to differ considerably dependent on the correlation structures used (see Table 10.3).

There appears to be no effect for age or WHO PS. Under all correlation structures, a significant effect for chronic disease CD can be detected. No differences in treatments at 6-12 weeks compared to baseline, nor at 36-52 weeks compared to baseline are observed. Treatment differences at the joint time points 18, 24, 30 weeks turn out to be significantly different from baseline treatment differences under UN and EXCH ($p = 0.033$ and 0.041, respectively). Under AR(1), only marginal evidence is found ($p = 0.072$).

10.5.3 Generalized Estimating Equations

In Section 10.4, evidence was found against MCAR and thus, strictly speaking, GEE is not valid. The GEE models fitted here are contrasted with the WGEE models of Section 10.5.2 to assess the sensitivity of inference on the missingness mechanism. Results of the GEE analysis are given in Table 10.4. Relying on earlier considerations, we focus on a first order autoregressive and an unstructured (working) correlation. For completion we compared the results obtained with those under the exchangeable working assumption. It is seen from Table 10.4 that the fixed effects parameter estimates are quite comparable over the various working assumptions, in contrast with the WGEE approach where much less similarity was detected. Deviations between the WGEE and GEE analysis might be expected, since the latter applied to incomplete data only leads to unbiased estimates if the missingness process is MCAR (which is a stronger assumption than MAR).

There is no significant effect for age, WHO PS or for treatment effect. It should be noted though that for WHO PS, similar p-values as in models 2 and 3 of section 10.5.1 are obtained (0.076, 0.088 and 0.094 respectively for AR(1), UN or EXCH). The effect for chronic disease at baseline turned out to be more significant under all working assumptions, compared to the random-effects models in Section 10.5.1. No differences in treatments at 6-12 weeks compared to baseline, nor at 36-52 weeks compared to baseline, are observed. The odds of

Table 10.4: *EORTC Trial 30893. GEE Estimates.*

		Empirical- based			Model -based		
Parameter	Estimate	Std Err	Z	Pr> $\|Z\|$	Std Err	Z	Pr> $\|Z\|$
GEE - *exchangeable*							
INTERCEPT	0.8485	0.3174	2.6731	0.0075	0.3222	2.6339	0.0084
Trt at baseline	-0.4625	0.3809	-1.2140	0.2246	0.3767	-1.2280	0.2195
Age	-0.2833	0.2827	-1.0020	0.3162	0.2692	-1.0520	0.2927
WHO PS	-0.7399	0.4419	-1.6740	0.0941	0.3961	-1.8680	0.0617
Chronic disease	-0.7724	0.2799	-2.7600	0.0058	0.2724	-2.8360	0.0046
T(36-52)	0.2300	0.3485	0.6601	0.5092	0.3340	0.6887	0.4910
T(18-30)	0.7635	0.3781	2.0195	0.0434	0.3227	2.3663	0.0180
T(6-12)	0.6138	0.3265	1.8797	0.0601	0.3042	2.0177	0.0436
Trt*T(36-52)	-0.4045	0.4675	-0.8652	0.3869	0.4735	-0.8542	0.3930
Trt*T(18-30)	-0.9227	0.4636	-1.9900	0.0466	0.4305	-2.1430	0.0321
Trt*T(6-12)	-0.3525	0.4086	-0.8627	0.3883	0.4127	-0.8541	0.3931
GEE - *AR(1)*							
INTERCEPT	0.8517	0.3185	2.6740	0.0075	0.3219	2.6461	0.0081
Trt at baseline	-0.5561	0.3810	-1.4590	0.1445	0.3824	-1.4540	0.1459
Age	-0.2410	0.2851	-0.8456	0.3978	0.2578	-0.9349	0.3498
WHO PS	-0.7867	0.4436	-1.7740	0.0761	0.3809	-2.0660	0.0389
Chronic disease	-0.7630	0.2821	-2.7050	0.0068	0.2618	-2.9140	0.0036
T(36-52)	0.3123	0.3583	0.8716	0.3834	0.3912	0.7983	0.4247
T(18-30)	0.7536	0.3702	2.0360	0.0418	0.3644	2.0683	0.0386
T(6-12)	0.5723	0.3227	1.7733	0.0762	0.2999	1.9084	0.0563
Trt*T(36-52)	-0.4448	0.4751	-0.9363	0.3491	0.5508	-0.8075	0.4194
Trt*T(18-30)	-0.8350	0.4581	-1.8230	0.0683	0.4866	-1.7160	0.0862
Trt*T(6-12)	-0.2357	0.4117	-0.5724	0.5671	0.4068	-0.5793	0.5624
GEE - *Unstructured*							
INTERCEPT	0.8547	0.3185	2.6836	0.0073	0.3235	2.6424	0.0082
Trt at baseline	-0.4890	0.3795	-1.2890	0.1975	0.3748	-1.3050	0.1919
Age	-0.2807	0.2829	-0.9921	0.3211	0.2777	-1.0110	0.3121
WHO PS	-0.7502	0.4393	-1.7080	0.0877	0.4088	-1.8350	0.0665
Chronic disease	-0.7731	0.2807	-2.7540	0.0059	0.2810	-2.7510	0.0059
T(36-52)	0.1914	0.3423	0.5591	0.5761	0.3221	0.5942	0.5524
T(18-30)	0.7448	0.3747	1.9879	0.0468	0.3364	2.2140	0.0268
T(6-12)	0.6073	0.3246	1.8708	0.0614	0.2823	2.1514	0.0314
Trt*T(36-52)	-0.3959	0.4579	-0.8646	0.3872	0.4553	-0.8696	0.3845
Trt*T(18-30)	-0.8907	0.4597	-1.9380	0.0527	0.4497	-1.9810	0.0476
Trt*T(6-12)	-0.3369	0.4101	-0.8215	0.4113	0.3829	-0.8799	0.3789

Note:
INTERCEPT: Orchidectomy effect at baseline
Trt at baseline: Adjuvant chemotherapy effect at baseline

having a higher physical functioning score PF in treatment arm 2 now tends to decrease by approximately 25% compared to the first treatment arm, for the joint time points 18, 24, 30 weeks.

As in Section 10.5.2, no correlation structure can be preferred above the other on the basis of efficiency in the parameter estimates nor on closeness between model-based and empirical-

based correlations and/or covariances (not shown).

10.5.4 Maximum Likelihood Estimation

In this section we fit a multivariate Dale model relating the same covariates as before to the PF as binary response variable. The fitting programs can only handle a relatively small number of assessment times. The relationship between response and the covariates CATAGE, chronic disease status CD and WHO PS, are held constant across time points. We allow different effects for Trt. For reasons of comparison we fix the three-order associations as well as the four-way association to unity ($\ln \psi_{123} = \ln \psi_{124} = \ln \psi_{134} = \ln \psi_{234} = \ln \psi_{1234} = 0$). The association structure is completed as follows:

$$\begin{array}{lll} \ln \psi_{12} & = \ln \psi_{23} & = \ln \psi_{34} \quad = \beta_8 \\ \ln \psi_{13} & = \ln \psi_{24} & = \beta_9 \\ \ln \psi_{14} & = \beta_{10} & \end{array} .$$

We fit

$$\text{logit}(p_j) \quad = \alpha_j + \beta_1 X_{\text{CATAGE}} + \beta_2 X_{\text{WHO PS}} + \beta_3 X_{\text{CD}} + \beta_{4+j} X_{\text{Trt}},\ j = 0, \cdots, 3.$$

The symbol p_0 refers to the marginal probability of having a PF score > 60 at the first of the four considered time categories (this is at baseline). The indices 1, 2, 3 refer to T(6-12), T(18-30) and T(36-52) respectively.

Studying the parameter estimates listed in Table 10.5 shows no apparent effect for CATAGE. Note the highly significant effect for CD and the borderline significant effect of having a relatively low WHO PS at baseline. The odds of having a better physical functioning score under the orchidectomy plus MMC treatment at T(18-30) weeks, is significantly decreased by a factor of exp(-1.381) $\approx$ 25% compared to treatment 1. We further notice the significant p-value with respect to T(6-12) and T(36-52) weeks ($p = 0.035$ and 0.021, respectively). There appears to be a highly significant treatment difference associated with T(18-36). Also the two-way associations are highly significant. The associations between the responses appear to be smaller between assessments which are more than 1 time point apart.

Table 10.5: *EORTC Trial 30893. Multivariate Dale Model Estimates.*

Parameter	Estimate	Std Err	Z	Pr> $\|Z\|$
INTERCEPT baseline	0.8656	0.3217	2.6907	0.0071
INTERCEPT T(6-12)	1.4160	0.3066	4.6187	0.0000
INTERCEPT T(18-30)	1.5949	0.3255	4.9002	0.0000
INTERCEPT T(18-30)	1.1202	0.3335	3.3592	0.0008
Age	-0.2865	0.2769	-1.0346	0.3008
Chronic disease	-0.7771	0.2808	-2.7675	0.0056
WHO PS	-0.7590	0.4034	-1.8816	0.0599
Trt*baseline	-0.4688	0.3721	-1.2598	0.2078
Trt*T(6-12)	-0.7294	0.3451	-2.1134	0.0346
Trt*T(18-30)	-1.3811	0.3641	-3.7937	0.0001
Trt*T(36-52)	-0.9316	0.4040	-2.3058	0.0211
2-Way Ass Dist 1	2.6503	0.3278	8.0860	0.0000
2-Way Ass Dist 2	1.7008	0.3882	4.3812	0.0000
2-Way Ass Dist 3	2.1387	0.6071	3.5227	0.0004
Diff 6-12 and Base	-0.2606	0.3776	-0.6901	0.4901
Diff 18-30 and Base	-0.9123	0.4354	-2.0953	0.0361
Diff 36-52 and Base	-0.4328	0.4549	-1.0174	0.3090

Trt*baseline= treatment effect at baseline; 2-Way Ass Dist 1= 2-way associations between responses that are consecutive in time; Diff 6-12 and Base= comparison of treatment difference at 6-12 with treatment difference at baseline (similar definitions for Trt*T(6-12), Trt*T(18-30), Trt*T(36-52), 2-Way Ass Dist 2, 2-Way Ass Dist 3, Diff 18-30 and Base, Diff 36-52 and Base).

10.6 Remarks

We used various approaches to analyze this QL data, preceded by a graphical exploratory analysis. The main purpose of this detective work is to structure the huge amount of information generally contained in a longitudinal data set. More specifically, the exploratory analysis is used to formulate ideas with respect to measurement and covariance structures.

There are two viewpoints within the exploration. The first looks at the individual level, the second considers averages. For each view, specific techniques for continuous responses are

Table 10.6: *EORTC Trial 30893. Comparison of Treatment Differences (Baseline Reference).*

Type	Parameter	Estimate	Std error	Z	Pr> |Z|
Diff at baseline	WGEE - AR(1)	-0.5571	0.3803	-1.4650	0.1429
	GEE - AR(1)	-0.5561	0.3810	-1.4590	0.1445
	DALE	-0.4688	0.3721	-1.2598	0.2078
Diff 6-12 and Base	WGEE - AR(1)	-0.2344	0.4077	-0.5749	0.5653
	GEE - AR(1)	-0.2357	0.4117	-0.5724	0.5671
	DALE	-0.2606	0.3776	-0.6901	0.4901
Diff 18-30 and Base	WGEE - AR(1)	-0.8161	0.4545	-1.7960	0.0726
	GEE - AR(1)	-0.8350	0.4581	-1.8230	0.0683
	DALE	-0.9123	0.4354	-2.0953	0.0361
Diff 36-52 and Base	WGEE - AR(1)	-0.4345	0.4856	-0.8948	0.3709
	GEE - AR(1)	-0.4448	0.4751	-0.9363	0.3491
	DALE	-0.4328	0.4549	-1.0174	0.3090

readily available to study mean trends, variance and covariances (see Section 8.2). Physical functioning is strictly speaking not a continuous variable, but it has at least 6 levels, and hence continuous methods seem justified for exploration. Individual profile plots may be useful to distinguish cross-sectional from longitudinal patterns. Care has to be taken when heterogeneous populations are involved. For larger data sets (leading to 'busy' individual profiles) and in the presence of many important covariates, average profile plots may be more straightforward to assess average trends.

Residual profiles can be studied to see whether there is constant variability over time, in which case we would not include other random effects than intercepts in our model. Standardized residual plots may highlight the importance of certain variance components. Note that the total variability can be split into a subject-level component, a serial component and measurement error variability. Based on the variance function, it seemed plausible to assume a constant variance over time. We therefore studied the variogram which showed that the most important part of the process variance should be ascribed to a decaying serial correlation. A scatter plot matrix of residuals with lowess estimated trends, may also be helpful in studying the correlation structure.

Relying on the exploratory analysis, we assumed that every individual profile could be modeled with time as a linear effect. Data reduction from the pool of potential covariates was

performed using a backward selection procedure. To this end, the model fit statistics of the SAS procedure MIXED were used.

We emphasized the importance of studying the underlying missing data processes. An advantage of likelihood methods is that they can handle various types of incomplete data. When the mechanism of missingness is MCAR both likelihood and frequentist methods can be applied. With MAR missingness likelihood methods can be applied, while this does not necessarily hold true for frequentist methods. It is therefore important to gather as much information as possible about the missingness mechanisms in the data.

Compound symmetry is seldom an appropriate covariance structure in longitudinal data derived from cancer clinical trials. Nevertheless, we retained it in most of the non-likelihood analyses to assess the sensitivity of specifying different covariance structures (compound symmetry, first-order autoregressive, unstructured covariance matrix) on the results. Note that a random-effects model allows us to ascribe part of the variability in the data to random effects. Therefore, generalized linear mixed models may also contribute to such a sensitivity analysis.

Several estimation procedures (likelihood based or not) were considered and evaluated in a second type of sensitivity analysis. Note that the performed GEE analyses must be seen as sensitivity analyses, since evidence against MCAR was found. The GEE method is attractive because, rather than having to make full distributional assumptions, it suffices to specify the marginal expectation of the repeated measures (as in a cross-sectional study, using a generalized linear model). The loss in efficiency (Liang and Zeger 1986) caused by replacing the true correlation matrix by a working correlation matrix is, in most cases, negligible. This may be explained by the use of the sandwich estimator which provides empirical estimates of the standard errors and results in consistent estimates even under mis-specification of the working correlation matrix.

Although the full data set was rather small, a reasonable consistency was observed between the various models. However, minor shifts were detected in the significance valuation of certain covariates and/or the precision by which they could be estimated, when comparing estimation approaches. In particular, the significance of the treatment differences at the various time points deviated from those obtained with other modeling approaches. This could partially be explained by the fact that the estimation procedure for convergence had to be weakened. Choosing one of the several existing approaches is primarily based on the scientific objective of the study, combined with the need to describe the data adequately. Where primary interest lies in inferences about the marginal parameters, as is normally the

case in clinical trials, methods such as GEE and weighted GEE are appropriate. However, a weighted GEE approach should be applied for analyzing data with dropouts not missing completely at random. When interest lies in prediction and classification, good estimates of the joint probabilities are required and a likelihood-based approach (such as the multivariate Dale model) is to be preferred.

Where the correlation structure is of primary interest, a random-effects model might be a more powerful tool. Although random-effects modeling can be extended to non-normal responses, some uncertainty exists as to the proper interpretation of the random effects estimated in such a setting (Breslow and Clayton 1993). If the response is non-normal and the correlation or covariance parameters *are* of primary interest, a range of alternative approaches exist which may be preferable (Breslow and Clayton 1993, Neuhaus 1992). These include solutions based upon Markov chain Monte Carlo methods including Gibbs sampling (Zeger and Karim 1991). A feature of the technique, especially in the light of longitudinal QL data, is that it leads to valid estimates only if the missing data process is MAR.

The WGEE approach only requires specification of the missing data mechanism and the marginal mean (not the associations), rather than the joint distribution of the missing data indicators and the response vector of interest, which is an advantage over maximum likelihood estimation. Used on incomplete data, the method yields consistent estimators when the responses are MAR.

The multivariate Dale model, as used in Section 10.5.4, is a marginal model, based on cumulative probabilities of (latent) continuous variables. Dependence between the outcomes is taken into account via (generalized) global cross-ratios. As not only the marginal distributions and the bivariate cross-ratios are taken into account, the model gives rise to a complete specification of the joint probabilities. Another strength of the multivariate Dale model is its ability to deal with multi-categorical responses.

Software Used

The exploratory analysis of Section 10.3 was mainly performed in S-plus. The continuous longitudinal analysis that followed was performed using PROC MIXED in SAS. It allows for unbalanced data and can fit a wide range of models and covariance structures. The additional GLIMMIX macro in SAS was used to fit a generalized linear mixed model. Note

however that GLIMMIX produces model statistics such as the deviance and the scaled deviance (defined as the deviance divided by the extra-dispersion parameter) that treat all outcomes as if they were independent! In addition, the model-fitting information provided by the PROC MIXED output is based on a linear transformation of the original data. Its use is therefore questionnable. A fast and user-friendly implementation of the GEE method is provided by the SAS procedure GENMOD. Model fit can be assessed using the scaled deviance. Also the WGEE approach is easy to implement. We used PROC GENMOD again and specified the missing data mechanism by calling the SAS option SCWGT w_i. It has the effect of multiplying the contributions of the log-likelihood function, the gradient, and the hessian matrix by w_i. Fixed effect parameters are estimated using estimating equations, whereas estimates for the correlations are moment-based. Fitting a multivariate Dale model is computationally somewhat more cumbersome and special attention has to be given to specifying the association structure and/or starting values for the estimation procedure. Computations are in general more time consuming than for instance the GEE approach of Liang and Zeger (1986) or the weighted GEE approach. For a proper performance of the available programs written in the statistical package GAUSS, a limited number of covariates and time points is recommended. Additional software is available to combine the multivariate Dale model with a logistic regression model for dropout. The latter however still needs further development in order to be accessible to a wider audience.

Chapter 11

Discussion

Successful integration of QL endpoints into clinical trials requires a comprehensive approach to research design, study implementation and statistical analysis. Much attention has been devoted to the latter two issues, however study design has received less focus. In Chapter 4 several QL studies were described. In most of these studies it was planned to assess QL until progression of disease, treatment failure or death, which ever occurred first. As such, most dropouts were design driven. If the primary objective of a study is to compare the effect of several palliative treatments on QL without expecting to prolong progression-free survival then collecting QL while patients remain on treatment or progression-free may be entirely appropriate. However, if the main objective of the trial is to extend the time to progression or duration of survival it is vital that QL is assessed until death in order to examine the overall impact of treatment on QL.

Further attention needs to be devoted to handling cases that are missing due to death. These data are 'missing' in a very special way; they are not actually missing but are no longer available since the subject is no longer alive. It is not sensible to treat their 'missing' values in the same way as those missing because of other reasons. Various studies using utility measures have shown that on average patients are only willing to give up small amounts of survival time in exchange for much improved QL (Rosendahl et al 1999). If survival times are similar between treatment groups then the bias in treatment comparisons due to death will be small. If survival times are significantly different between two groups then QL is generally considered to be a secondary issue. Thus, performing a conditional analysis conditioning on survival status appears to be a possible solution.

As with the design of the QL component of a study, sufficient care and attention should be taken at the beginning of a study to ensure an adequate infrastructure, including appropriate personnel and material to carry out the study. No matter how well the analysis is thought out and how accurate assumptions are about missing data mechanisms, inferences in the presence of incomplete data are not as convincing as inferences based on a complete dataset.

Although, summary measures and summary statistics have been used extensively in QL research to-date, they are limited for various reasons. Longitudinal data analyses make more use of the available data. For example, they allow one: to examine the correlation structure between repeated assessments during model fitting; to describe the between-patient variability and within-patient variability; to take into account that patients with poorer scores may be more likely to dropout earlier, and therefore produce potentially less biased results. In addition, the dropout rate may be conditional on covariates such as treatment group. It is not necessary to assume linear change of QL scores over time and analyzing the data longitudinally can circumvent the problems that arise when the treatment schedule, and thus the QL assessment schedule, is not synchronized for treatment arms. Thus more sophisticated techniques provide a clearer overall picture of the impact of disease and treatment on the QL of patients. Although longitudinal techniques require a higher level of statistical sophistication from the analyst the results can easily be disseminated and interpreted by a non-statistical audience.

The use of selection models in QL settings appears to be intuitive because the dropout process is thought of as being dependent on the measurement process. In contrast, the interpretation of pattern-mixture models is not so obvious since it implies that the QL scores for an individual are dependent on the time that patient will drop out. With selection models assumptions need to be made concerning the dependence of the dropout process on measurements which have not been obtained. Similarly, in Chapter 7 we observed that pattern-mixture models are underidentified. The missing data taxonomy is usually presented in the selection modeling framework rather than in the pattern-mixture context. Using the terminology developed by Rubin (1976) the dropout process may be classified as MCAR, MAR or MNAR. In Chapter 9 we showed that pattern-mixture models can be classified similarly, and further that the intermediate MAR category is equivalent to the ACMV restriction in the case of monotone missingness. Using identifying restrictions all the parameters in the model may be identified and so estimates for these parameters and the marginal probabilities may be obtained. This provides a way to compare ignorable selection models with their counterpart in the pattern-mixture setting.

Although selection models and pattern-mixture models are considered to be probabilistically

equivalent, they shed different light in the context of a real data analysis. For example, in pattern-mixture models the overall distribution of the longitudinal measurements is a mixture of the conditional distributions, given the pattern of missingness. Therefore, the overall distribution may not necessarily be multivariate normal.

In Chapter 9 we outlined a general framework for identifying restrictions in Section 9.4, with particular attention being given to CCMV, ACMV and NCMV as special cases. We illustrated that ACMV and NCMV may be of particular interest in QL settings. The CCMV strategy borrows information from the the 'best' group in the sense that it groups patients who stay longer in the study and hence have on average a better prognosis. ACMV, which compromises between all strategies may be more realistic, but NCMV may be even better since information is borrowed from the nearest pattern, which is then based on the nearest patients in terms of dropout time and perhaps prognosis and quality of life evolution.

Using the procedure described in Chapter 9 the analysis is performed in a series of steps. This is advantageous as useful information such as the reasons for missingness, the patterns of missingness and time dependent-covariates may be incorporated into the imputation process. Multiple imputation accounts for *sampling uncertainty* leading to unbiasedness in terms of both point estimates and measures of precision. Once the data is imputed inference may be performed in any number of ways provided the *proper* nature of the imputation is preserved (Rubin 1987). For example, one could conduct a per-pattern global analysis using pattern as a covariate or even use selection modeling.

By contrasting these strategies on a single set of data, one obtains a range of conclusions rather than a single one, which provides insight into the sensitivity to the assumptions made. The identifying restrictions strategy provides further opportunity for sensitivity analysis, beyond what has been presented here. Indeed, since CCMV and NCMV are extremes for the $\boldsymbol{\omega}_s$ vector in (9.3), it is very natural to consider the idea of *ranges* in the allowable space of $\boldsymbol{\omega}_s$. Clearly, any $\boldsymbol{\omega}_s$ which consists of non-negative elements that sum to one is allowable, but also the idea of extrapolation could be useful, where negative components are allowed, given they provide valid conditional densities.

As with many areas in statistical research we focused on the continuous data setting. However, an extension of the identifying restrictions strategies developed here to the categorical data situation deserves further research. The approach presented here was developed into a SAS macro. One of the priorities for analysis of QL measurements with missing data must be improvement of the computational aspects of the methodology and widespread dissemination of usable software. Without tools of this sort, even the most effective method will be

rendered useless—if a method is too difficult to implement, it will not be routinely used and will eventually fail. Development of faster and easier software will facilitate the accumulation of expereince with models for MNAR missing data, and will make feasible more extensive testing of the methods, especially to determine the effects of model mis-specification.

The problem of missing data in quality of life research in clinical trials research is by no means a trivial one, and while some solutions have begun to appear, a great deal of work remains to be done. The work performed in this thesis constitute a step in the journey toward a valid and complete analysis of quality of life data.

Chapter 12

Nederlandse Samenvatting

Dit werk richt zich op longitudinale gegevens uit data die levenskwaliteit bestuderen. Zulke studies zijn onderhevig aan onvolledigheid in het algemeen en uitval ter wille van een verscheidenheid aan redenen in het bijzonder. In Hoofdstuk 1 wordt een overzicht gegeven van de eigenheid van studies m.b.t. levenskwaliteit. Onvolledigheid komt niet alleen vaak voor, ze heeft ook verstrekkende gevolgen voor data manipulatie en analyse. Gedurende de laatste decennia werden een aantal eenvoudige *ad hoc* methoden voorgesteld om het probleem van onvolledige datasets te ondervangen. Een aantal hiervan worden overlopen en de gevaren ervan worden onderstreept. Naast eenvoudige technieken, zoals een analyse van de volledige gegevens en *available case* analyse, wordt er recent meer werk gemaakt van het expliciet modelleren van onvolledigheid. Sectie 1.4 geeft hiervan een overzicht.

In Hoofdstuk 2 wordt notatie en terminologie ingevoerd om het spreken over onvolledige longitudinale studies te vergemakkelijken, in de context van levenskwaliteit. De familie van missing data mechanismen zoals ingevoerd door Rubin (1976) worden voorgesteld, alsook begrippen zoals ignorability, separabiliteit en vertekening.

Hoofdstuk 3 behelst een literatuurstudie. Verscheidene methoden voor het omgaan met onvolledige gegevens worden ingevoerd, te beginnen met de klassieke methoden, zoals volledige en beschikbare gegevens, om te vervolgen met imputatietechnieken. Zowel enkelvoudige als meervoudige imputatie worden beschreven. Hierop aansluitend geven we een overzicht van likelihood methoden, zowel voor continue als voor discrete respons variabelen.

Hoofdstuk 4 introduceert de longitudinale datasets welke doorheen dit werk gebruikt worden. Hoofdstuk 5 legt nadruk op het manipuleren van vragenlijsten die één of meer onvolledige items bevatten. Methoden om de analyse hiervan mogelijk te maken worden besproken: (1) case deletion (2) eenvoudige imputatie van het gemiddelde en (3) algemene imputatiemethoden. Eenvoudige imputatie van het gemiddelde is de meest verbreide methode en is gebaseerd op traditionale psychometrische methoden voor het ontwerpen van schalen en de analyse ervan. We geven voorbeelden van situaties waar deze aanpak niet aangewezen is en alternatieve imputatiemethoden dienen beschouwd.

Hoofdstuk 6 bestudeert verscheidene technieken, voorgesteld in de literatuur, gekend onder de termen summary measures en summary statistics. Deze methoden worden geïllustreerd m.b.v. data, verzameld in EORTC klinische studie 10921, in lokaal geavanceerde borstkanker. In het bijzonder tonen we aan dat verschillende methoden tot verschillende conclusies leiden m.b.t. kwaliteit van het leven. De beperkingen van deze methoden worden aangegeven: (1) informatie gaat verloren omdat niet alle observaties gebruikt worden en (2) ze kunnen vertekening veroorzaken omdat ze geeen rekening houden met ontbrekende gegevens en de mechanismen die hiervoor verantwoordelijk zijn.

In Hoofdstuk 7 worden twee methoden voor het identificeren van het type van onvolledigheid in levenskwaliteit onderzocht. De eerste aanpak is gebaseerd op het verzamelen van informatie over waarom de gegevens niet volledig verzameld werden. Dit kan als basis dienen om een onderscheid te maken tussen verschillende mechanismen. De tweede methode is erop gericht het missing data mechanisme te modelleren en, hierop gebaseerd, een onderscheid te maken tussen mechanismen via het toetsen van hypothesen. Twee methoden voor het onderzoeken of de onvolledige gegevens missing completely at random (MCAR) zijn worden voorgesteld en toegepast op onvolledige levenskwaliteit gegevens uit internationale multicentrische kankerstudies. De eerste methode (Ridout 1991) is gebaseerd op logistische regressie en de tweede methode (Park en Davis 1993) vertrekt vanuit een modificatie van gewogen kleinste kwadraten. In één toepassing (geavanceerde borstkanker) was het MCAR zijn niet plausibel. In de tweede applicatie (vroeg stadium borstkanker) hing het al of niet MCAR zijn af van de gebruikte schaal (haarverlies; anxiety). MCAR en MAR (missing at random) hebben verschillende implicaties voor data analyse. Het is daarom van belang er een onderscheid tussen te maken. Discriminatie tussen MAR en missing not at random (MNAR) is niet evident en vereist veronderstelling die fundamenteel niet kunnen getoetst worden (Glynn, Laird en Rubin 1986).

Hoofdstukken 8 en 9 bestudeert continue respons variabelen. Twee alternatieve kaders voor het modelleren van onvolledige longitudinale gegevens zijn in omloop: selectiemodellen (Dig-

gle en Kenward 1994) en pattern-mixture modellen (Little 1993, Little 1995 en Hogan en Laird 1997). Zij benaderen het probleem van dropout op een verschillende manier: in een selectiemodel wordt de dropout kans beschreven, conditioneel op het meetproces. In een pattern-mixture model worden de responsen gemodelleerd, conditioneel op dropout. Selectiemodellen kunnen gebruikt worden in een sensitiviteitsanalyse om de invloed van verscheidene veronderstellingen op, bijvoorbeeld, het effect van behandeling te onderzoeken. Selectiemodellen vereisen veronderstellingen die fundamenteel niet kunnen getoetst worden. In een pattern-mixture model wordt het dropout mechanisme gewoonlijk eenvoudig gemodelleerd, eventueel via een multinomiale verdeling om de proportie van de patiënten in een bepaald patroon te beschrijven. De vereiste hierbij is dat er voldoende subjecten per patroon zijn om efficiënt schatten mogelijk te maken. Het aanpassen van een selectiemodel kan tamelijk complexe vormen aannemen. Een pattern-mixture model leidt, eens aangepast, op een voor de hand liggende manier tot marginale grootheden, zoals het marginale effect van behandeling. In Hoofdstuk 8 worden beide methoden vergeleken a.h.v. twee datasets: de milk protein content trial (Diggle en Kenward 1994) en levenskwaliteit in een EORTC klinische studie.

De natuurlijke parameters in selectiemodellen en pattern-mixture modellen hebben een verschillende betekenis en het ene kader in het andere vertalen is niet voor de hand liggend, zelfs niet voor normaal verdeelde gegevens. Zoals eerder aangehaald, dienen niet verifieerbare veronderstellingen gemaakt te worden m.b.t. de relatie tussen missing data en respons (discussie van Diggle en Kenward 1994, Molenberghs, Kenward, en Lesaffre 1997). In pattern-mixture modellen is het expliciet zichtbaar welke parameters niet identificeerbaar zijn. Little (1993) suggereert het gebruik van identificerende verbanden tussen identificeerbare en niet-identificeerbare parameters. Dus, ondanks het feit dat zulke relaties zelf niet kunnen geverifieerd worden (Little 1995), het voordeel is een grote duidelijkheid over welke informatie vervat is in de data en welke niet. In Hoofdstuk 9 stellen we een nieuwe strategie voor voor het aanpassen van pattern-mixture modellen die natuurlijk leiden tot sensitiviteitsanalyse. We verkennen het idee van het exploreren van onvolledige patronen onder verscheidene assumpties. Dit idee werd eerder gesuggereerd door Hogan (1999).

Levenskwaliteit wordt niet alleen longitudinaal opgetekend maar ook gebruik makend van ordinale schalen. In de recente literatuur wordt meer en meer aandacht gegeven aan dit soort van gegevens wanneer ze ook onderhevig zijn aan onvolledigheid. Het doel van Hoofdstuk 10 is het nagaan van verschillen in statistische conclusies wanneer verschillende modellen gebruikt worden. Dit wordt geïllustreerd aan de hand van een EORTC faze III studie. Hoofdstuk 10 legt de klemtoon op selectiemodellen. Voor informatie m.b.t. pattern-mixture modellen verwijzen we naar Michiels, Molenberghs, en Lipsitz (1999). In Hoofdstuk 10 fit-

ten we eerst een random-effects model om een binaire longitudinale respons (afgeleid van de fysische functioneringsschaal van de QLQ-C30) te linken aan verscheidene covariaten. In een tweede aanpak worden marginale modellen aangepast gebaseerd op de eerder afgeleeide gemiddelde structuur. Het aangepaste marginale model verschilt m.b.t. de schattingsmethode: generalized estimating equations (GEE), weighted generalized estimating equations (WGEE) en maximum likelihood (ML).

Appendix

A.1 Variogram

Specializing (8.1) to random intercept only, D simplifies to a scalar, d say, and it is easy to show (Diggle 1990) that the variogram equals

$$V(u) = \sigma^2 + \tau^2(1 - \rho(u)),$$

where $u = t_{ij} - t_{ik}$ is the time lag between both measurements and $\rho(u)$ is the serial correlation between two measurements with the specified lag, calculated for example from

$$\text{Corr}(t_{ij}, t_{ik}) = \exp\left(-\frac{|t_{ij} - t_{ik}|}{\phi}\right) = \rho^{|t_{ij}-t_{ik}|}, \tag{1}$$

where $\rho = \exp(-1/\phi)$ or from the Gaussian counterpart:

$$\text{Corr}(t_{ij}, t_{ik}) = \exp\left(-\frac{(t_{ij} - t_{ik})^2}{\phi^2}\right) = \rho^{(t_{ij}-t_{ik})^2}, \tag{2}$$

with $\rho = \exp(-1/\phi^2)$.

Note that $V(0) = \sigma^2$ and $V(\infty) = \sigma^2 + \tau^2$. Plotting the process variance,

$$\text{Var}(Y_{ij}) = \nu^2 + \sigma^2 + \tau^2,$$

as a horizontal line and the variogram as a curve, the three components of variability are easy to retrieve. The measurement error is $V(0)$, the random intercept variance is the difference between the process variance and $V(\infty)$, and the variance of the serial process is seen as the band, occupied by the variogram, which increases from $V(0)$ to $V(\infty)$. With irregularly spaced data, it is usually necessary to smooth the variogram. The shape of the variogram conveys information about the structure of the serial correlation function.

A.2 WHO Performance Status

The WHO Performance Status is determined by a medical doctor when conducting a patient examination.

Grade	Performance status
0	Able to carry out all normal activity without restriction.
1	Restricted in physical strenuous activity but ambulatory and able to carry out light work.
2	Ambulatory and capable of all self-care but unable to carry out any work: up and about more than 50% of waking hours.
3	Patient is up and about <50% of waking hours.
4	Complete disabled; cannot carry on any self-care, totally confined to bed or chair.

A.3 Prostate (ICD-O 185); T, N, M and G Categories

Rules for Classification
The classification applies only to carcinoma. There should be histological confirmation of the disease. The following are the procedures for assessment of the T, N and M categories:

T categories	Physical examination, imaging, endoscopy and biopsy
N categories	Physical examination and imaging
M categories	Physical examination, imaging, skeletal studies and biochemical tests

Regional Lymph Nodes
The regional lymph nodes are the nodes of the true pelvis which essentially are the pelvic nodes below the bifurcation of the common iliac acteries. Laterality does not effect the N classification.

TNM Clinical Clasification

T - primary tumour

TX	primary tumour
T0	no evidence of primary tumour
T1	Tumour is incidential histological finding
	T1a 3 or fewer microscopic foci of carcinoma
	T1b More than 3 microscopic foci of carcinoma
T2	Tumour present clinically or grossly, limited to the gland
	T2a Tumour 1.5 cm or less in greatest dimension with normal tissue on at least three sides
	T2b Tumour more than 1.5 cm in greatest dimension or in more than one lobe
T3	Tumour invades into the prostatic apex or into or beyond the prostatic capsule or bladder neck or semonal vesical, but is not fixed
T4	Tumour is fixed or invades adjacent structures other than those listed in T3

N-regional lymph nodes

The definitions of the N categories apply to all urological sites except penis. There are:

NX Regional lymph nodes cannot be assessed
N0 No regional lymph node metastasis
N1 Metastasis in a single lymph node 2 cm or less in greatest dimension
N2 Metastasis in a single lymph node more than 2 cm but
not more than 5 cm in greatest dimension,
or multiple lymph nodes, none more than 5 cm in greatest dimension
N3 Metastasis in a lymph node more than 5 cm in greatest dimension

M-distant metastasis
The definitions of the M categories for all urological tumours are:

MX Presence of distant metastasis cannot be assessed
M0 No distant metastasis
M1 distant metastasis

The categories M1 and pM1 may be further specefied according to the following notation:

Pulmonary	PUL	Bone marrow	MAR
Osseous	OSS	Pleura	PLE
Hepatic	HEP	Peritoneum	PER
Brain	BRA	Skin	SKI
Lymph nodes	LYM	Other	OTH

pTNM Pathological Classification

The pT, pN and pM categories correspond to the T, N and M categories.

G Histopathological Grading

GX	Grade of differentiation cannot be assessed
G1	Well differentiated, slight anaplasia
G2	Moderately differentiated, moderata anaplasia
G3-4	Poorly differentiated-undifferentiated, marked anaplasia

References

Aaronson, N. K., Ahmedzai, S., Bergman, B., et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, (1993), **85**, 365 – 376.

Baker, S. G., and Laird, N. M. Regression analysis for categorical variables with outcome subject to non-ignorable non-response. *Journal of the American Statistical Association*, (1988), **83**, 62 – 69.

Baker, S. G., Rosenberger, W. F., and DerSimonian, R. Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, (1992), **11**, 643 – 657.

Beacon H. J., Thompson S. G. Multi-level models for repeated measurement data: application to quality of life data in clinical trials. *Statistics in Medicine*, (1996), **15**, 2717 – 2732.

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

Buck, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B*, (1960), **22**, 302 – 306.

Cella, D. F. Tulsky, D. S. Gray, G., et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *Journal of Clinical Oncology*, (1993), **11**, 570 – 579.

CHIRWA T. F. Comparative study on the performance of some tests for discrete longitudinal data under different types of missingness patterns. *Thesis, Master in biostatistics*, Belgium: LUC, (1996).

CHOI, S. AND LU, I. L. Effect of non-random missing data mechanisms in clinical trials. *Statistics in Medicine*, (1995), **14**, 2675 – 2684.

CONAWAY, M. R. The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, (1992), **87**, 817 – 824.

CONAWAY, M. R. Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics*, (1993), **42**, 105 – 115.

COX D. R. *Analysis of binary data*, London: Chapman and Hall, (1970).

CURRAN, D., MOLENBERGHS, G., FAYERS P. M., AND MACHIN, D. Incomplete quality of life data in randomized trials: missing forms. *Statistics in Medicine*, (1998a), **17**, 697 – 709.

CURRAN, D., BACCHI, M., SCHMITZ, F. H., MOLENBERGHS, G., AND SYLVESTER, R. J. Identifying the types of missingness in quality of life data from clinical trials, *Statistics in Medicine*, (1998b), **17**, 739 – 756.

CURRAN D., FAYERS P., MOLENBERGHS G., MACHIN D. 'Analysis of incomplete quality-of-life data in clinical trials'in *Staquet M, Hays R and Fayers P. (eds) Quality of Life assessment in clinical trials: Methods and practice*, London: Oxford University Press, (1998c).

CURRAN D., VAN DONGEN J. P., AARONSON N., KIEBERT G., FENTIMAN I. S., MIGNOLET F., BARTELINK H. Quality of life of early breast cancer patients treated with mastectomy or breast conserving procedures: Results of EORTC trial 10801. *European Journal of Cancer*, (1998d), **34**, 307 – 314.

CURRAN D., AARONSON N., STANDAERT B., MOLENBERGHS G., THERASSE P., RAMIREZ A., KOOPMANSCHAP M., ERDER H., AND PICCART M. Summary measures and summary statistics in the analysis of quality of life data: an example from an EORTC-NCIC-SAKK locally advanced breast cancer study. (2000a), *Submitted for publication.*

CURRAN D., PIGNATTI F., MOLENBERGHS G. Milk protein trial: missing data or stratified analysis. (2000b), *Submitted for publication.*

CURRAN D., MOLENBERGHS G., AARONSON N., FOSSA S. D., SYLVESTER R. J. Analysis of longitudinal quality of life data with dropout. (2000c), *Submitted for publication.*

CULLIS, B. R. Discussion to Diggle, P. J. and Kenward, M. G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, (1994), **43**, 79 – 80.

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

DAVIS C. S. 'Analysis of incomplete categorical repeated measures' in *Proceedings of the 17th Annual SAS Users Group International Conference*, North Carolina, SAS Institute Inc., (1992).

DAWSON J. D. Stratification of summary statistic tests according to missing data patterns. *Statistics in Medicine*, (1994), **13**, 1853 – 1863.

DE HAES J. C. J. M., VAN KNIPPENBERG, F. C. E. AND NEIJT, J. P. Measuring psychological and physical distress in cancer patients: Structure and application of the Rotterdam Symptom Checklist. *British Journal of Cancer*, (1990), **62**, 1034 – 1038.

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

DEMPSTER, A. P., AND RUBIN, D. B. *Overview, in Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*, New York: Academic Press, (1983).

DE REIJKE, T. M., KEUPPENS, F. I., WHELAN, P., KLIMENT, J., ROBINSON, M. R. G., REA, L. A., AND SYLVESTER, R. J. Orchidectomy and orchidectomy plus Mitomycin C in patients with poor prognosis metastatic prostate cancer: The final results of an EORTC-GU group trial. *Journal of Urology*, (1999), **5**, 1658 – 64;

DIGGLE P. J. Testing for random dropouts in repeated measurements data. *Biometrics*, (1989), **45**, 1255 – 1258.

DIGGLE, P. J. *Time Series: A Biostatistical Introduction*, Oxford: Oxford University Press, (1990).

DIGGLE, P. J., AND KENWARD M. G. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, (1994), **43**, 49 – 93.

DIGGLE, P. J., LIANG, K. Y., AND ZEGER, S. L. *Analysis of Longitudinal Data*, Oxford: Clarendon Press, (1994).

EFRON, B. Missing data, imputation, and the bootstrap (with discussion). *Journal of the American Statistical Association*, (1994), **89**, 463 – 479.

EKHOLM, A., AND SKINNER, C. The muscatine children's obesity data reanalysed using pattern mixture models. *Applied Statistics*, (1998), **47**, 251 – 263.

THE EUROQOL GROUP. Euroqol-a facility for the measurement of health related quality of life. *Health Policy*, (1990), **16**, 199 – 228.

FAIRCLOUGH, D. L., AND GELBER R. D. 'Quality of Life: Statistical Issues and Analysis' in *Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, B.*, Spilker Lippincott-Raven Publishers, (1996).

FAIRCLOUGH D. L. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. *Statistics in Medicine*, (1997), **16**, 1197 – 1209.

FAIRCLOUGH D. L., PETERSON H. F., CHANG V. Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine*, (1998a), **17**, 667 – 677.

FAIRCLOUGH, D. L., PETERSON, H. F., CELLA, D., BONOMI, P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Statistics in Medicine*, (1998b), **17**, 781 – 796.

FAYERS, P. M., CURRAN, D. AND MACHIN, D. Incomplete quality of life data in randomized trials: missing items. *Statistics in Medicine*, (1998), **17**, 679 – 696.

FAYERS, P. M., AARONSON, N. K., BJORDAL, K., CURRAN D., AND GROENVOLD, M. *EORTC QLQ-C30 Scoring Manual: 2nd Edition*, Brussels: EORTC, (1999).

FITZMAURICE G. M., LAIRD N. M. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, (1993), **80**, 141 – 151.

FITZMAURICE, G. M., LAIRD, N. M. AND LIPSITZ, S. R. Analysing incomplete longitudinal binary responses: a likelihood-based approach. *Biometrics*, (1994), **50**, 601 – 612.

FOSSA, S. D., CURRAN, D., AARONSON, N. K., KEUPPENS, F., KLIMENT, J., ROBINSON, M. R. G., DE REIJKE, T. M., HETHERINGTON, J., KIL, P. J. M., AND REA, L. A. Quality of life of patients with newly diagnosed poor prognosis M1 prostate cancer

undergoing orchiectomy without or with mitomycin C: results from EORTC phase III trial 30893, *European Urology*, (2000), *in press.*

GLYNN R. J., LAIRD N. M., RUBIN D. B. Selection modelling versus mixture modelling with nonignorable nonresponse, *in* Wainer, H. (ed.), *Drawing inferences from self selected samples*, Springer Verlag, New York, (1986), 115 – 142.

GLONEK, G. F. V. AND MCCULLAGH, P. Multivariate logistic model. *Journal of the Royal Statistical Society, Series B*, (1995), **57**, 533 – 546.

GRALLA, R. J., HOLLEN, P. J., EBERLEY, S. AND COX, C. Quality of life score predicts both response and survival in patients receiving chemotherapy for non-small cell lung cancer. *Supportive Care Cancer*, (1995), **3**, 378 – 379.

GREENLEES, W. S., REECE, J. S., AND ZIESCHANG, K. D. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, (1982), **77**, 251 – 261.

GRIZZLE, J. E., STARMER, C. F. AND KOCH, G. G. Analysis of categorical data by linear models. *Biometrics*, (1969), **25**, 489 – 504.

HEDEKER, D. AND GIBBONS, R. D. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, (1997), **2**, 64 – 78.

HEITJAN D. F. AND BASU S. Distinguishing "Missing at Random" and "Missing Completely at Random". *The American Statistician*, (1996), **50**, 207 – 213.

HOCHBERG Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrica*, (1988), **75**, 800 – 802.

HOGAN, J. W. AND LAIRD, N. M. Mixture models for the joint distribution of repeated measures and event Times. *Statistics in Medicine*, (1997), **16**, 239 – 257.

HOGAN, J. ENAR, *Biometrika*, (1999), **63**, 581 – 592.

HOLLEN P. J., GRALLA R. J., COX C., EBERLY S. W., KRIS M. A dilemma in analysis: issues in serial measurement of quality of life in patients with advanced lung cancer. *Lung Cancer*, (1997), **18**, 119 – 136.

HOPWOOD, P., STEPHENS, R. J., AND MACHIN, D. Approaches to the analysis of quality of life data: experiences gained from a Medical Research Council Lung Cancer Working Party. *Quality of Life Research*, (1994), **3**, 339 – 352.

KENWARD M. G., LESAFFRE E., MOLENBERGHS G. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics*, (1994), **50**, 945 – 953.

KIEBERT, G. M., CURRAN, D., AND AARONSON N. K. Quality of Life as an endpoint in EORTC Clinical Trials. *Statistics in Medicine*, (1998), **17**, 561 – 569.

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

LAIRD, N. M., LANGE, N., AND STRAM, D. Maximum likelihood computations with repeated meausres: application of the EM algorithm. *Journal of the American Statistical Association*, (1987), **82**, 97 – 105.

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

LESAFFRE, E. AND MOLENBERGHS, G. A sensitivity analysis of two multivariate response models. *Computational statistics and Data Analysis*, (1994), **17**, 363 – 391.

LESAFFRE E., MOLENBERGHS G., DEWULF L. Effect of dropouts in longitudinal study: an application of a repeated ordinal model, *Statistics in Medicine*, (1996), **15**, 1123 – 1141.

LI, K. H., RAGHUNATHAN, T. E., AND RUBIN, D. B. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distributions. *Journal of the American Statistical Association*, (1991), **86**, 1065 – 1073.

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

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

LIPSITZ S. R., LAIRD N. M., HARRINGTON D. P. Weighted Least Square analysis of repeated categorical measurements with outcomes subject to nonresponse, *Biometrika*, (1994), **50**, 11 – 24.

LIPSITZ S. R., KIM K. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, (1994), **13**, 1149 – 1163.

LIPSITZ S. R., KIM K., AND ZHAO L. Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine*, (1994), **13**, 1149 – 1163.

LITTLE, R. J. A. AND RUBIN, D. B. *Statistical analysis with missing data*, New York: John Wiley, (1987).

LITTLE, R. J. A. A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association*, (1988), **83**, 1198 – 1202.

LITTLE, R. J. A. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* (1993), **88**, 125 – 134.

LITTLE, R. J. A. A class of pattern-mixture models for normal incomplete data. *Biometrika*, (1994), **81**, 471 – 483.

LITTLE, R. J. A. Modeling the drop-out mechanism in repeated-measures studies, *Journal of the Americal Statistical Society*, (1995), **90**, 1112 – 1121.

LITTLE, R. J. A. AND WANG, Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, (1996), **52**, 98 – 111.

MADOW, W. G., NISSELSON, H. AND OLKIN, I. Incomplete data in sample surveys, *Report and case studies*, New York: Academic press, (1983).

MATTHEWS J. N. A refinement to the analysis of serial data using summary measures, *Statistics in Medicine*, (1993), **15**, 27 – 37.

MCARDLE, J. J. AND HAMAGAMI, F. Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research*, (1992), **18**, 145 – 166.

MCCULLAGH, P., NELDER, J. A. *Generalized linear models*, London, Chapman and Hall, (1989).

MCIVER J. P., CARMINES E. G. *Unidimensional Scaling*, London, Sage Publications Ltd., (1981).

MENG X. L. Multiple-imputation inference with uncongenial sources of input, *Statistical Science*, (1994), **9**, 538 – 573.

MICHIELS, B., MOLENBERGHS, G., AND LIPSITZ, S. R. Selection models and pattern-mixture models for incomplete categorical data with covariates. *Biometrics*, (1999), **55**, 000 – 000.

MICHIELS, B., MOLENBERGHS, G., BIJNENS, L., AND VANGENEUGDEN, T. Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Submitted for publication.* (1998)

MOLENBERGHS G., LESAFFRE E. Marginal modelling of correlated ordinal data using an n-way Plackett distribution, *Journal of the American Statistical Association*, (1994), **89**, 633 – 644.

MOLENBERGHS G., KENWARD M. G., LESAFFRE E. The analysis of longitudinal ordinal data with non-random dropout, *Biometrika*, (1997), **84**, 33 – 44.

MOLENBERGHS, G. AND KENWARD, M. G. 'Calculating the appropriate information matrix for log-linear models when data are missing at random', in *Lecture Notes in Statistics, Proceedings of the Nantucket conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, Gregoire, T. (ed), New York, Springer-Verlag, (1997).

MOLENBERGHS G., GOETGHEBEUR E. J. T., LIPSITZ S. R. Non-random missingness in categorical data: strengths and limitations, (1997), *Submitted for publication.*

MOLENBERGHS, G., MICHIELS, B., AND KENWARD, M. G. Pseudo-likelihood for combined selection and pattern-mixture models for missing data problems. *Biometrical Journal*, (1998), **40**, 557 – 572.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G., AND DIGGLE, P. J. Missing data mechanisms and pattern-mixture models. (1998) *Submitted for publication.*

MORRIS J., COYLE D. Quality of life questionnaires in cancer clinical trials: imputing missing values, *Psycho-oncology*, (1994), **3**, 215 – 222.

NEUHAUS, J. Statistical methods for longitudinal and clustered designs with binary responses, *Statistical Methods in Medical Research*, (1992), **1**, 249 – 273.

NUNNALLY J. C., BERNSTEIN I. H. *Psychometric Theory, 3rd ed.*, New York, McGraw-Hill, (1994).

OLSCHEWSKI, M., SCHULGEN, G., SCHUMACHER, M., AND ALTMAN, D. G. Quality of Life Assessment in Clinical Cancer Research, *British Journal of Cancer*, (1994), **70**, 1 – 5.

PARK, T. AND DAVIS, C. S. A test of the missing data mechanism for repeated categorical data, *Biometrics*, (1993), **49**, 631 – 638.

PARK, T. AND BROWN, M. B. Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, (1994), **89**, 44 – 52.

POCOCK S. J., GELLER N. I., TSIATIS A. A. The analysis of multiple endpoints on clinical trials, *Biometrics*, (1987), **43**, 487 – 498.

PLACKETT, R. L. A class of bivariate distributions, *Journal of the American Statistical Society*, (1965), **60**, 516 – 522.

PREGIBON, D. Typical survey data: estimation and imputation, *Survey Methodology*, (1977), **2**, 70 – 102.

PULKSTENIS, E. P., TEN HAVE, T. R. AND LANDIS, J. R. Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication, *Journal of the American Statistical Society* (1998), **93** 442.

RIDOUT M. Testing for random dropouts in repeated measurement data, *Biometrics*, (1991), **47**, 1617 –1621 .

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

ROSENDAHL K. I., KIEBERT G., CURRAN D., COLE B., WEEKS J. C., DENIS L. J., HALL R. R. A quality-adjusted survival (Q-Twist) analysis of EORTC trial 30853 comparing maximal androgen blockade (MAB) with orchiectomy in patients with metastatic prostate cancer. *The prostate*, (1999), **38**, 100 – 109.

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

RUBIN, D. B. Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*, U.S. Department of Commerce, pp. 1 – 23, (1978).

RUBIN, D. B. AND SCHENKER, N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, (1986), **81**, 366 – 374.

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

RUBIN, D. B. Multiple imputation after 18+ years, *Journal of the American Statistical Association*, (1996), **91**, 473 – 489.

SABBIONI, M., HURNY, C., BERNHARD, J., ET AL. Interaction between psychosocial factors and immunity in breast cancer patients, *Annals of Oncology*, (1996), **7**, 13 – 13.

SAS INSTITUTE. *SAS/STAT User's Guide*, North Carolina, SAS Institute Inc., (1989).

SCHAFER J. L. *Analysis of incomplete multivariate data.* London: Chapman and Hall, (1997).

SCHLUCHTER M. D. Methods for the analysis of informatively censored longitudinal data, *Statistics in Medicine*, (1992), **11**, 1861 – 1870.

SEYMOUR M. T., SLEVIN M. L., KERR D. J. ET AL. Randomized trial assessing the addition of interferon alpha 2a to fluorouracil and leucovorin in advanced colorectal cancer. Colorectal Cancer Working Party of the United Kingdom Medical Research Council, *Journal of Clinical Oncology*, (1996), **14**, 2280 – 2288.

SHEINER, L. B., BEALE, S. L., AND DUNNE, A. Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. *Journal of the American Statistical Association*, (1997), **92**, 1235 – 1244.

SMITH, D. M., ROBERTSON, B. AND DIGGLE, P. J. 'Object-oriented Software for the Analysis of Longitudinal Data' in *SAS Technical Report MA 96/192. Department of Mathematics and Statistics*, University of Lancaster, (1996).

STANISH, W. M., GILLINGS, D. B. AND KOCK, G. G. An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data, *Biometrics*, (1978), **34**, 305 – 317.

STASNY, E. A. Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, (1986), **81**, 42 – 47.

TANNER, M. A. AND WONG, W. H. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, (1987), **82**, 528 – 550.

TANNOCK, I. F., OSOBA, D., STOCKLER., ET AL. Chemotherapy With Mitoxantrone Plus Prednisone or Prednisone Alone for Symptomatic Hormone-Resistant Prostate Cancer: A Canadian Randomized Trial With Palliative End Points, *Journal of Clinical Oncology*, (1996), **14**, 1756 – 1764.

THERASSE P., MAURIAC L., WELNICKA M. ET AL. Neo-adjuvant dose intensive chemotherapy in locally advanced breast cancer (LABC): An EORTC-NCIC-SAKK randomized phase III study comparing FEC (5FU, Epirubicin, Cyclophosphamide) vs high dose intensity EC + G-CSF (Filgrastim), *Journal of Clinical Oncology*, (1998), **17**, 124 – 124.

THIJS H., MOLENBERGHS G., AND VERBEKE G. The Milk Protein Trial: Influence Analysis of the Dropout Process. *Submitted for publication.* (1999).

THIJS H., MOLENBERGHS G., MICHIELS B., VERBEKE G. AND CURRAN D. Strategies to fit Pattern-Mixture Models. *Submitted for publication.* (2000).

TROXEL A. B., LIPSITZ S. R., TROYEN A. B. Weighted estimating equations with nonignorably missing response data, *abstract*, (1996).

TROXEL A. B., FAIRCLOYGH D. L., CURRAN D., HAHN E. A. Statistical analysis of quality of life with missing data in cancer clinical trials, *Statistics in Medicine*, (1998), **17**, 653 – 666.

TROXEL A. B. A comparative analysis of quality of life data from a Southwest Oncology Group randomized trial of advanced colorectal cancer, *Statistics in Medicine*, (1998), **17**, 767 – 779.

VAN STEEN K., CURRAN D., MOLENBERGHS G. Sensitivity Analysis of Longitudinal Binary Quality of Life Data with Dropout: An example using the EORTC QLQ-C30. *Submitted for publication.* (1999)

VERBEKE G., MOLENBERGHS G. *Linear mixed models in practice*, Springer-Verlag New York, (1997).

VERBEKE, G., LESAFFRE, E., AND SPIESSENS, B. The practical use of different strategies to handle dropout in longitudinal studies. *Submitted for publication.* (1998)

VERBYLA, A. P. AND CULLIS, B. R. Modelling in repeated measures experiments. *Applied Statistics*, (1990), **39**, 341 – 356.

WANG-CLOW, F., LANGE, N., LAIRD, N. M., AND WARE, J. H. A simulation study of estimators for rate of change in longitudinal studies with attrition. *Statistics in Medicine*, (1995), **14**, 283 – 297.

WARE J. E., SHERBOURNE C. D. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection, *Medical Care*, (1992), **30**, 473 – 83.

WARE J. E., JR., SNOW K. K., KOSINSKI M. AND GANDEK, B. *SF-36 Health Survey Manual and Interpretation Guide*, Boston, New England Medical Centre, (1993).

WEI L. J., LACHIN J. M. Two sample asymptotically distribution-free tests for incomplete multivariate observations, *Journal of the American Statistical Association*, (1984), **79**, 653 –669 .

WEI L. J., JOHNSON W. E. Combining dependent tests with incomplete repeated measurements, *Biometrika*,(1985), **72,2**, 359 – 364.

WOOLSON, R. F. AND CLARKE W. R. Analysis of categorical incomplete longitudinal data, *Journal of the Royal Statistical Society, Series A*, (1984), **147**, 87 – 99.

WU M. C., CARROLL R. J. Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring mechanism, *Biometrics*, (1988), **44**, 175 – 188.

WU M. C., BAILEY K. R. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, (1988), **7**, 337 – 346.

WU M. C., BAILEY K. R. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics*, (1989), **45**, 939 – 955.

WU M C., HUNSBERGER S., ZUCKER D. Testing for differences in changes in the presence of censoring: parametric and non-parametric methods, *Statistics in medicine*, (1994), **13**, 635 –646.

ZEE B., PATER J. Statistical analysing of trials assessing quality of life, in *Effects of Cancer on Quality of Life*, (1991), CRC Press Florida

ZEE B. C. Growth curve model analysis for quality of life data, *Statistics in Medicine*, (1998), **17**, 757 – 766.

ZEGER, S. L. AND LIANG, K. Y. Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, (1986), **42**, 121 – 130.

ZEGER S. L., LIANG K. Y., ALBERT P. S. Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, (1988), **44**, 1049 – 1060.

ZEGER, S. L. AND LIANG, K. Y. An overview of methods for the analysis of longitudinal data, *Statistics in Medicine*, (1988), **11**, 1825 – 1839.

ZEGER, S. L. AND KARIM, M. R. Generalized linear models with random effects; a Gibbs sampling approach, *Journal of the American Statistical Association*, (1991), **86**, 79 – 86.

ZHAO L. P., PRENTICE R. L. Correlated binary regression using a quadratic exponential model, *Biometrika*, (1990), **77**, 642 – 648.

ZWINDERMAN A. H. Statistical analysis of longitudinal quality of life data with missing measurements, *Quality of Life Research*, (1992), **1**, 219 – 224.