The Statistical Analysis of Self-Assessed Quality of Life Data in Cancer Clinical Trials

THESIS

presented for the

DEGREE

of

DOCTOR OF PHILOSOPHY

in the Faculty of Science

(Field of Study: Statistics)

by

Heather J Beacon

Medical Statistics Unit

Department of Epidemiology and Population Sciences

London School of Hygiene and Tropical Medicine

University of London

October 1996

Abstract

The assessment of quality of life as a primary outcome in cancer clinical trials is now almost universal. Such data are necessarily longitudinal and multidimensional, and are often severely unbalanced by missing values or early patient death. However, to date, their reporting in the applied literature has generally used simple descriptive summaries that ignore many of these complexities. Not only can these be misleading, but they generally do not allow firm conclusions to be drawn about a major endpoint. The aim of this thesis is to assess the practical application of recent developments in statistical methodology for the analysis of quality of life data collected using self assessment questionnaires within cancer clinical trials. Its emphasis is on the use of relatively simple and flexible tools that will allow more reliable and powerful inferences to be drawn from the data than is done at present.

The principal statistical tools considered are random coefficient and marginal models. It is shown that these can be successfully used for the analysis of continuous, binary and ordinal responses. In particular, they offer a simple approach to the analysis of repeated multivariate outcomes and can be very easily extended to model the complex patterns of response that are often seen in following cancer treatment.

In relation to the problem of censored quality of life as a result of patient death, analyses that attempt to combine the survival and quality of life endpoints in a single variable are contrasted with those that consider the two endpoints as a multivariate problem. It is shown how this latter model can provide a summary of the quality of life response conditional on

patient survival that with further work should have great application to such quality of life data.

Finally, the problem of intermittent missing data is reviewed. The implications of missing data for some of the analyses presented in the thesis are assessed, and two models that attempt to determine the nature of intermittent missing data are developed. It is concluded that the problem of non-ignorable intermittent missing data presents a very challenging area of further research.

Acknowledgements

First of all I would like to thank my supervisor Professor Simon Thompson for all his support, patience and invaluable guidance throughout the past three years. I wish him the very best in his new appointment. Also thanks to Chris Frost for his helpful discussion and advice.

I am very grateful to a number of scientific collaborators who made their data available for this work. For the CRC non small cell lung cancer study I would like to thank Dr Joan Houghton and Kathryn Monson of the Cancer Research Campaign clinical trials centre and their clinical collaborators (Dr A Timothy and Dr F Calman). For the CRC Hepatic artery pump trial, I am extremely grateful to UK hepatic artery pump trial group, (T G Allen-Mersh, Sally Earlham and Carol Fordy). Finally, for the MRC non small cell lung cancer study, my thanks to the MRC lung cancer working party. From these groups, I am particularly grateful to Sally Earlham for her clinical input, and to Richard Stephens, Dr Max Parmar and Peter Fayers for their statistical discussion.

I would also like to thank all the staff and PhD students of the Medical Statistics Unit. In particular, Mosty, Giota, Peter and Ian for their scientific advice, not forgetting Tim for his moral support and Mark for his efficient computing support. To Mike, who kept me fit over the summer, and was foolish enough to volunteer to proof read my thesis, merci beaucoup. On on a personal level, thanks to Craig and Siân for their understanding and help with my many personal traumas over the past three years (and more).

Acknowledgements

This research has been funded by the Education and Psychological research studentship grant number CE1142 from the Cancer Research Campaign. I am grateful to the Cancer Research Campaign for this funding, thereby giving me the opportunity of performing this research. Also to the British Rail pensions administration for their continued financial assistance throughout my long student career.

Finally, a big thank you to Anthony for putting up with my highs and lows over the past two years and for his incredible patience and support over the last six months.

Table of Contents

List of Tables

List of tables

List of Figures

List of figures

Abbreviations

The following abbreviations are used throughout the text of the thesis.

Abbreviation	Definition
ALR	alternating logistic regression
AR*x*	auto-regressive of order *x*
CI	confidence interval
cov	covariance
CRC	Cancer Research Campaign
CRC NSCLC study	CRC non small cell lung cancer study
CRC HAP trial	CRC Hepatic artery pump trial
E-M algorithm	estimation-maximisation algorithm
FM	multiple fraction radiotherapy (MRC LU07 study)
FUDR	fluro-deoxyuridine
F2	two fraction radiotherapy (MRC LU07 study)
GEE	generalised estimating equation
Gy	Gray (unit of radiation)
HAD scale	Hospital Anxiety and Depression scale
HAI	hepatic artery implant (CRC HAP trial)
K-M	Kaplan-Meier
log lh	log likelihood
MAR	missing at random
MCAR	missing completely at random
ML	maximum likelihood
MRC	Medical Research Council
MRC LU07 study	MRC non small cell lung cancer study
MQL	marginal quasi-likelihood
NMAR	not missing at random

Abbreviations

PQAS	partitioned quality adjusted survival
PQL	penalised quasi-likelihood
PSE	present state examination
QALY	quality adjusted life year
QAS	quality adjusted survival
Q-TWiST	quality adjusted TWiST
OR	odds ratio
REML	restricted maximum likelihood
(R)IGLS	(restricted) iteratively generalised least squares
RSCL	Rotterdam Symptom Checklist
SD	standard deviation
SE	standard error
SIP	sickness impact profile
TNQOL	time with normal quality of life
TWiST	time without symptoms and toxicity
WLS	weighted least squares
UWLS	unweighted least squares
var	variance

Notation

The following general notation is used consistently throughout the thesis. Any additional notation required, this is explained in more detail in the relevant section of work.

Definition	Symbol	Dimension
<u>Indices</u>		
Subject	i	$1,\ldots,n$
Measurement occasion (fixed for all subjects)	j	$1,\ldots,m$
Measurement occasion (different across subjects)	t	$1,\ldots,m_i$
Quality of life dimension	l	$1,\ldots,L$
Ordinal response categories	k	$1,\ldots,K$
<u>Variables</u>		
Response variable subject i at fixed occasion j	y_{ij}	real
Response variable subject i at time t	y_{it}	real
Response vector for subject i over fixed occasions $j=1,\ldots,m$	$\mathbf{y}_i$	$1\times m$
Response vector for subject i over variable occasions $t=1,\ldots,m_i$	$\mathbf{y}_i$	$1\times m_i$
Response vector for all subjects over fixed occasions	$\mathbf{Y}$	$1\times nm$
Response vector for all subjects over variable occasions	$\mathbf{Y}$	$1\times M$ $(M=\sum_{i=1}^{n} m_i)$
Matrix of p explanatory variables for subject i	$\mathbf{x}_i$	$m\times p$

To my mum and dad

1 Introduction

The use of quality of life assessment in clinical trials has become increasingly common and 'quality of life' is now included as a primary endpoint in the protocol of many clinical trials. Based on a search of the literature and previous work by Fayers and Jones (1983), Schumacher *et al.* (1991) reported on the steadily increasing number of articles which use the term 'quality of life' in the title, or as a keyword, from 1982 through to 1989. Since 1989, this trend has continued and is highlighted by the number of special conferences and symposia that have recently taken place, as well as the introduction of the quarterly journal '*Quality of Life Research*' in 1992.

This increase in use illustrates a realisation that the quality, as well as the quantity, of survival is important in the evaluation of treatment efficacy, and is particularly relevant in cancer clinical trials because of the often very aggressive and invasive treatment regimens that patients face. Moreover, there remain many cancers - for example, non small cell bronchial and gastrointestinal - for which no decisive chemotherapeutic treatment has been found. Therefore, palliation is the primary concern of treatment in these areas, and alongside symptomatic relief, patient quality of life is the main outcome of interest.

The definition of the term 'quality of life' was up until recently vague and was applied in many different contexts including the level of patient side-effects or toxicity, the degree of their symptoms, as well as their overall well-being. A consensus has now been reached that the outcome of interest is *health related* quality of life, which should be considered as

> *"1. multidimensional, comprising important elements of a patient's emotional, social and physical well-being;*
>
> *2. subjective, relying primarily on the patient's own judgements; and*
>
> *3. non-static and subject to changes over a patient's lifetime."*

(Olschewski *et al.*, 1992). It is also suggested that, dependant on the nature of disease and treatment regimens under research, it may be appropriate to include symptoms and side effects of treatments - such as vomiting and pain - as well as patient satisfaction with the treatment received within a quality of life assessment (Nayfield *et al.*, 1992, Girling *et al.*, 1994).

If quality of life is best assessed by the patient, ideally this will be done by interview or present state examination (PSE) (Fallowfield, 1990). However, time constraints on both clinician and patients, as well as cost, mean that in clinical trials this is rarely feasible. The solution has been *self assessment questionnaires* which can be easily completed by the patient with or without supervision.

There exist a large number of such quality of life measuring instruments. Some of these are classed as *generic* and question general aspects of quality of life, whereas others are more *disease specific* and relate to particular symptoms or problems posed by the disease in question. For example, the Hospital Anxiety and Depression (HAD) scale (Zigmond and Snaith, 1983) is a generic instrument, whereas the Rotterdam Symptom Checklist (RSCL) (de Haes *et al.*, 1990) is disease specific and deals primarily with items relating to issues concerned with the treatment of cancer.

The questionnaires are comprised of a number of questions or *items* which address different aspects or dimensions of patient quality of life. Responses to these items are either binary (yes/no) or on a given ordinal scale. As an example, a typical item relating to patient anxiety

Box 1.1

Typical item from the HAD scale

"I get sudden feelings of panic

(3) Very often (2) Quite often (1) Not very often (0) Not at all"

from the HAD scale is given in box 1.1.

Using factor analysis or subjective reasoning by the questionnaire designers, many of these instruments then allow weighted or simple summations of these individual items to give summary scores for overall, or dimension specific, quality of life. For example, the HAD scale consists of fourteen items all measured on a four point ordinal scale. By summing the ratings of seven of these items, a summary depression score is obtained. The remaining seven items can be similarly aggregated to give an anxiety score. In contrast, the Sickness Impact Profile (SIP) (Bergner *et al.*, 1981) involves weighted summaries of binary responses for 136 items relating to 12 dimensions of patient quality of life. For some instruments, recommended boundaries for a 'normal score' in each dimension are also available. For example, with the anxiety and depression scores on the HAD scale, a score ≤ 7 is deemed to be normal, a score between 8 and 10 suggests a possible case of clinical anxiety or depression, whereas a score ≥ 11 is considered to identify definite cases.

All of the instruments used in medical research have undergone assessment of validity, reliability and responsiveness. For a particular instrument, this means that: (i) it measures what it was designed to measure (validity); (ii) on repeated use on the same subject under identical conditions it produces the same result (reliability); and (iii) it will be able to exhibit changes in underlying response that do occur (responsiveness). Rather than developing new

instruments, it has been advocated that in study design, use should be made of existing instruments where possible. This not only the saves time and resources involved in reinventing (and validating) tools which already exist, but it should make it easier to compare quality of life across studies (Aaronsen, 1989). As the focus of this work is on the analysis of quality of life data, the evaluation and validation of the quality of life measurement instruments which have been used for data collection are not considered further. Details of validation techniques have been given by Bowling (1983), Bergner *et al.* (1981), Chinn and Burney (1987) and Guyatt *et al.* (1991). Reviews of the practical use of available instruments have been done by Fallowfield (1990), Bowling (1983) and, on behalf of the MRC cancer therapy committee working party on quality of life, by Maguire and Selby (1989).

A comprehensive review of the analysis of quality of life data and associated problems was given by Cox *et al.* (1992). This review and subsequent discussion, examined the important issues which need to be addressed for analysis of quality of life data. With the focus on the practical application of recent developments in statistical methodology, the work of this thesis will address four particular areas of concern raised by these authors. Namely, the analysis of repeated measurement data, multiple dimensionality, patient dropout due to death, and missing data. Each successive chapter will tackle a distinct problem and hence the relevant literature is reviewed therein. The objective of the thesis is to provide a detailed account of the practical use and extension of new statistical methods for the analysis of self assessed quality of life data, for cancer clinical trials in particular. For this aim, the work is presented in terms of examples using quality of life data collected in three recent cancer clinical trials. Although concise details of each trial are given in Appendix 1, a brief outline of each is included here. Descriptions of the quality of life measuring instruments which were used in each study are given in Appendix 2.

The first study is the Cancer Research Campaign, Clinical Trials Centre, non-small cell lung cancer trial (CRC NSCLC). This is a randomised trial designed to compare the results of palliative radiotherapy in patients with previously untreated non-small cell lung cancer. In the first randomised group (denoted split course) patients were given an initial intensive dose of radiotherapy. They were then re-assessed at 4 weeks. If they were considered well enough, they underwent a second randomisation which determined whether a second intensive dose was administered. Patients in the second randomised group (denoted continuous course) received the standard 4 week continuous radiotherapy course. Quality of life was measured using the HAD scale and the RSCL. This was done before the start of treatment and then weekly for an eight week period. 82 patients were entered into the study, 42 to receive the continuous 4 week course and 40 the split course. Within the examples used in this thesis, the added complication introduced by the second randomisation in the split course group will be ignored, and the data analysed according to the two main randomisation groups. No difference in patient survival between these two groups was seen.

The second study (CRC HAP) also comes from the Cancer Research Campaign, Clinical Trials Centre, in collaboration with the Charing Cross and Westminster Medical School. The objectives of the study were to assess the survival, quality of life and tumour response in patients with colorectal hepatic metastases who were treated by intra-hepatic arterial fluro-deoxyuridine (FUDR) infusion (denoted HAI) compared with that in patients receiving the conventional symptomatic treatment (denoted control). Quality of life in the study was measured prior to randomisation, and monthly at the same time as clinical follow-up. Three measurement instruments were used in the study: the SIP, the RSCL, and the HAD scale. Only the data from the RSCL and the HAD scale are used in this work. 100 patients were entered and randomised into the study, 51 to the HAI group and 49 to the control group. A survival advantage for the HAI group with no apparent difference in quality of life was reported by the

study investigators (Allen-Mersh *et al.*, 1994).

The final data set comes from the MRC lung cancer working party (MRC LU07). Once again it is concerned with the palliative care of patients with non-small cell lung cancer. The aim of this study was to assess whether radiotherapy given in two fractions one week apart (denoted F2) gave equally good palliation as a conventional multiple fraction course (denoted FM). Self assessed quality of life in the study was measured daily using a diary card (Fayers and Jones, 1983) from the start of treatment and for six months thereafter. 369 patients were entered and randomised into the study, 184 to receive the shorter course (F2) and 185 to receive the conventional treatment (FM). The results of the study (Bleehan *et al.*, 1991) showed no evidence of a survival difference between the two treatment arms. Descriptive analyses of quality of life data highlighted transient dysphagia following treatment in both groups, but no evidence of a palliative gain of the conventional longer dose to the shorter dose was reported.

Further details of the quality of life in each of these studies is given in Chapter 2 using descriptive analyses that are typical of those which have generally been used in the reporting of quality of life in the literature. The use of such descriptive techniques is reviewed in the chapter, along with a discussion of their relative merits in addressing the four issues of concern outlined by Cox *et al.* (1992).

Chapters 3, 4 and 5 concentrate on the analysis of continuous, binary and ordinal repeated measurement data respectively. In Chapter 3, the use of random coefficient (hierarchical) models (Goldstein, 1986, Goldstein, 1995, Longford, 1995) for the analysis of unbalanced univariate repeated continuous outcomes is demonstrated. This is then extended for the analysis of multidimensional outcomes. These analyses use the overall summary scores from the HAD scale and RSCL in the CRC NSCLC study. In Chapter 4 random coefficient and

marginal models, in the form of generalised estimating equations (Liang and Zeger, 1986), are used for the analysis of binary repeated measurement data for the univariate case. Once again, this work is then extended to the multivariate case. Also covered in this chapter, is the analysis of complex patterns of response which are often seen in quality of life data in cancer trials as a result of the invasive nature of the treatment (Girling *et al.*, 1994). The data used in the chapter come from a dichotomisation of the ordinal responses obtained from individual items of the RSCL in the CRC NSCLC study, and the daily diary card in the MRC LU07 study. In Chapter 5, this work is extended further for the analysis of repeated ordinal data (Ware *et al.*, 1988, Zeger, 1988). Once again this work uses the responses obtained from individual items of the RSCL in the CRC NSCLC study.

In Chapter 6, the problem of patient dropout due to death is addressed. As a result of the severity of patient disease in cancer clinical trials, patient death during the study is often inevitable. This makes interpretation of the available quality of life data difficult, as well as being a possible source of bias. In this chapter, statistical methods for the analysis of longitudinal data subject to dropout (Diggle and Kenward, 1994, Little, 1995, Wu and Carroll, 1988, Schlucter, 1992) are reviewed, and their relevance in solving the problems faced in the analysis of quality of life data are assessed. As an alternative approach to the same problem, analyses which combine quality of life and survival to a single endpoint are also reviewed. The appropriateness of such *quality adjusted survival* analyses have been a source of conflict in the literature but have been rarely used in practice for the analysis of self assessed quality of life data (Cox *et al.*, 1992, Schumacher *et al.*, 1991). The data used for this work are that of the CRC HAP trial. As full survival and quality of life data are now available for this study, the problems faced as a result of patient death during follow-up are easier to combat. In order to recreate the more realistic scenario often faced in the light of patient death - that is, individuals for whom survival is censored - this full data set has been restricted to contain only information

on patient survival and quality of life available on June 1st 1993. Patients still alive at this time were considered censored and any quality of life data they provided beyond this date was ignored. When used, this data set is referred to as the 'restricted' data, whereas the complete data set is referred to as the 'full' data.

The repeated assessment of quality of life data over a long period of time, in a group of patients who often become too ill to complete questionnaires, means that quality of life data is often subject to large amounts of missing responses. For example, Hurny *et al.* (1992) reported compliance rates varying between 37% and 58% at each measurement occasion, and reports on behalf of the MRC lung cancer working party of studies gave daily compliance rates of about 70% (MRC lung cancer working party, 1989, 1991a, 1991b, 1992). In the work presented in Chapters 3 to 6, it is assumed that the problem of missing data can be ignored, and the occurrence of missing data considered only as an issue which generates unbalanced data. It has been well documented in the literature however, that the bias implications of missing data may not be ignorable and will depend on the underlying reasons for data being missing (Rubin, 1976, Laird, 1988). In Chapter 7, the issues raised by the assumption of ignorable missing data made in Chapters 3 to 6 are formally addressed using data from the CRC NSCLC study.

Each chapter is concluded with a summary of its main results and a discussion of the issues raised. Finally, in Chapter 8, the implications of the work as a whole are considered in the context of issues that may be useful for further research.

It should be noted that the emphasis throughout the thesis is on the practical application and interpretation of statistical methodology for the analysis of quality of life data, rather than the future treatment implications of the quality of life results presented in each example. These examples should therefore be seen as a demonstration tool, as opposed to a means of drawing

conclusions for the purpose of directing the treatment of patients. This is particularly important in the case of the CRC NSCLC study for which the precise treatment schedules have not been considered, and also for the CRC HAP trial which is presented in an incomplete state. It is also recognised that the small sample size and extent of missing data in the CRC NSCLC study, mean that these data do not justify the depth of analysis which are presented here.

All data analyses presented use MLn (Rasbash and Woodhouse, 1995) and S-Plus statistical software (Becker *et al.*, 1988). OSWALD (Smith and Diggle, 1994), an additional library of S-Plus functions for the analysis of longitudinal data is also used. Any further statistical programming that is required for data analysis is presented in Appendix 3.

2 The Exploratory Data Analysis of Quality of Life Data

2.1 Introduction

Although most of the work in this thesis concentrates on more formal model based analyses of quality of life data, the multidimensional, as well as longitudinal, nature of quality of life data means exploratory data analysis is an important stage of the analysis process. In fact, a large proportion of reported studies in quality of life research have relied solely on such analyses. Anderson *et al.* (1993) plotted median quality of life scores over time. Reports on behalf of the MRC lung cancer working party (1992, 1993b) gave summaries of the number of patient days spent with improved quality of life compared with baseline and plotted the proportion of patients over time recording symptoms of a particular grade or above. Fallowfield (1986) and MRC lung cancer working party (1991b) reported quality of life data results similarly. The primary aim of the chapter is to review these and other approaches which have advocated for the exploratory data analysis of continuous, binary and ordinal quality of life data. This is done with practical examples using data from the three studies outlined in Chapter 1. The chapter's secondary aim is to give a clear description of many aspects of these data which are presented in the more formal statistical analyses developed in subsequent chapters.

The chapter is structured to discuss exploratory data analysis for repeated continuous outcomes in Section 2.2 and repeated categorical outcomes in Section 2.3. The presence of missing data and patient death during follow-up are major issues for the analysis of quality of life data. In Section 2.4, a number of approaches for the informal examination of each are

reviewed as means to determine their implications for inference. The advantages and disadvantages of the different types of display are discussed within the relevant sections with a general discussion highlighting areas of particular concern given in the final section of the chapter.

2.2 Repeated continuous data

This section of work examines different approaches for exploratory data analysis for quality of life responses that can be considered as continuous - that is the validated summary scores which are generally obtained by summing individual item responses on standard quality of life measuring instruments. Although these scores may only take some integer value within a limited range - for example, the summary HAD scores take integer values within the range 0 to 21 - assuming the scores to be continuous is the approach generally taken in the literature and is probably the most accessible solution for practical analysis. The work is divided into four areas of exploratory data analysis: individual patient profiles, summary statistics, population average profiles over time and associations across dimensions.

2.2.1 Individual patient profiles

Although rarely reported, an important first step in exploratory data analysis is the examination of individual patient profiles (Diggle *et al.*, 1994). These not only allow an informal examination of the consistency of the response across patients, but can help highlight errata and outlying individuals in the data, as well as patterns of missing responses during the follow-up period. Further, if it is feasible to display more than one dimension, an informal examination of the relationship between dimensions is then also possible.

The obvious problem with individual profiles is that the large number of patients in a study

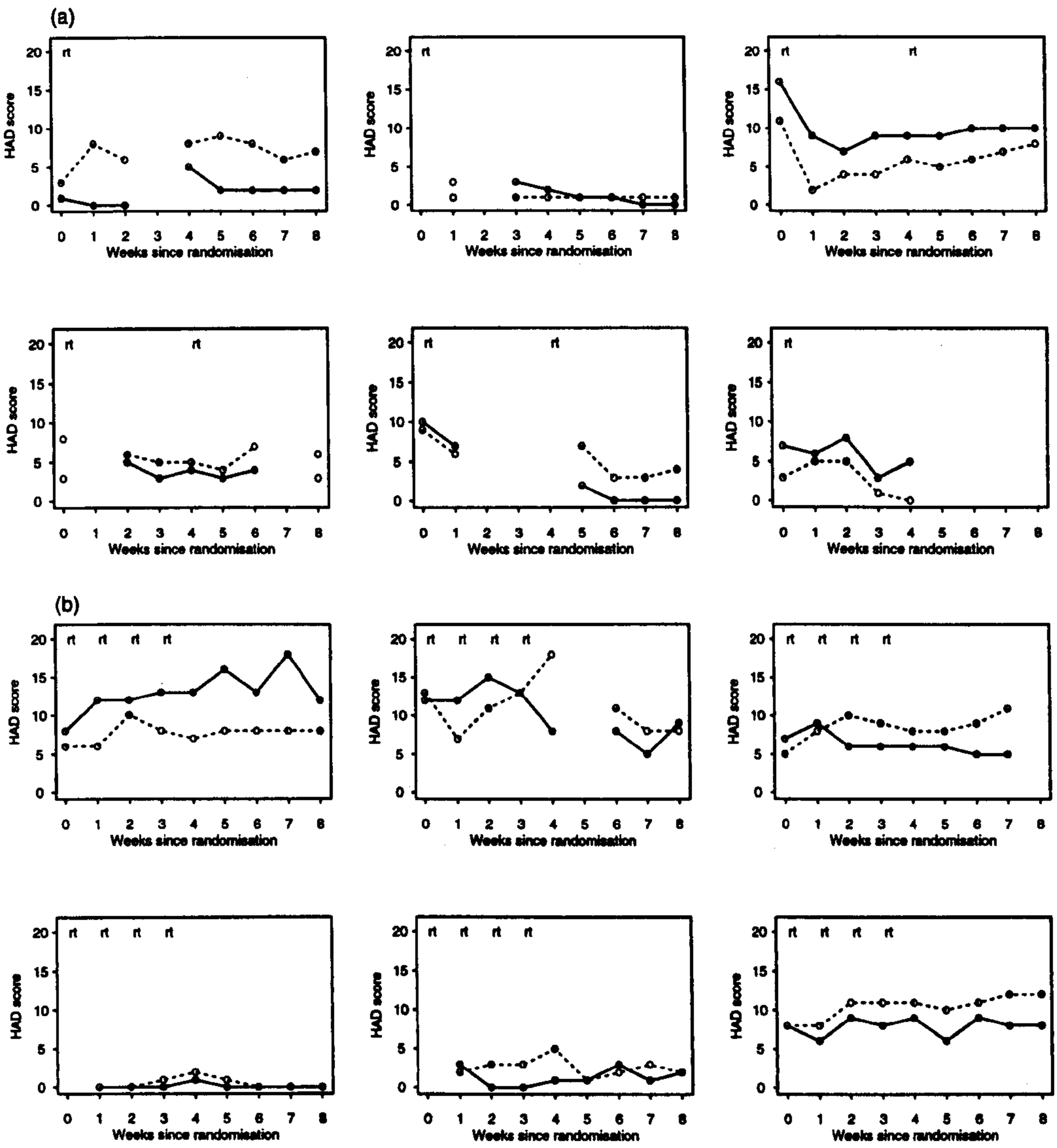

Figure 2.1: Random sample of individual patient profiles of HAD anxiety and depression scores for the CRC NSCLC trial. The scores are plotted over time from baseline (0) through the eight week follow-up: (a) split course radiotherapy; (b) continuous course radiotherapy. Discontinuities in the connecting line indicate missing responses. The timing of radiotherapy for each patient is shown at the top of the figure. Anxiety responses: ————; depression responses: --------.

will often make it impractical to display concisely the behaviour of all patients in a single study and random noise within individuals makes overall patterns difficult to determine. A simple solution is to focus on a simple random sample of the individuals with these again plotted separately or overlayed over a scatter plot of all the data. Alternatively, with moderately sized

studies, if the scores can be sensibly categorised - for example, with the use of 'normal' scores - a lexis diagram can be a useful way of displaying large amounts of data in a very concise way.

Such ideas are demonstrated in figures 2.1, 2.2 and 2.3. Figure 2.1 shows the anxiety and depression scores over time for a random sample of subjects in the CRC NSCLC study. For clarity, subjects who provided at least four quality of life responses over the eight week follow-up were selected. The timing of radiotherapy treatment is marked on the figure to make it possible to identify any obvious trends in patient responses as an immediate result of treatment as advocated by the MRC lung cancer working party (1991a).

These profiles give a good overview of the data and the typical behaviour of patient scores during the follow-up period. They clearly highlight the differences in patient experience both in terms of the behaviour over time and in the underlying level. In the sample shown, they also show some possible relationship between the anxiety and depression responses over time.

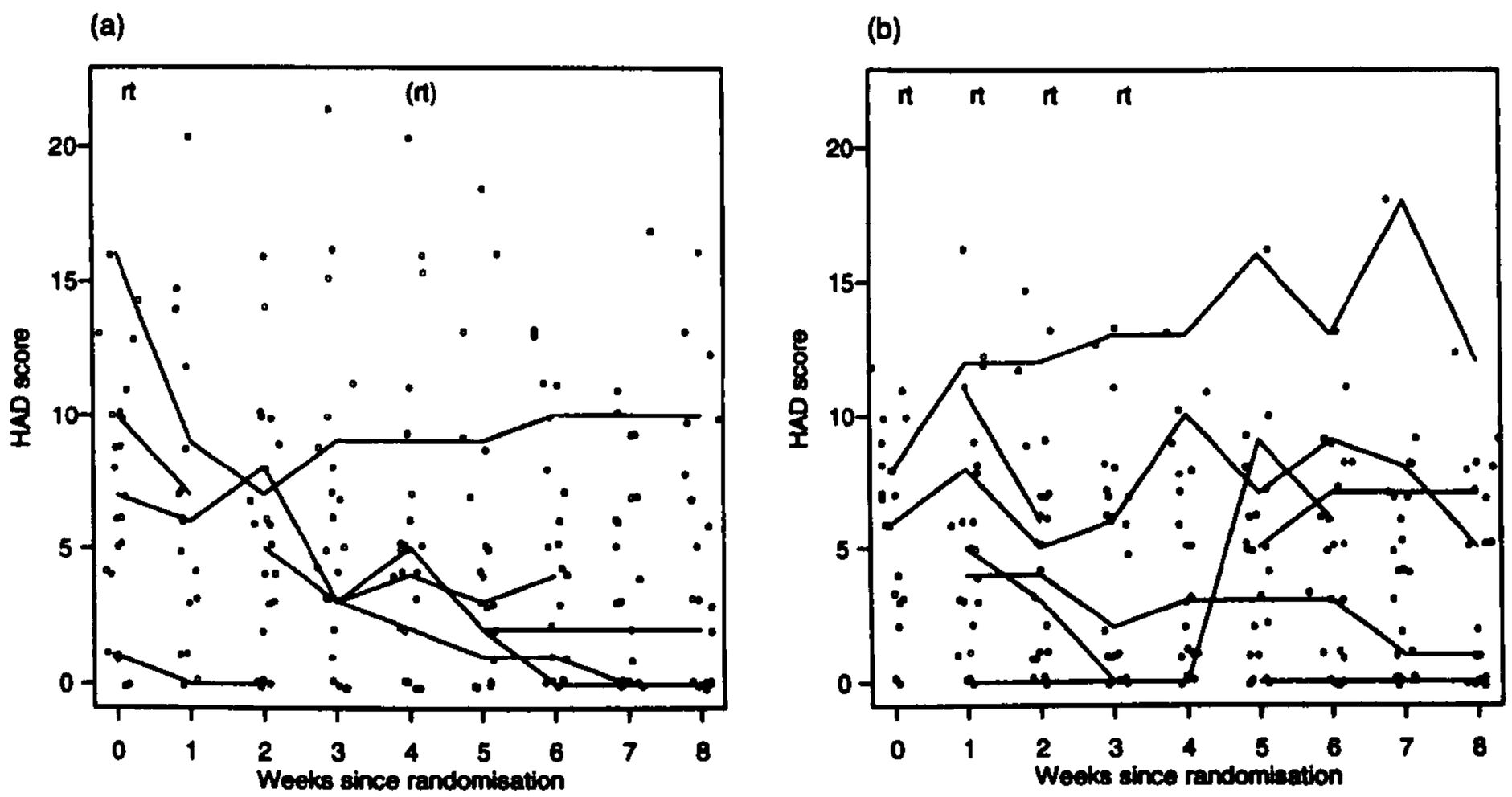

Figure 2.2: HAD anxiety scores over time from baseline (0) through the eight week follow-up for patients in the CRC NSCLC study with random profiles of a selection of patients overlaid: (a) split course radiotherapy; (b) continuous course radiotherapy.

However, only viewing a small sample of the subjects in the data set it is unclear whether the patterns are consistent throughout. In particular, the sample displayed here is very selective as it relied on patients having at least four responses.

In figure 2.2, the anxiety scores for all individuals are plotted against time with the profiles highlighted for the same individuals shown in figure 2.1. Since the scores are not strictly continuous and can take a limited number of values overall, there will naturally be some repetition of the same score by different individuals. To avoid scores being superimposed when this occurs, the responses in figure 2.2 have been 'jittered' - that is, a small degree of random noise has been added to observed measurement occasion and score.

The problem with the figure is that with the exception of the patients whose observations have been connected, it is impossible to see patterns of individual profiles, and even then, if the connected profiles are subject to missing data, the overall profile is not clear. It does, however, allow the extent of the variation in the data to be assessed, although it is not clear whether this derives from variation between or within subjects.

In figure 2.3, the RSCL physical responses for all subjects in the CRC NSCLC study are plotted in a lexis diagram. The data are reclassified into 'normal' and 'abnormal' scores according to the RSCL guidelines (de Haes *et al.*, 1990). Normal responses are shown as solid lines, abnormal scores (20 units or more) by a dashed line. Discontinuities in the lines represent missing responses. An asterix at the end of the patient response line denotes the time of death for that patient to the nearest week of follow-up. Patients are ordered from top to bottom by date of entry into the study and again the timing of radiotherapy is shown at the top of the figure.

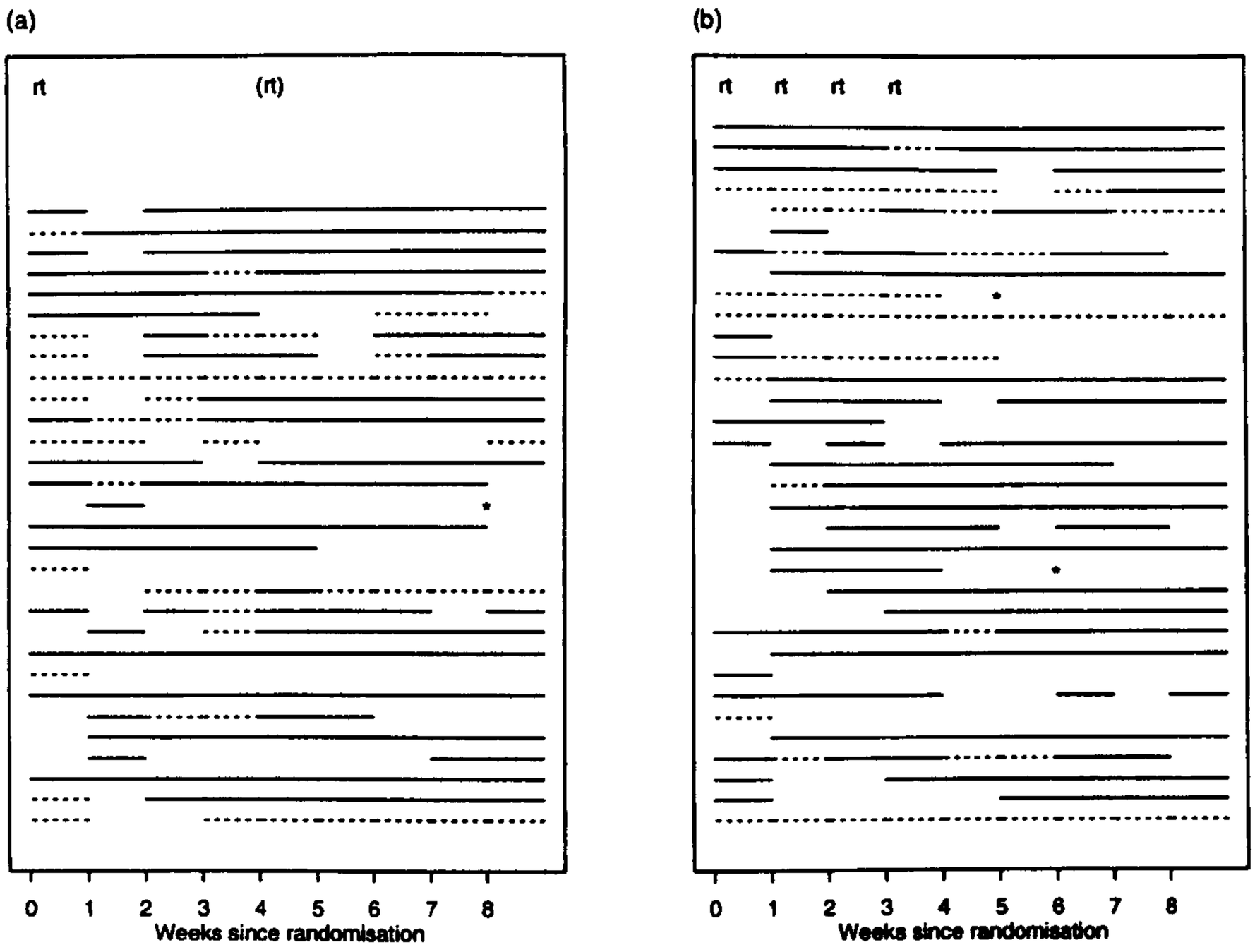

Figure 2.3: Lexis diagrams of patient physical quality of life profiles over time (from baseline 0) in terms of the RSCL normal score classifications: (a) split course radiotherapy; (b) continuous course radiotherapy. Normal scores: ————; abnormal scores: --------; patient death: *.

In this example, the figure highlights the degree of missing baseline responses for patients on the continuous course radiotherapy who were recruited in the middle of the study (figure 2.3(b)). It also highlights missing data prior to death for each of three patients who died during the course of quality of life follow-up. The prevalence of abnormal physical score seems similar between the two treatment groups, and no particular patterns over time are revealed.

On the whole, although it is impossible to obtain clear inferences from any plots of individual profiles over time, they can highlight consistencies (and inconsistencies) in patterns over time in terms of the underlying level of response, the occurrence of missing data, and how this relates to previous quality of life responses or patient death. In addition, they allow examination of the total variability of the data and, with treatment schedules clearly shown in

the figures, the immediate effects of treatment on patient quality of life can also be highlighted.

2.2.2 Individual patient summaries

Since longitudinal data has repeated assessments on the same subject, the total variability of the data can be partitioned into that which is derived between subjects and that which derives within, that is, differences in the underlying responses between subjects and the variability of individual responses around this underlying response for a particular subject. Although individual patient profiles allow some examination of the overall variation in the data, it is generally difficult to determine how this variation partitions into that between and within subjects. Such information is important however as it indicates whether subjects tend to be consistent in their underlying responses or not, which then has substantial implications for the generalisation of conclusions to be later drawn from an analysis. The examination of individual patient summaries during exploratory data analysis is therefore important.

Such analyses have been advocated by Matthews *et al.* (1990) not only for exploratory data analysis, but for formal statistical analysis of longitudinal data. Here they are presented only for descriptive purposes. This is because the unbalanced nature of quality of life data somewhat complicates their use for formal analysis (Matthews, 1993).

For continuous data, an often reasonable summary of a patients response is an overall mean, or a fitted regression line over time. In figure 2.4, separate regression lines fitted for each subject over time are shown for the HAD anxiety responses in weeks 1-8 from the CRC NSCLC study. These are plotted for each treatment group separately. Also shown on the figure are the distributions of the fitted intercepts and slopes for the two patient groups.

The figure shows an underlying fall in the level of anxiety over the period which appears

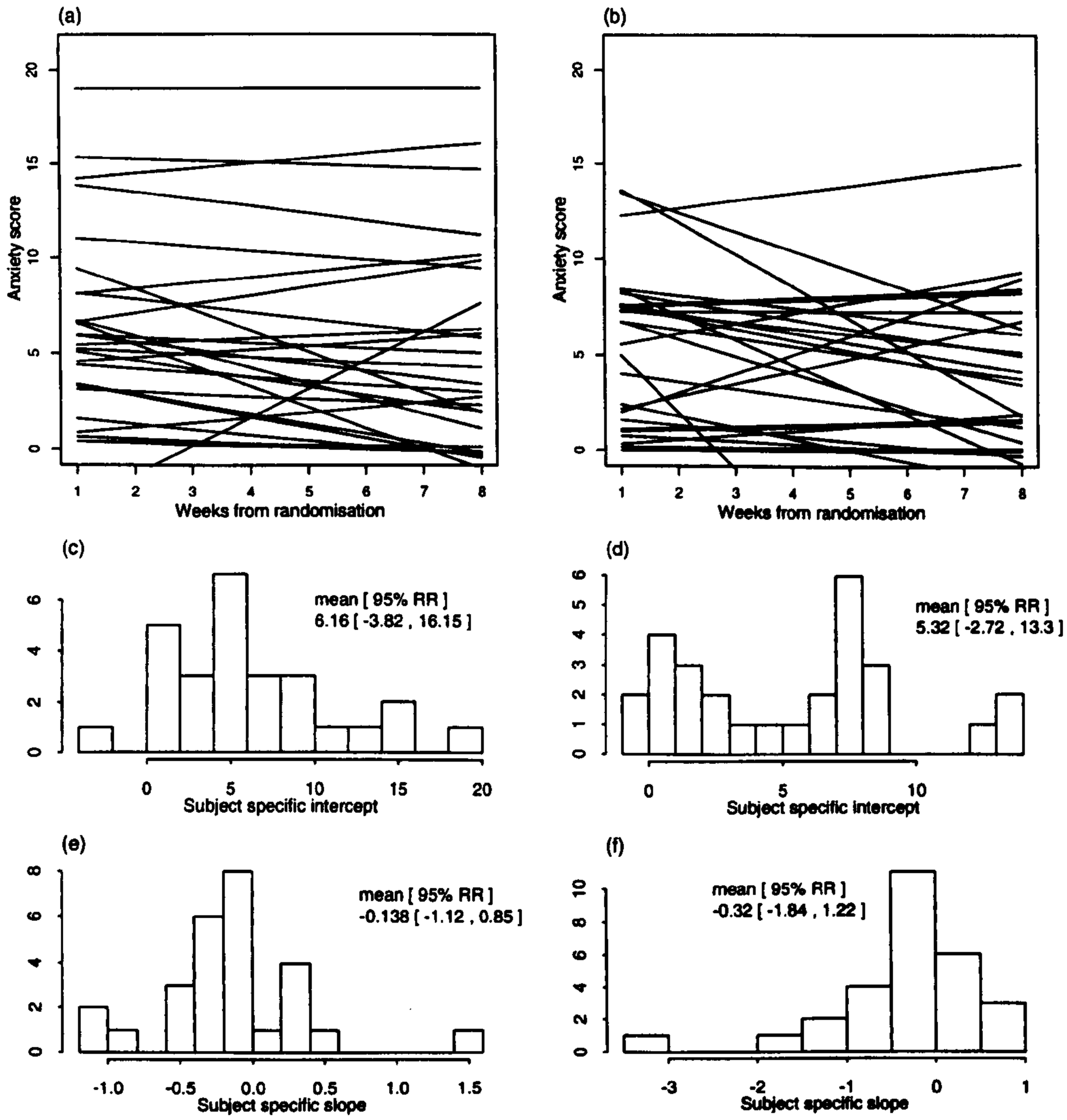

Figure 2.4: Subject specific regression analyses for HAD anxiety scores from the CRC NSCLC study: (a) split course radiotherapy; (b) continuous course radiotherapy. Distribution of subject specific intercept (c) split course; (d) continuous course. Distribution of subjects specific slopes (e) split course; (f) continuous course.

more consistent across patient in the split course than for those on the continuous course. In contrast, fitted responses for the underlying level of response at one week seem more consistent between subjects in the continuous course group. A number of obvious outliers are seen in the figures. In particular, one subject in the split course showed a marked increase in their response over time, and one on the continuous course showed a marked decrease. Examination

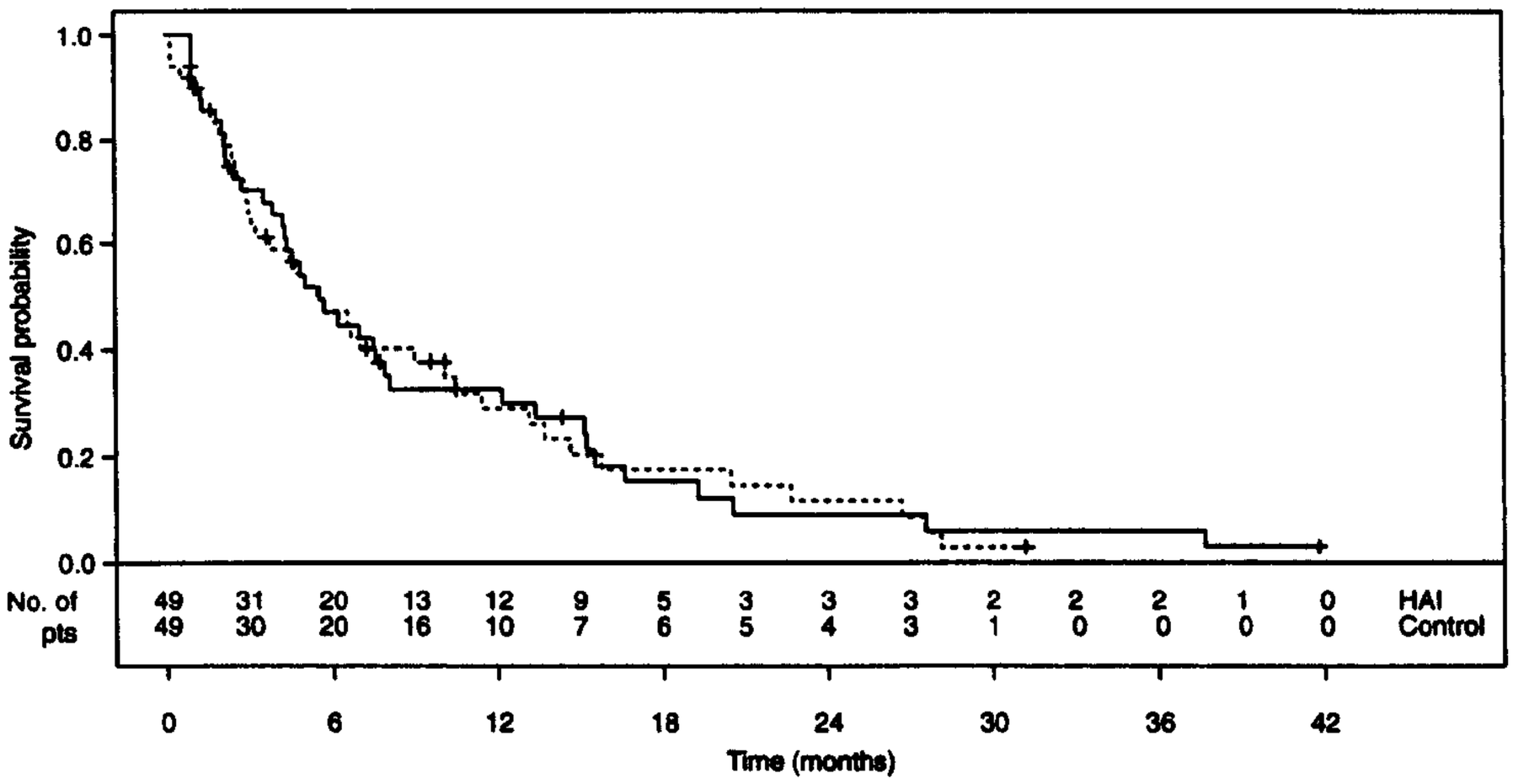

Figure 2.5: Kaplan-Meier representation of the time to abnormal physical quality of life or death in the CRC HAP trial. HAI:————; control: --------; censored observations: +.

of the data for these individuals showed that they had only three and two responses respectively.

Matthews *et al.* (1990) discuss alternative summary statistics for continuous data. One which may be of scientific interest for quality of life data analysis, is the *time to an event of interest*. An example which is shown in figure 2.5 is the time to the first occurrence of 'abnormal' quality of life. Here, the RSCL 'normal' score classification has been used to determine the time to the first occurrence of abnormal quality of life or death for the CRC HAP trial. Because some patients may be censored, this is presented using a Kaplan-Meier survival curve. In this example, no difference between the two groups is observed. Unfortunately, this event definition makes it is impossible to separate the quality of life and survival experience of patients and may therefore be best restricted to situations where patient quality of life is expected to deteriorate progressively, and in cases with a short follow-up period, so that patient

death occurs rarely during the period of quality of life assessment.

2.2.3 Summaries over time

Having examined the variation of responses across subjects, the most obvious method of summarizing a continuous response over time is to plot patient group mean scores at each measurement occasion against time. For an indication of the precision of the mean, confidence intervals can be displayed, calculated with either a specified confidence level at each time point or a more conservative overall confidence level for the period as a whole using a Bonferroni correction (Armitage and Berry, 1987). Using both of these intervals will give limits for the range of confidence, with the pointwise intervals being too narrow and the Bonferroni intervals being too wide as they assume that all time points are independent and therefore over adjust.

An example is shown in figure 2.6 where the mean HAD anxiety score at each weekly measurement occasion in the CRC NSCLC study is plotted against time by treatment group. Both pointwise and Bonferroni 95% confidence intervals are shown on the figure given by the inner and outer horizontal bars respectively. Also shown is the number of patients contributing to the estimated mean at each time point.

In this case the figure highlights a slight fall in the mean anxiety score over the period in both treatment groups. The mean response for the patients on the continuous course of radiotherapy is also shown to be consistently lower than that for those treated on the split course although all confidence intervals throughout the period overlap.

A problem with displaying data in this way is the tendency for over interpretation of confidence intervals which fail to recognise that the data derive from repeated assessments of the same individuals. Further, in studies where quality of life follow-up is long relative to the

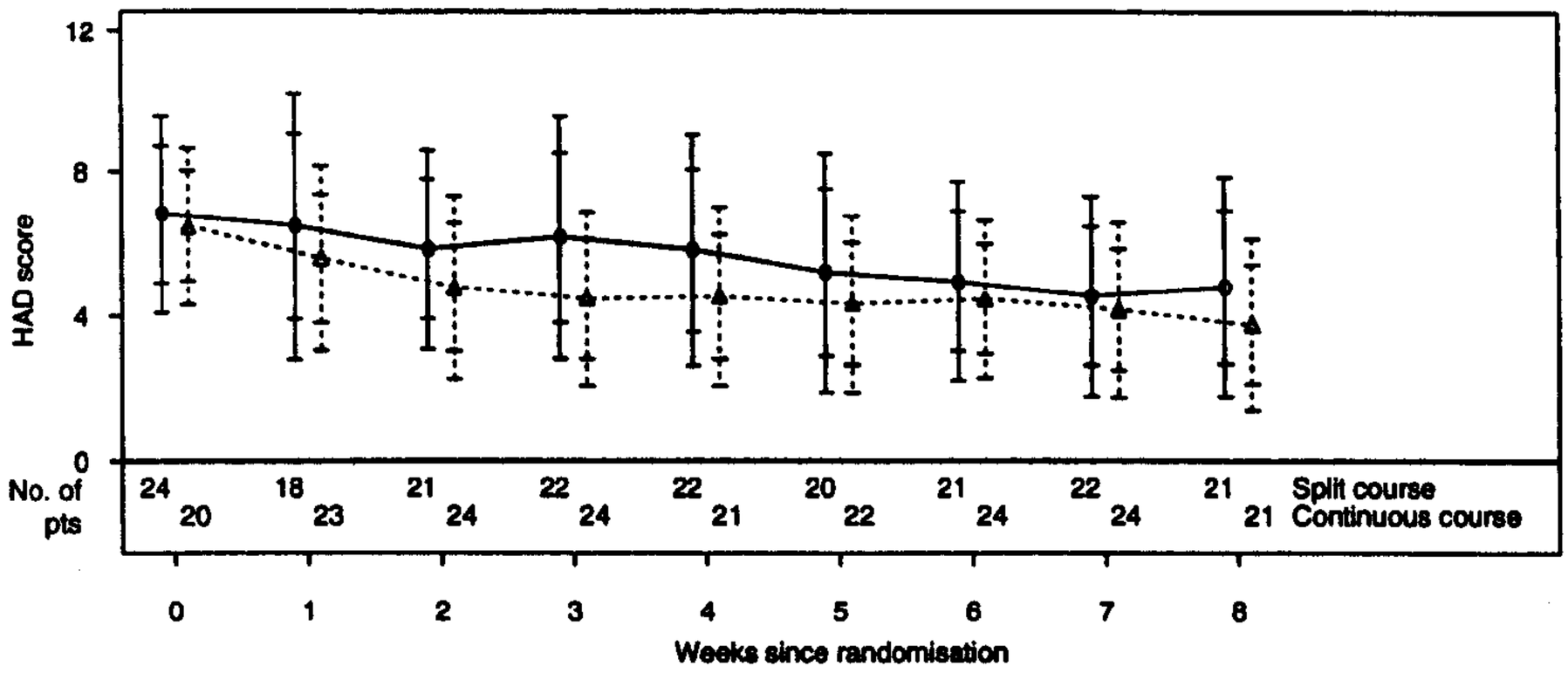

Figure 2.6: Mean anxiety scores over time from baseline (0) through the eight week follow-up for the CRC NSCLC trial with pointwise and overall 95% confidence intervals given by the inner and outer horizontal bars respectively. Split course radiotherapy:————; continuous course radiotherapy: --------.

expected patient survival, patient attrition due to death makes the figure difficult to interpret, particularly in terms of the overall trend. This problem has been addressed by several authors (Stephens *et al.*, 1992, Hopwood *et al.*, 1994) and is discussed in more detail in Section 2.4.

A further problem with the presentation of group mean quality of life scores is the lack of an intuitive interpretation of the scores which makes it difficult to convey the meaning of results to people with no knowledge of the measurement instrument used, and difficulties in comparing results across studies (Fayers and Jones, 1983, Cox *et al.*, 1992). Cox *et al.* (1992) suggest that a solution is to present results on a transformed scale representing the "percentage out of the maximum" for the instrument. Although such transformations may have more intuitive appeal than the raw scores, comparisons across studies may still not be possible across studies using different measurement instruments as the sensitivity of different instruments to quality of life changes might vary greatly.

A representation which may improve an analysis for easier comparison across studies is to present results in terms of the instrument defined 'normal' scores which, it is hoped, would be more consistent in their classification of responses. The use of summaries of 'normal' scores over time is exemplified in figure 2.7 for the CRC NSCLC RSCL physical data previously shown in figure 2.3. In the figure, the proportion of patients recording 'abnormal' scores at each follow-up is plotted over time. Again the numbers of patients contributing data at each time point are given. In the example presented a slight decrease in the quality of life response over time is seen. There is however an indication of a discrepancy between the proportion of patients with 'abnormal' scores in the two groups at baseline. This may be related to the excess of missing data at baseline in the continuous course which was shown in figure 2.3, and highlights the importance of profiles of individual patient data.

For the examples in figures 2.6 and 2.7, quality of life was measured at very short intervals

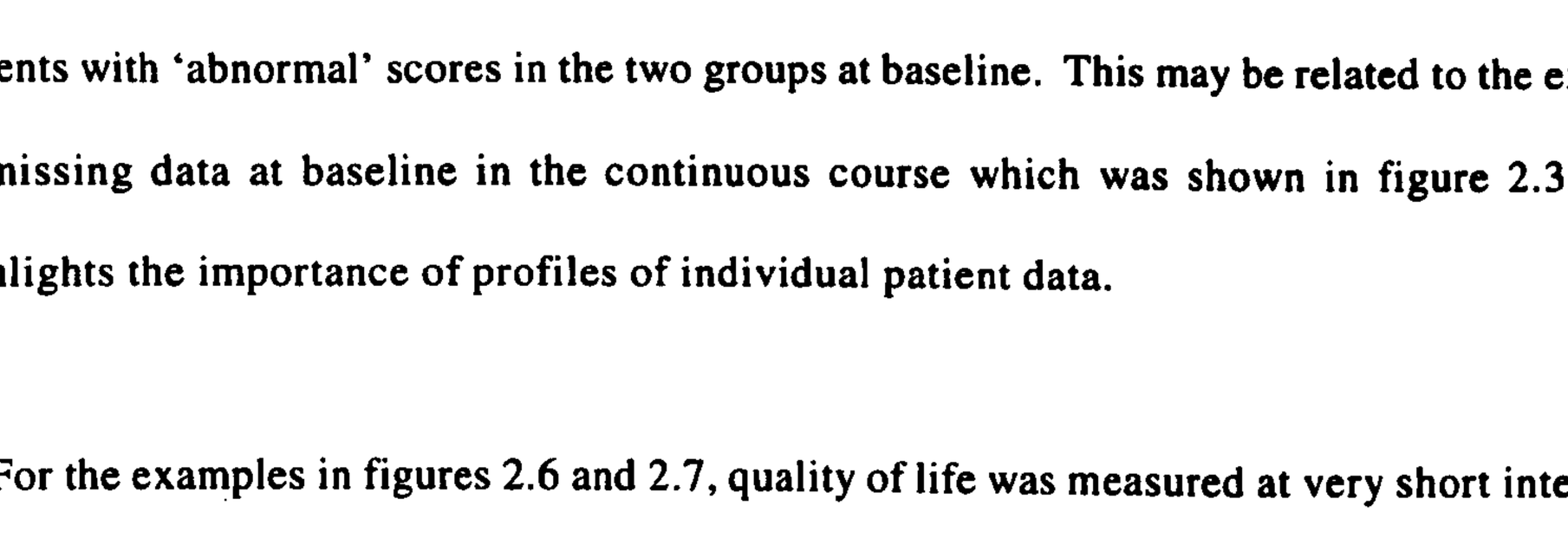

Figure 2.7: Proportion of patients in the CRC NSCLC trial giving 'normal' scores for the RSCL physical dimension plotted over time from baseline (0) and throughout the eight week follow-up where normal scores are classified as responses greater than 20. Split course radiotherapy: ------; continuous course radiotherapy: --------.

which meant that the exact timing of measurement was as planned in the study protocol. When assessments are less frequent - for example, monthly - time constraints on a clinician or patients being unable to attend follow-up appointments, typically the actual timing of follow-up will not always be as planned. Summaries of mean response over time then become more difficult as there are no longer distinct time points at which the required means can be calculated. The most obvious solution is to group follow-up times, for example, to the nearest month. This has been done in figure 2.8 which shows the mean RSCL physical quality of life scores over time for the restricted data set for HAP trial. It shows very little difference between the pump and the control groups over the first year of follow-up. Beyond the first year, although some difference seems apparent, examination of the number of patients contributing data stresses the problem of falling patient numbers which make these apparent differences difficult to interpret. Missing data and patient death during follow-up can have serious implications for the interpretation of these mean summaries. It is again therefore very important that patient

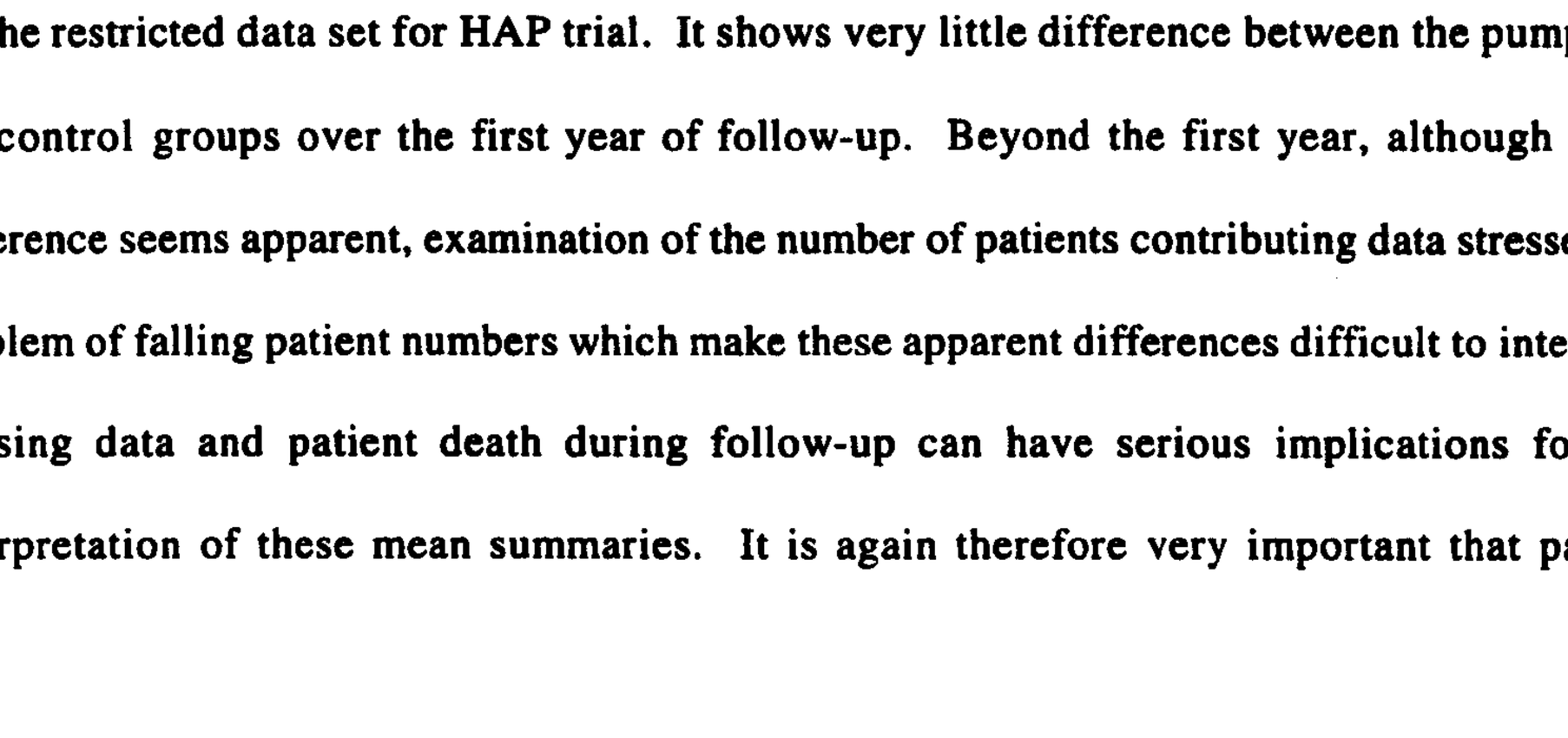

Figure 2.8: Mean RSCL physical scores over time from baseline (0) for two years follow-up for the restricted HAP trial data. Follow-up is grouped to the nearest month. HAI: ————; control: --------.

numbers are given on a figure.

An alternative to grouping data together in fixed time points, is a *kernel smoother* (Hastie and Tibsharani, 1990). The basic principle of kernel smoothing is to obtain a smooth representation of the data with time by grouping data within a 'moving window' of a given width. For each successive window, some summary of all the data points within the window is calculated. These summaries are then joined over time. The wider the width of the window, the smoother the resulting summary over time will be. For example, a lowess kernel smoother (Cleveland, 1979) is used in figure 2.9 again for the restricted data from the CRC HAP data. A lowess smoother is a particular kernal smoother which is insensitive to outliers. Within each moving window, a weighted least squares regression line is fitted with weights determined by the distance of each point from the centre of the window. The residual of each observation from this fitted line is then calculated, outlying observations are down weighted and the line re-fitted and the process repeated a number of times. The value of the lowess curve for each window is then simply the predicted value for the line at the centre of the window. Figure 2.9 shows little difference between the HAI and control groups, although the response of the control group is shown to lie consistently below that of the HAI group for most of the follow-up period. Although for the control group, the line does begin an upward turn towards the end of the follow-up period it is much less sensitive to the outlying values observed for this group at the end of follow-up than the profile given by simply grouping data together as in figure 2.8. As it is now impossible to give the precise number of subjects contributing data at each time point, the raw data is shown on this figure as an indication of the amount of available data. This clearly shows the depletion of data towards the end of follow-up. Full details of lowess and other kernal smoothers are given in Hastie and Tibsharani (1990).

Using mean profiles over time is perhaps the most important part of exploratory analysis for

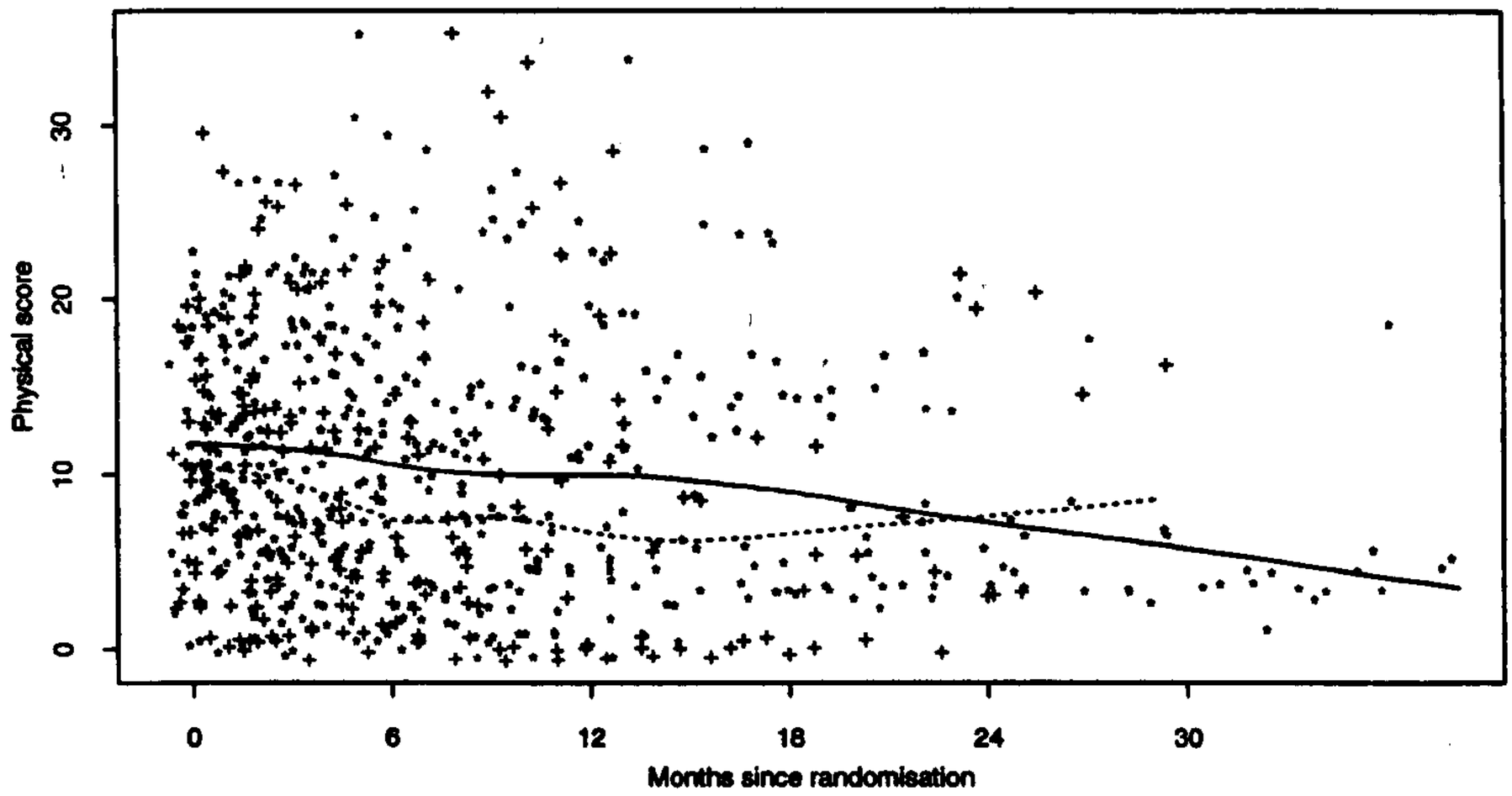

Figure 2.9: RSCL physical scores for the restricted data taken from the CRC HAP trial: HAI: *; control: +. The underlying response is highlighted using a lowess kernel smoother: HAI: ————; control: --------.

longitudinal quality of life data as it generally addresses the primary questions of interest in the data - the difference between patient groups and the behaviour of response over time. Because of problems of missing data, irregularly spaced follow-up assessments and patient attrition due to death, it is important that either patient numbers are given within a figure, or the overall profiles are overlayed on the raw data. In addition, as measures of confidence and the profiles themselves ignore the dependency of observations within the data, such profiles should not be over interpreted for formal analysis.

2.2.4 Associations between dimensions

Although many different dimensions are measured as part of a quality of life assessment, an aspect of the data which has generally not been reported in the literature is the correlation (or association) between responses in the different dimensions. Since the data are longitudinal

and the total variation can be partitioned into that between and that within subjects, the associations across dimensions need also to be partitioned in the same way. That is, the correlation between dimensions between subjects, and the correlation between dimensions within subjects need to be evaluated separately.

When data are balanced, the most simple estimate of the association between subjects is obtained by calculating the subject specific mean scores for each dimension in turn over the follow-up period and then to examine the associations across dimensions between these subject specific means. This investigates whether subjects who tend to have high scores on average in one dimension tend also to have high scores in other dimensions. This is shown in a scatter plot matrix in figure 2.10 for all four quality of life dimensions of quality of life measured in the CRC NSCLC study. It shows a high positive association between subject specific means in the

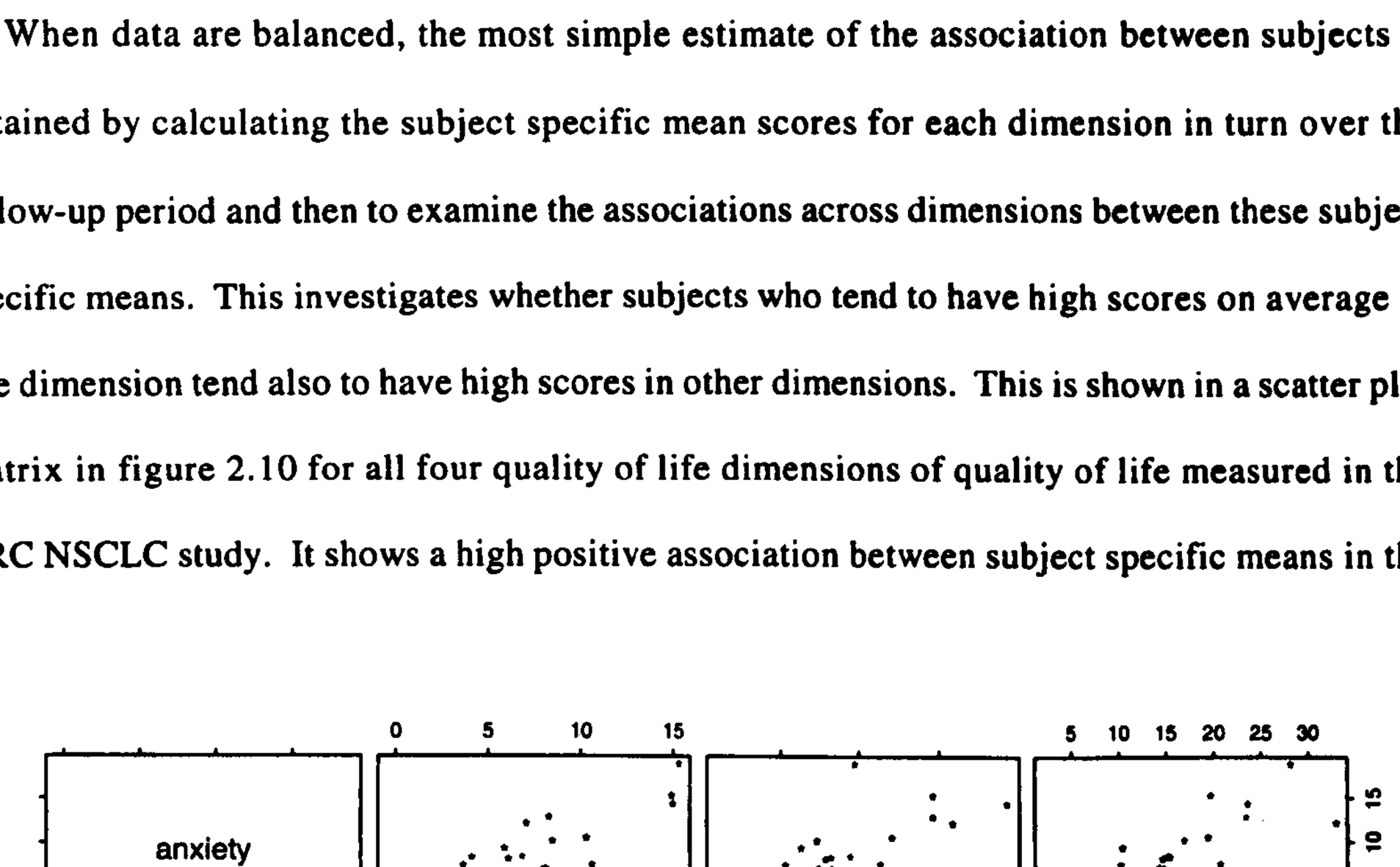

Figure 2.10: Scatter plot matrix of the subject specific means over time for each of the four quality of life dimensions measured in the CRC NSCLC study for a representation of between subject, between dimension correlations. The HAD anxiety and depression scores have values in the range 0-21, those for the RSCL physical and psychological scores are in the range 0-40.

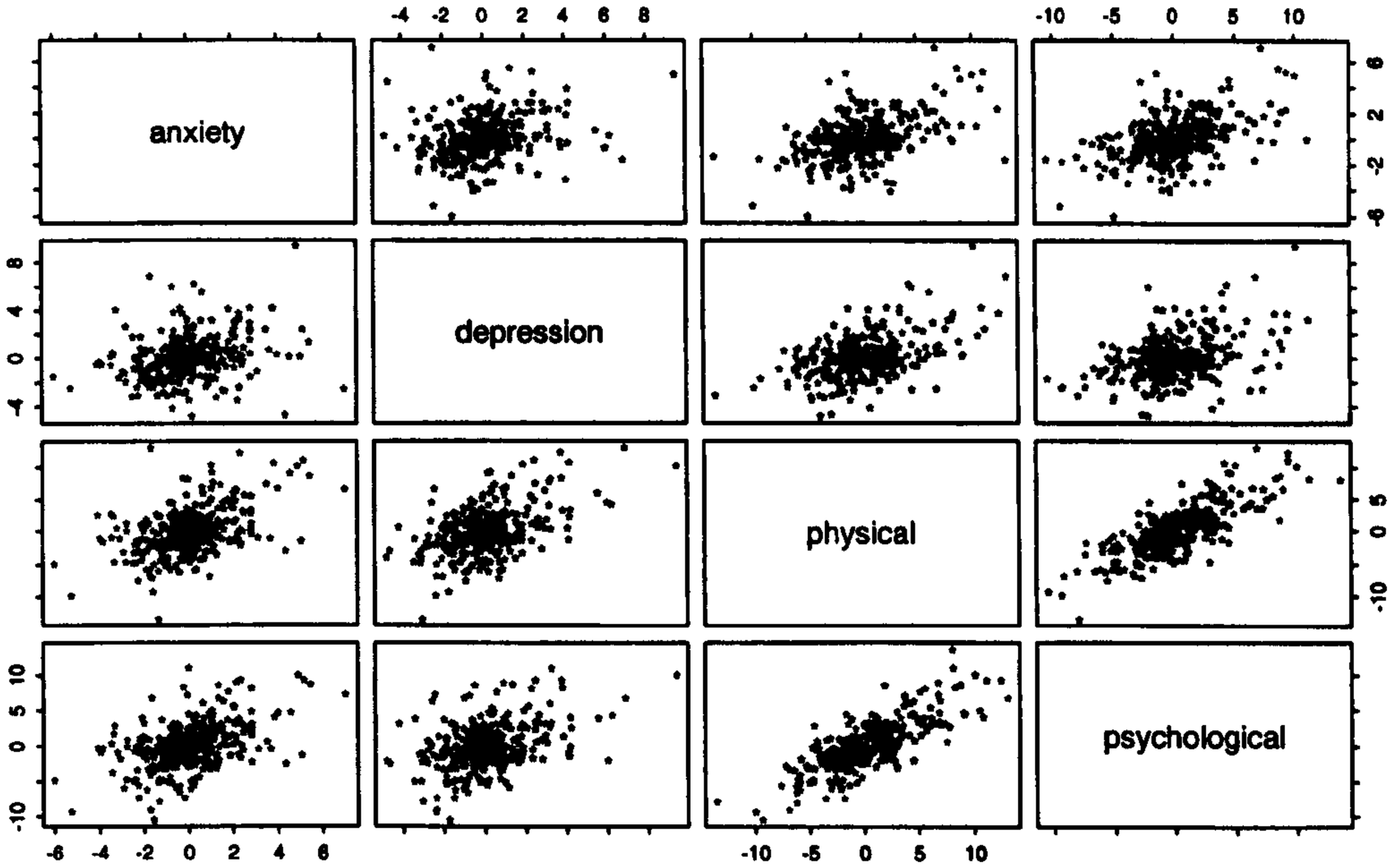

Figure 2.11: Scatter plot matrix for the within subject between dimension correlations for the quality of life data for the CRC NSCLC study.

physical and psychological dimensions of the RSCL. Although positive, the associations seen between other dimensions are generally weaker.

The within subject across dimension associations address the question whether, on a particular occasion, subjects who have higher than expected observed responses in one dimension also have higher than expected responses in a second dimension. For reasonable estimation of such correlation, Bland and Altman (1995b) suggest the residuals from subject specific regression analyses be used. For descriptive purposes, it is perhaps sufficient to simply subtract the subject specific mean responses from their respective observed responses and examining the pattern of association of these 'residuals' across all subjects. This is shown in figure 2.11 for the NSCLC quality of life data. This clearly shows that although there was a high degree of association between the two dimensions of the RSCL within subject, those between the two dimensions of the HAD scale were very much weaker. Some positive

association between depression and physical well being is also shown.

Each of these methods, however, depend on the data being balanced. When data are unbalanced, although visual displays of association for descriptive purposes may still be used with caution, estimation of correlation coefficients is not recommended without the varying degrees of precision on each subject specific mean taken into account (Bland and Altman, 1995a, 1995b).

2.3 Repeated categorical data

Although many of the most commonly used quality of life instruments have a validated way of summarizing individual items on the questionnaire to give a simple summary score, which may then be treated as continuous responses, there remain some instruments - for instance, the daily diary card - for which such a summation is not feasible or relevant. This section focuses on exploratory data analysis for the binary and ordinal data which are obtained from such instruments. It also considers the analysis of individual items within questionnaires in general. This enables individual aspects of a patient's life to be examined and effects investigated which may be masked by the calculation of an overall summary measure. The binary responses presented are generally dichotomies of the ordinal responses of the individual items on the measurement instruments such as the example item given in box 1.1 for the HAD anxiety scale. As for the previous section, this work is separated into four subsections addressing the use of individual summaries, summary statistics, profiles over time and across dimensional associations.

2.3.1 Individual summaries

As with a continuous response, examination of the individual patient behaviour in order to examine the consistency of responses across subjects is as important as the more commonly reported population summaries for patient groups. For ordinal, and in particular, binary data, this is best done using a lexis diagram. Such diagrams allow large proportions of the data to be displayed, and the small number of possible response categories can generally be quite easily distinguished on a single figure. This is shown in figure 2.12 in which the prevalence of dysphagia in the MRC LU07 study is plotted over time for a random sample of patients in each treatment group. A solid line indicates that no symptoms of dysphagia were reported, the dashed line indicates some symptoms - that is, a response in category 2 or above. Again,

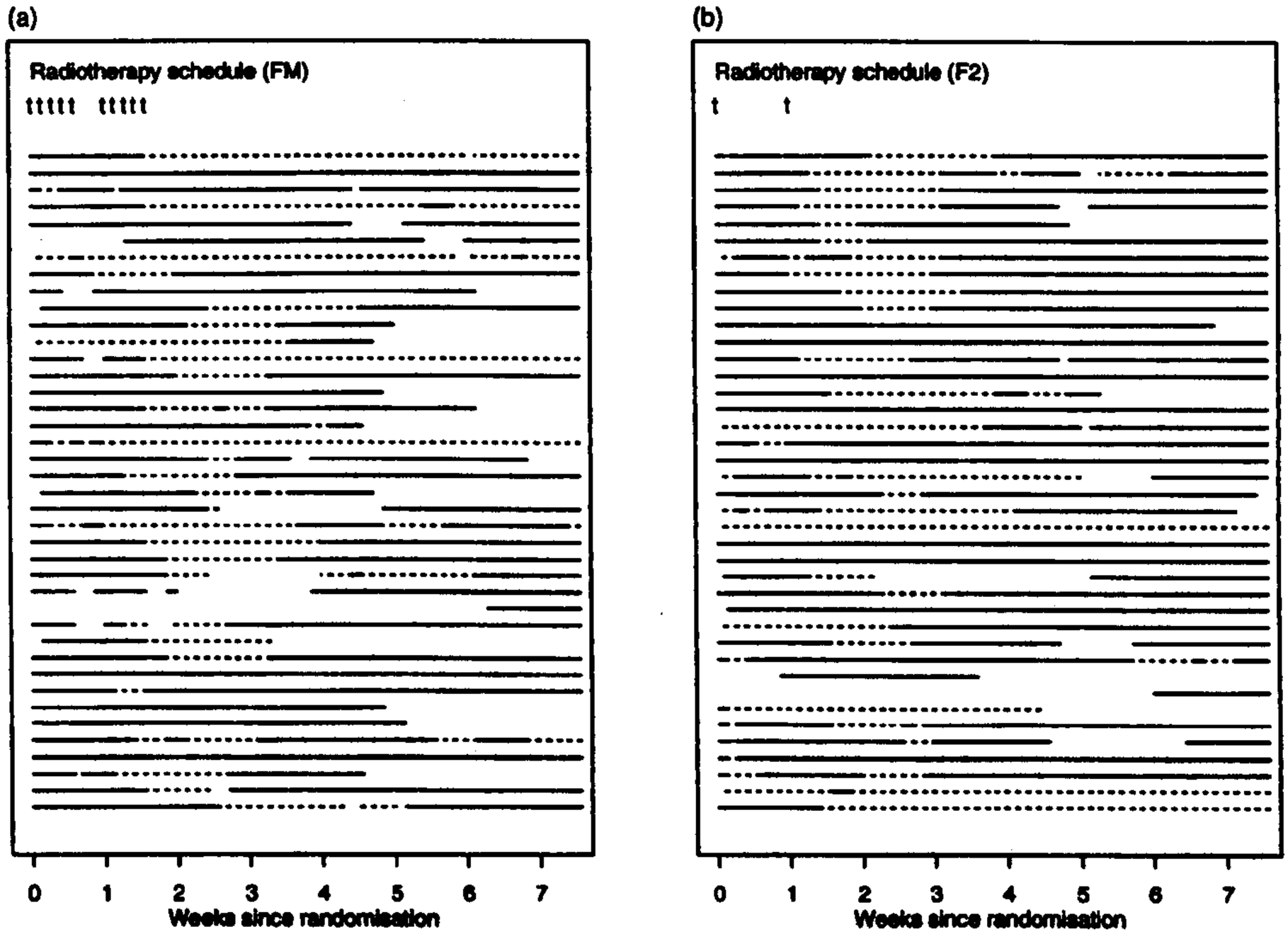

Figure 2.12: Random sample of subjects and their reported symptoms of dysphagia measured daily from the start of treatment (0) through the following eight week period in the MRC LU07 study for (a) multiple fraction radiotherapy (FM); (b) two fraction radiotherapy (F2). No symptoms (category 1):————; some symptoms (category 2 or above): --------. Discontinuities in the lines indicate missing responses.

periods of radiotherapy are shown on the figure. The figure reveals a large proportion of patients in both treatment groups reporting some symptoms of dysphagia towards the end of the radiotherapy for a few days in both treatment groups, although the exact timing of the onset and relief of symptoms varies greatly across the patients shown.

2.3.2 Summary statistics

The use of summary statistics for the analysis of binary and ordinal responses is less common than for continuous data. It is however just as useful in highlighting whether variation derives primarily from differences between or within subjects. The natural summary statistic for such data is the proportion of patient time with a response in each category, where this is calculated for each subject as the number of responses given in each category taken as a proportion of the total number of responses given by that subject. Two possible summaries of the proportions are shown in figures 2.13 and 2.14 for summary statistics calculated for the MRC LU07 activity quality of life scores from the daily diary card. These scores were measured on a five points scale ranging from 1 (normal work/housework) to 5 (confined to bed). Further details of the item are given in box A2.3 in Appendix 2. Figure 2.13 shows the distributions of the proportions in terms of box plots. The middle 50% of the data is shown by the shaded block with the median give by the white bar. The tails of the box go out to the 10th and 90th percentiles with other outlying points shown as individual lines. The figure here shows the distributions in all but the middle category to be positively skewed with a large proportion of patients spending none of their time in some of the categories. This is particularly marked for categories 1 and 5 for which the 90th centile is zero. It also highlights the large degree of variability across subjects with a large spread for the middle 50% of the data in the middle categories. No obvious differences between the treatment groups are seen.

These data are also summarised in figure 2.14. Mean and median proportions of responses

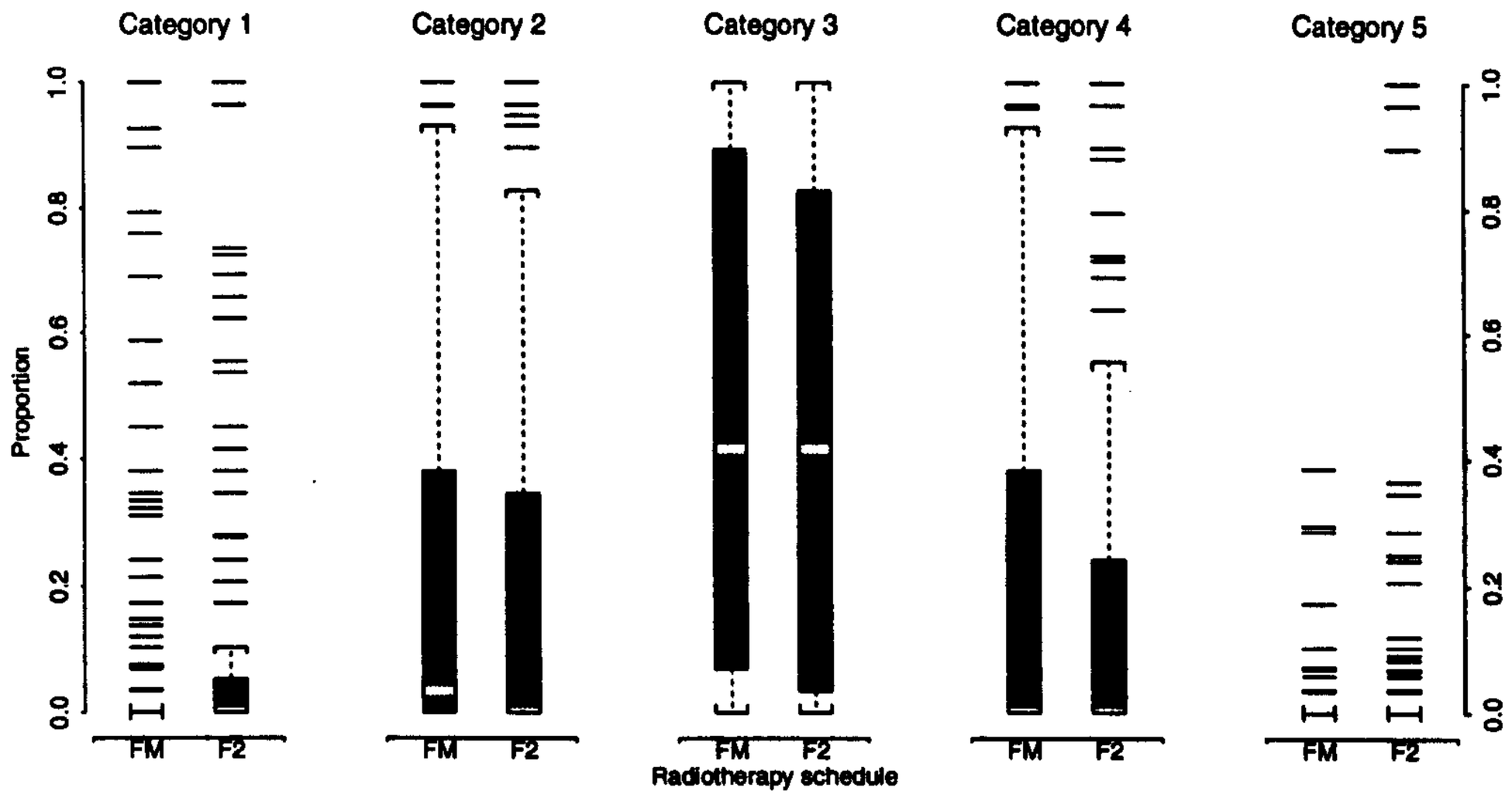

Figure 2.13: Distribution of the proportion of responses each patient gives in each category (1 to 5) for the activity rating of quality of life on the daily diary card in the MRC lung cancer study during the four week treatment period. The middle 50% of the data is shown by the shaded block with the median given by the white bar. The tails of the box go out to the 10th and 90th percentiles with other outlying points shown with individual lines. Multiple fraction radiotherapy (FM) and two fraction radiotherapy (F2).

in each category are plotted with their respective 95% confidence intervals. Given the skewed nature of the distributions, the summaries based around the median will be preferable to those around the mean. No apparent differences between the two treatment groups are highlighted.

Although showing a clear summary of patient experience over the period as a whole, and allowing a simple treatment comparison between the treatment groups, the problem with figures 2.13 and 2.14 is that they do not allow examination of patterns of change over time. Thus patterns of change over time, such as that highlighted in figure 2.12 for the dysphagia data would be missed. Assuming a linear trend in log odds, for binary data, overall trends could be identified in a similar way to that for continuous data, with subject specific rates of change in the in log odds of symptoms obtained from a logistic regression analysis and plotted in the same way as in figure 2.4. In such a case, the underlying subject specific model is then perhaps too sophisticated for simple exploratory data analysis.

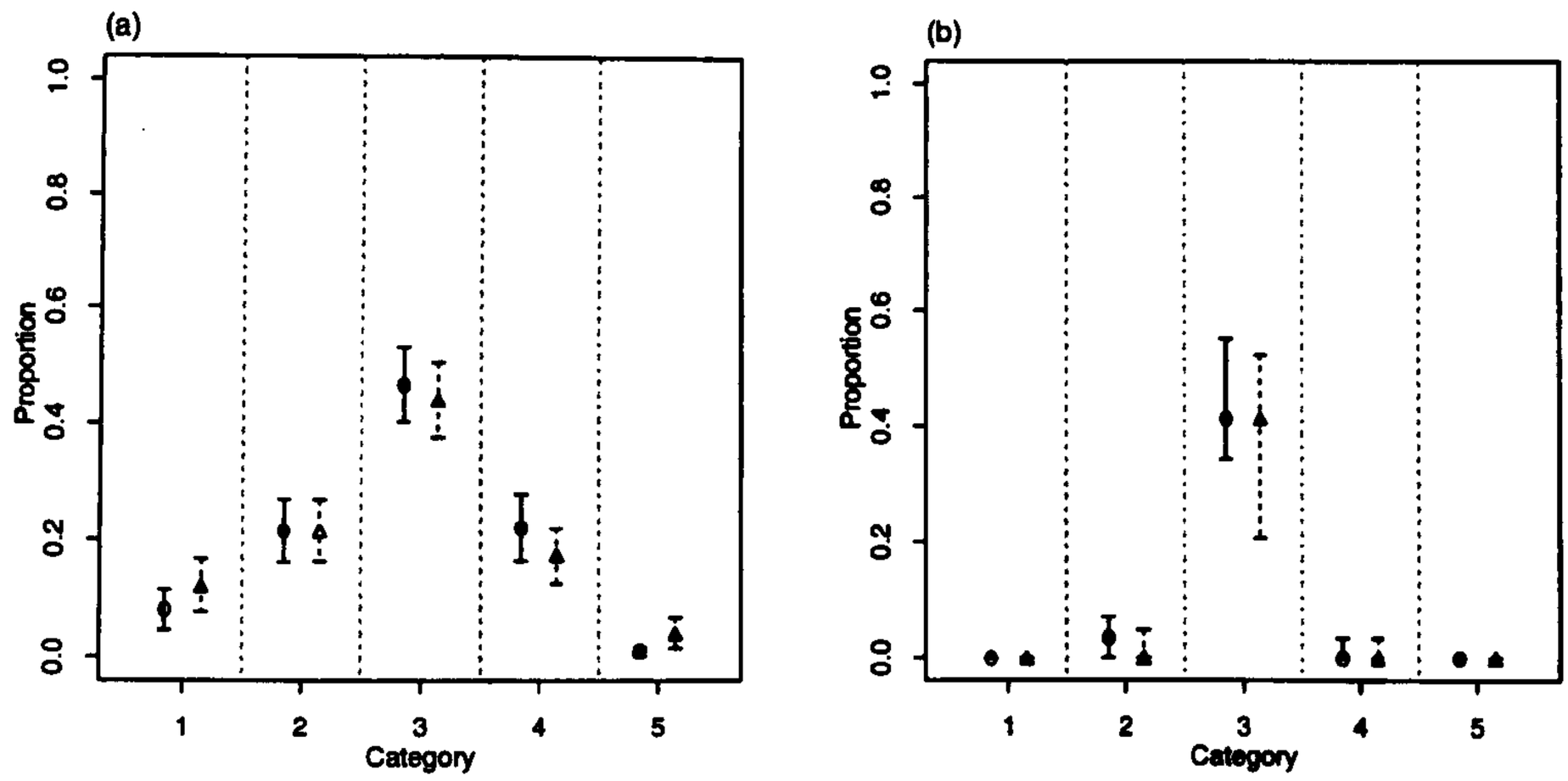

Figure 2.14: (a) Mean and (b) median proportion of responses given in each category (1 to 5) with 95% confidence intervals for the activity scores on the daily diary card. Confidence intervals for the mean are truncated at 0 and 1. Intervals given for the median are based on critical values of the sign test. Multiple fraction radiotherapy (FM):————; two fraction radiotherapy (F2): --------.

2.3.3 Summaries over time

The most commonly used way to summarize patient quality of life measured on a binary or ordinal scale is to consider the proportion of patients falling into each category over time. These proportions can be plotted separately for each category, cumulatively over categories by considering the proportions in category k or below (or above) for all k, or for a simple dichotomy of the response. Such figures summarize the data to give an impression of the level of quality of life at each specific time in the study and can be useful in highlighting changes in the distribution of patient responses at particular times during the study follow-up. They have been used by many authors (MRC lung cancer working party, 1991a, 1991b, 1992, 1993b, Fallowfield *et al.* 1986). Two such examples are presented here.

Figure 2.15 shows the proportion of responses in category k or below (k=0,1,2) recorded on

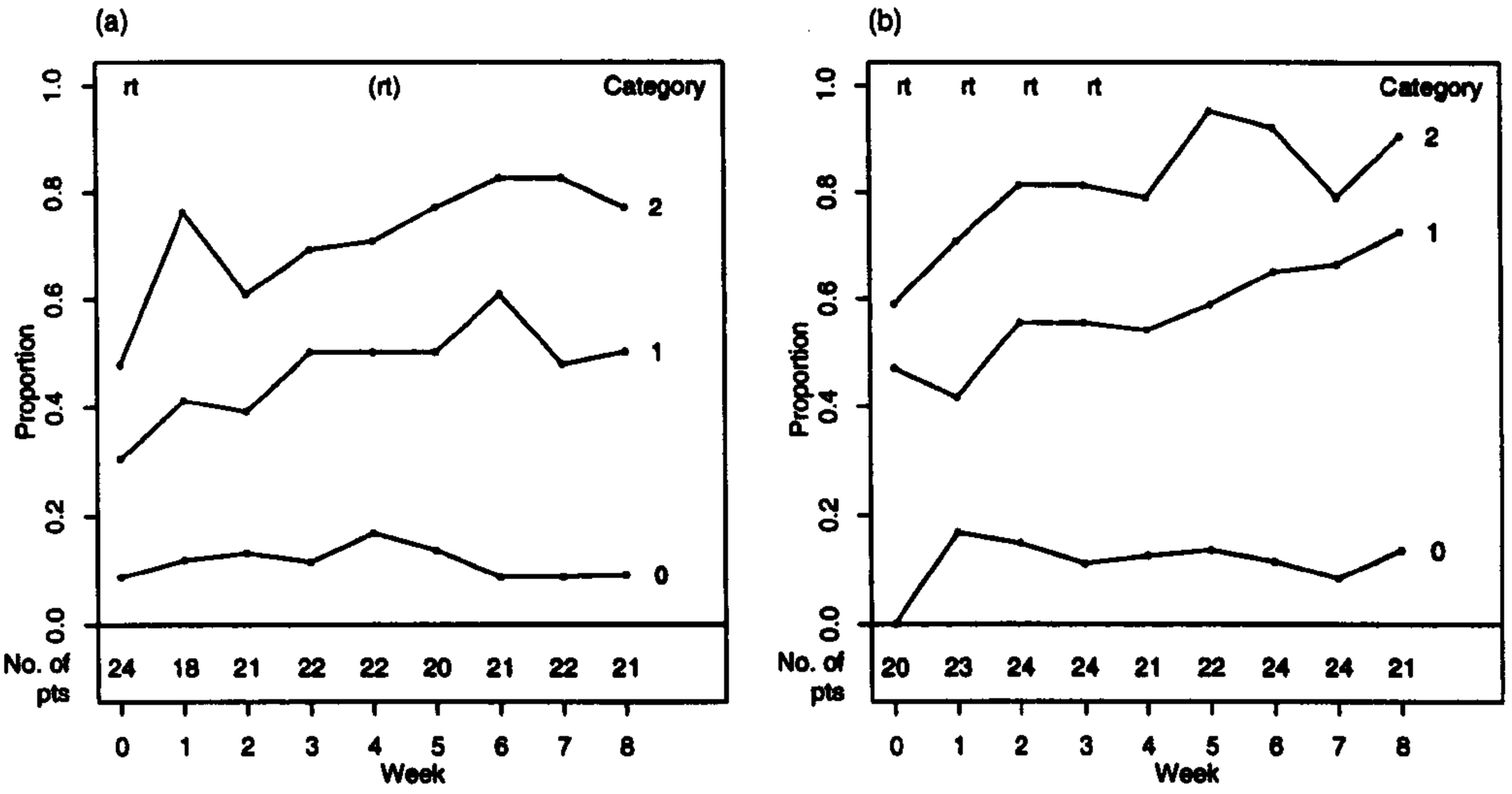

Figure 2.15: Proportion of patients recording responses in category *k* or below over time from baseline (0) through the eight week follow-up for the '*shortness of breath*' item on the RSCL questionnaire used in the CRC NSCLC trial for patients on the (a) split course radiotherapy; (b) continuous course radiotherapy.

the RSCL shortness of breath item in the CRC NSCLC study, where a score of 0 reflects no symptoms to a score of 3 that implies very restrictive symptoms. It shows an increase in the proportion of patients reporting symptoms of grade one or below (and therefore grade two or below) over time. This increase is perhaps more apparent in the continuous course radiotherapy group.

Figure 2.16 plots the proportion of patients over time reporting some symptoms of dysphagia (difficulty in swallowing) - a response in category 2 or above - in the MRC LU07 study. It clearly highlights that patients in both treatment arms experienced a transient period of dysphagia during the immediate period following radiotherapy treatment as was suggested from the random selection of individual profiles in figure 2.12. There was some difference in the pattern of response between the two groups with a lower proportion of patients affected in the F2 radiotherapy course. A similar picture was observed by the MRC lung cancer working

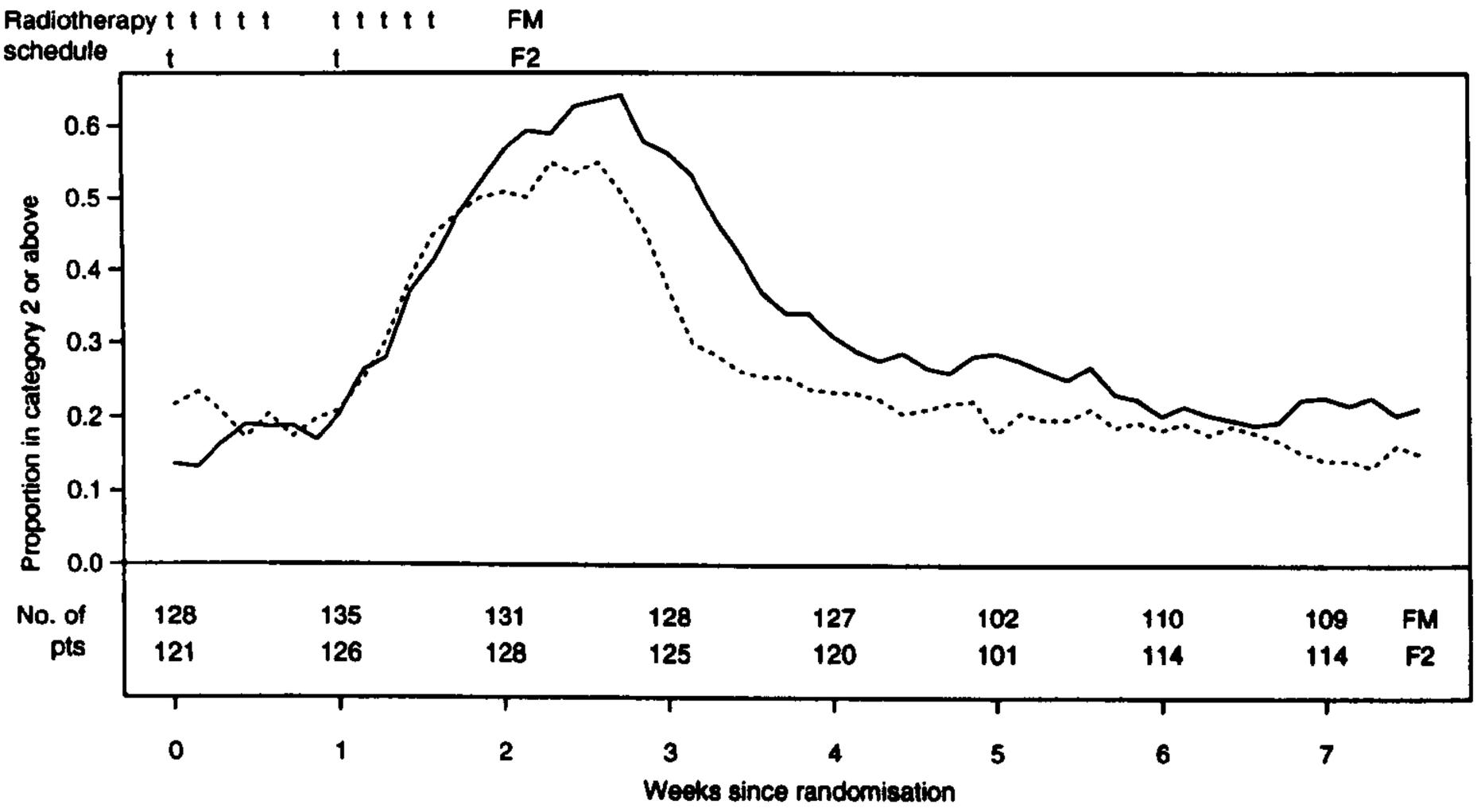

Figure 2.16: Proportion of patients reporting some symptoms of dysphagia (category 2 or above) over time from the start of treatment (0) through a subsequent eight week daily follow-up. Multiple fraction radiotherapy (FM):————; two fraction radiotherapy (F2): ----------.

party (1991b) with the same data set although in their analysis they concentrated of symptoms of category 3 or above. A formal statistical comparison of each these examples is presented in Sections 4.4 for the MRC LU07 dysphagia data and 5.3 for the CRC NSCLC shortness of breath data.

2.3.4 Associations between dimensions

The examination of associations between dimensions for binary and ordinal repeated measurement data is just as important as for the continuous case and it is again important to distinguish between correlations between and within subjects. Methods for the analysis are, however, more difficult and have not been discussed in the literature. As with the continuous case, it is clear that to estimate associations across dimension between subjects, a subject level summary of the data is needed, with associations between these summaries then examined. For ordinal data with sufficient ordered categories, the simplest approach for a crude representation

is to treat the responses as continuous and continue as in Section 2.2.4. For binary data, a solution is less clear. An attempt is made in figure 2.17 using the dichotomised responses - no symptoms (0/1) versus some symptoms (2/3) - for six items on the RSCL in the CRC NSCLC study. For each subject, for each dimension, the proportion of positive responses was calculated and these subject specific proportions are then plotted in a scatter plot matrix.

For the within subject correlation (figure 2.18) the subject specific proportions are treated as an average for that subject, and the deviations from this average for subject's individual responses were calculated and displayed in the same way as figure 2.11. Neither of these figures show a clear indication of any association between scores in the different dimensions either between or within subjects. In chapter 4, a formal statistical model is used to investigate this further and determine whether there really was little association between the different

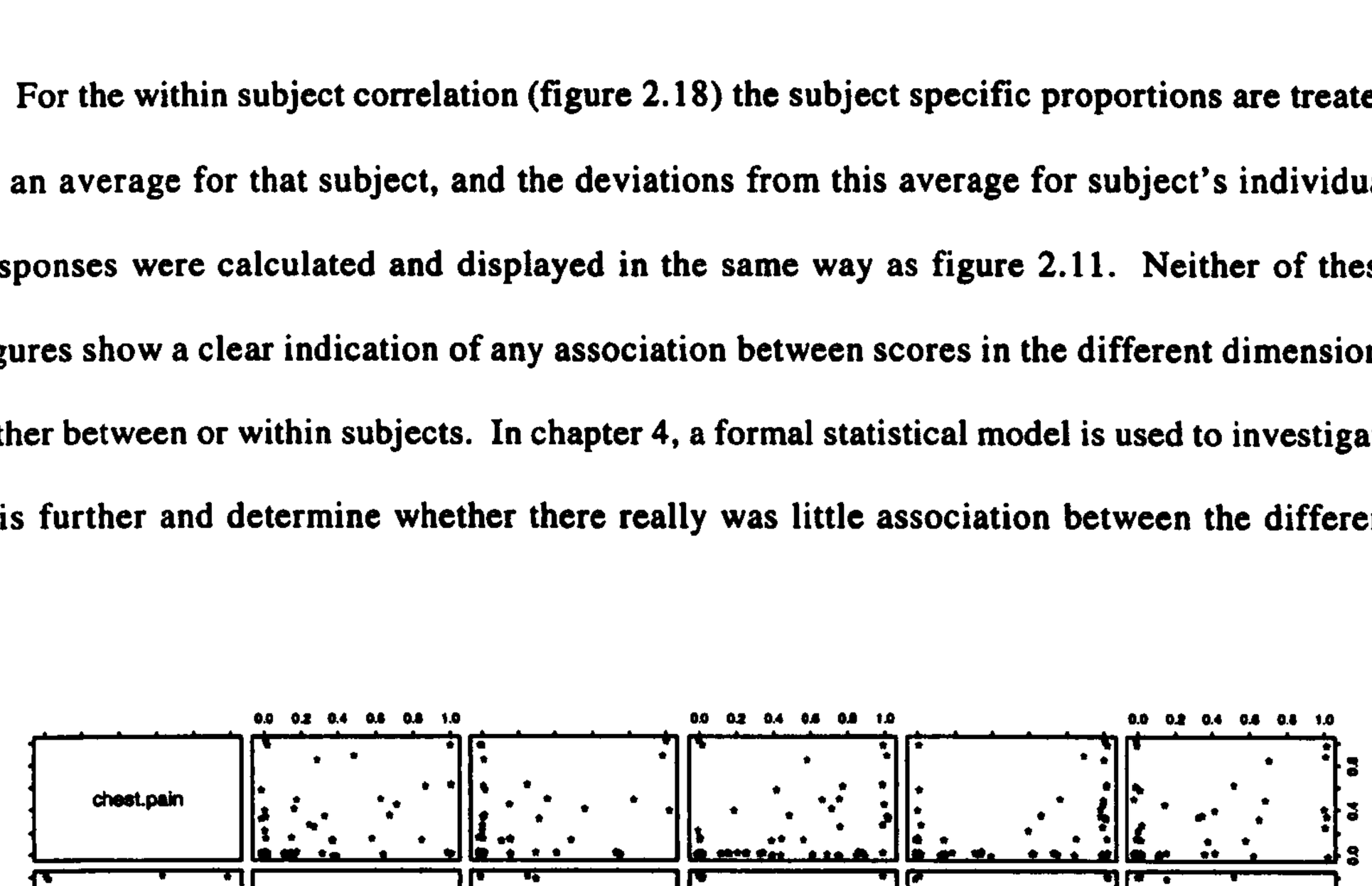

Figure 2.17: Scatter plot matrix of the proportion of days recorded with symptoms in six items of the RSCL for subjects on the CRC NSCLC study showing the between subject between dimension associations.

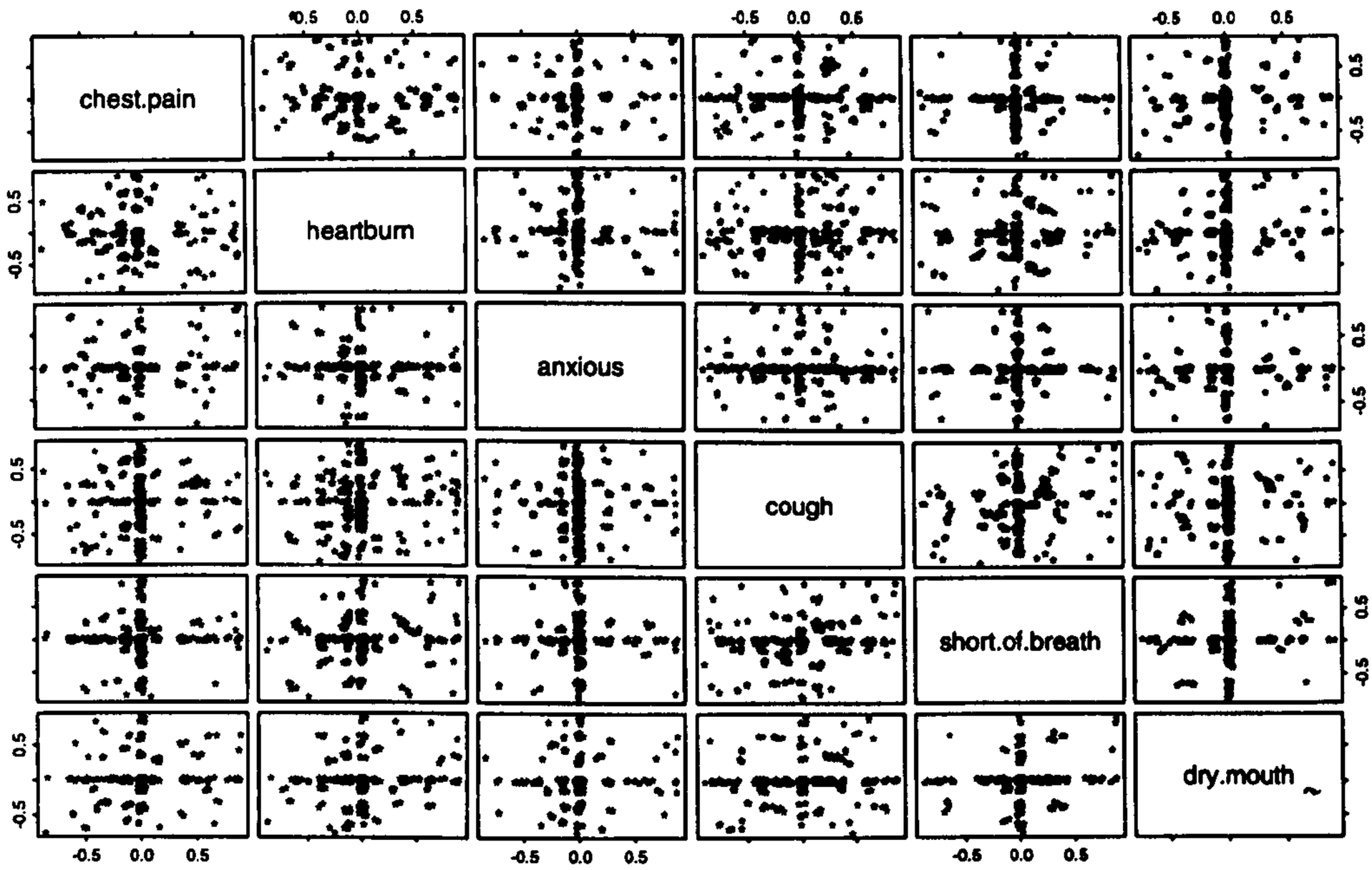

Figure 2.18: Scatter plot matrix showing the within subject between dimension associations for a binary response for individual item response from the CRC NSCLC study.

dimensions, or whether such simple representations are particularly unclear.

2.4 Missing data and patient death

A major problem for any analysis of quality of life data is incomplete data. This may be due simply to patients failing to return a questionnaire at a specific time, or as a result of patient death during follow-up. Whatever the source, incomplete data causes concern as to whether the available data are representative of the study population as a whole. For example, if subjects miss assessments because they are generally unwell and have a poor quality of life, any analysis based solely on the available data will be subject to some bias. An examination of the patterns of missing data in the study is therefore needed to assess the possibility of such a bias. A similar problem is patient death during follow-up and it is important to determine whether the quality of life experience of those who die early in follow-up is the same as those who have

a longer survival. Although alike in nature, the two issues raise different problems for analysis and need to be examined separately. In this section informal examination of the patterns of non-response is addressed. More formal analyses are discussed in Chapter 6 for patient death and in Chapter 7 for missing data.

2.4.1 Missing data

Missing data have been reported as a particular problem in quality of life studies. If the data available are to be assumed representative of the study population there should be no evidence that a group of non-responders are systematically different from those who do respond. In particular, if comparisons of different groups of patients in the study are to be made - such as, treatment comparisons - the extent of missing data should be reasonably equal within these groups over time, as well as in its relationship to patient quality of life.

The obvious starting point for examining missing data is to tabulate compliance rates over time and by patient, that is, the proportion of responses available at each time point, and the proportion of total responses given by each patient (Fayers and Jones, 1983). Such summaries are presented for the CRC NSCLC study in tables 2.1 and 2.2. Table 2.1 shows the proportion of available responses at baseline and for the eight weeks following treatment, whereas in table 2.2, a summary of the number of questionnaires completed by each patient is given. Both tables

Table 2.1:Proportion of available data at each weekly assessment for the CRC NSCLC study.

	Week								
	baseline	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Continuous course	0.48	0.55	0.57	0.57	0.50	0.52	0.57	0.57	0.50
Split course	0.60	0.45	0.53	0.55	0.55	0.50	0.53	0.55	0.53

Table 2.2: Proportion of the total possible responses (maximum=9) given by each patient in CRC NSCLC study.

	Proportion of questionnaires returned			
	0	**0.11 - 0.44**	**0.56 - 0.89**	**1**
Split course	10	6	18	6
Continuous course	8	9	20	5

show the compliance rates overall and as stratified by treatment group and show no difference between the two. Since the focus of investigation in this section is missing data due to non-compliance and not due to death, all proportions are given out of the possible total given patients' survival.

To determine whether compliance is related to quality of life or underlying patient condition, the MRC lung cancer working party (1993b) considered compliance rates (in terms of the proportion of completed responses per subject) and their relationship with baseline patient characteristics. They gave their results in tabular form. Here similar results are shown in figure 2.19 as scatter plots of baseline physical and psychological scores as well as baseline Karnofsky and FEV1 against the number of available questionnaires for each subject. The data used here are from the RSCL in the CRC NSCLC study. To avoid points being superimposed, a small amount of random noise has been added to the data points. For this example, no striking relationships between the number of responses given by individual patients and their baseline response are shown.

A further suggestion from Hopwood *et al.* (1994), that is not shown here, is a comparison of mean quality of life profiles over time for groups of patients giving different numbers of quality of life responses. To be effective, this relies on quality of life being measured over a fairly long period and more importantly a reasonably sized patient sample. It would therefore

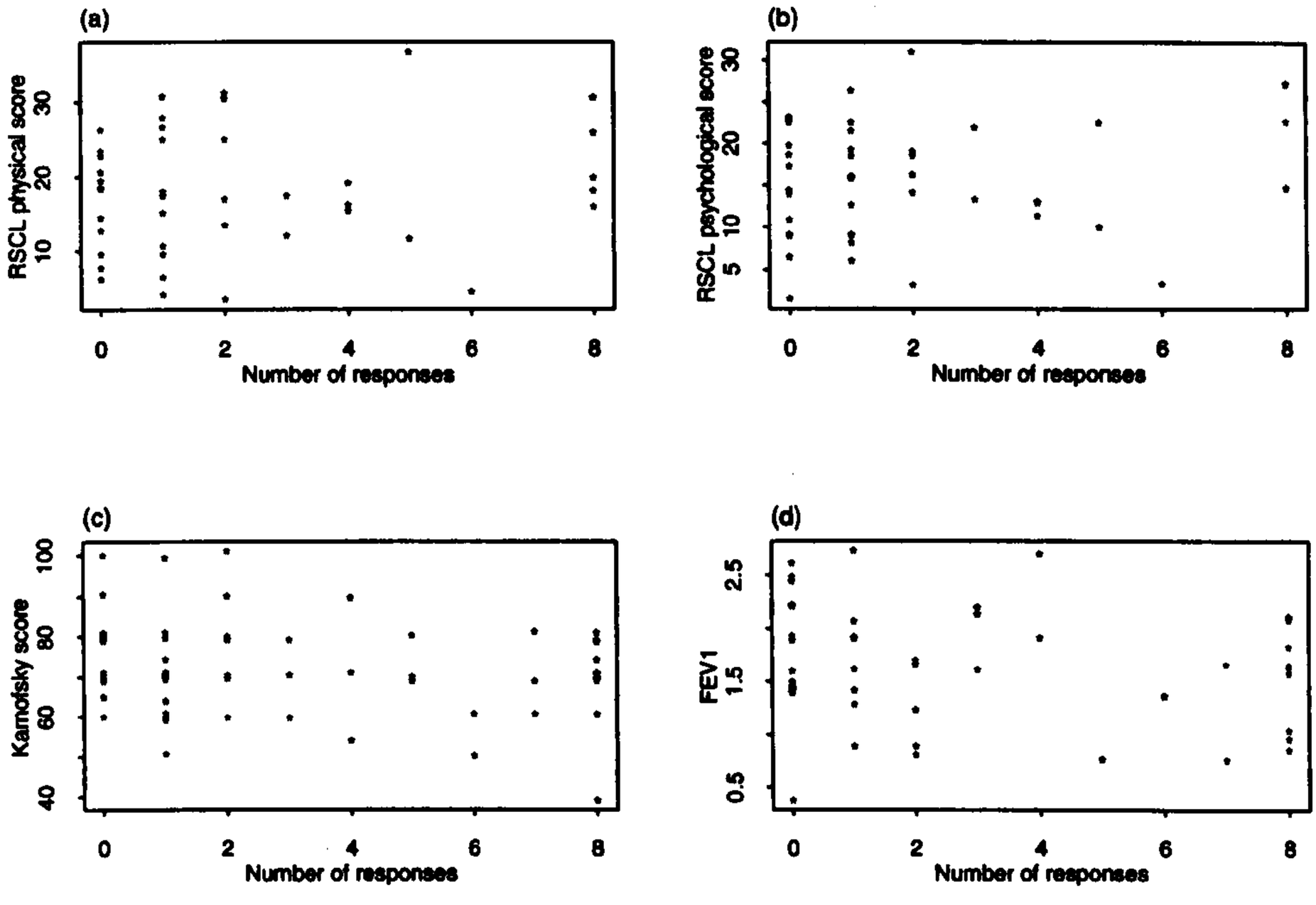

Figure 2.19: Scatter plots of baseline patient data against the number of available quality of life responses to the RSCL questionnaire in the CRC NSCLC study: (a) RSCL physical score; (b) RSCL psychological score; (c) Karnofsky score; (d) FEV1.

be more suitably applied to the MRC LU07 data than to the CRC NSCLC study data.

An added complication with the examination of patterns of missing data is the irregular timing of patient follow-up since this makes it difficult to determine whether data are missing or simply measured at an earlier or later occasion.

2.4.2 Patient death during quality of life assessment

Although both missing data and patient death during follow-up can both be classed very generally as issues of non-compliance, they are very different. The problem of missing data is whether the data available are representative of the quality of life experience of the sample as a whole. On the other hand, patient death raises the question of whether there is agreement in

the overall quality of life behaviour of patients with different lengths of survival. As full examination of this is not possible unless all patients have been followed until death, in the next section quality of life differences in patients for whom survival is known is investigated. More formal methods of analysis for which the full data set, including censored individuals, can be used are considered in Chapter 6.

Two different approaches are described here. The first is based on the idea of Hopwood *et al.* (1994) for examining patterns of missing data. Mean quality of life profiles for patients grouped by their observed survival time are estimated and plotted over time. An example is given in figure 2.20 for the HAP trial RSCL physical scores. In this example, the mean profiles were estimated by data grouped to the closest month. Kernel smoothers could also have been

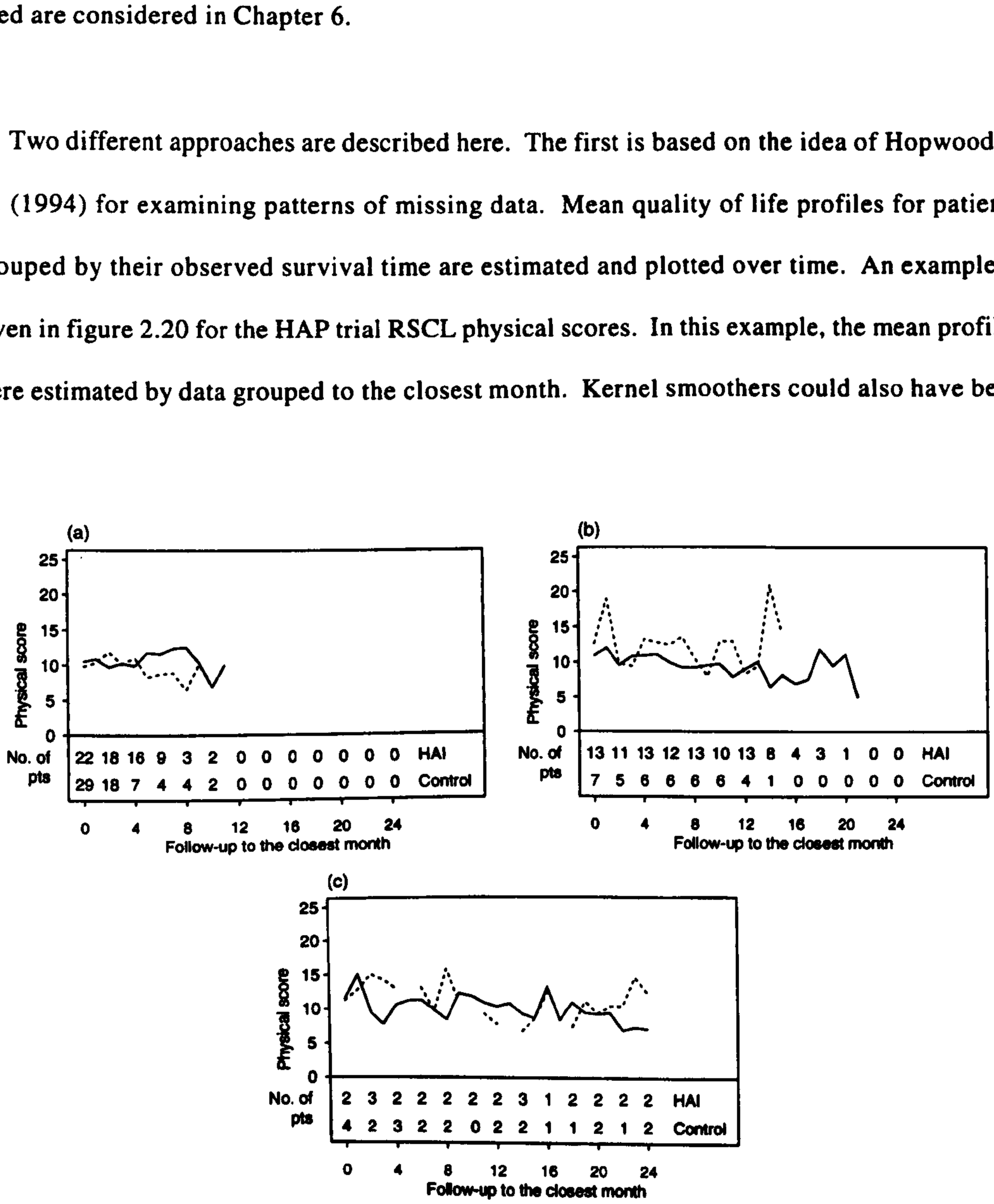

Figure 2.20: Mean RSCL physical scores over time from baseline (0) for a period of two years for the CRC HAP trial grouped by survival time: (a) < 1 year; (b) between 1 and two years; (c) > 2 years. Follow-up in each case in grouped to the nearest month. HAI: ————; control: ----------.

used. Very similar patterns of response over time for patients treated with a HAI in all three survival groups are seen. This is similarly so for patients in the control group. Unfortunately, unless samples sizes and numbers of observed deaths are large, these figures are of limited use. This is shown figure 2.20(c) in which the mean profile shown is estimated using a maximum of four subjects.

As an alternative, Morris *et al.* (1986) plotted mean quality of life over time measured backward from death to examine whether the behaviour of quality of life differed across different survival times and in particular if there was a noticeable change in quality of life towards the time of death. Such an analysis is shown in figure 2.21, with quality of life responses now grouped to the nearest month *prior* to death overall. Figure 2.21(a) shows this

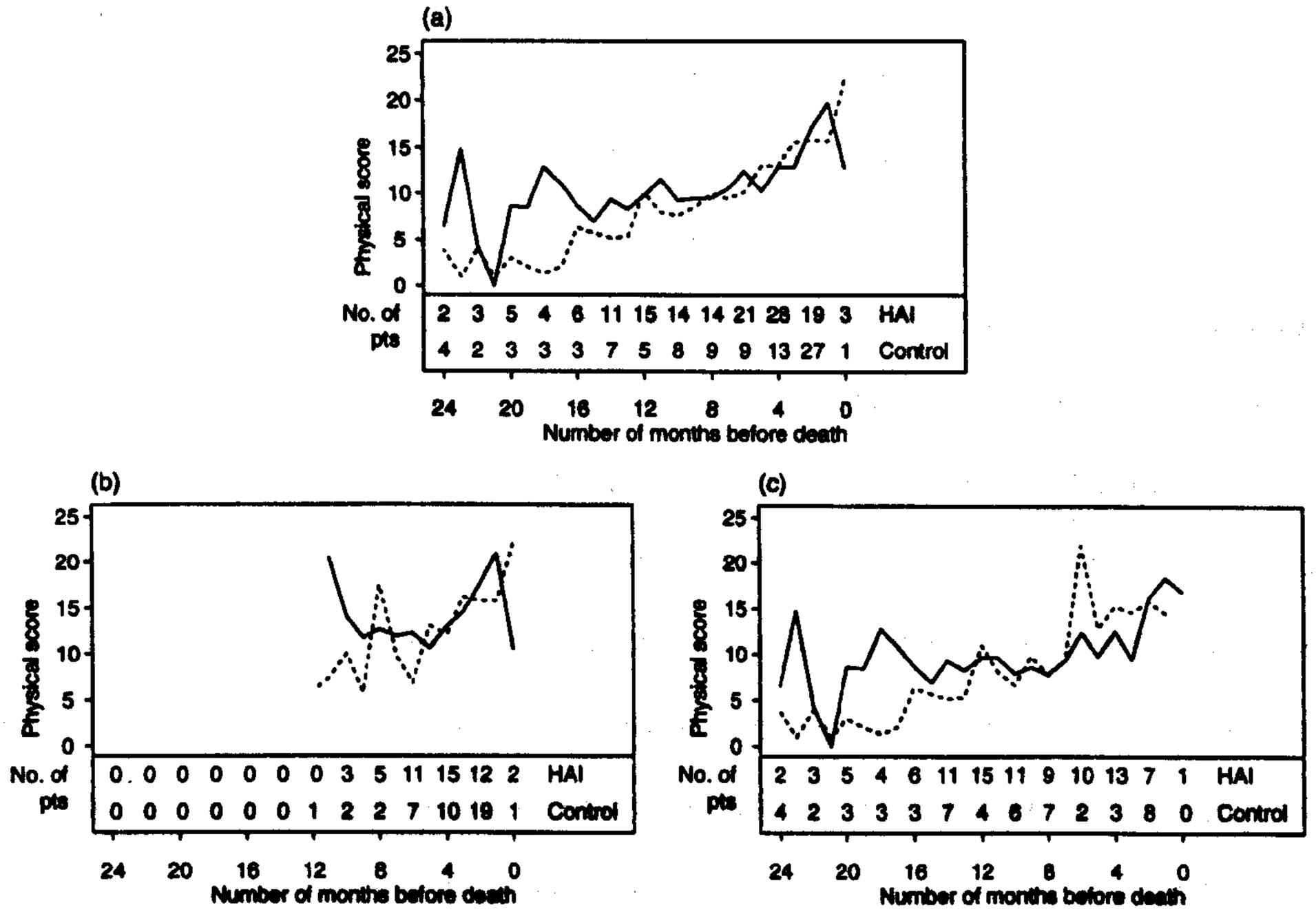

Figure 2.21: Mean RSCL quality of life over time from death going backwards for a maximum of two years for the restricted data of the CRC HAP trial: (a) Overall; and grouped by survival (b) ≤ 1 year; (c) > 1 year. HAI: ————; control: --------.

for the whole sample, whereas figures 2.21(b) and 2.21(c) show the data grouped by observed survival. Again, subjects for whom time of death is not known are not included. For the CRC HAP study, this figure shows an increase in the average level of quality of life towards death which was not apparent when the data were plotted from the time of entry into the study. Given this evident change in the quality of life experience, for reliable inferences from these data, it is clear that patient survival should be incorporated into the formal data analysis in some way. It is important to note however, that grouping the data in this way (from time of death) means that it is not possible to make unbiased treatment comparisons as we are no longer guaranteed to have patient groups which can be considered the same at the start of time (in this example 24 months prior to death).

2.5 Summary and discussion

This chapter has reviewed some methods of descriptive analysis that have been used for the analysis of quality of life data in the literature, and other relevant exploratory data analyses that may be used for future analyses. The different methods can be classified into four main areas: individual profiles; summary statistics; population averages over time; and associations between dimensions. In turn these areas allow the examination of typical patient responses over time; underlying behaviour of patients and how this varies across patients; an overall population summary of response; and how responses in the different dimensions are related to each other. To fully understand the behaviour of repeated measurement quality of life data collected in many dimensions, an exploratory data analysis should include at least the first three of these. Further, to determine their implications to be assessed not only in terms of the interpretation of descriptive displays, but also for more formal analyses which may be performed at a later stage, it is vital that a comprehensive analysis of the missing data in a study is also carried out.

For analyses based on individual profiles, a selection of random profiles is perhaps the best way of indicating typical responses over time when quality of life is measured on a continuous scale. For binary and ordinal data, lexis diagrams with different response categories highlighted by different line styles were shown to be very informative and allowed much more data to be displayed. These may also be used for continuous responses categorised in terms of normal responses and can also serve as a useful tool for examining patterns of missing data or the nature of quality of life response prior to death. Diggle *et al.* (1994) have also suggested that, rather than showing random profiles, individuals are ordered in terms of some analysis factor of interest in the data. Profiles are then shown for selected quantiles of this ordering statistic. Again, such a presentation may be of particular use in studies where patient death during follow-up was an issue with profiles shown for different quantiles of survival.

Summary statistics have been recommended in the literature for formal statistical analysis of longitudinal data in general (Matthews *et al.*, 1990). Not only do these recognise the longitudinal data structure, but also are intuitively appealing and require only that a summary measure of individual patient data of scientific interest can be defined. Unfortunately, their use for formal statistical analysis is somewhat complicated if data are unbalanced. However, for descriptive purposes, they are such a flexible and simple tool that they should form an integral part of exploratory data analysis for quality of life data.

The most commonly used method of exploratory analysis of quality of life data in the literature is a plot of means over time. The examples in the literature have tended to have had observations spaced equally for all subjects, making calculation of means at each time trivial. If this is not the case, population summaries can still be presented either by explicitly grouping observations by time of measurement in an appropriate manner, or by the use of kernel smoothers. The main problem of displaying mean profiles over time, is that they are

susceptible to over interpretation of apparent differences. To avoid this, it is recommended that some attempt is made to show the degree of confidence on a figure. This may be by confidence limits, the number of patients contributing data at each time point, or by overlaying mean profiles on the raw data.

A final part of the exploratory data analysis for quality of life data which was discussed in this work, is the examination of associations across dimensions. With many dimensions of quality of life invariably measured in a study, a sensible scientific question relates to how these dimensions relate to one another. As the data come from repeated assessments of subjects, simple scatter plots across dimensions are not appropriate. Instead, the variation in the data has to be partitioned into that between and that within subjects. Associations are then investigated within each partition. This is easily done for continuous data and was demonstrated in Section 2.4. Binary and ordinal data pose more of a problem however, and further work is needed in this area to determine simple descriptive tools to display cross dimension associations for such types of response.

To summarise, there exist many different approaches to exploratory data analysis within the four general areas of exploratory data analysis discussed in this chapter, as well as for the examination of patterns of missing data. Although most of the work in this thesis concentrates on formal statistical models for data analysis, the quantity of data which is generated in a quality of life study requires that comprehensive exploratory data analysis is carried out before undergoing more formal analysis. In order to fully understand the data, at least, this should include examination of individual profiles, summary statistics as well as population averages over time. Finally, patterns of missing data and the relationship between underlying quality of life and patient survival also need to be thoroughly examined.

3 Hierarchical Models for Repeated Continuous Outcomes

3.1 Introduction

Many approaches exist for the formal statistical analysis of repeated continuous outcomes. These range form the very simple to the more complex. Concise reviews of these methods are given by Crowder and Hand (1993), Everitt (1995) and Diggle *et al.* (1994). Unfortunately, many of the simpler analyses require that all subjects are measured at the same times in follow-up, and each have full sets of responses. In the light of missing data or irregularly spaced follow-up times, this can lead to the analysable data set being very much reduced. The more complex models, which on the whole are due to recent advances in statistical methodology have provided a number of techniques which do not have such restrictions, and can easily handle the analysis of unbalanced longitudinal data, (Zeger and Liang, 1992, Goldstein and MacDonald, 1988, Longford, 1995). This chapter demonstrates the use of hierarchical (or multilevel) models which is one approach for which software is available.

Multilevel models are random coefficient models (Longford, 1995, Goldstein, 1995) suitable for the analysis of data with some underlying hierarchical structure where it may reasonably be assumed that units within each level of hierarchy are randomly drawn from some underlying population. They have been used extensively in educational and social research to model child attainment data for children nested within classes nested within schools (Goldstein *et al.*, 1993). For example, a study which assesses the reading ability of children over many different schools in a particular district has two levels of hierarchy. The level one units are the students which are nested within the different schools. These schools form the level two units.

At level two it is assumed that the individual mean scores for each school are randomly distributed around the underlying mean score for the district as a whole. This gives level two, or between school, variation. Similarly, the scores of individual students are assumed to be randomly distributed around the underlying mean score for their school, constituting the level one, or within school variation.

Repeated measurements fall naturally into this framework. At level one, observations taken over time are nested within subjects. These subjects are drawn from some population to give the level two units. The variation at level two derives from the differences in individual responses between subjects. At level one, it is simply the deviation of individual observations from the subject's individual response. A possible model for such a scenario is given in equation (3.1), where y_{ij} denotes the response for subject i on occasion j made at time x_{ij}.

$$y_{ij}=\alpha+\beta x_{ij}+(u_i+v_i x_{ij}+e_{ij}) \quad \text{where} \quad \begin{matrix}(u_i,v_i)\sim\ \mathrm{N}(0,\Sigma_b)\\ e_{ij}\sim\ \mathrm{N}(0,\sigma_e^2)\end{matrix} \quad i=1,\ldots,n,\ j=1,\ldots,m_i \tag{3.1}$$

In this model, the underlying response over time for the population of interest is given by the parameters (α,β). The variation at level two derived from differences in individual response profiles are given by the subject level residuals on the population intercept and slope (u_i,v_i) for subject i. Combining these two components, gives a response profile for subject i,

$$\mathrm{E}(y_{ij}|x_{ij},u_i,v_i)=(\alpha+u_i)+(\beta+v_i)x_{ij} \tag{3.2}$$

In Section 2.2.2, similar subject specific regression lines were used to model the HAD anxiety data of the CRC NSCLC study. The differences between that model and the multilevel model of equation (3.1), are the distributional constraints on the subject specific components and the focus of estimation. In Section 2.2.2, the subject specific components were assumed fixed and were estimated explicitly and separately for each subject. In the multilevel (or random

coefficient model) they are assumed to be random variables from an underlying multivariate Normal distribution with variance Σ_b and, rather than the residual pairs (u_i, v_i) $(i=1,...,n)$, it is this variance that is of interest in estimation. As estimation of the residual variance, σ_e^2, is also required, the fixed analysis of Section 2.2.3 uses $(2n+1)$ parameters, whereas the random coefficient model uses only six. This makes the model much more flexible both in terms of modelling the fixed (or population) parameters, but also the components of variance. Their parameterisation and estimation procedures also make them very easy to extend to more than two levels of hierarchy.

In Section 3.2 the model, its assumptions and estimation are described in more detail. The subsequent work in the chapter then demonstrates the application of the model to the quality of life data from the CRC NSCLC study. In Section 3.3, the most simple models are used to analyse the HAD anxiety data. Section 3.4 then extends the basic two level model used so far to a three level model, and shows how hierarchical models can be used to analyse the multidimensional endpoints of continuous quality of life outcome data. In Section 3.5, these analyses are extended to demonstrate the modelling of the variance components in the data.

3.2 Hierarchical models for continuous outcomes

The following section describes the most general hierarchical statistical model, its assumptions and the estimation of its parameters. The notation for the general model with $h=1,...,H$ levels, is first described. Throughout, the description of the model is translated into the more familiar notation already introduced for the basic two level model for repeated measurements given in equation (3.1).

3.2.1 Notation

Y is the $(N \times 1)$ vector of responses which is ordered in its hierarchical structure such that, at the hth level, Y can be naturally partitioned into $n^{(h)}$ subvectors each consisting of the $n_r^{(h)}$ observations on the rth unit at the hth level where,

N is the total number of observations (or level one units);

$n^{(h)}$ is the number of units at the hth level;

and $n_r^{(h)}$ is the number of observations in the rth h-level unit.

For the two level repeated measurement structure, the highest level of hierarchy (H=2) is the subject level. Consequently, in the notation of equation (3.1), where the subjects (or level two units) are indexed by i, for i=r ,the rth level H (or level two) unit corresponds to the ith subject and $n^{(H)}$, the number of units at this level, is equivalent to the total number of subjects n. $Y=(y_1^T,\ldots y_n^T)^T$ where $y_i=(y_{i1},\ldots,y_{im_i})$, the vector of observations for subject i, at occasions $j=1,\ldots,m_i$. The total number observations measured for the rth H-level unit, $n_r^{(H)} \equiv n_i^{(H)} \equiv m_i$.

3.2.2 The model

In its most general form, the hierarchical (or multilevel) model is written

$$Y=X_0\beta_0+\sum_h X^{(h)}\beta^{(h)} \qquad (3.3)$$

where β_0 is a $(p_0 \times 1)$ vector of fixed parameters;

$\beta^{(h)}=(\beta_1^{(h)T},\ldots,\beta_{n^{(h)}}^{(h)T})^T$ is a $(n^{(h)}p^{(h)} \times 1)$ vector containing the $n^{(h)}$ subvectors of the $p^{(h)}$ random parameters or coefficients at level h where each of these subvectors have dimension $(p^{(h)} \times 1)$;

X_0 is the $(N \times p_0)$ design matrix for the p_0 fixed parameters;

and $X^{(h)}$ is the $(N \times n^{(h)}p^{(h)})$ design matrix for the $n^{(h)}p^{(h)}$ random parameters (or residuals) at the hth level. The matrix is block diagonal with the blocks corresponding to the hth level partition of Y.

In terms of the two level repeated measurement model of equation (3.1),

$\beta_0=(\alpha,\beta)^T$ is the population intercept and slope;

$\beta^{(2)}=\left((u_1,v_1),\ldots,(u_n,v_n)\right)^T$ is the $(2n \times 1)$ vector containing the n, (2×1) subvectors of the intercept and slope residuals for each subject;

and $\beta^{(1)}=\left(e_1^T,\ldots,e_n^T\right)^T$, is the $(N \times 1)$ vector of level one residuals where $e_i=(e_{i1},\ldots,e_{im_i})^T$ for subjects $i=1,\ldots,n$.

The design matrices X_0, $X^{(2)}$ and $X^{(1)}$ are given by,

$$X_0=\begin{pmatrix} 1 & x_{11} \\ . & . \\ . & . \\ . & . \\ 1 & x_{nm_n} \end{pmatrix}, \qquad X^{(2)}=\mathrm{diag}\begin{pmatrix} 1 & x_{i1} \\ . & . \\ . & . \\ . & . \\ 1 & x_{im_i} \end{pmatrix} \quad i=1,\ldots,n, \qquad X^{(1)}=I_N$$

They have dimension $(N \times 2)$, $(N \times 2n)$, and $(N \times N)$ respectively.

3.2.3 Model assumptions

In the general notation, at the hth level, it is assumed that the $\beta_r^{(h)}$ $(r=1,\ldots,n^{(h)})$ are independent, identically distributed, multivariate Normal random vectors with zero mean and variance $\Omega^{(h)}$. $\mathrm{Var}(\beta^{(h)})$ is then an $(N \times N)$ block diagonal matrix with $\Omega^{(h)}$ on the diagonal. This can be written, $\mathrm{var}(\beta^{(h)})=I_{n^{(h)}} \otimes \Omega^{(h)}$, where $\otimes$ denotes the Kronecker product, the matrix operation which multiples every element of the left hand matrix by the right hand matrix. I_n is the $(n \times n)$ identity matrix. The elements of $\Omega^{(h)}$ are referred to as the *variance components*. The random coefficients across levels are also assumed independent. That is, $\beta^{(h)} \perp \beta^{(h')}$ for

$h \neq h'$.

From equation (3.3) it follows that

$$V=\text{var}(Y)=\sum_{h=1}^{H} X^{(h)}\text{var}(\beta^{(h)})X^{(h)\mathrm{T}} \qquad \textbf{(3.4)}$$

implying that for each level h partition, V is block diagonal. Specifically, at the level H, this reflects the fact that observations taken on different top level units are assumed independent. At this level, these diagonal blocks correspond to

$$\text{var}(y_r)=\sum_{h=1}^{H} x_r^{(h)}\text{var}(\beta_r^{(h)})x_r^{(h)^{\mathrm{T}}} \text{ for the } r=1,\ldots,n^{(H)}\ H\text{-level units} \qquad \textbf{(3.5)}$$

Put in the context of the simple repeated measurement model of equation (3.1), $\Omega^{(1)}=\sigma_e^2$ and $\Omega^{(2)}=\Sigma_b=\begin{pmatrix}\sigma_u^2 & \sigma_{uv}\\ \sigma_{uv} & \sigma_v^2\end{pmatrix}$, where $\sigma_u^2=\text{var}(u_i)$, $\sigma_v^2=\text{var}(v_i)$ and $\sigma_{uv}=\text{cov}(u_i,v_i)$. Also $\text{var}(Y)=\text{diag}\{\text{var}(y_i)\}$ where $\text{var}(y_i)=$

$$\begin{pmatrix}
\sigma_u^2+2\sigma_{uv}x_{i1}+\sigma_v^2x_{i1}^2+\sigma_e^2 & \sigma_u^2+\sigma_{uv}(x_{i1}+x_{i2})+\sigma_v^2x_{i1}x_{i2} & \cdots & \sigma_u^2+\sigma_{uv}(x_{i1}+x_{im_i})+\sigma_v^2(x_{i1}x_{im_i})\\
\vdots & \sigma_u^2+2\sigma_{uv}x_{i2}+\sigma_v^2x_{i2}^2+\sigma_e^2 & & \vdots\\
\sigma_u^2+\sigma_{uv}(x_{im_i-1}+x_{im_i})+\sigma_v^2(x_{im_i-1}x_{im_i}) & \cdots & \ddots & \vdots\\
\sigma_u^2+\sigma_{uv}(x_{i1}+x_{im_i})+\sigma_v^2(x_{i1}x_{im_i}) & \cdots & \cdots & \sigma_u^2+2\sigma_{uv}x_{im_i}+\sigma_v^2x_{im_i}^2+\sigma_e^2
\end{pmatrix} \qquad \textbf{(3.6)}$$

For a specific case with 2 subjects, measured on $m_1=2$, $m_2=3$ occasions, $\text{var}(Y)=\begin{pmatrix}V_1 & 0\\ 0 & V_2\end{pmatrix}$

where $V_1=\begin{pmatrix}\sigma_u^2+2\sigma_{uv}x_{11}+\sigma_v^2x_{11}^2+\sigma_e^2 & \sigma_u^2+\sigma_{uv}(x_{11}+x_{12})+\sigma_v^2x_{11}x_{12}\\ \sigma_u^2+\sigma_{uv}(x_{11}+x_{12})+\sigma_v^2x_{11}x_{12} & \sigma_u^2+2\sigma_{uv}x_{12}+\sigma_v^2x_{12}^2+\sigma_e^2\end{pmatrix}$

and $$V_2=\begin{pmatrix} \sigma_u^2+2\sigma_{uv}x_{21}+\sigma_v^2x_{21}^2+\sigma_e^2 & \sigma_u^2+\sigma_{uv}(x_{21}+x_{22})+\sigma_v^2x_{21}x_{22} & \sigma_u^2+\sigma_{uv}(x_{21}+x_{23})+\sigma_v^2x_{21}x_{23} \\ \sigma_u^2+\sigma_{uv}(x_{21}+x_{22})+\sigma_v^2x_{21}x_{22} & \sigma_u^2+2\sigma_{uv}x_{22}+\sigma_v^2x_{22}^2+\sigma_e^2 & \sigma_u^2+\sigma_{uv}(x_{22}+x_{23})+\sigma_v^2x_{22}x_{23} \\ \sigma_u^2+\sigma_{uv}(x_{21}+x_{23})+\sigma_v^2x_{21}x_{23} & \sigma_u^2+\sigma_{uv}(x_{22}+x_{23})+\sigma_v^2x_{22}x_{23} & \sigma_u^2+2\sigma_{uv}x_{23}+\sigma_v^2x_{23}^2+\sigma_e^2 \end{pmatrix}$$

and the block diagonal nature of var(Y) reflects the fact that observations on different individuals are assumed independent.

3.2.4 Estimation of the model parameters

A number of estimation procedures exist for such models. That used here is based on generalised least squares (GLS). A review of the alternative algorithms is given by Goldstein (1995) and Longford (1995).

If $V=\text{var}(Y|X_0\beta_0)$ is known, the generalized least squares estimators for $\hat{\beta}_0$ and var($\hat{\beta}_0$) are

$$\hat{\beta}_0=(X_0^TV^{-1}X_0)^{-1}X_0^TV^{-1}Y \quad \text{var}(\hat{\beta}_0)=(X_0^TV^{-1}X_0)^{-1} \tag{3.7}$$

Similarly, since $V=E\left((Y-X_0\beta_0)(Y-X_0\beta_0)^T\right)$, if β^*, the parameters of V, are the focus of interest and β_0 is known, the generalised least squares estimate of β^* is

$$\hat{\beta}^*=(X^{*T}V^{*-1}X^*)^{-1}X^{*T}V^{*-1}Y^*, \quad V^*=V\otimes V \tag{3.8}$$

where $Y^*=(y_1^{*T},\ldots y_{n^{(H)}}^{*T})^T$ for y_r^{*T}, a vector of squares and products of residuals for the rth H-level partition, $r=1,\ldots,n^{(H)}$, which is given by the upper triangular elements of the rth diagonal block of $(Y-X_0\beta_0)(Y-X_0\beta_0)^T$. $V^*=\text{var}(Y^*)$ and X^* is the design matrix which links Y^* to V. After some algebra, Goldstein (1995) shows $\text{var}(\hat{\beta}^*)=2(X^{*T}V^{*T}X^*)^{-1}$.

In terms of the repeated measurement example of equation (3.1), $\beta^* = (\sigma_u^2, \sigma_{uv}, \sigma_v^2, \sigma_e^2)^T$. From equation (3.6), it follows that the design matrix $\boldsymbol{X}^*$ which links $\boldsymbol{Y}^*$ to $\boldsymbol{V}$ can be written

$$\boldsymbol{X}^* = \text{diag}\{\boldsymbol{x}_i^*\} \text{ where } \boldsymbol{x}_i^* = \begin{pmatrix} 1 & 2x_{i1} & x_{i1}^2 & 1 \\ 1 & x_{i1}+x_{i2} & x_{i1}x_{i2} & 0 \\ 1 & x_{i2}+x_{i3} & x_{i2}x_{i3} & 0 \\ & \vdots & & \vdots \\ 1 & x_{i1}+x_{im_i} & x_{i1}x_{im_i} & 0 \\ 1 & 2x_{i2} & x_{i2}^2 & 1 \\ & \vdots & & \vdots \\ 1 & x_{im_i-1}+x_{im_i} & x_{im_i-1}x_{im_i} & 0 \\ 1 & x_{im_i} & x_{im_i}^2 & 1 \end{pmatrix} \text{ and } \boldsymbol{Y}^* = \text{diag}\{\boldsymbol{y}_i^*\} \text{ where } \boldsymbol{y}_i^* = \begin{pmatrix} y_{i1}^2 \\ y_{i1}y_{i2} \\ y_{i1}y_{i3} \\ \vdots \\ y_{i1}y_{im_i} \\ y_{i2}^2 \\ \vdots \\ y_{im_i-1}y_{im_i} \\ y_{im_i}^2 \end{pmatrix}$$

Goldstein (1986) demonstrates that when neither $\boldsymbol{V}$ nor β_0 are known, equations (3.7) and (3.8) can be solved iteratively with a starting value for $\boldsymbol{V}$ given by its ordinary least squares estimate. He shows that the *iteratively generalised least squares* (IGLS) estimates obtained from this routine are asymptotically efficient. Under the further assumption of multivariate Normality of response, they are equivalent to maximum likelihood (ML) estimates. It is known, however, that ML estimates of the random parameters are biased. This is because the sampling variability in the fixed parameters, that are used in their estimation, is ignored (Patterson and Thompson, 1971). Unbiased estimates can be obtained using restricted maximum likelihood (REML). In a further paper (Goldstein, 1989a) the IGLS procedure is modified to give *restricted* IGLS (RIGLS) estimates which are equivalent REML estimates under the assumptions of multivariate Normality of response. The distinction between the results of the two procedures is most important for smaller sample sizes. Full details of the both these estimation processes are given by Goldstein (1995); computational details are given by Goldstein and Rasbash (1992).

Finally, the log likelihood of the estimated model is obtained by evaluating the model log likelihood (log lh) for the parameter estimates of the fixed and variance components. Nested models can then be compared with a comparison of deviances given by -2 log lh.

Although standard errors can be obtained for both the fixed and variance components, it is not recommended that those on the variance components be used for inference. This is because they are based on asymptotic Normal properties which are unlikely to hold except in particularly large samples. Instead it is recommended that inference about the variance components in a model is based on a comparison of model deviances. In the examples presented in subsequent work, although standard errors will be given for estimated variance components, they are presented only as a guide.

3.2.5 Estimation of residuals

Although the random components are not of primary interest in this analysis, it is still useful to be able to obtain estimates for them. This may be for diagnostic purposes, or to show estimated individual profiles as typical response profiles. Although they are not estimated explicitly in model estimation, using the parameter estimates of the multilevel model, it is possible to obtain *conditional* (or *shrunken*) estimates for them, where the conditioning is upon the estimated fixed and variance components (Goldstein, 1986, 1995). The implication of this is that the residuals of extreme (or outlying observations) will tend to be lower than expected. In the context of the two level repeated measurement example, this will result in fitted values which are closer to the population mean than would be obtained from the subject specific analysis presented in Section 2.2.3. Estimation uses simple linear regression.

Re-arranging equation (3.3) gives the overall residual component

$$\sum_{h=1}^{H} X^{(h)}\beta^{(h)} = Y - X_0\hat{\beta}_0 \qquad (3.10)$$

where the elements of $\beta^{(h)}$, the residuals at level h, are the parameters of interest. Since the residuals at each level are independent, this is a linear combination of the total residuals at each level. For each level, h=1,...,H, predicted values for $\beta^{(h)}$ can therefore be obtained from a linear regression of $\beta^{(h)}$ on the overall residual component $(Y - X_0\hat{\beta}_0)$,

$$\hat{\beta}^{(h)} = W^{(h)^T} V^{-1}(Y - X_0\hat{\beta}_0) \qquad (3.11)$$

where $W^{(h)} = \text{cov}(Y - X_0\hat{\beta}_0, \beta^{(h)}) = X^{(h)}(I_{n^{(h)}} \otimes \Omega^{(h)})$, and $V = \text{var}(Y)$. It is because these estimated regression coefficients, given by $W^{(h)^T} V^{-1}$, are based on the estimated variance components from the model, that the predicted $\hat{\beta}^{(h)}$ are conditional and have smaller variance than those that would be estimated unconditionally from a regression model based (hypothetically) on observed $\left\{\beta^{(h)}, (Y - X_0\hat{\beta}_0)\right\}$.

3.3 Application of the two level repeated measurement model

Within this section, the basic two level model for the analysis of repeated continuous outcomes is demonstrated with an analysis of the CRC NSCLC study, HAD anxiety data. The basic objective of the analysis was to determine the strength of evidence for a change in the level of anxiety over time and also to determine whether this response was different in the two treatment arms. The models have been fitted using MLn software (Rasbash and Woodhouse, 1995) and RIGLS.

The analysis is presented in three parts. Section 3.3.1 focuses on modelling the change in response over time. This analysis uses responses on the 57 patients in the study who gave at

least one HAD anxiety score during the eight week period following radiotherapy. In Section 3.3.2, this model is extended to consider differences between the two treatment groups. For this analysis, an adjustment for baseline anxiety is made. As baseline data was not available on some of the patients used in the initial analysis, this second analysis uses a smaller data set containing only 37 patients who were divided 15:22 between the continuous and the split course radiotherapy groups. Section 3.3.3 examines the assumptions of the models.

All parameter estimates from these models are presented with standard errors (SE). As these are based on asymptotic Normality assumptions, for the variance components of the models they are considered only as a guide and not used as the basis of significance testing of the individual estimates.

3.3.1 Modelling a change over time

With anx_{ij} representing the jth anxiety response of the ith subject and occ_{ij} the timing of response measured in weeks following the start of treatment, a model for a change in the level of anxiety over time was formulated in three stages. The three models used are described below.

Null model

An overall mean response α was assumed constant over all occasions and treatment groups.

$$anx_{ij} = \alpha + u_i + e_{ij} \qquad (3.12)$$

Model one - occasion as a fixed effect

A linear trend in response over time with the measurement occasion fit as a continuous covariate in the interval [0,7], where occasion 0 denotes 1 week following the start of treatment,

$$anx_{ij}=\alpha+\beta occ_{ij}+u_i+e_{ij} \quad (3.13)$$

Here the parameter α gives an estimate of the intercept (or mean response at 1 week after the start of treatment) for the population. The intercept for patient i is then given by $(\alpha+u_i)$. The slope, β, gives an estimate of the rate of change in response over the period which is assumed constant across subjects.

Model two - occasion as a random effect

In model two, the occasion effect was allowed to vary over individuals, such that the rate of change in score over the follow-up period is different for each patient. This corresponds to the basic model of equation (3.1).

$$anx_{ij}=\alpha+\beta occ_{ij}+u_i+v_i occ_{ij}+e_{ij} \quad (3.14)$$

Again the parameter α gives an estimate of the intercept (or mean response at one week) for the population. The intercept for patient i is then given by $(\alpha+u_i)$. With the slope now allowed to vary over patients. The rate of change in response for subject i is $(\beta+v_i)$, with β estimating the average rate for the population in general.

As described in Section 3.2, it is the variance of the subject specific effects, u_i and v_i, in each of these models which are estimated, rather than the effects themselves. However, conditional estimates of these effects can be obtained as described in Section 3.2.5. A comparison of the goodness of fit of these models was made using a comparison of scaled deviances, where the scaled deviance was given as the difference in the -2 log lh of the two models.

The results of sequentially fitting the above models are given in table 3.1. The fixed parameter estimates gave some evidence of a fall in the level of anxiety over the period with a reduction in the deviance from the null model to model one of 15.6 on 1 df. The estimated rate of the fall was 0.17 units per week (95% CI=[0.09, 0.25]). By allowing a random slope

Table 3.1: Parameters estimates (SE) for modelling the rate of change over time for the CRC NSCLC HAD anxiety scores.

		Parameter estimates (SE)		
Model		**Null**	**One**	**Two**
Fixed parameters				
α	(*cons*)	4.98 (0.57)	5.54 (0.59)	5.63 (0.59)
β	(*occ*)		-0.17 (0.04)	-0.19 (0.06)
Random parameters				
Level two	σ_u^2	17.90 (3.47)	17.80 (3.44)	18.35 (3.72)
	σ_v^2	-	-	0.11 (0.04)
	σ_{uv}	-	-	-0.28 (0.28)
Level one	σ_e^2	3.03 (0.25)	2.88 (0.24)	2.32 (0.21)
-2 log lh		1582.0	1566.4	1547.8

These results correspond to a set of 57 patients who gave at least one HAD anxiety response following the start of treatment. the time of measurements is measured in weeks from 1 to 8 coded *occ*=0,...,7.

across subjects, a slight increase in the estimated average rate of the fall in anxiety was seen (-0.19, 95% CI=[0.07, 0.31]). The standard error of this estimate was slightly increased compared to that of the previous model. This was expected given the increased variability across subjects which has been allowed by the introduction of the random slope. In conjunction with the larger rate of fall in the response level over the period, was an increase in the intercept parameter.

The estimates of the random components indicated a large degree of variation across subjects. Within all three models, the estimated variance of the intercept, σ_u^2, was very high relative to the fixed estimate. Consistent with the reduction in precision of the fixed parameter estimates, it increased further on inclusion of a random slope. The negative covariance between intercept and slope, given by σ_{uv}, suggested that subjects who witness a more marked reduction in their level of response tended to have relatively higher initial responses than those

with a more modest fall in response. This could be an artefact of the bounded nature of the response - subjects with relatively low initial scores have less scope for change in their subsequent responses over time. There was strong evidence in favour of slope variation across subjects shown by a reduction in deviance of 18.6 on 2 df. On this basis, with model two as the preferred model, subject specific predicted response profiles were estimated using the conditional estimates of the subject specific residuals, (u_i, v_i), $i=1,...,n$. These are shown in figure 3.1. Contrasted against those of figure 2.4, they show less variation in overall response. In particular, the outlying profiles highlighted in Section 2.2.2 did not stand out in figure 3.1. This was because, in the earlier section, the estimated profiles for these subjects were based on very few observations, whereas, in figure 3.1 their prediction was based also on the distribution of response for all subjects. In the multilevel model, fitted profiles for these subjects were shrunk towards the population average.

Given the variance estimates of table 3.1, expected ranges for both the intercept (initial level of response after treatment) and the slope (rate of change in response) can be calculated. Under the assumption that the model is correct, these ranges will give boundaries in which about 95% of subject specific parameters for the population of interest are expected to lie. They will be similar to the reference ranges given on figure 2.4, but since they are based on the full variance structure they will be slightly narrower. Calculated as $\hat{\beta} \pm 1.96\,\sigma_v$, the expected range for subject specific slopes was [-0.84, 0.46] around a mean of -0.19. That is, given that the model was true, the experience of individual subjects in terms of a rate of change in anxiety score in the immediate week following radiotherapy, may be expected to range between a fall of nearly one unit per week to an increase of just under half a unit. This is in contrast to an overall reference range for subject specific slopes of [-1.51, 1.07] for the profiles of figure 2.4 around a mean of -0.22.

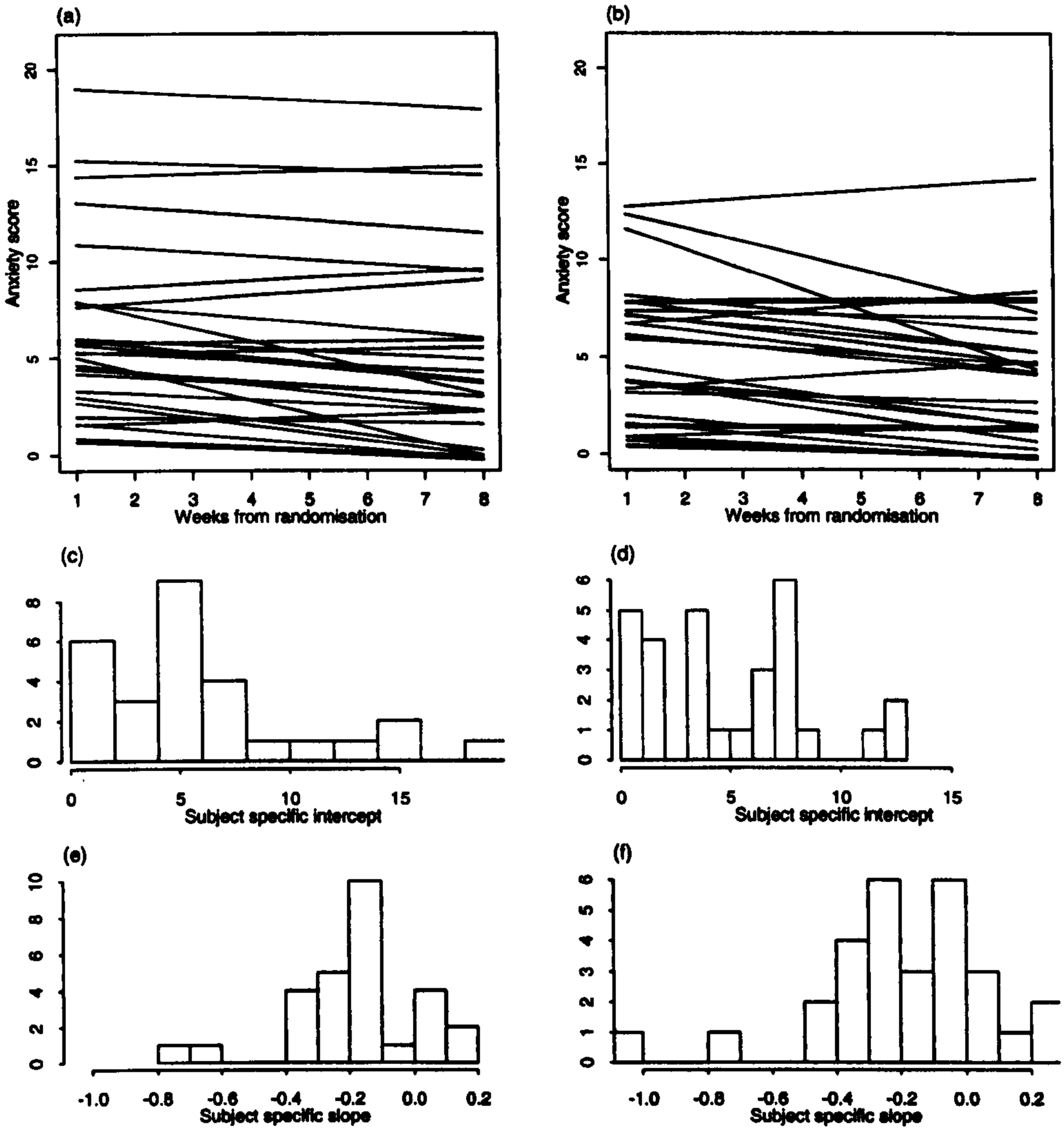

Figure 3.1: Plots of the predicted subject specific profiles from Model two (a) for the split course radiotherapy group; (b) for the continuous course radiotherapy group. The distributions of residuals for the split course and continuous course respectively are given in (c) and (d) for the intercept residuals; and (e) and (f) for the slope residuals.

From each of these models, the estimated within subject residual variance at level one was much smaller than that between subjects. The particular aspect to notice about the estimates of the subsequent models is the fall in residual variance on inclusion of the random slope. This is to be expected in the light of allowing the estimated subject specific responses more freedom to adhere closer to the observed responses for each subject.

3.3.2 Modelling the effects of treatment

The second objective of this analysis was to obtain a comparison of patient anxiety over time in the two treatment arms. This was done by extending the current model for a constant treatment group difference over time (parallel lines), a diverging difference over time (non-parallel lines with the same intercept), and a general difference over time (non-parallel lines with different intercepts). To adjust for any pre-treatment differences, patient baseline responses (subtracting the overall depression response mean of 7.89) denoted $base_i$, were included into the model. Treatment was modelled with rt_i=0 for the split radiotherapy course, 1 for the continuous course.

Model three - adjusting for baseline

$$anx_{ij}=\alpha+\beta occ_{ij}+\gamma base_i+u_i+v_i occ_{ij}+e_{ij} \quad (3.15)$$

Model four - constant treatment difference

Model four considered a constant treatment difference throughout the eight week follow-up.

$$anx_{ij}=\alpha+\beta occ_{ij}+\gamma base_i+\xi rt_i+u_i+v_i occ_{ij}+e_{ij} \quad (3.16)$$

Model five - unconstrained treatment difference

Model five extended model four by allowing a difference between the rate of change in response over the period between the two treatment groups. This was done by introducing an "occasion by treatment interaction", denoted $occ.rt_{ij}$.

$$anx_{ij}=\alpha+\beta occ_{ij}+\gamma base_i+\xi rt_i+\delta occ.rt_{ij}+u_i+v_i occ_{ij}+e_{ij} \quad (3.17)$$

Model six - diverging treatment difference

Within model six the intercepts within both treatment groups were forced to be equal thus giving a diverging treatment difference.

$$anx_{ij} = \alpha + \beta occ_{ij} + \gamma base_i + \delta occ.rt_{ij} + u_i + v_i occ_{ij} + e_{ij} \qquad (3.18)$$

A summary of the results of the analysis is given in table 3.2. Since the inclusion of baseline anxiety data caused a reduction in the size of the available data set from 57 to 37, the results of model two fitted to this smaller data set are also given in the table.

The first thing noted from this analysis was the strong effect of the centred baseline giving a reduction in the model deviance of 27.3 on 1 df. Because the variable was centred relative

Table 3.2: Parameter estimates (SE) for difference in HAD anxiety scores over time between continuous and split course radiotherapy in the CRC NSCLC study.

		Parameter estimates (SE)				
Model		**Two**	**Three**	**Four**	**Five**	**Six**
Fixed parameters						
α (*cons*)		6.24 (0.73)	6.42 (0.49)	6.26 (0.62)	6.18 (0.63)	6.43 (0.49)
β (*occ*)		-0.24 (0.08)	-0.25 (0.08)	-0.25 (0.08)	-0.18 (0.11)	-0.19 (0.10)
γ (*base*)		-	0.72 (0.11)	0.72 (0.11)	0.72 (0.11)	0.72 (0.11)
ξ (*rt*)		-	-	0.42 (0.97)	0.62 (0.99)	-
δ (*occ.rt*)		-	-	-	-0.17 (0.17)	-0.15 (0.16)
Random parameters						
Level two	σ_u^2	18.07 (4.58)	7.07 (2.16)	7.87 (2.20)	7.82 (2.19)	7.68 (2.15)
	σ_v^2	0.16 (0.06)	0.16 (0.06)	0.16 (0.06)	0.16 (0.06)	0.17 (0.06)
	σ_{uv}	-0.30 (0.38)	-0.01 (0.26)	0.01 (0.26)	0.02 (0.26)	0.02 (0.26)
Level one	σ_e^2	2.40 (0.27)	2.39 (0.27)	2.39 (0.27)	2.38 (0.27)	2.39 (0.27)
-2 log lh		1033.1	1005.8	1005.6	1004.6	1005.0

These results correspond to a set of 37 patients who gave at least one HAD anxiety response following the start of treatment and a response at baseline.

to the overall mean, little change in the intercept parameter - interpretable as the estimated underlying week one response for a subject with mean depression response - was seen. The inclusion of the baseline responses in the model helped to explain a substantial amount of the variation between subjects. This was shown by the reduction in the standard error on α, the population intercept, and in particular the reduction in the estimated variance of subject specific random intercepts, σ_u^2.

Model four gave no evidence to suggest a constant difference in the underlying level of response between the two treatment groups throughout the period. Subjects in the split course radiotherapy arm had on average a slightly lower level of response than those in the continuous arm with an estimate of the difference of 0.42 and a 95% CI=[-1.48,2.32]. Similarly, there was no evidence that the rate of change over time may have differed across the groups. The estimated difference for this rate of change of -0.17 (95% CI=[-0.50,0.16]) indicated that the level of anxiety of subjects in the continuous group tended on average to fall more than for subjects on the split course of radiotherapy although this was consistent with a difference which could have occurred by chance.

Apart from the expected fall in the level two variance parameter u_i on inclusion of the baseline level of response, there was very little change to the random parameters of the model throughout.

3.3.3 Testing the model assumptions

The assumptions of the hierarchical model were described in Section 3.2.3. These are independent identically multivariate Normality of the random effect at each level, with these effects independent across levels.

As a test for k-variate Normality, it is suggested (Johnson, 1988) that it is generally sufficient to check univariate Normality of the k variates, as well as bivariate Normality of the $\binom{k}{2}$ variate pairs, although a formal test for multivariate Normality does exist.

For multivariate k-dimensional Normal random vectors x_i, $i=1,\ldots,n$, with mean $\bar{x}$, and variance Σ, the squared (Mahalanobis) distances,

$$d_i^2=(x_i-\bar{x})^T\Sigma^{-1}(x_i-\bar{x}) \qquad (3.19)$$

will follow a Chi-squared distribution on k degrees of freedom. In terms of the random components of the hierarchical model, this implies that, under the assumed multivariate Normal distribution for the $r=1,\ldots,n^{(h)}$ units at level h, the squared distances

$$d_r^2=\beta_r^{(h)T}\Omega^{(h)-1}\beta_r^{(h)} \qquad (3.20)$$

should be approximately Chi-squared random variables with $p^{(h)}$ degrees of freedom, although, the exact distribution of these quantities is unknown. Since the residuals have been estimated on the basis of the model and the covariance matrix, it is expected that in reality, it will be less dispersed than the $\chi^2_{p^{(h)}}$. For a particular model, this can be assessed by using a *Chi-squared* or *Gamma* plot, in which the squared distances, calculated using the shrunken residuals and estimated covariance structure of the model, are ordered from smallest to largest, and plotted against the appropriate centiles of the Chi-squared distribution with $p^{(h)}$ degrees of freedom,

This is shown in figure 3.2 for the level two (subject) random effects of model six. Also shown in the figure are the univariate Normal plots and a bivariate plot of the intercept and slope residuals. Although some skewness to the tails of the distributions are seen, each these plots show reasonable accordance to the Normality assumptions.

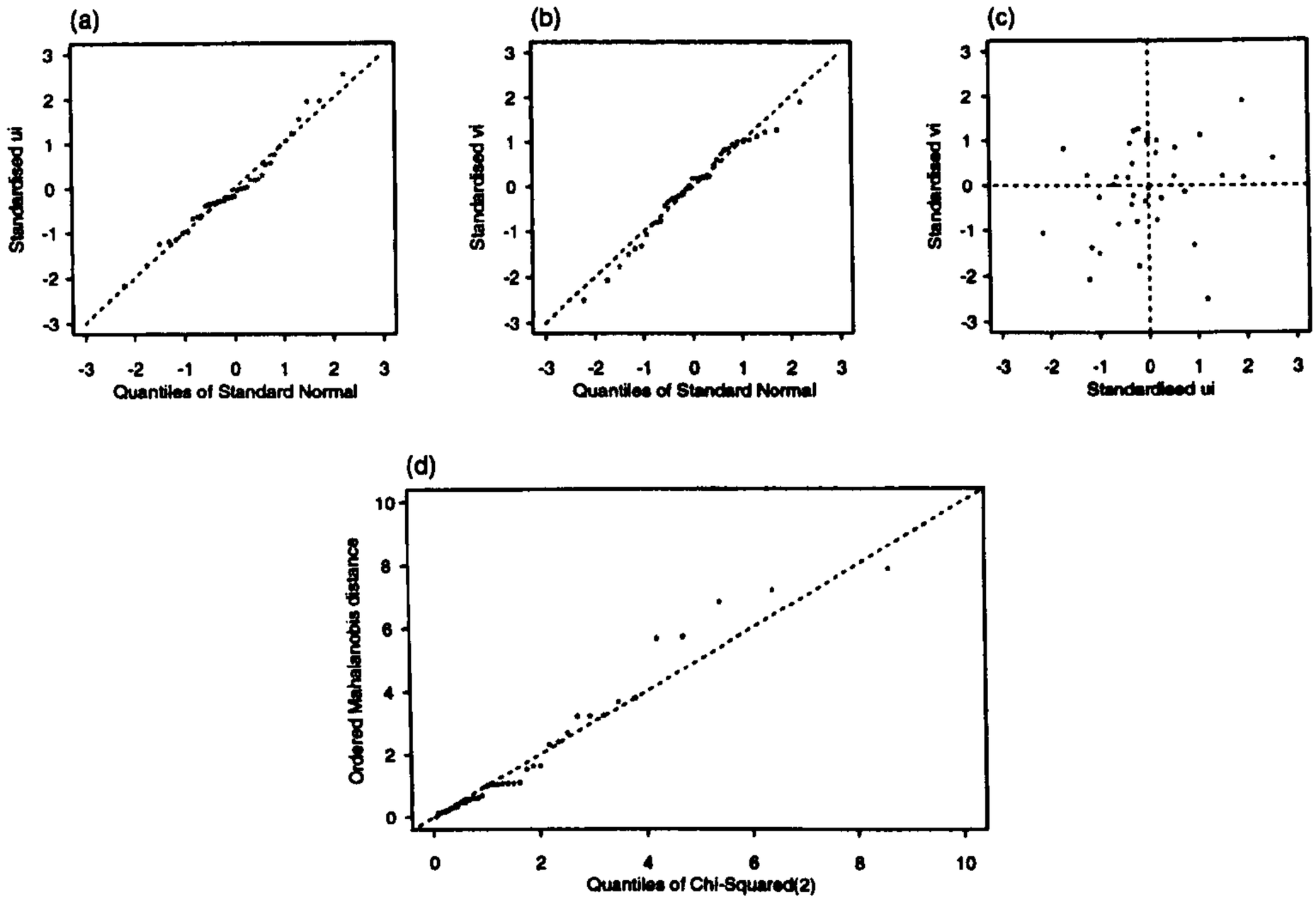

Figure 3.2: Residual diagnostics for the level two residuals for Model six using (a) Normal plot of intercept residuals; (b) Normal plot of slope residuals; (c) bivariate scatter plot of intercept and slope residuals; (d) Gamma plot of the Mahalanobis distances.

A common feature of longitudinal data is serial correlation. Therefore, as well as the need to check the Normality assumptions of the level one residuals, some check for independence of the residuals is also needed. This can easily be done by plotting the level one residuals at occasion *j* against those at occasion *j*-1, or at lag one. This is shown in figure 3.3 along with a Normal plot. The same assumption can also be assessed by modelling an auto-correlated structure for the level one residuals (Goldstein *et al.*, 1994), that is $e_{ij}=\rho e_{ij-1}+v_{ij}$. The resultant change in -2 log lh was 0.1, and the estimated auto-correlation coefficient 0.08, giving no evidence of a violation of the model assumptions.

3.3.4 Conclusions

The overall conclusion from the modelling attempted so far was that although there was

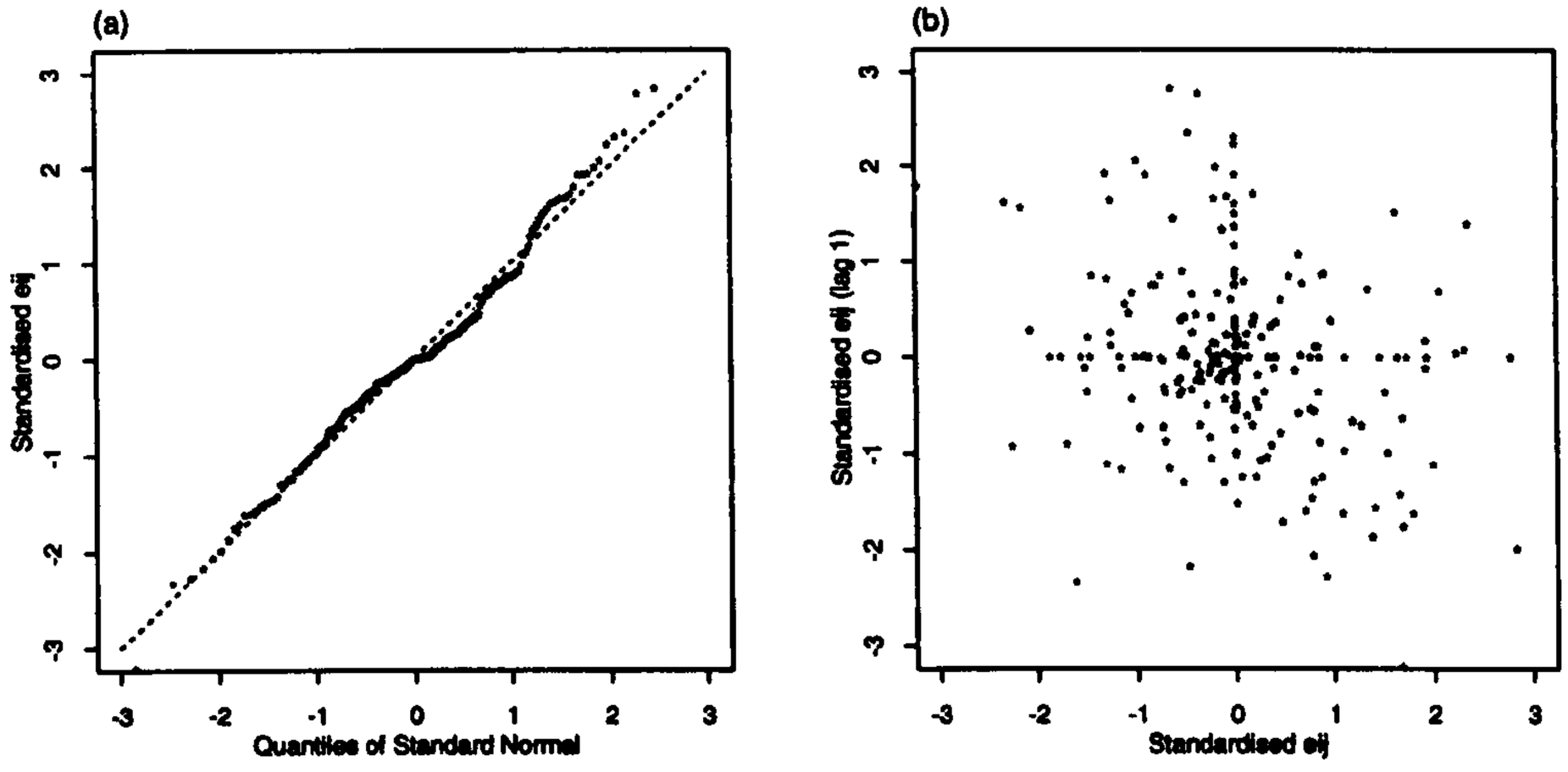

Figure 3.3: Residual diagnostics for the level one residuals for Model six showing (a) Normal plot of residuals; (b) scatter plot of residuals against residuals at lag one.

some evidence of a fall in the level of anxiety over the follow-up period, there was no evidence that the extent of this fall was related to the treatment received (model six). Similarly there was no evidence of a difference in the underlying level of response between treatment groups (model four).

There was an apparent inconsistency in the results of the two models, with the average profile - given by $(\hat{\alpha}, \hat{\beta})$ - for subjects on the split radiotherapy course lying above that of those on the continuous course in model six, and below in model four. In Normal linear regression modelling, this would not happen. However, in the hierarchical model, as the total residual is formed of a subjects component, (u_i, v_i), as well as the residual component, e_{ij}, such results may occur.

One of the advantages of multilevel modelling over alternative models for repeated measures data is the ability to explicitly model the variance structure both within and between

subjects. In this example the variance estimates obtained highlighted a large degree of variation particularly between subjects. This was seen both in terms of the underlying response level (intercept term) and the estimated trend over time. However, introduction of baseline response (centred around the overall mean) did greatly reduced the variance between subjects on the intercept term (model three).

Despite the lack of evidence for a treatment difference it was decided for completeness that a treatment component should be retained in the model. On the basis of the model deviances, and an examination of the residuals from models four and six, the latter was taken as the preferred model to be used in further examples of model extensions.

3.4 Multiple dimensions in response - a three level model

One of the major issues in the analysis of quality of life data arises from its multi-dimensionality with responses often taken in many distinct areas of quality of life. Although the most commonly used measuring instruments have scoring systems which give overall summary scores in appropriate dimensions, to combine these dimensions further to give a single overall score is inappropriate. Therefore problems of multiple comparisons of scores in several dimensions can arise in the assessment of treatment comparisons. Cox *et al.* (1992) suggest Bonferroni type corrections be used as a simple solution. An alternative solution is presented by Tandon (1990) using global statistics (O'Brien, 1984). This approach, which has also been advocated by Pocock *et al.* (1987) combines the results of multiple analyses into a weighted global score with weights determined by the precision of the individual estimates. This gives an overall test statistic for the hypothesis of an overall treatment benefit in favour of one treatment which can be used to complement the results for each individual item. Unfortunately, the approach does not allow estimation of a combined covariate effect, simply a test statistic.

It is also *directional*, meaning it is only appropriate if the direction of effect is consistent over all variables.

Although a global statistic could be formed from the results of a series of two level models, repeated measurement data with multivariate outcomes can be modelled within the framework of a three level model. Such a model will have advantages over both Bonferroni corrections and global statistics because it gives dimension specific estimates of covariates of interest, as well as enabling full estimation of the covariance structure between dimensions, between and within subjects. This not only enables estimation of the correlation between dimensions, but also makes it possible to make a direct comparison of the ways in which the covariate effects vary across dimensions. For example, it becomes possible to test directly whether the response to treatment is the same in all dimensions, whether a patient's physical morbidity is affected to a greater extent than their psychological morbidity, or whether the pattern of change in response over time the same in all dimensions. Given homogeneity of a parameter of interest across dimensions, a combined estimate over all dimensions can then be estimated directly from the model. Like the global statistic, this appropriately weights for the precision of the dimension specific effects. The analysis also has a technical advantage when complete responses for all dimensions are not available. Unlike the global statistic which requires observations available in all dimensions at a single time, the three level multivariate model can cope with unbalanced data across dimensions as it draws upon information about the covariance structure of the available data to give information about those which are missing. It will therefore have an increased power over the respective univariate analyses and global statistic.

Within the three level model, the individual dimensions are regarded as level one units nested within measurement occasions at level two, nested within subjects at level three. The within and between variance components (at level two and level three) are considered in exactly

the same way as for the two level analysis. At level one, the differences between dimensions are assumed fixed, that is differences between dimensions are not regarded to derive from random variation. The assumptions of the model are the same as those of the univariate two level model. In this section, after giving details of the parameterisation of the model, it is used for the analysis of the responses from the HAD scale in the CRC NSCLC study which is measured in two dimensions. The results of the analysis are contrasted against those given by the corresponding univariate analyses.

3.4.1 Parameterisation of the model

To illustrate the parameterisation of the three level model for multivariate repeated measurement data, the two level model of equation (3.1) is extended to three levels to incorporate a multidimensional response. Letting y_{ijl} denote the lth dimensional response (l=1,...,L) for the jth measurement of the ith patient, the underlying model is written

$$y_{ijl}=\sum_{l=1}^{L}\left\{\alpha_l+\beta_l occ_{ij}+u_{il}+v_{il}occ_{ij}+e_{ijl}\right\}z_{ijl}^{(l)} \qquad (3.21)$$

where $z_{ijl}^{(l)}=1$ when y_{ijl} is a response in dimension l, 0 otherwise. It is assumed here that the covariates of the model for each dimension is the same. This is done for simplicity and is not a necessary restriction of the model.

The (fixed) parameters α_l and β_l represent the population average effects for response in the lth dimension for a given mean response in the other dimensions. Subject specific responses are assumed distributed randomly around these population averages with the extent to which subject i deviates from (α_l,β_l) given by (u_{il},v_{il}). Over all dimensions, as in the univariate case, these residuals, $(\boldsymbol{u}_i,\boldsymbol{v}_i)=\{(u_{i1},...,u_{iL}),(v_{i1},...,v_{iL})\}$, are assumed to be independent identically distributed multivariate Normal random vectors with mean $\mathbf{0}$ and, for a simple two dimensional case, variance, $\Omega^{(3)}=\begin{pmatrix}\Sigma_u & \Sigma_{uv}\\ \Sigma_{uv} & \Sigma_v\end{pmatrix}$ where

$$\Sigma_u = \begin{pmatrix} \sigma^2_{u1} & \sigma_{u12} \\ \sigma_{u12} & \sigma^2_{u2} \end{pmatrix}, \quad \Sigma_{uv} = \begin{pmatrix} \sigma_{uv1} & \sigma_{u1v2} \\ \sigma_{u2v1} & \sigma_{uv2} \end{pmatrix}, \quad \Sigma_v = \begin{pmatrix} \sigma^2_{v1} & \sigma_{v12} \\ \sigma_{v12} & \sigma^2_{v2} \end{pmatrix} \tag{3.22}$$

Here, the variance components σ^2_{ul}, σ^2_{vl} and σ_{uvl}, l=1,2 are specific to the lth dimension and correspond to the variance components σ^2_u, σ^2_v and σ_{uv} estimated in the univariate model. The covariance terms σ_{u12} and σ_{v12} indicate how the intercept and slopes in the two dimensions co-vary across subjects. The terms σ_{u1v2} and σ_{u2v1} give information about the relationship between a subject specific intercept in one dimension and the slope in the second dimension. In the subsequent example, these latter terms are constrained to be zero, and Σ_{uv} is assumed diagonal.

At level two, the within subject residuals, $e_{ij}=(e_{ij1},...,e_{ijL})$, are assumed to have a multivariate Normal distribution with mean **0** and variance (for the two dimensional case) $\Omega^{(2)} = \Sigma_e = \begin{pmatrix} \sigma^2_{e1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e2} \end{pmatrix}$ where σ_{e12} describes how, on individual occasions, the responses within subjects deviate from the mean for that subject. For example, on a particular occasion, when a response in one dimension is a long way from its expected value, whether responses in other dimensions tend also to be a long way from their expectation.

For the general case with L dimensions, Σ_u, Σ_v, Σ_{uv} and Σ_e will be of the same structure with dimension $(L \times L)$. By constraining the inter-dimensional covariances at levels three and two to be equal to zero, the responses at each level can be treated as independent. The results from this model will be the same as those obtained from the L corresponding univariate analyses.

3.4.2 Extension of the current analysis

The analyses so far for the anxiety data gave the following basic model,

$$anx_{ij} = \alpha + \beta\, occ_{ij} + \gamma\, base_i + \delta\, occ.rt_{ij} + u_i + v_i occ_{ij} + e_{ij} \qquad \textbf{(3.23)}$$

For simplicity, the same underlying model for the depression scores was also used. This gave a multivariate model written,

$$qol_{ijl} = \sum_{l=1}^{2} \{\, \alpha_l + \beta_l occ_{ij} + \gamma_l base_i + \delta_l occ.rt_{ij} + u_{il} + v_{il} occ_{ij} + e_{ijl} \,\} z_{ijl}^{(l)} \qquad \textbf{(3.24)}$$

At level three (between subjects), the variance components are given by $\Omega^{(3)} = \begin{pmatrix} \Sigma_u & \Sigma_{uv} \\ \Sigma_{uv} & \Sigma_v \end{pmatrix}$, where Σ_u and Σ_v are as given in equation (3.22) and $\Sigma_{uv} = \begin{pmatrix} \sigma_{uv1} & 0 \\ 0 & \sigma_{uv2} \end{pmatrix}$. At level two (within subjects) $\Omega^{(2)} = \Sigma_e = \begin{pmatrix} \sigma_{e1}^2 & \sigma_{e12} \\ \sigma_{e12} & \sigma_{e2}^2 \end{pmatrix}$.

The results from this model, alongside those of the corresponding univariate models are shown in table 3.3 for the fixed parameters and table 3.4 for the random parameters. Very slight differences were apparent. These are attributable to the additional covariance structure

Table 3.3: Fixed parameter estimates (SE) for the multivariate and appropriate univariate models for the anxiety and depression outcomes of the HAD scale measured in the CRC NSCLC study.

		Parameter estimates (SE)		
		Univariate	**Multivariate**	
Anxiety	α (*cons*)	6.43 (0.49)	α_1 (*cons*)	6.42 (0.59)
	β (*occ*)	-0.19 (0.10)	β_1 (*occ*)	-0.20 (0.11)
	γ (*base*)	0.72 (0.11)	γ_1 (*base*)	0.72 (0.09)
	δ (*occ.rt*)	-0.15 (0.16)	δ_1 (*occ.rt*)	-0.12 (0.17)
Depression	α (*cons*)	6.83 (0.52)	α_2 (*cons*)	6.79 (0.55)
	β (*occ*)	-0.17 (0.11)	β_2 (*occ*)	-0.15 (0.10)
	γ (*base*)	0.72 (0.12)	γ_2 (*base*)	0.62 (0.10)
	δ (*occ.rt*)	0.19 (0.16)	δ_2 (*occ.rt*)	0.10 (0.14)

which has been introduced giving a slightly different interpretation to the parameters. They now reflect population average effects in one dimension for given population effects in the second dimension. Some difference was also due to two outlying depression responses given on separate occasions by two subjects. Examination of the anxiety scores of these subjects at these occasions did not reflect the same outlying behaviour. Given the full covariance structure estimated in the multivariate analysis, the outlying depression responses had less leverage. It is believed that the large differences between σ^2_{u2} and σ^2_{u} estimated from the univariate depression model (table 3.4) was also attributable of these outlying responses.

Since the three level model estimates the full covariance structure of the data, a direct comparison of parameter estimates across dimensions is possible. Given no evidence of a difference between the dimensions, combination of the scores to give a simple overall summary

Table 3.4: Dimension specific random parameter estimates (SE) for the multivariate and appropriate univariate models for the simultaneous analysis of the anxiety and depression responses of the HAD scale in the CRC NSCLC study.

		Parameter estimates (SE)		
		Univariate		**Multivariate**
Anxiety				
Level two	σ^2_u	7.67 (2.15)	σ^2_{u1}	7.51 (2.08)
	σ^2_v	0.17 (0.06)	σ^2_{v1}	0.17 (0.06)
	σ_{uv}	0.02 (0.26)	σ_{uv1}	-0.01 (0.21)
Level one	σ^2_e	2.39 (0.27)	σ^2_{e1}	2.39 (0.27)
Depression				
Level two	σ^2_u	7.76 (2.27)	σ^2_{u2}	9.11 (2.40)
	σ^2_v	0.15 (0.06)	σ^2_{v2}	0.13 (0.06)
	σ_{uv}	-0.19 (0.28)	σ_{uv2}	-0.36 (0.25)
Level one	σ^2_e	3.08 (0.34)	σ^2_{e2}	3.10 (0.34)

may then be obtained. Naturally this estimate will have more precision than several dimension specific estimates. This is of particular interest for the estimated treatment effect. In this example, although the treatment effects in the two dimensions had different signs, there was no statistical evidence of a difference in the treatment by occasion interaction for each dimension (p=0.17). By fitting a single parameter in place of the two dimension specific parameters for the different slopes, a summary estimate of the difference in rate of change in anxiety and depression over the follow-up period of 0.01(SE 0.14) units per week was obtained, giving no evidence of a difference in the improvement of quality of life over time between the two treatment groups.

The further gain of the three level model for the analysis of multidimensional repeated measurement data over other analyses suggested in the literature stems from the ability to estimate the inter-dimension covariance structure. Given this information, it is then possible to obtain an estimate of the correlation between dimensions which allows for its repeated measures aspect. Bland and Altman (1995a, 1995b) discuss estimation of such quantities when data are balanced, and in Section 2.2.4 this was presented as part of an exploratory data analysis. Since the data from the CRC NSCLC study are unbalanced, the calculation of correlation estimates is complicated and they were not presented. Given the covariance estimates from the multivariate model, appropriately adjusted estimates of association, both within and between subjects, can be estimated from this model despite the unbalanced nature of the data. These are presented in table 3.5. Within each 2x2 block of the table, the estimated covariance (SE) between dimensions are given below the diagonal, the estimated variances, as given in table 3.4, are given in bold type on the diagonal, and the estimated correlation between dimensions, as calculated from the appropriate covariance and variances, are given above the diagonal.

Table 3.5: Covariance / correlation estimates between dimensions as given from the multivariate model.

	Covariance (SE) / Correlation			
	Intercept		Slope	
Dimension	**Anxiety**	**Depression**	**Anxiety**	**Depression**
Between subject				
Anxiety	**7.51**	0.65	**0.17**	0.47
Depression	5.39 (1.72)	**9.11**	0.07 (0.04)	**0.13**
Within subject				
Anxiety	**2.39**	0.15		
Depression	0.46 (0.22)	**3.10**		

Within each 2x2 block of the table, the estimated covariance (SE) between dimensions are given below the diagonal, the estimated variance of responses, as given in table 3.4, are given in bold on the diagonal, and the estimated correlation between dimensions, as calculated from the appropriate covariance and variance estimates, are given above the diagonal.

The correlation estimates between subject give the level of association of average levels of responses across subjects. The estimate of $5.39/\sqrt{(7.51 \times 9.11)}=0.65$ for the intercept term suggested that subjects who on average had a high anxiety response also had relatively high depression scores. For the association of the subject specific slopes in the two dimensions, the positive correlation of 0.47 suggested that changes in the level of response over the follow-up period in one dimension tended to be accompanied by changes in the same direction in the second dimension. These relationships are reflected in the scatter plot matrix of residuals shown in figure 3.4.

The within subject correlation estimates correspond to estimation of the association between dimensions within a subject on each occasion. That is, whether a higher than average response for a subject on a particular occasion (given their underlying profile) in one dimension is matched by a similarly higher than average response in a second dimension. The very low

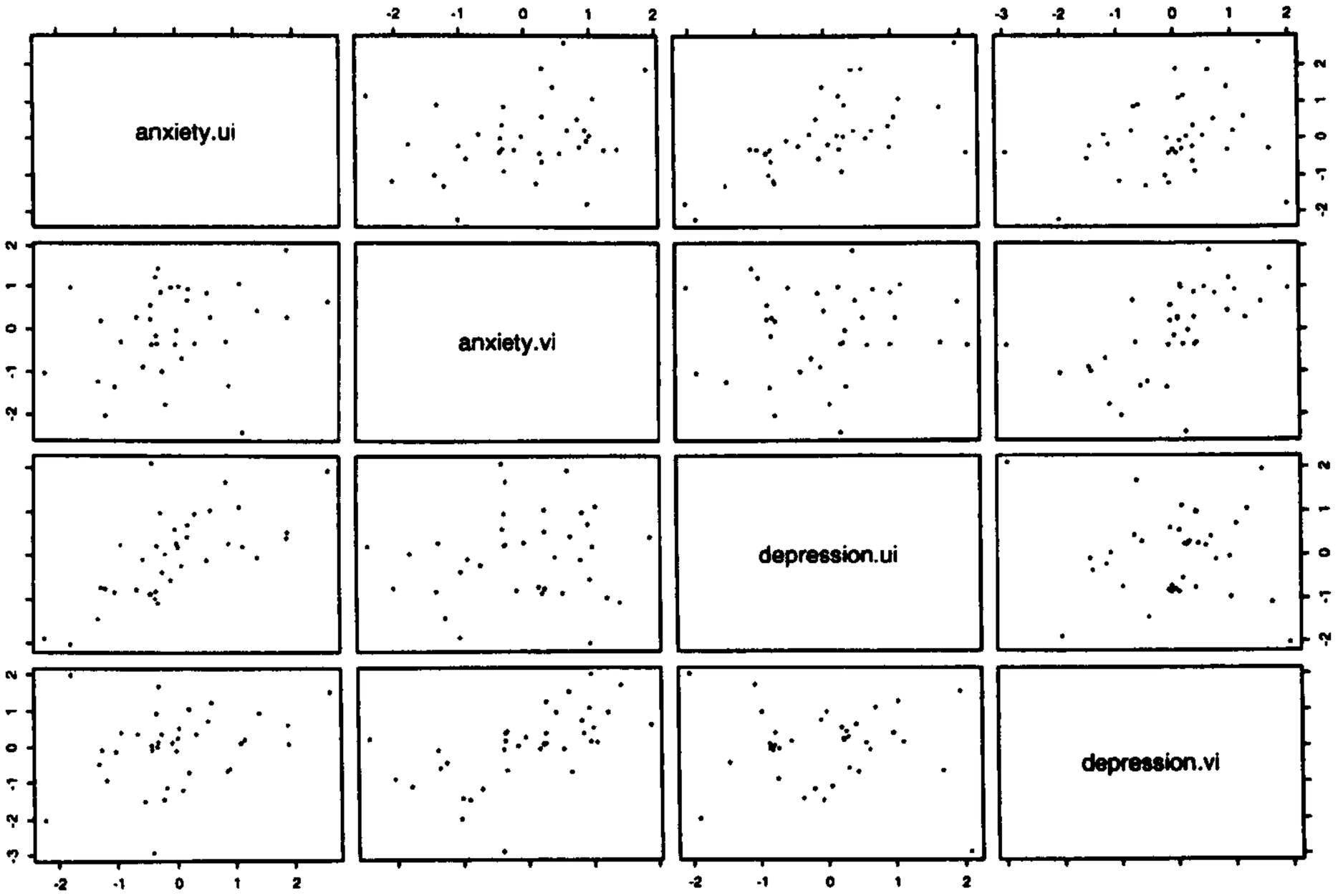

Figure 3.4: Scatter plot matrix of the between subject residuals on the intercept and slope showing the associations between dimensions at a subject level.

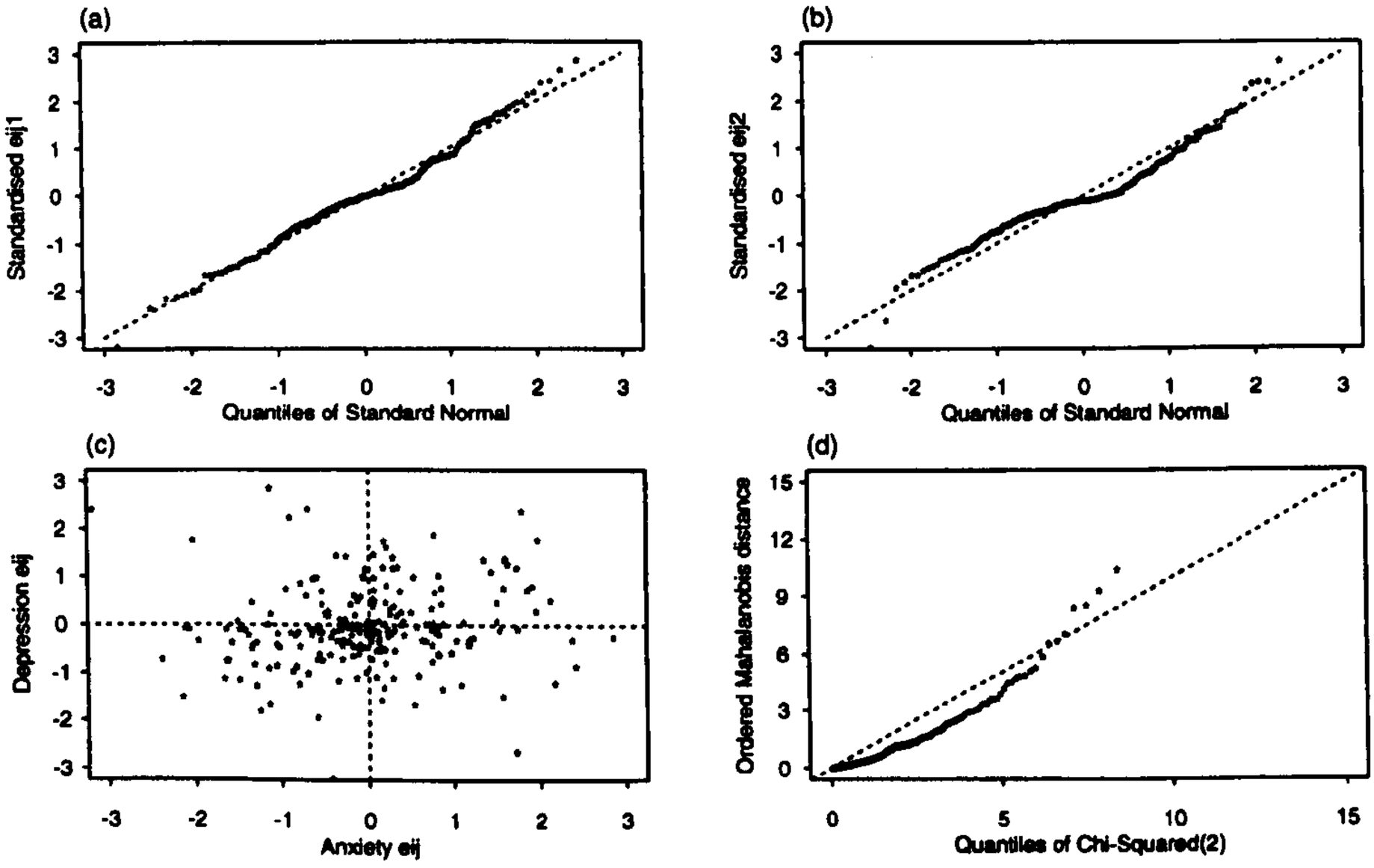

Figure 3.5: Scatter plot matrix of the within subjects residuals across dimensions.

estimate of 0.15 for these data indicates that at individual occasions there was little association between the responses in terms of differences from their expectation. This lack of relationship is shown in figure 3.5 in terms of a scatter plot of predicted level two residuals for the two dimensions. This figure also shows univariate Normal plots of the residuals and a Gamma plot for a test of bivariate Normality of the residuals. Similar plots for checking the assumptions of the level three residuals are given in figure 3.6. These show satisfactory univariate Normal distributions in all cases.

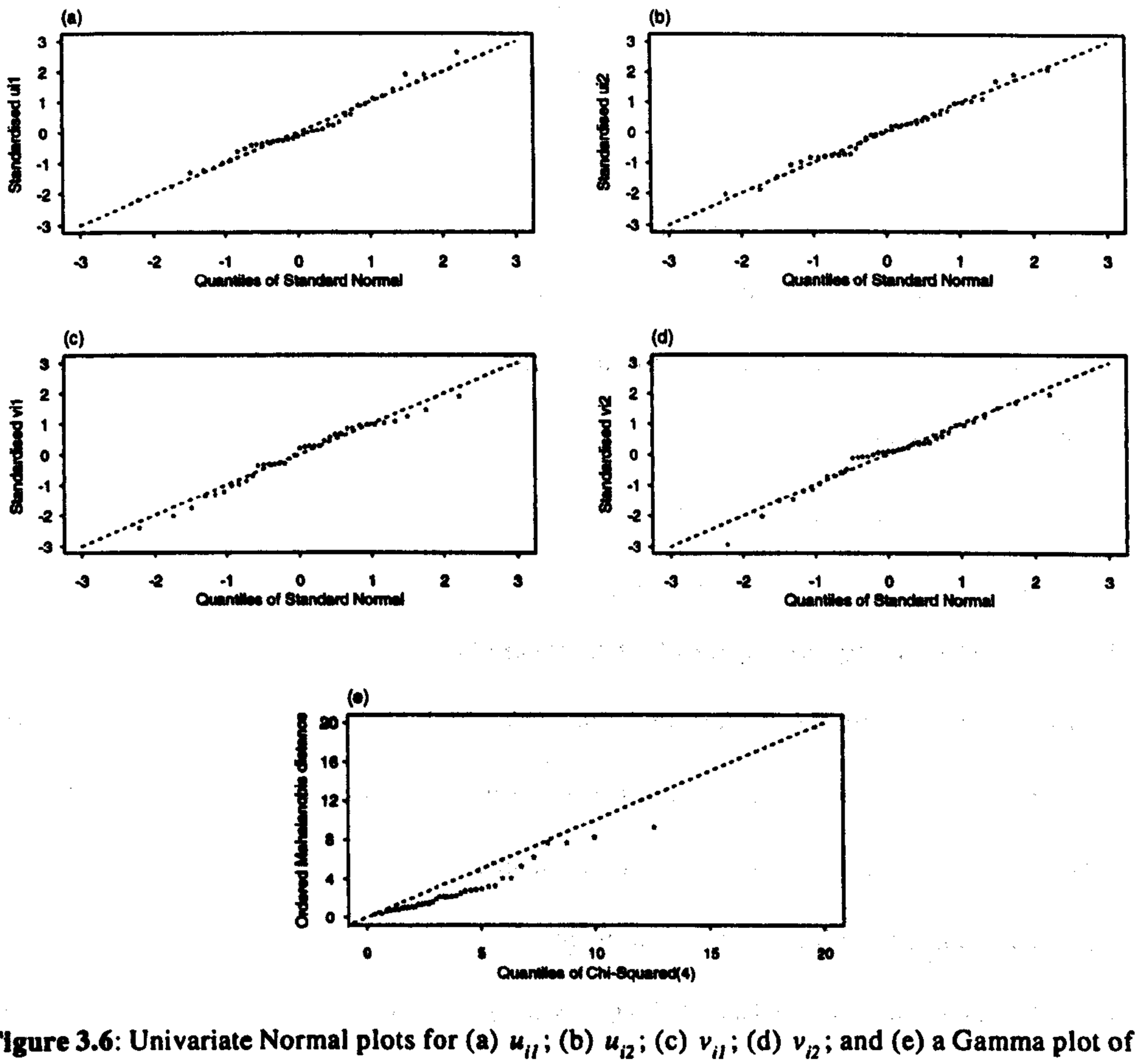

Figure 3.6: Univariate Normal plots for (a) u_{i1}; (b) u_{i2}; (c) v_{i1}; (d) v_{i2}; and (e) a Gamma plot of the Mahalanobis distances for multivariate Normality.

3.4.3 Conclusion

This work has demonstrated the extension of the simple two level model for repeated measurements to three levels for the analysis of multidimensional repeated measurement data. The analysis has many advantages over univariate analyses, with Bonferroni corrections to adjust for multiple comparisons, and global statistic methods that have been suggested in the literature. These all stem from the ability to model the data from all dimensions simultaneously and thus obtain estimates for the inter and intra dimension covariance structure both within and between subjects, even with highly unbalanced data. Given these estimates, it is possible to test homogeneity of estimated covariate effects in each dimension, and then if appropriate, to combine these estimates to a single effect. This not only gives an global test statistic, it also gives a global estimate for the effect of interest. Such an estimate may be of particular interest in quality of life studies, when the bulk of available information to convey often causes problems.

Although the example presented here was very simple - including only two dimensions with the same underlying model - this was only done for ease of presentation and is not a restriction of the models themselves.

3.5 Modelling complex patterns of level one variation

An assumption within the modelling framework which has not yet been discussed is that of constant residual variance across occasions within subjects. In the presence of variance heterogeneity, the conventional modelling approach would be to try and eliminate it using a variance stabilising transformation. A multilevel model however allows the heterogeneity to be modelled explicitly (Goldstein, 1995). As well as overcoming the problem by forcing an underlying structure on the residuals, this variance modelling also enables the specific sources

of heterogeneity to be investigated which itself could be of interest. For instance, subjects on one treatment regimen may tend to exhibit more variation in their responses than subjects on another regimen. Alternatively, residual variation may be related to the underlying level of response. Although variance modelling can be done at any level, modelling of the level one residual components is illustrated here. For simplicity, the univariate model for the HAD depression scores given in tables 3.3 and 3.4 is extended for complex residual variation. The residual variance is modelled in terms of two covariates, both as a function of treatment received and level of baseline response. The extensions to the model parameterisation are first laid out.

3.5.1 Re-parameterisation of the model

In the basic model of equation (3.1), the level one variance is assumed constant across subjects, that is $\text{var}(e_{ij})=\sigma_e^2$ for all i. Assuming we can partition the population into two subgroups (1 and 2) with the level one residual in each subgroup denoted e_{1ij} and e_{2ij} respectively, the level one residual can be written,

$$e_{ij}=e_{1ij}z_{1ij}+e_{2ij}z_{2ij} \qquad (3.25)$$

where z_{1ij} and z_{2ij} are dummy variables defined so $z_{1ij}=1$ for subgroup 1, 0 for subgroup 2, and $z_{2ij}=0$ for subgroup 1, 1 for subgroup 2. Denoting the residual variance in each of these subgroups σ_{e1}^2 and σ_{e2}^2 the level one residual variance then becomes

$$\text{var}(e_{ij})=\sigma_{e1}^2 z_{1ij}+\sigma_{e2}^2 z_{2ij} \qquad (3.26)$$

In terms of the general model notation of Section 3.2, this is just a simple extension of the level one design matrix, $\boldsymbol{X}^{(1)}$ to the $(N\text{x}2)$ matrix $(\mathbf{Z}_1 \ \ \mathbf{Z}_2)$, where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ contain the elements z_{1ij} and z_{2ij} corresponding to $\boldsymbol{Y}$. The variance components, σ_{e1}^2 and σ_{e2}^2, form $\Omega^{(1)}=\begin{pmatrix} \sigma_{e1}^2 & 0 \\ 0 & \sigma_{e2}^2 \end{pmatrix}$.

An alternative and more flexible way of modelling the variation is to concentrate on the difference between the two variances directly. In this model a single dummy variable is used to define subjects in the second subgroup. The level one residuals e_{ij} can then be written,

$$e_{ij}=e_{1ij}+e_{2ij}z_{2ij} \qquad (3.27)$$

where e_{2ij} is the additional residual for subjects in the second subgroup with the level one variance then modelled as,

$$\text{var}(e_{1ij}+e_{2ij}z_{2ij})=\sigma^2_{e1}+2\sigma_{e12}z_{2ij} \qquad (3.28)$$

The interpretation of σ^2_{e1} is unchanged with $2\sigma_{e12}$ giving the additional level one variation for subjects in the second subgroup and may be positive or negative.

In terms of the more general notation, $X^{(1)}$ is the $(N\text{x}2)$ matrix $\left(\mathbf{1}_N \;\; Z_2\right)$, where $\mathbf{1}_N$ is an $(N\text{x}1)$ vector of ones and $\Omega^{(1)}=\begin{pmatrix}\sigma^2_{e1} & \sigma_{e12}\\ \sigma_{e12} & 0\end{pmatrix}$.

The conclusions from the two model specifications will be identical. The latter, however, gives more flexibility in terms of further modelling, and makes the model easy to generalise to more than two subgroups, or to the case in which the residuals are structured in terms of continuous explanatory variables. For example, for a continuous covariate, x_{ij}, for subject i at occasion j,

$$e_{ij}=e_{0ij}+e_{1ij}x_{ij} \quad \text{and} \quad \text{var}(e_{ij})=\sigma^2_{e0}+2\,\sigma_{e01}x_{ij}$$

where $2\sigma_{e01}$ is the additional level one variance due to a one unit change in x_{ij}.

3.5.2 Extension of the current model for the study data

The basic model which is extended in this example using the CRC NSCLC HAD scale

depression data has the same parameterisation as model six (table 3.2). For dep_{ij}, the depression response for the ith subject at the jth measurement occasion,

$$dep_{ij}=\alpha+\beta\, occ_{ij}+\gamma\, base_i+\delta\, occ.rt_{ij}+u_i+v_{ij}occ_{ij}+e_{ij} \qquad (3.29)$$

The residual components of this model were modelled in terms of two covariates already defined - treatment and baseline response.

Model seven - residual variation modelled as a function of treatment

Allowing the residual variation to be a function of treatment group tested the hypothesis that, within subjects, patient depression within one treatment group was more variable than in the second group.

$$dep_{ij}=\alpha+\beta\, occ_{ij}+\gamma\, base_i+\delta\, occ.rt_{ij}+u_i+v_i occ_{ij}+e_{0ij}+e_{1ij}rt_i \qquad (3.30)$$

where e_{0ij} denotes the within subject residual for all subjects, e_{1ij} is the additional within subject residual for the continuous group and rt_i is the treatment covariate which equals 0 for patients in the split course, 1 for the continuous group. The variance at level one was modelled as

$$\text{var}(e_{0ij}+e_{1ij}rt_i)=\sigma^2_{e0}+2\sigma_{e01}rt_i \qquad (3.31)$$

where σ^2_{e0} is the variance of the within subject residual components for patients in the split course group, and $2\sigma_{e01}$ is the additional variation experienced by subjects in the continuous radiotherapy arm.

Model eight - residual variation as a function of baseline depression

Modelling the residual variation as a function of baseline depression, tested the hypothesis that subjects with low baseline scores tended to exhibit different amounts of residual variation than those with a higher baseline score. This difference was modelled as a linear trend in increasing baseline score

$$dep_{ij}=\alpha+\beta\, occ_{ij}+\gamma\, base_i+\delta\, occ.rt_{ij}+u_i+v_i occ_{ij}+e_{0ij}+e_{2ij}base_i \qquad (3.32)$$

where, as in the fixed part of the model, each subject's baseline depression score was modelled as a difference from the overall mean depression score. The variance at level one was parameterised as,

$$\text{var}(e_{0ij}+e_{2ij}base_i)=\sigma^2_{e0}+2\sigma_{e02}base_i \qquad (3.33)$$

where σ^2_{e0} is the expected residual variation for subjects with a mean baseline depression score, and $2\sigma_{e01}$ is the expected increase in this variance for each unit increase in baseline score.

The results of extending model six for the depression scores, to examine the level one variance components, are given in table 3.6. Inferences for these analyses were made on the basis of changes in the model deviances because of the difficulties in interpreting the asymptotic standard errors of the variance components, as discussed in Section 3.2.4.

Table 3.6: Parameter estimates (SE) for modelling level one heterogeneity for depression scores.

		Parameter estimate (SE)		
		Model six	**Model seven**	**Model eight**
Fixed parameters				
α	(*cons*)	6.83 (0.52)	6.88 (0.53)	6.91 (0.50)
β	(*occ*)	-0.17 (0.11)	-0.19 (0.11)	-0.18 (0.11)
δ	(*base*)	0.72 (0.12)	0.72 (0.12)	0.70 (0.12)
γ	(*occ.rt*)	0.19 (0.16)	0.19 (0.18)	0.15 (0.16)
Random parameters				
Level two	σ^2_u	7.76 (2.27)	8.44 (2.39)	7.15 (2.03)
	σ^2_v	0.15 (0.06)	-0.36 (0.32)	-0.17 (0.26)
	σ_{uv}	-0.19 (0.28)	0.20 (0.07)	0.16 (0.06)
Level one	σ^2_{e0}	3.08 (0.34)	1.96 (0.29)	3.50 (0.42)
	σ_{e01}	-	1.20 (0.40)	0.31 (0.04)
-2 log lh		1044.3	1032.3	986.2

Evidence of variance heterogeneity related to treatment was seen with a change in deviance of 12.0 on 1 df. The estimated difference in variance suggested that subjects receiving the continuous treatment course tended to have a larger degree of within subject variation (1.96 versus 4.36). The evidence of a linear relationship with the level of depression response at baseline was even more convincing with a reduction in the model deviance of 58.1 on 1 df. The estimated relationship showed an increasing within subject variance with an increasing level of baseline response.

3.5.3 Conclusions of the model

The model extensions demonstrated here were used in an attempt to explain the possible heterogeneity of the within subject level one variance. It was seen that there was some evidence that the extent of within subject residual variance was related to treatment received (model seven), and also to baseline response for which it was estimated that the within subject residual variation changed from 3.50 for mean baseline depression, at a rate 0.62 units change per 1 unit change on the mean baseline response. However, when these results were examined further using plots of the level one residuals in model six for the depression data by the two covariates of interest in this analysis (figure 3.7),it was seen that, particularly in terms of the relationship with adjusted baseline response, these observed relationships were attributable to two outlying observations. Both of these patients were on the continuous radiotherapy course, and had baseline depression scores above the overall mean. Omitting these patients and repeating the heterogeneity analyses yielded no evidence of variance heterogeneity.

These observations, emphasised that, although modelling of the variance components is possible, the results may not be robust to non-Normality or outliers. This will be particularly important in small samples when the results are more likely to be influenced by a few subjects with rather extreme results. Careful inspection of these aspects is therefore important before

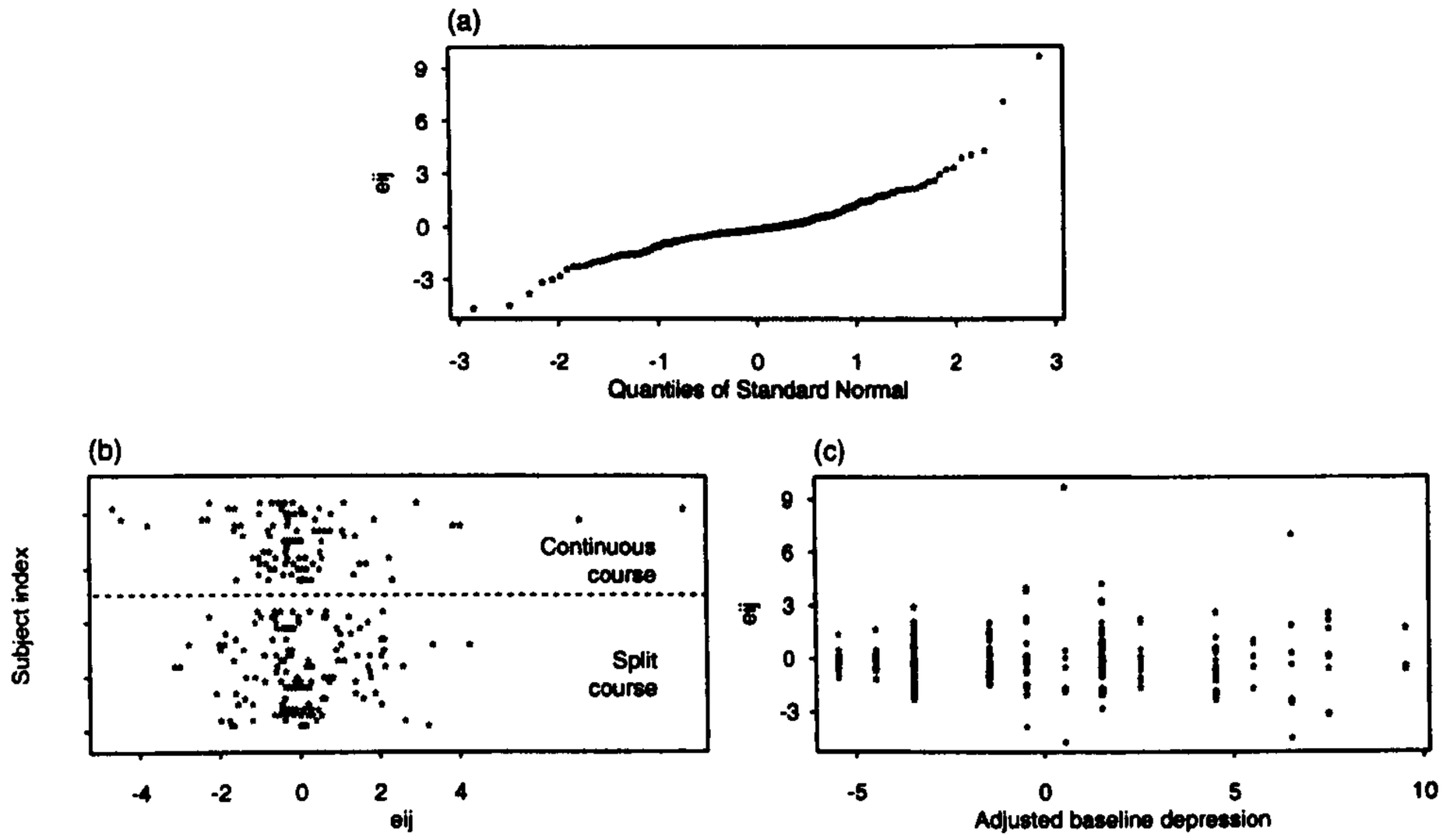

Figure 3.7: Within subject (non-standardised) residuals from model six for the CRC NSCLC HAD depression scores: (a) Normal plot for Normality; and plotted (b) by treatment group; and (c) by baseline depression from which the overall mean depression score has been subtracted.

claiming evidence of variance heterogeneity, and it was concluded that heterogeneity of the level one residuals was not a problem in these data.

3.6 Summary and discussion

The work presented here has demonstrated the use of hierarchical (or multilevel) models for analysis of quality of life data given as repeated continuous outcomes. Their flexibility for the analysis of longitudinal data which is severely unbalanced (in this case due to missing data) as well as incorporating multiple dimensional outcomes was demonstrated. The models require a number of critical assumptions, but although not well developed within the multilevel literature, model checking was shown to be relatively straightforward using estimated shrunken residuals. Further work is needed in this area in order to determine the exact properties of these

shrunken residuals for the assessment of multivariate Normality. Although shown to be non-robust to outlying observations, modelling of residual variances is also feasible to look for evidence of variance heterogeneity. A further problem with longitudinal data, which has not been discussed here in detail, is serial correlation. In terms of the two level model, this would occur at level one and result in the non-independence of the level one residuals. It has been argued by Jones (1990) that, after incorporating random subject effects, further serial correlation of the within subject residuals may be unlikely. However, the flexibility of the hierarchical model structure makes inclusion of a component of serial correlation to test this assertion straightforward (Goldstein *et al.*, 1994).

Because of the flexibility of the models, it is important that an appropriate analysis strategy is pre-defined. This should involve exploratory data analysis as outlined in Chapter 2, and outlining questions of scientific interest pertaining to the population (or fixed) effects and perhaps more importantly, those to be addressed in terms of the random components representing the between and within subject variation. This importance was demonstrated in Section 3.5, where the results of the analysis modelling of the level one variance were seen to be heavily dependent on a few outlying observations. Further work is perhaps needed in this area to determine whether this is a problem relating to examples with small samples or a more general concern of the modelling strategy.

Perhaps the most important use of hierarchical models for the analysis of quality of life data is the ability to incorporate data from many dimensions of quality of life measured into one analysis, thus obtaining overall covariate effects, as well as appropriately adjusted estimates of inter-dimension associations within and between subjects. The multivariate example given here was the most simple case with only two dimensions with an equal number of responses in each dimension. However, the model does not require such restrictions. For instance, as often

occurs in quality of life studies, a subject may have a response in one dimension where their response in another dimension is missing. In this situation the multivariate model in fact gives a technical advantage over univariate approaches in that, given the reasons underlying the data being missing can be ignored, the analysis draws upon information available about the correlation structure between dimensions, improving the precision of dimension specific estimates.

Within this analysis, the substantial amount of missing data (detailed in tables 2.1 and 2.2) were treated simply as causing an unbalanced data problem. In terms of the introduction of bias due to missing data, they were *ignored*. The implications of this, along with a more detailed consideration of the missing data problem, are discussed in Chapter 7.

In terms of the example presented, another concern was the small sample size on which the analysis was based. Of the 82 subjects in the study, 57 contributed at least one post randomisation response. With the inclusion of baseline responses only 37 cases out of these 57 were included in the main analyses presented. For the use of these types of models in practice, larger samples than this would generally be recommended, not least from a scientific point of view - little can be inferred from 37 patients who exhibit a large degree of variability. Although a simulation study carried out to investigate the robustness of the analysis in such a small sample gave no evidence of bias in any of the fixed or random parameter estimates when a restricted iteratively generalised least squares procedure (RIGLS) was used, further work is still needed in this area to help determine methods for sample size calculations when such models are to be used for data analysis.

Alternative analyses for repeated measurement data are possible. These are reviewed by Crowder and Hand (1993) and Everitt (1995). For balanced data these include modifications

to a split plot analysis of variance or, for more complex correlation structures, multivariate analysis of variance (MANOVA). With unbalanced data and missing values these approaches become infeasible. In addition, MANOVA has been shown to lack power when the number of measurement occasions is large as is often the case in quality of life studies in cancer research. Antedependence models (Kenward, 1987) achieve greater parsimony and hence precision than MANOVA by restricting the correlation structure to take various sensible forms. However, they do not cope well with unequally spaced data and their results may be difficult to present simply for clinical purposes. Simple methods based on constructing summary statistics for each subject (Matthews *et al.*, 1990) which are intuitively simple become more difficult when subjects have different numbers of measurements, yielding varying precision, or if some subjects have very few measurements (Matthews, 1993). Moreover, such analyses ignore information available from other subjects in terms of the distribution of subject specific profiles. Marginal modelling approaches such as weighted least squares and generalised estimating equations (GEE) (Zeger and Liang, 1992) provide an alternative framework which can cope flexibly with unbalanced data and may allow more general correlation structures than the simple autoregressive model. Within such models, the focus of the analysis is solely on the underlying mean process. The parameters defining the variance structure are regarded as nuisance parameters necessarily included in the analysis to adjust the precision of the estimates of this mean process. Unlike the hierarchical model, they do not therefore allow any inferences to be drawn about the nature of this variance structure. They do perhaps have an advantage over the hierarchical model, in that, being based very closely to models for cross-sectional analyses, their results are perhaps easier to communicate to non statisticians. Such models are discussed in more detail in Chapter 4, where their use and interpretation are contrasted with the hierarchical model for the analysis of repeated binary outcomes.

4 The Analysis of Repeated Binary Outcomes

4.1 Introduction

With a few exceptions, quality of life measuring instruments consist of a series of questions to which patients respond with a binary outcome - "do you suffer from this symptom, yes/no?" - or on an ordered categorical scale - "rate the severity of symptom on a scale of 0-3". Generally, the number of positive responses or the ordinal ratings are then summed to give a 'continuous' score which may then be analysed using the methods discussed in the previous chapter. In some instances, this summation of items may not be recommended (for example, with the daily diary card) and in all cases it will result in a loss of detail about the prevalence and severity of particular symptoms which may be of interest to a clinician or patient. Thus the ability to analyse the repeated data on the original categorical scale may be advantageous. The current and subsequent chapter are devoted to this topic. In this chapter, two different approaches to the analysis of binary outcomes are discussed. This work is then extended for repeated ordered categorical outcomes in Chapter 5. Although other analysis options do exist, the use of marginal and random effect models are discussed. For reviews of these and other models see Diggle *et al.* (1994), Agresti (1989), Landis *et al.* (1988).

In Chapter 3 the general theory of random effect models for the specific case of variance components analysis was outlined. In Section 4.2, this model is extended for use with binary outcomes, and the theory behind marginal models given. The basic assumptions and formulation of each model is reviewed, and their estimation and interpretation discussed. The use of the models, and more importantly the differences between them, are demonstrated using

two examples with data from the MRC LU07 diary card and the RSCL in the CRC NSCLC study in Section 4.3. Sections 4.4 and 4.5 present two extensions of these basic models to tackle two very different problems faced in the analysis of quality of life data in practice, those of complex patterns of response over time and multidimensional outcomes.

4.2 Marginal and random effect models for repeated binary data

Cox (1972) outlined a number of possible approaches for analysing multivariate binary outcomes many of which needed further research to be of practical use. Since then, with much research and improved computing facilities, many of the ideas which Cox proposed have become practicable for data analysis. Two classes of model that have received much attention, and which are particularly useful for the analysis of longitudinal data, are those Cox referred to as the *logistic* and *latent variable* models, more generally now referred to as *marginal* and *random effects* models. It is these two classes of model that will be discussed here. Although the structure of each model is discussed in greater detail in subsequent sections, a simple illustration of the differences between them and their interpretations for binary data is outlined below. Similar discussions are given by Zeger and Liang (1992), Neuhaus *et al.* (1991), and Diggle *et al.* (1994).

The fundamental difference between the marginal and random effect models is the way in which they incorporate the dependence between repeated observations on the same subject. For a binary response vector, $y_i=(y_{i1},...,y_{im_i})$, measured for subject i, over measurement occasions $j=1,...,m_i$, the marginal model has a generalised linear model for its expectation $\theta_i=(\theta_{i1},...,\theta_{im_i})$, incorporating the dependence between observations as nuisance parameters in the residual error component of this model. In contrast, the random effect model assumes that the dependence between observations comes as a result of some subject specific random effect, conditional

upon which, the observations within subjects are assumed independent. This is demonstrated in equations (4.1) and (4.2) which give the most basic form of each model to investigate the effect of a treatment covariate, x_i. As for a standard generalised linear model, $g(.)$ is a link function, with e_{ij} the binomial error with $E(e_{ij})=0$.

Marginal model

$$y_i=g^{-1}(\alpha^M+\beta^M x_i)+e_i^M, \quad \text{var}(e_i^M)=V_i \tag{4.1}$$

Random effect model

$$y_i=g^{-1}(\alpha^{RE}+\beta^{RE}x_i+u_i)+e_i^{RE}, \quad u_i\sim N(0,\sigma_u^2), \quad \text{var}(e_i^{RE})=\text{diag}\{\text{var}(e_{ij})\} \tag{4.2}$$

This difference in incorporating the dependence between observations leads to a very different interpretation of the parameters of the two models. In the marginal model they are *population average* effects and where $g(.)$ is the logit link function, $\exp(\beta^M)$ is the odds ratio of symptoms for subjects who undergo treatment versus those who do not. In the random effect model they are *subject specific* effects conditional upon a subject's underlying response determined by the random effect, u_i. $\text{Exp}(\beta^{RE})$ is then the odds ratio of a positive response for subject i if he has treatment versus if he does not.

Although this difference in interpretation exists for the analysis of all types of outcome - continuous or discrete - by taking the expectation over the distribution of the random effects, it is possible to transform the parameters of the random effect model to give a marginal (population average) interpretation,

$$E(y_i|x_i)=\int_{u_i} g^{-1}(\alpha^{RE}+\beta^{RE}x_i+u_i)f(u_i)\,du_i \tag{4.3}$$

If $g(.)$ is the identity link, this integration is trivial,

$$E(y_i|x_i)=(\alpha^{RE}+\beta^{RE}x_i)+\int_{u_i} u_i f(u_i)\,du_i \qquad (4.4)$$

and since $E(u_i)=0$, shows an equivalence between the marginal and random effect parameters for this case. It is for this reason that the distinction between the interpretation of the two models is often overlooked for continuous outcomes where the identity link tends to be used. However, for other link functions, the distinction can be crucial. In particular, with the logit link function commonly used for the analysis of binary outcomes, taking expectation over the random effects, equation (4.3) cannot easily be reduced further,

$$E(y_i|x_i)=\int_{u_i}\frac{\exp(\alpha^{RE}+\beta^{RE}x_i+u_i)}{1+\exp(\alpha^{RE}+\beta^{RE}x_i+u_i)}f(u_i)\,du_i \qquad (4.5)$$

However, Neuhaus *et al.* (1991) show that equation (4.5) implies that $|\beta^{RE}|\geq|\beta^{M}|$, with equality only in the trivial case when $u_i=0$ for all i. More specifically, they show that if the covariate x_i has no effect, $\beta^{M}\sim\beta^{RE}[1-\rho(0)]$ where $\rho(0)$ is the intra subject correlation amongst y_i, defined as

$$\rho(0)=\frac{\mathrm{var}(\theta_{ij})}{E(\theta_{ij})E(1-\theta_{ij})} \qquad (4.6)$$

where $\mathrm{logit}(\theta_{ij})=\alpha^{RE}$. In practical application, as this relationship is based on the assumed distribution of u_i as well as the assumption that the covariate has no effect, the ratio $\frac{\beta^{M}}{\beta^{RE}}$ would be expected to approximate to $[1-\rho(0)]$. Further, Zeger and Liang, (1992), show that if the random effects, u_i, $i=1,\ldots,n$, are assumed to follow a Normal distribution with mean 0 and variance σ_u^2,

$$\frac{\beta^{M}}{\beta^{RE}}\sim(0.346\sigma_u^2+1)^{-\frac{1}{2}} \qquad (4.7)$$

4.2.1 Marginal models for binary longitudinal outcomes

The simplest form of the marginal model for binary longitudinal outcomes was proposed by Liang and Zeger (1986) and uses a standard generalized linear model for the expected response θ_{ij}, for subject i at occasion j, conditional on a set of p covariates, which are given by the $(m_i x p)$ design matrix $\boldsymbol{x}_i$,

$$E(\boldsymbol{y}_i) = \theta_i = g^{-1}(\boldsymbol{x}_i \beta) \tag{4.8}$$

The var($\boldsymbol{y}_i$) is an $(m_i x m_i)$ matrix made up of two components: an $(m_i x m_i)$ diagonal variance matrix which defines the Binomial variance of y_{ij}, as a known function of the mean parameters θ_{ij}, $\boldsymbol{A}_i$=diag$\{\theta_{ij}(1-\theta_{ij})\}$, j=1,...,m_i; and an $(m_i x m_i)$ 'working' correlation matrix, denoted $\boldsymbol{R}_i$, which defines the correlation structure of the data. Given these two components, var($\boldsymbol{y}_i$)=$\boldsymbol{V}_i^*$ is simply

$$\boldsymbol{V}_i^* = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}} \tag{4.9}$$

Estimates for the marginal parameters β can be obtained by solving the multivariate analogue of the score equations for generalised linear models, or generalised estimating equations (GEE1),

$$S_\beta(\beta) = \sum_{i=1}^{n} \left(\frac{d\theta_i}{d\beta} \right)^T \boldsymbol{V}_i^{*-1} (\boldsymbol{y}_i - \hat{\theta}_i) = 0 \tag{4.10}$$

If $\boldsymbol{R}_i$, and therefore $\boldsymbol{V}_i^*$, is correctly specified, these score equations will be the optimal estimating equations for β as given by Godambe (1960) and

$$\text{var}(\hat{\beta}) = I_0 = \sum_{i=1}^{n} \left(\frac{d\theta_i}{d\beta} \right)^T \boldsymbol{V}_i^{*-1} \left(\frac{d\theta_i}{d\beta} \right) \tag{4.11}$$

However, Liang and Zeger (1986) showed that even if $\boldsymbol{R}_i$ is incorrectly specified, estimates for β obtained from the score equations of equation (4.10), although not fully efficient, will still

be consistent. Further, they show that a robust estimate for their variance is given by $\text{var}(\hat{\beta})=\mathbf{I}_1^{-1}\mathbf{I}_0\mathbf{I}_1^{-1}$ where

$$\mathbf{I}_1=\sum_{i=1}^{n}\left(\frac{d\theta_i}{d\beta}\right)^T \boldsymbol{V}_i^{*-1}(y_i-\hat{\theta}_i)(y_i-\hat{\theta}_i)^T\boldsymbol{V}_i^{*-1}\left(\frac{d\theta_i}{d\beta}\right) \qquad (4.12)$$

Although the 'working' correlation matrix $\boldsymbol{R}_i$, can take any form, good specification is required to maximise efficiency. Liang and Zeger (1986) suggested that this is best obtained by assuming a general form of the matrix with unknown parameters ρ. For example, for longitudinal data with fixed measurement occasions, $j=1,\ldots,m$, exchangeable or auto-regressive general forms can be assumed for some underlying lag one correlation ρ. That is,

$$\boldsymbol{R}_i^{EX}(\rho)=\begin{pmatrix} 1 & \rho & \rho & \cdots \\ \rho & 1 & \rho & \cdots \\ \vdots & & \ddots & \\ \cdots & \rho & \rho & 1 \end{pmatrix} \text{ and } \boldsymbol{R}_i^{AR}(\rho)=\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \\ \rho^{m-2} & \cdots & \rho & 1 & \rho \\ \rho^{m-1} & \cdots & \rho^2 & \rho & 1 \end{pmatrix}$$

respectively.

Estimates for ρ are obtained via some function of the Pearson residuals following estimation of β. The specific form of the function used depends on the general form of the correlation structure. Examples are given by Liang and Zeger (1986). GEE1 is then an iterative process where at the first iteration, starting values for ρ or an independence model are assumed.

Although GEE1 performs well in estimation of β, estimation of ρ relies on the basic assumption underlying the general form of $\boldsymbol{R}_i$ which is generally arbitrary. If information about the correlation is of interest, it is therefore suggested GEE1 is not used (Liang *et al.*, 1992). Prentice (1988) suggested that when such inferences about the correlation structure are required, a general model for the two-way cross products and their expectations is incorporated

into the analysis to give direct estimation of correlation parameters, ρ. This analysis (GEE2) uses an iterative procedure with a second set of estimating equations for ρ using the cross-products and their expectation and of the same form as those of equation (4.10). Fitzmaurice and Laird (1993) presented a similar model with the two-way associations modelled in terms of log odds ratios. Although giving more efficient estimates for the association parameters than GEE1, Carey *et al.* (1993) show that GEE2 relies on the correct specification of a model for the association parameters - whether in terms of correlation or odds ratios - and thus may only be an improvement over GEE1 if there exists some *a priori* knowledge of the association structure. Further, the increased complexity with the additional set of estimating equations makes estimation computationally intensive and impractical even when the number of repeated measurements is relatively small (for example, m_i=5). For these reasons GEE2 will not be practical for many quality of life problems when the number of measurement occasions is large, and is therefore not used here.

Other specifications of the marginal model are also possible. For example, Carey *et al.* (1993) have suggested a model which improves the estimation of association parameters and avoids the computational difficulties of GEE2. They call the approach *alternating logistic regression* (ALR) because both the marginal distribution and the association parameters are modelled using logistic regression. For ALR it is assumed that the association between pairs of observations, (y_{ij},y_{ik}), $j \neq k$, may be represented by a log odds ratio, which in the most simple case is assumed to be some constant, ρ. In this case, the odds ratio of positive response between y_{ij} and y_{ik}, denoted Ψ_{ijk}, is equal to $\exp(\rho)$. Estimation is an iterative two stage process. In stage one, GEE1 is used along with a current estimate for α to give estimates for β. Given these estimates, the estimate for ρ is updated using a logistic regression of each y_{ij} on each y_{ik} $(j<k)$, with an offset derived from the current estimates of β and ρ. Formally, denoting $E(y_{ij})=\theta_{ij}$, $E(y_{ij}y_{ik})=\nu_{ijk}$ and assuming a constant log odds ratio for all pairs (y_{ij},y_{ik}),

$j<k$, the stage two logistic regression for ρ is

$$\text{logit } \Pr(y_{ij}=1 \mid y_{ik}) = \rho\, y_{ik} + \textbf{offset} \qquad \textbf{(4.13)}$$

where $\textbf{offset} = \log\left(\dfrac{\theta_{ij} - \nu_{ijk}}{1 - \theta_{ij} - \theta_{ik} + \nu_{ijk}} \right)$, the log odds of $y_{ij}=1$ for $y_{ik}=0$. This can be evaluated using current estimates of the marginal expectations $\hat{\theta}_{ij}$ and $\hat{\theta}_{ik}$ from the stage one solution to GEE1 and an estimate for ν_{ijk} which itself can be evaluated from the current estimate of ρ and the stage one marginal parameters using the expression derived by Lipsitz *et al.* (1991):

$$\nu_{ijk} = \begin{cases} \dfrac{f_{ijk} - \{f_{ijk}^2 - 4\rho(\rho-1)\theta_{ij}\theta_{ik}\}^{\frac{1}{2}}}{2(\rho-1)} & \text{for } \rho \neq 1 \\ \rho\theta_{ij}\theta_{ik} & \text{for } \rho = 1 \end{cases} \qquad \textbf{(4.14)}$$

where $f_{ijk} = \{1 - (1-\rho)(\theta_{ij} + \theta_{ik})\}$.

To demonstrate the general use of these models, only GEE1 is used.

4.2.2 Random effect models for longitudinal binary outcomes

The random effect model for repeated binary outcomes accounts for the correlation between observations on the same subject by assuming that each subject has their own underlying propensity of a positive response which is incorporated into the model as a random effect. Given each subject specific effect, observations within subject are then assumed to be independent binomial random variables.

For subject i at occasion j the effect of a covariate x_{ij} on the outcome y_{ij} can be represented in the logistic regression model,

$$y_{ij} = \theta_{ij} + e_{ij} \qquad \textbf{(4.15)}$$

where $y_{ij}=1$ for a positive outcome, 0 otherwise and

$$\theta_{ij}=\frac{\exp(\alpha+\beta x_{ij}+u_i)}{1+\exp(\alpha+\beta x_{ij}+u_i)} \tag{4.16}$$

The subject specific random effect is denoted u_i and is assumed Normally distributed with zero mean and variance σ_u^2. Given u_i, the residual e_{ij}, $j=1,\ldots,m_i$, for subject i are assumed to be independent binomial random variables with variance $v_{ij}=\sigma_e^2\theta_{ij}(1-\theta_{ij})$ where σ_e^2 is an over dispersion parameter. Although other appropriate link functions may be used for binary outcomes, the work here focuses on the commonly used logit link given in equation (4.16).

Re-writing this simple model of equation (4.16) in terms of its linear predictor for $\text{logit}(\theta_{ij})$,

$$\log\frac{\theta_{ij}}{1-\theta_{ij}}=\alpha+\beta x_{ij}+u_i \tag{4.17}$$

the interpretation of the parameters is clear: $(\alpha+u_i)$ can be interpreted as the log odds of a positive response for subject i; and β as the effect of a unit change in the value of the covariate, x_{ij}, on this subject specific log odds. For example, if $y_{ij}=1$ for subject i, reporting symptoms at occasion j, and x_{ij} is a treatment group indicator, the interpretation of $\exp(\beta)$ is the odds ratio of symptoms for subject i if he undergoes treatment.

As for continuous outcomes, the model can be considered as a hierarchical model with two levels. At level two is the variation between subjects given by the random effect u_i, and at level one is the independent binomial errors for observations within subjects. Written in the general notation of the hierarchical model, equation (4.16) becomes

$$\theta_{ij}=g^{-1}(x_i\beta_\bullet+x_i^{(2)}\beta^{(2)})=g_{ij} \tag{4.18}$$

where $g(.)$ is the logit link function, $\beta_\bullet$ is a $(p_0 \times 1)$ vector of fixed parameters acted on by a $(m_i \times p_0)$ design matrix x_i and $\beta_i^{(2)}$ is the vector of random parameters acted on by the design matrix $x_i^{(2)}$.

A number of procedures are available for estimation of the fixed parameters β_0 and the variance components var($\beta^{(2)}$) (Stiratelli *et al.*, 1984, Longford, 1988, Goldstein, 1991, 1995, Zeger and Karim, 1991, Breslow and Clayton, 1993). As it is easily implemented within software for practical use, that attributable to Goldstein (1991, 1995) involving marginal and penalised quasi-likelihood and (restricted) IGLS, is considered here.

This is a two stage procedure involving the linearisation of the g_{ij} in stage one followed by parameter estimation of this linear function in stage two. The linearisation of g_{ij} uses a Taylor series expansion where it is assumed that parameter estimates from the previous iteration are known. For the $(t+1)$th iteration, a first order Taylor series expansion of g_{ij} is written

$$g_{ij}(H_{t+1})=g_{ij}(H_t)+x_i(\beta_{0t+1}-\beta_{0t})g'_{ij}(H_t)+(x_i^{(2)}\beta_i^{(2)})g'_{ij}(H_t) \qquad (4.19)$$

where H_t denotes the realised expectation of g_{ij} from the tth iteration. It is the definition of H_t that forms the distinction between the marginal and penalised quasi-likelihood procedures. Under marginal quasi-likelihood (MQL), the Taylor expansion is carried out about the fixed part of the model and hence $H_t=x_i\beta_{0t}$. The penalised (or predictive) quasi-likelihood (PQL) incorporates the estimated residuals from the previous iteration, $\hat{\beta}_t^{(2)}$, into this expression, that is $H_t=x_i\beta_{0t}+x_i^{(2)}\beta_{it}^{(2)}$ such that the expansion is carried out around the predicted value for the ith subject. In this case the last term of equation (4.19) becomes $x_i^{(2)}(\beta^{(2)}-\hat{\beta}_i^{(2)})g'_{ij}(H_t)$.

Whatever the chosen expression for H_t, by defining $x_i^*=x_ig'_{ij}(H_t)$ and $x_i^{*(2)}=x_i^{(2)}g'_{ij}(H_t)$, equation (4.15) can be written as a linear function of the parameters of interest:

$$y_i^*=x_i^*\beta_{0t+1}+x_i^{*(2)}\beta_i^{(2)}+v_ie_i \qquad (4.20)$$

where $y_{ij}^*=y_{ij}-\{g_{ij}(H_t)-x_i^*\beta_{0t}\}$, $z_{ij}=\sqrt{\hat{\theta}_{ij}(1-\hat{\theta}_{ij})}$ for $v_i=(v_{i1},\ldots,v_{in_i})$ with σ_e^2 constrained to equal one. The IGLS (or RIGLS for small samples) procedure can then be applied to this function

to obtain estimates for β_{0t+1} and $\text{var}(\beta_i^{(2)})_{t+1}$. Binomial variation at level one is assured by the definition of v_{ij} and the constraint on $\sigma_e^2=1$. If this constraint is removed, σ_e^2 can be estimated as an over dispersion parameter. For an improved approximation, a second order Taylor series expansion may be used although this obviously adds to convergence time. For details of see Goldstein (1995).

In simulation studies, Rodriguez and Goldman (1995) show that the MQL procedures can lead to a large degree of bias. PQL however has been shown to perform well (Goldstein, 1995, Goldstein and Rasbash, 1996). A potential problem with the models is that parameter estimation is non-robust to the failure of the assumed distribution of random effects, therefore diagnostic checking of residuals is particularly important.

4.3 Marginal versus random effect models in practice

Within this section, the use of marginal and random effect models for the analysis of binary quality of life data is compared to demonstrate both the practical application and the inferential differences of the two models. The first example is based on data from the activity item on the daily diary card within the MRC LU07 study. The second analysis uses the individual shortness of breath item on the RSCL within the CRC NSCLC study. In both cases, the binary response was created by dichotomising the score on the original ordinal scale. For the marginal models, GEE1 with independence and exchangeable 'working' correlation matrices were used. Corresponding random effect models were fitted using RIGLS first order MQL and first order PQL. The second order PQL model did not converge. An over dispersion parameter was also fitted within the random effect models. As this is not possible for the marginal model, the comparison between the two classes of model is made between the marginal model with exchangeable working correlation matrix and the PQL model with the binomial variance constrained at the lowest level.

4.3.1 MRC LU07 daily activity scores

For the first example data from the activity item on the daily diary card for the first four weeks of follow-up within the LU07 study was analysed. This corresponds to the data presented in figure 2.13. Responses were dichotomised from the original five point scale as follows: 1/2= few or no symptoms (y_{ij}=0); 3-5 = limitations (y_{ij}=1). Details of the scoring of the original scale are given in box A2.3 in Appendix 2. The focus of the analysis was to assess the prevalence of limitations in activity in the two treatment groups controlled for age was estimated. The results of these models are shown in table 4.1. The top half of the table gives the results for the marginal models, the lower half give those for the random effect models. The basic form of the two models are given below.

<u>Marginal</u> <u>Random effects</u>

$$\log\frac{\theta_{ij}}{1-\theta_{ij}}=\alpha^{M}+\beta^{M}age_i+\delta^{M}rt_i \qquad \log\frac{\theta_{ij}}{1-\theta_{ij}}=\alpha^{RE}+\beta^{RE}age_i+\delta^{RE}rt_i+u_i \qquad \textbf{(4.21)}$$

Statistically, the example demonstrates the importance of the robust standard error for a poorly specified correlation structure for the marginal model. Based on the naive standard errors from the independence model, gross errors of interpretation would result. In contrast, it seemed that the exchangeable correlation structure was reasonable for these data, shown by the robust standard error being little changed from the naive estimate. Comparing the robust standard errors from the two models demonstrated that even when the correlation structure was poorly specified, reliable inferences could have been made using the robust standard error which in the independence model was very close to that of the exchangeable model. Similarly, the example demonstrates elements of the earlier discussion for the random effects model with the parameter estimates of the 1st order MQL model very close to those of the marginal model whereas those from the first order PQL were much greater. The exchangeable marginal model and the first order PQL model with constrained binomial variance (PQL (1)) are used for

inferential purposes.

Concentrating first on the results from the exchangeable marginal model, it was estimated that the odds of reduced activity in the FM group was 1.25 times that in the F2 group (95% CI=[0.79, 1.88]). There was no evidence to suggest that this difference was not due to chance.

Table 4.1: Coefficients and SE for logistic regression model of daily diary card LU07 activity data.

		Marginal (GEE1)					
		Independence			**Exchangeable**		
		Estimate	**Naive SE**	**Robust SE**	**Estimate**	**Naive SE**	**Robust SE**
α	(*cons*)	0.81	(0.04)	{0.15}	0.87	(0.15)	{0.15}
β	(*age*)	-0.02	(0.003)	{0.014}	-0.019	(0.014)	{0.014}
δ	(*rt*)	-0.18	(0.05)	{0.22}	-0.22	(0.22)	{0.21}
Estimates for the correlation structure							
		-			ρ=0.68		
		Random effect (MLn)					
		1st order MQL		**1st order PQL (1)**			**1st order PQL (2)**	
		Estimate	**SE**	**Estimate**	**SE**	β^M/β^{RE}	**Estimate**	**SE**
α	(*cons*)	0.88	(0.15)	2.11	(1.90)	0.46	2.22	(0.36)
β	(*age*)	-0.019	(0.014)	-0.051	(0.03)	0.37	-0.06	(0.03)
δ	(*rt*)	-0.20	(0.22)	-0.42	(0.45)	0.52	-0.49	(0.52)
Estimates for the variance structure								
		σ^2_u=3.10 (0.28) σ^2_e=1		σ^2_u=11.7 (1.16) σ^2_e=1		0.45	σ^2_u=16.2 (1.55) σ^2_e=0.45 (0.008)	

age=patient age - 68.2, where 68.2 is the mean age of the sample in years

rt=1 for F2 course, 0 for FM course

For MQL and PQL (1), σ^2_e is constrained to equal 1 to give binomial variation at level one.

The ratio β^M/β^{RE} is calculated using β^M from the GEE1 exchangeable model, and β^{RE} from the PQL model.

The reporting of symptoms was also shown to fall with increasing age. This may be due to a lower expectation of activity levels in the older patients. The estimated difference in the odds of symptoms for a one year difference in age was -2%, 95% CI=[-4%, +1%]. Estimated from a marginal model, this represented the odds ratio for the population as a whole. So for the population as a whole, the odds of reporting adverse levels of activity in a group of patients aged 50, would be expected to be 25% higher than that in a group of patients aged 60. Again there was no evidence to suggest the observed age effect was not simply due to chance.

The first order PQL estimates for the random effects model showed a large degree of variation between the reporting of symptoms across subjects. This was consistent with that seen in the summaries of subject specific responses shown in figure 2.13. All coefficients in this model were consistent in sign with those of the marginal model, although as expected they were much larger. In terms of the treatment covariate, the subjects specific treatment ratio was 1.52 for the multiple fraction course of radiotherapy (FM) over the two fraction course (F2) (95%CI=[0.63, 3.68]). Given a subject's underlying odds of reporting adverse symptoms of activity, this reflects a 52% increased odds if they were treated with the multiple fraction course rather than the two fraction course. For the age covariate, the estimated change in the subject specific odds of reporting adverse activity for a one year increase in age was -5% (95% CI=[-10%, +1%]). Thus given a subject's underlying odds, and all other variables remaining constant, at the age of 50 their odds of reporting problems with activity would be expected to be 67% higher than that at age 60. As for the marginal model, there was no evidence to suggest that any of the observed effects were not due to chance.

Table 4.1 also shows the ratio of the marginal and random effect parameters. The theoretical estimate according to the result of Zeger and Liang (1992) in equation (4.7) is given in the final row of the table amongst the estimates for the variance parameters. Some

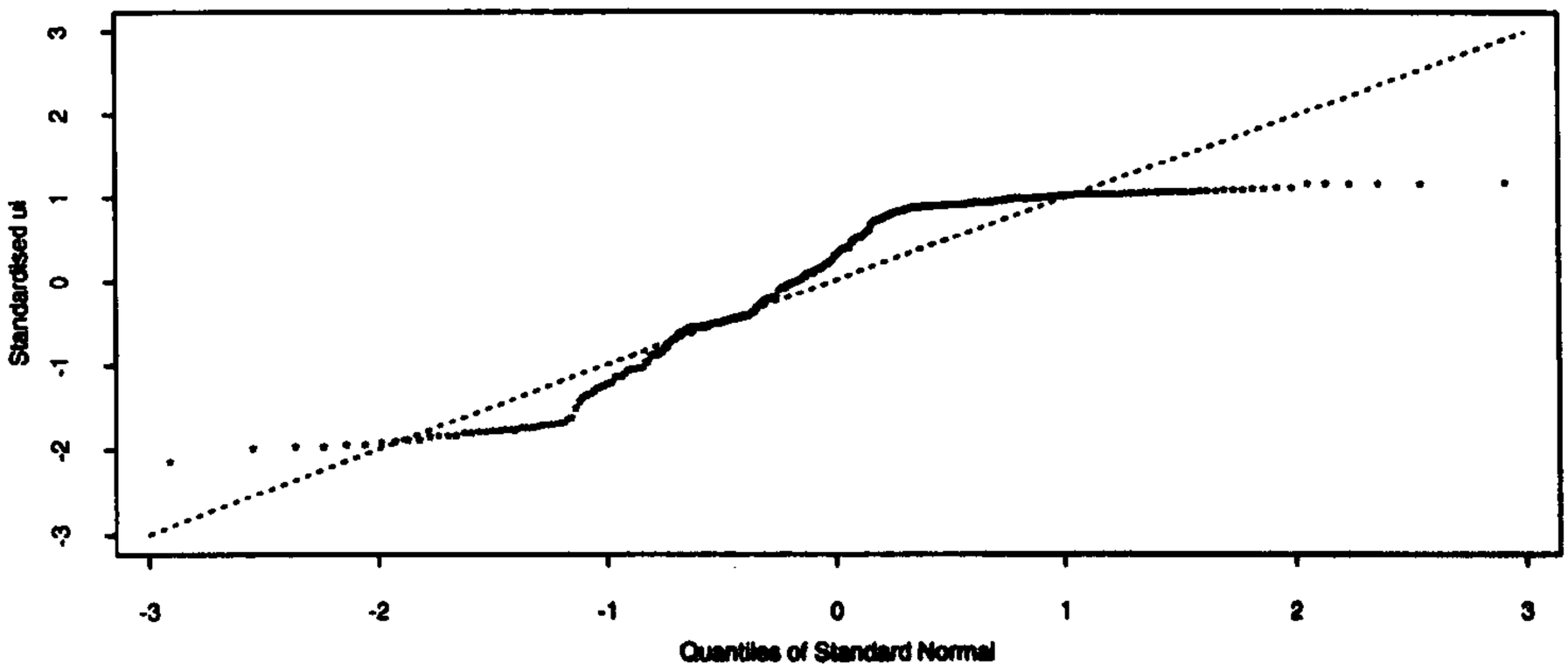

Figure 4.1: Normal plots of the standardised level two residuals for the random effects model for MRC LU07 activity data with distributionally constrained variance of level one residuals.

differences between those observed and the theoretical estimate were seen. This is because the theoretical estimate was based on not only on the assumption of Normality of the level two residuals, but also that the working correlation matrix was correctly specified. In terms of the residuals of the model with distributional constraint at level one, these showed a strong indication of a lack of Normality (figure 4.1(a)). The striking 'S' shape of the distribution occurs as a result of patients responding positively or negatively throughout the whole follow-up period.

There was also some evidence of under dispersion with this model shown when the constraint on σ_e^2 was removed (PQL (2) in table 4.1). This signifies that within this model, the level one residuals were less dispersed than expected given their expectation. This was because the subject specific effects, u_i, allow the expectation for each observation to realise a value close to that observed, thus leaving the residuals, e_{ij} too small for the binomial variation. This was also signified by an increase in the estimated variance of subject specific effects, σ_u^2, in the

unconstrained model. This will once again be largely explained by those subjects reporting positively and negatively throughout. Given their random effect, the expected value at each occasion with be very close to that observed.

4.3.2 CRC NSCLC shortness of breath

The second example uses the individual item 'shortness of breath' taken from the RSCL used in the NSCLC study. Again the response was dichotomised to give a binary score: 0/1=no symptoms (y_{ij}=0); 2/3=symptoms (y_{ij}=1). The analysis here focused on a constant treatment difference and a linear trend in log odds over time. The resulting models are given in equation (4.22). A marginal profile is shown is figure 4.2 and shows a clear downward trend in the odds of reporting shortness of breath over time and a consistently lower odds for patients in the continuous radiotherapy course.

Marginal	Random effects
$\log\frac{\theta_{ij}}{1-\theta_{ij}}=\alpha^{M}+\beta^{M}occ_{ij}+\delta^{M}rt_{i}$	$\log\frac{\theta_{ij}}{1-\theta_{ij}}=\alpha^{RE}+\beta^{RE}occ_{ij}+\delta^{RE}rt_{i}+u_{i}$

(4.22)

The results of this analysis (table 4.2) show some evidence of a linear trend in log odds equating to a fall in the prevalence of symptoms over the period. Within the marginal model, this is given by an odds ratio of 0.86, 95% CI=[0.79, 0.94], for each additional week of follow-up, giving an estimated 14% reduction in the odds of symptoms each subsequent week, (95% CI=[6%, 21%]). From the random effects model, the corresponding reduction in subject specific odds was 31%, 95% CI=[16, 43%]. No evidence of a treatment difference was found, although, consistent with figure 4.2, the odds of shortness of breath were 27% lower (95% CI=[76% lower, 124% higher]) in the group of patients treated with the continuous course of radiotherapy. In terms of the effect on the subject specific odds, given a patient's underlying odds of reporting shortness of breath, given the continuous radiotherapy course they had a 48%

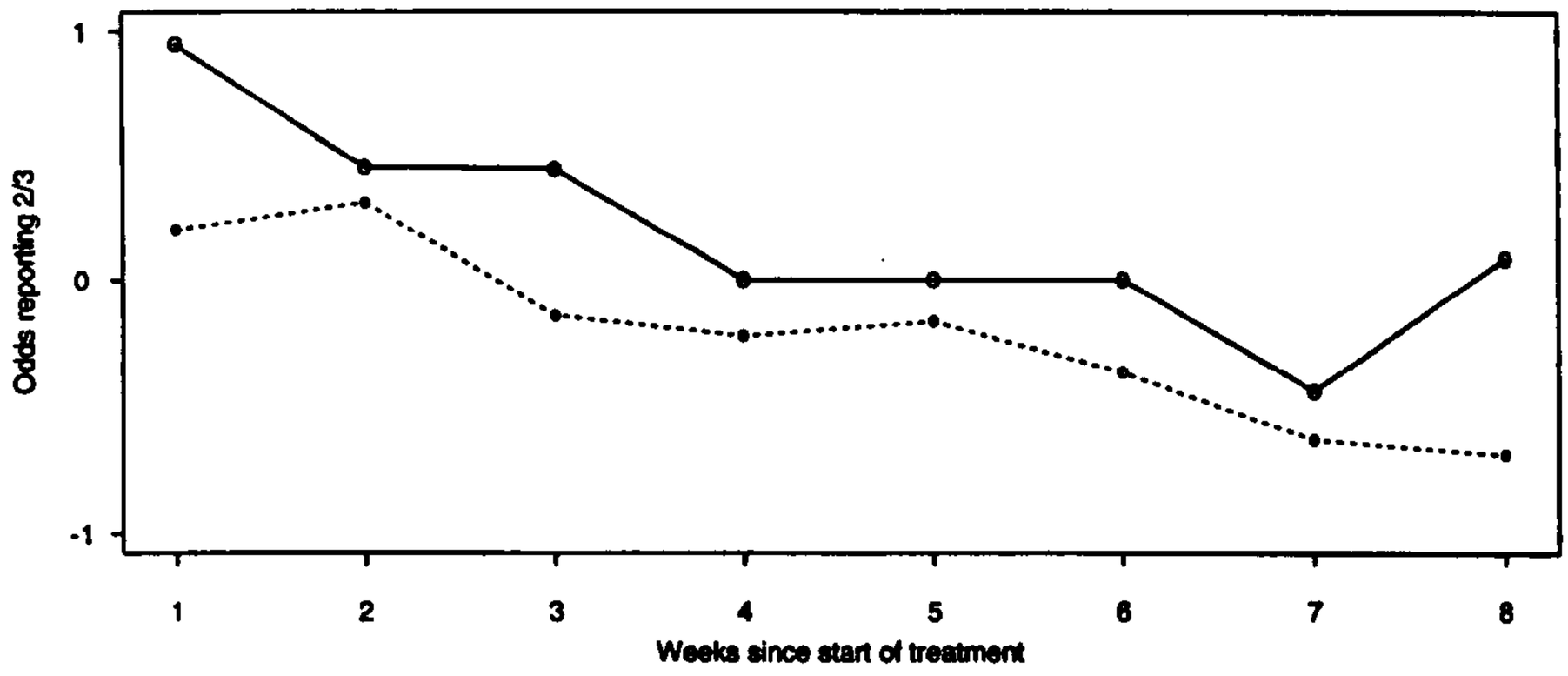

Figure 4.2: Marginal profile for the odds of reporting symptoms of shortness of breath on the RSCL in the CRC NSCLC study over the eight week follow-up for patients on the split course of radiotherapy: ————; and the continuous course: --------.

lower odds (95% CI=[93% lower, 278% higher]) of symptoms than if given the split course.

There was a large degree of variability in the subject specific logs odds resulting in a 95% reference range for the probability of symptoms of [0.01, 0.99], demonstrating that some subjects reported no symptoms at all whereas others reported them consistently throughout the follow-up.

The ratio of the marginal and random effects coefficients again showed some difference from the expected ratio using the results of Liang and Zeger (1992). A normal plot of the level two residuals (figure 4.3) again showed a heavy tailed distribution which corresponds to some subjects reporting no symptoms throughout the follow-up thus having large negative residuals against the large positive residuals of subjects reporting symptoms at each occasion. There was also some evidence of under dispersion noted from removing the constraint on σ_e^2.

Within this study, quality of life was measured on patients pre-treatment. It was felt that including these data into the model as a covariate might help control for those subjects reporting positive or negative throughout and so improve the distributional assumptions of the random effects. The results are given in table 4.3 with baseline dichotomised in the same way as the response, $base_i$=1 for subject i with symptoms graded 2/3 at baseline; 0 otherwise.

Table 4.2: Coefficients and SE for RSCL shortness of breath item.

	Marginal (GEE1)					
	Independence			**Exchangeable**		
	Estimate	**Naive SE**	**Robust SE**	**Estimate**	**Naive SE**	**Robust SE**
α (*cons*)	0.55	(0.26)	{0.39}	0.49	(0.38)	{0.37}
β (*time*)	-0.15	(0.06)	{0.05}	-0.15	(0.03)	{0.05}
δ (*rt*)	-0.26	(0.26)	{0.58}	-0.31	(0.56)	{0.57}
Estimates for the correlation structure						
	-			ρ=0.69		
	Random effect (MLn)					
	1st order MQL		**1st order PQL (1)**			**1st order PQL (2)**
	Estimate	**SE**	**Estimate**	**SE**	β^M/β^{RE}	**Estimate**	**SE**
α (*cons*)	0.51	(0.43)	1.27	(0.73)	-	1.79	(0.89)
β (*time*)	-0.15	(0.06)	-0.36	(0.10)	0.41	-0.50	(0.08)
δ (*rt*)	-0.29	(0.58)	-0.65	(1.01)	0.48	-0.96	(1.32)
Estimates for the variance structure						
	σ^2_u=2.53 (0.72) σ^2_e=1		σ^2_u=7.76 (2.18) σ^2_e=1		0.52	σ^2_u=14.5 (3.72) σ^2_e=0.42 (0.04)	

occ=week of measurement coded 0-7 for weeks 1-8
rt=1 for the continuous, 0 for the split course radiotherapy
For MQL and PQL (1), σ^2_e was constrained to equal 1 to give Binomial variation at level one.
The ratio β^M/β^{RE} was calculated using β^M from the GEE1 exchangeable model, and β^{RE} from the PQL model.

The introduction of the baseline symptoms gave an obvious reduction in the variance between subjects, and gave strong evidence that patient symptoms at baseline had a great bearing on patient symptoms following treatment. From the marginal model, it was estimated that the group of patients who reported shortness of breath at baseline were 37.5 times more likely to report symptom following the start of radiotherapy (95% CI=[9.5, 141.5]). In terms of the effect of baseline on the subject specific odds, the constrained PQL model estimated that given their underlying propensity to report shortness of breath, if a subject reported symptoms at baseline, they were 113 times more likely to report symptoms following the start of radiotherapy than if they did not report symptoms (95% CI=[17.9, 758]).

The residuals from this model are shown in figure 4.4 and show a much better distribution in terms of Normality than those of the model without baseline. This indicates that the adjustment for baseline has gone someway to adjusting for the effect of subjects who respond positively or negatively throughout. Unfortunately, in terms of the ratio of parameter estimates

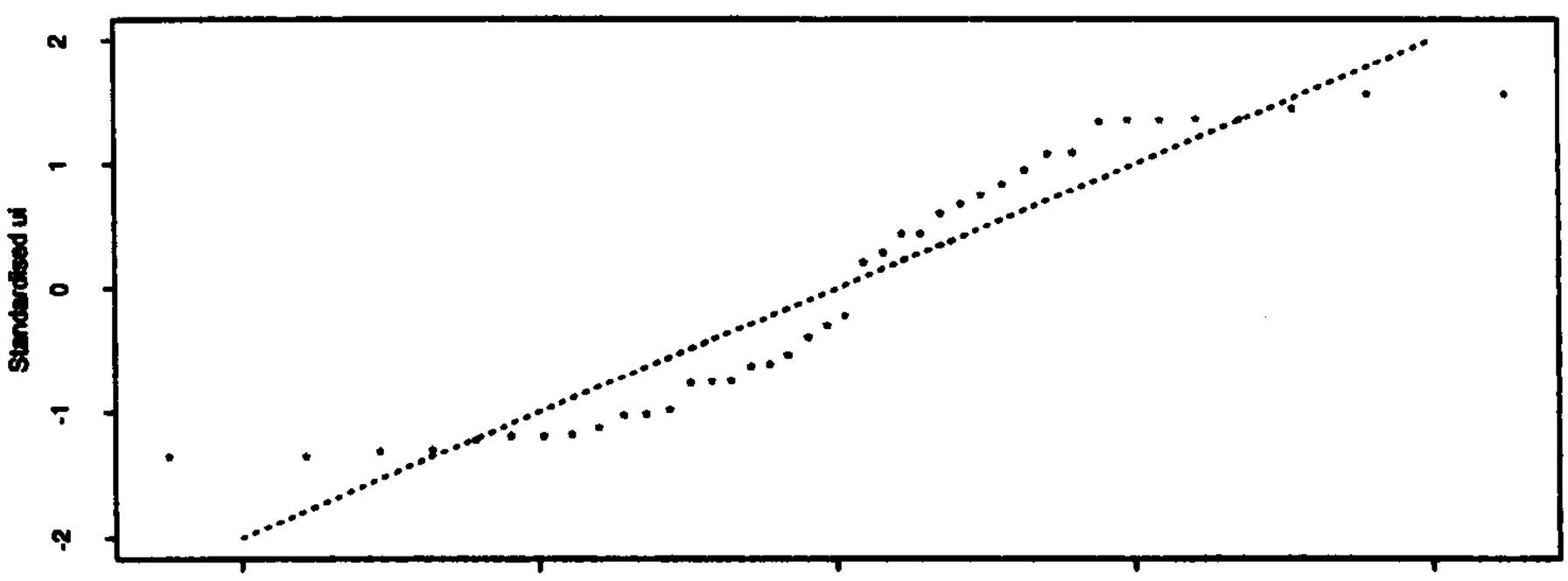

Figure 4.3: Normal plot of the standardised level two residuals for the shortness of breath item on the RSCL in the CRC NSCLC study.

Table 4.3: Parameter estimates for shortness of breath in CRC NSCLC study with an adjustment for baseline.

		Estimate (SE)			Ratio
		Marginal	**Constrained PQL**	**Unconstrained PQL**	β^{M}/β^{RE}
α	(*cons*)	-1.99 (0.62)	-2.07 (0.94)	-2.46 (1.11)	-
β	(*occ*)	-0.23 (0.07)	-0.39 (0.10)	-0.51 (0.08)	0.59
δ	(*rt*)	-0.007 (0.64)	-0.05 (0.90)	-0.07 (1.11)	0.14
γ	(*base*)	3.60 (0.69)	4.73 (0.97)	5.98 (1.81)	0.76
Estimates for the variance structure					
		α=0.46	σ_u^2=4.66 σ_e^2=1	σ_u^2=8.48 σ_e^2=0.42	0.62

The marginal model presented used an exchangeable working correlation matrix.

from the marginal and random effects models, these results were less convincing than those of the previous model. There was also some evidence of underdispersion of the level one residuals. This suggested that in the previous analysis, this may not have been due to the subjects responding positively and negatively throughout. Both this underdispersion and the

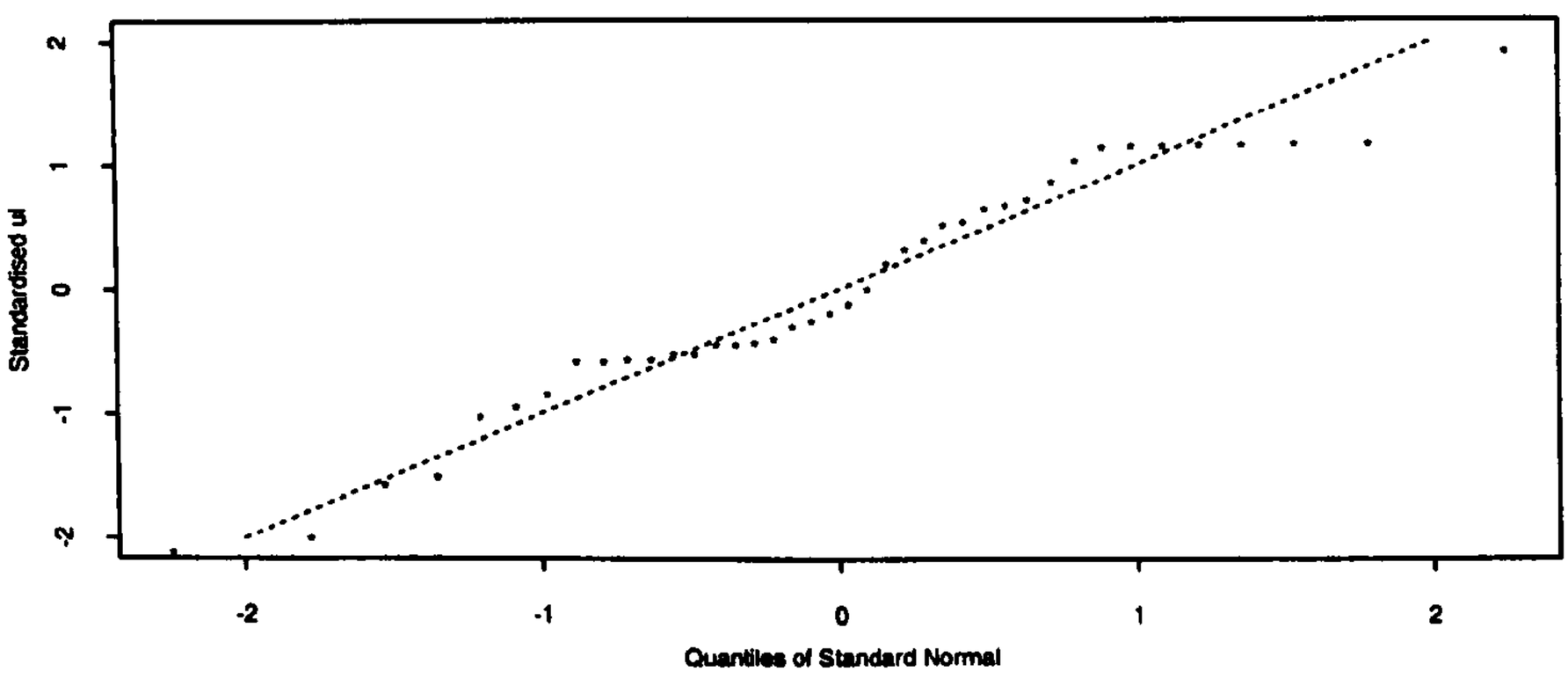

Figure 4.4: Normal plot of standardised level two residuals for analysis of reporting shortness of breath with an adjustment for baseline.

inconsistency of the ratio between the marginal and random effect parameter estimates, suggests that caution is needed with the use of random effect models for such analyses and more research is needed to determine what is causing the problems.

4.3.3 Conclusion

The work of this section detailed two classes of model for the analysis of binary response data: modelling of the marginal response with allowance for the data dependence via an appropriate working correlation matrix; or by assuming the dependence between observations on the same individual arises due to some latent process which is modelled as a random effect under some distributional assumptions. Each of the models have their advantages and disadvantages which have been discussed in the previous sections: the marginal model does not allow inferences to be made about the association parameters (the correlation structure of the data); perhaps more seriously, the random effect model is non-robust to failure of the assumptions about the distribution of its random effects which may be particularly problematic for repeated binary data. These factors aside, it should be recognised that the parameters estimated from the two models have very different interpretations and they should not be regarded as alternative ways of answering the same question.

4.4 The analysis of complex patterns of binary response

A problem that often occurs in the analysis of quality of life data relates to the potentially complex response functions over time which can occur when the prevalence of symptoms increases as a result of treatment. This was demonstrated by MRC Lung cancer working party (1989) reporting bouts of nausea and vomiting following chemotherapy. In such cases, plots of the marginal response over time give a very informative description but do not allow a formal treatment comparison or adjustment for covariates of interest. In this example a

marginal model is used in conjunction with a natural smoothing spline to give the same informative description of the data and in addition allows estimation and formal testing of a treatment difference. This is demonstrated by modelling the apparent treatment difference in the prevalence of dysphagia measured on the daily diary card following radiotherapy treatment in the MRC LU07 study, described in Chapter 2.

A summary of the data has already been given in Chapter 2 and involved daily reporting of dysphagia symptoms by patients over a 6 month period. For this example, the ordinal response of the diary card was dichotomised to give a binary response such that y_{ij}, the response of the *i*th subject on the *j*th day, takes the value 0 if the patient reported no symptoms (a score of 1 on the diary card), or 1 otherwise. A positive response can then be interpreted as the reporting of any symptoms of dysphagia. The marginal response profile shown in figure 2.16 highlighted a dramatic increase in the prevalence of symptoms following the start of treatment which fell back to baseline levels once radiotherapy had finished. The extent of this increase appeared different across the two treatment groups although no formal comparison was made. The objective was to obtain a model for these data which gave both a reasonable representation for the marginal response and allowed unbiased estimation of the apparent treatment difference. A generalised estimating equation for a logistic regression model was used to give robust standard errors taking account of the dependence of observations taken on the same subject, with a natural cubic spline to represent the complex shape in the marginal response. Due to computing constraints, only the first eight weeks of the follow-up period could be analysed. This period captured the entire period of treatment related symptoms and the return to baseline shown in figure 2.16. Before presenting the results of the analysis, the definition and use of cubic splines is first discussed.

4.4.1 Natural cubic splines

A cubic spline is a series of cubic functions which are joined together smoothly at a series of specified time points or *knots* in the follow-up period. The smoothness of the function is obtained by constraining the value of the 1st and 2nd derivatives of functions evaluated at their adjoining knots to be equal. A natural cubic spline has the additional restriction that it is linear outside the first and last knots. The advantage of splines is that a natural spline with p knots can be expressed with a design matrix with p parameters and can therefore be easily be incorporated into a multiple regression analysis.

By definition the natural cubic spline for a data series $y=(y_1,...,y_m)$ observed over a period $x_j, j=1,...,m$, with p knots at times $x=k_l,\ l=1,...,p$ is linear up to k_1, beyond which it is a series of cubic splines written,

$$y=\alpha_0+\alpha_1 x+\sum_{l=1}^{p}\beta_l\phi_l(x) \qquad (4.23)$$

where $\phi_l(x)=(x_j-k_l)^3$ for $x_j\geq k_l$, 0 otherwise. The additional constraints $\sum_{l=1}^{p}\beta_l=\sum_{l=1}^{p}\beta_l k_l=0$ ensure that the spline is linear for $x_j\geq k_p$ and mean that equation (4.23) can be re-written

$$y=\alpha_0+\alpha_1 x+\sum_{l=1}^{p-2}\beta_l\Phi_l(x) \qquad (4.24)$$

where the functions $\Phi_l(x)=\phi_l(x)-\frac{(k_p-k_l)}{(k_p-k_{p-1})}\phi_{p-1}(x)+\frac{(k_{p-1}-k_l)}{(k_p-k_{p-1})}\phi_p(x)$, for $l=1,...,p-2$, are straightforward to evaluate and give an $(m\times p)$ design matrix defining the spline which can be incorporated into a multiple regression analysis (Benjamin and Pollard, 1980). Alternative formulations of this design matrix are possible and are discussed in Hastie and Tibsharani (1990).

4.4.2 The analysis of transient dysphagia following radiotherapy

For this problem the natural cubic spline with knots at the limits of the data and at 7, 14, 21, 28 days were chosen to represent the shape of the marginal log odds of reporting symptoms.

This was assumed to be the same for each treatment group. The treatment difference was modelled as a constant log odds ratio. Defining $rt_i=1$ for the F2 treatment group, the marginal model for the $\theta_{ij}=E(y_{ij})$ can then be written

$$\text{logit}(\theta_{ij})=\alpha_0+\alpha_1 x_{ij}+\sum_{l=1}^{4}\beta_l\Phi_l(x_{ij})+\delta rt_i \qquad (4.25)$$

As a naive analysis for comparison purposes, equation (4.25) was also fitted excluding the spline and assuming an independence working correlation matrix, giving a model for a constant log odds and log odds ratio through time. This model will be referred to as the *constant* model. For the analyses including the spline, three different structures for the working correlation were used: independence; exchangeable; and autoregressive of order 1. These models will be referred to as the *independence*, *exchangeable* and *AR1*. Unfortunately, due to computing constraints, it was impossible to iteratively fit the AR1 model using GEE1 procedure described in Section 4.2.2. Instead, $\boldsymbol{R}_i(\rho)$ was assumed fixed with lag one correlation, $\rho=0.8$. This value was determined from the lag one correlation of the Pearson residuals taken from the independence model. All results are given in table 4.4 with the fitted marginal response over time plotted in figure 4.5.

Table 4.4: Estimated treatment difference in reporting symptoms of dysphagia using GEE1 with a natural spline.

Model	δ	Naive SE	Robust SE	OR [95% CI]	ρ
Constant	-0.25	0.04	0.18	0.78 [0.55, 1.12]	-
Independence	-0.25	0.04	0.20	0.77 [0.52, 1.14]	
Exchangeable	-0.77	0.17	0.34	0.46 [0.24, 0.91]	0.50
AR(1)	-0.21	0.11	0.20	0.81 [0.55, 1.19]	0.80

The estimates for ρ given in the last row of the table refer to the assumed correlation between successive units for the exchangeable and the lag one correlation for the AR(1) working correlation matrices. Due to computing constraints, that for the AR(1) model was fixed and therefore not updated during estimation.
The 95% CI for the odds ratio is given with a based on the robust standard error.

The estimated log odds ratios given from the constant, independence and AR(1) models were all very similar translating to about a 20% lower odds of dysphagia for patient receiving the F2 course of treatment as opposed to FM. There was no evidence to suggest that this was not due to chance. The exchangeable model, however, gave a very different picture, estimating a 54% lower odds in the F2 group with a 95% CI which excluded an odds ratio of 1. From figure 4.5 it can be seen, however, that this model did not fit the data very well, consistently over estimating the response in the FM group and underestimating the response in the F2 group.

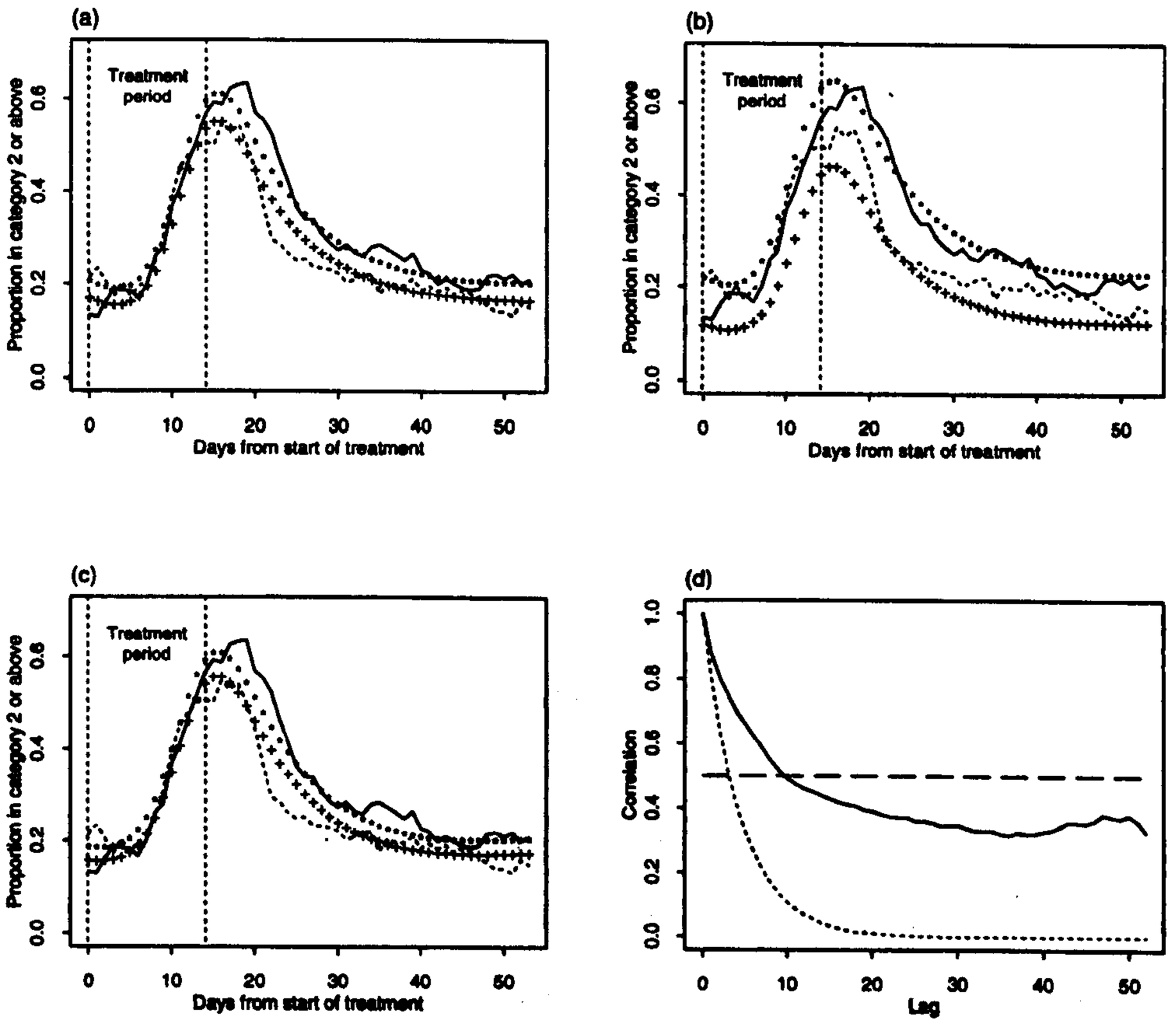

Figure 4.5: Fitted marginal profiles for the MRC LU07 dysphagia data using (a) independence; (b) exchangeable; (c) AR1 working correlation matrices. The observed lagged correlation of the Pearson residuals from the independence model: ————; and the correlation for the exchangeable: — — — —; and AR1:-------- working correlation matrices are shown in (d). Within (a), (b) and (c), the observed profiles for the multiple fraction radiotherapy (FM): ————; two fraction radiotherapy: ---------. The symbols * and + give the fitted profiles for the multiple fraction (FM) and two fraction (F2) radiotherapy groups respectively.

This was due to a misspecification of the correlation structure of the data, in particular between observations close together which formed the bulk of the data. This is shown in figure 4.5(d) which plots the observed correlation between Pearson residuals from the independence model for different lag times. In terms of the fit of the each model, determined by comparing the observed and fitted values shown in figure 4.5, the AR1 model was the best as it tended to an asymptote of zero slope at the limits of the data whereas the independence model showed a slight upward turn of the spline at the lower limit.

Having adjusted for the shape of the response function over time, this simple analysis gave no evidence of a constant difference in log odds over time in the reporting symptoms of dysphagia. However, examination of the data and knowledge about the nature of symptoms, would suggest that a constant treatment difference is unlikely. Two alternative extensions to model a non constant treatment difference were considered. The first was a quadratic treatment:time interaction. Like the constant treatment difference, it is addressing a very specific hypothesis about the nature of response. The second model used two different splines for each treatment group, and thus tested the more general hypothesis of some difference in response between the two groups. Significance in each case was assessed by constructing a Chi squared statistic for simultaneous contrasts on c df given by

$$(\boldsymbol{Q}\hat{\beta})^{T}(\boldsymbol{Q}\boldsymbol{V}_{\hat{\beta}}\boldsymbol{Q}^{T})^{-1}\boldsymbol{Q}\hat{\beta} \qquad (4.26)$$

where $\boldsymbol{Q}$ is a $(c \times p)$ matrix of c contrasts for p parameters, and $V_{\hat{\beta}} = \text{var}(\hat{\beta})$. For the quadratic treatment effect this had 2 df, for the more general hypothesis it had 5 df. Each model was fit using an AR1 working correlation structure with $\rho=0.8$. The results are summarised in table 4.5. Both naive and robust results are given based on the naive and robust estimated $\text{var}(\hat{\beta})$.

Based on the robust results, no evidence of a quadratic treatment difference was seen

Table 4.5: Chi-squared statistic (p value) for testing non constant treatment differences for the MRC LU07 dysphagia data.

	Naive		Robust	
	Chi-squared statistic	***p* value**	**Chi-squared statistic**	***p* value**
Quadratic	7.38	(0.025)	2.68	(0.262)
Different splines	**13.4**	**(0.020)**	**12.02**	**(0.035)**

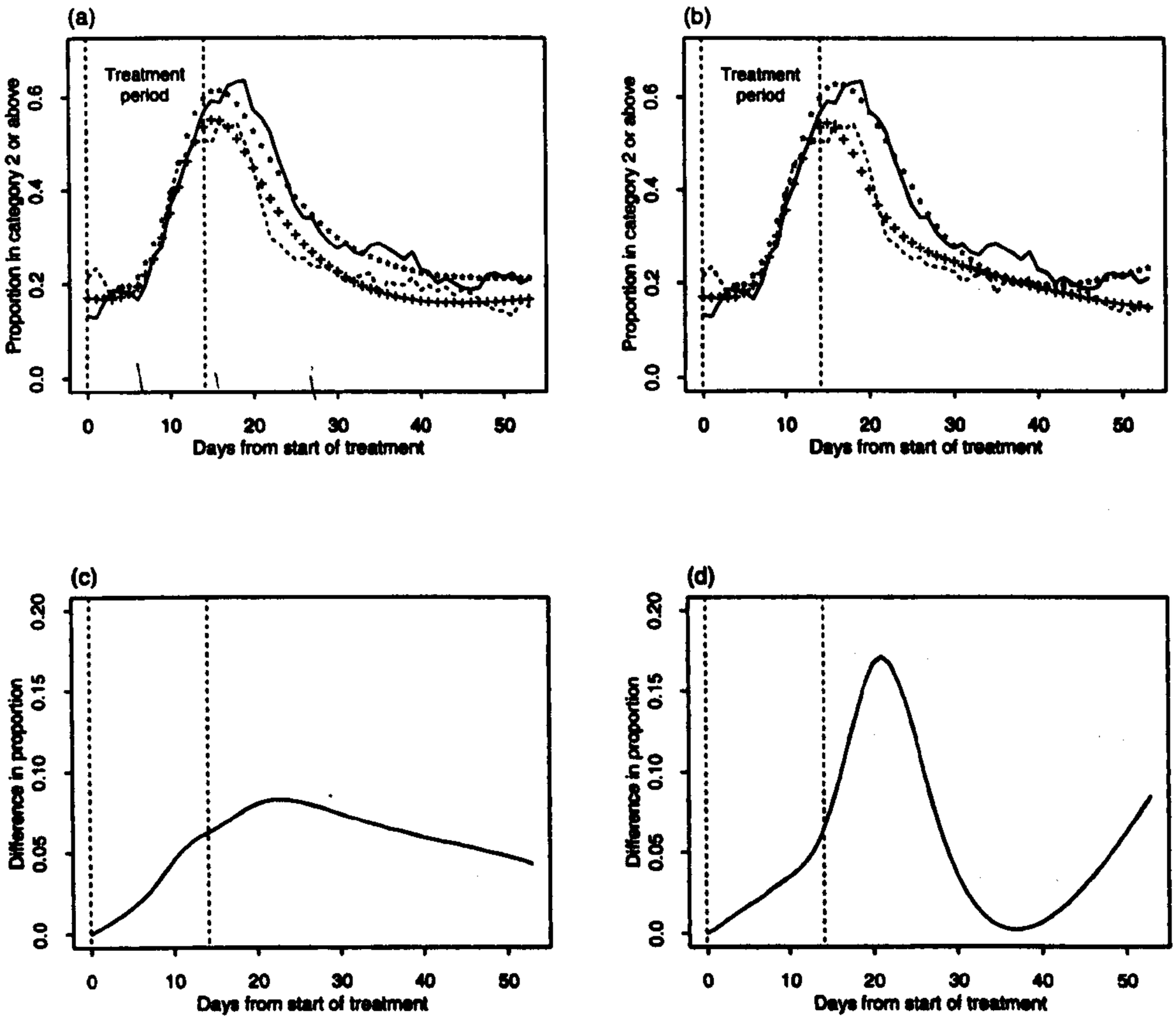

Figure 4.6: Estimated response profiles for MRC LU07 dysphagia for non constant treatment difference fitted using (a) quadratic; (b) different splines for each treatment group, (c) and (d) show the realised estimated treatment difference in proportions for each model respectively.

(p=0.262). However, there was some evidence of a difference in shape although this was not very convincing (p=0.035). The fitted profiles for each of these models are given in figure 4.6. Also shown on the figure is the shape of the estimated treatment difference. This is particularly important for the more general hypothesis because it is the only way to visualise the estimated effect. In terms of fitting the data, the quadratic treatment effect was the most satisfactory with some evidence of over fitting seen when different splines were used with the fitted profile for the multiple fraction group starting to upturn. The effect of this upturn is shown quite dramatically to affect the estimated treatment difference shown figure 4.6(d). As a result of this it was not felt that the results of the model with different splines could be relied upon in terms of assessing the difference between the two groups.

4.4.3 Conclusion

This analysis has demonstrated the use of cubic splines for accommodating complex patterns of response into a marginal model using a natural spline. It has been shown (Ford *et al.*, 1995) that omitted covariates in a logistic regression analysis can result in biased and inefficient estimation of other covariates of interest. Therefore, the purpose of this analysis was to incorporate the obvious behaviour of the marginal response over time into the analysis in order to obtain an unbiased and efficient estimate of the treatment effect having adjusted for the behaviour in response over time. On the whole the analysis proved successful, in particular, a good representation of the response over time was obtained. However, when assumed constant over time, little difference was seen in the estimated treatment effect from a model which ignored the response over time and that from the model incorporating a natural spline. This was the case both in terms of the treatment estimate and its precision.

The main drawback with the use of splines in an analysis such as this is that the knots have to be chosen. If the shape of response over time is simply a nuisance part of the analysis, as in

this case, then this is not of great concern. Otherwise, the sensitivity of the results to the choice of knots position should be assessed. In this example they were arbitrarily chosen at weekly intervals from the start of treatment. Further analyses with the knots chosen at more specific time points based on close examination of the data (k=0, 9, 16, 23, 30, 53 days) did not impact greatly on the results presented in table 4.4 although as would be expected the better fitting spline did slightly increase the precision on the estimated treatment effect. For example, for the AR(1) model the estimated log odds ratio was -0.22 (robust SE=0.20), translating to an odds ratio of 0.80, 95% CI=[0.55, 1.18]. The second analysis, when the treatment difference was modelled in terms of a difference in shape using two different splines, was not treating the splines as nuisance parameters however, and therefore inferences based on such analyses need careful examination. Generally it is perhaps better not to use such models and to concentrate on the more specific hypotheses about treatment differences.

It should be noted that this model could just as easily have been fitted using a random effects model, but for the objectives here - to give a clear description of the marginal response over time and an estimated treatment comparison - the marginal model was more appropriate. In other situations - for instance when the timing of treatment is variable for different subjects - a random effect model with different splines for each subject may be a suitable model to choose. Such models have been used to some degree by Kenward and Welham (1996). Further work would be needed to determine whether they may be of use for the analysis of quality of life data. Similarly, alternative models could have been applied to model the shape of the data, in particular fractional polynomials (Royston and Altman, 1994). For this analysis, little success was achieved with such models.

4.5 Multivariate binary outcomes - extensions to the multilevel model

As demonstrated in the continuous case, the simple two level model for repeated observations can be easily extended to a three level model to incorporate multivariate outcomes with the dimensions as level one, occasions at level two and subjects at level three. This extension may be particularly useful in the case of binary outcomes as it allows the analysis of individual item responses within a questionnaire instead of summarizing the questionnaire by a single summary score. This not only gives a greater insight into the individual aspects of quality of life, it may also give a more intuitive outcome measure - the odds of symptoms - rather than an arbitrary score. In addition, it allows the estimation of the degree of association between the different items thus aiding the understanding of the behaviour of quality of life throughout follow-up. This is particularly important in the binary data case when, as discussed in Section 2.3.4, it becomes much more difficult to display or summarize the data.

To demonstrate the use of the multivariate model, an analysis of the individual item responses of the RSCL taken from the NSCLC study was used. Each individual item response was again dichotomised from the original four point scale to a two point scale as follows: slight (0,1), moderate (2,3). For simplicity, a model with a constant odds of symptoms over time was assumed and only six out of the possible thirty-six items, chosen to be of particular interest to patients with lung cancer were analysed. These items were: *pain in the chest*, *heartburn*, *cough*, *shortness of breath*, *dry mouth* as well as *anxiety*. Baseline responses were also incorporated into the analysis. Therefore the data used were restricted to the 40 subjects who gave a baseline (pre-treatment) and at least one post treatment response to the RSCL. These were divided between the two treatment groups in the ratio 17:23. All analyses used first order PQL restricted IGLS estimation.

The aim of the analysis was to investigate the level of reported symptoms over the follow-up

period and how this related to the treatment course received, as well as how the response for different symptoms were related to each other.

4.5.1 Extension of the two level model for binary data

The two level model for repeated binary outcomes is extended to three levels for multivariate repeated binary outcomes in the same way as shown for the continuous case in Section 3.4. The dimensions for the level one units are clustered within occasion at level two, within subjects at level three. Again the differences between dimensions are assumed fixed, and the full covariance between dimensions at level three is estimated. It is at level two that the difference lies because the residuals at that level have binomial distribution and their variances therefore are known given the marginal parameters.

As for the continuous case, the model is constructed using a series of dummy variables for the response in each dimension. For example, to investigate the effect of covariate x_i in each dimension, $l=1,...,L$, a model for $E(y_{ijl})=\theta_{ijl}$ is

$$\text{logit}(\theta_{ijl})=\sum_{l=1}^{L}\{\alpha_l+\beta_l x_{il}+u_{il}\}z_{ijl}^{(l)} \qquad (4.27)$$

where $z_{ijl}^{(l)}=1$ for dimension l, 0 otherwise, $l=1,...,L$. At level three, $\text{var}(u_i)$ is an $(L\text{x}L)$ matrix with diagonal elements, $\text{var}(u_{il})=\sigma_{ul}^2$, and off diagonal elements, $\text{cov}(u_{il},u_{ik})=\sigma_{ulk}$, at row l, column k, for dimensions $l\neq k$. At level two, $\text{var}(e_{ij})$ has on its diagonal, $\text{var}(e_{ijl})=\sigma_{el}^2\theta_{ijl}(1-\theta_{ijl})$, with $\sigma_{elk}=\text{cov}(e_{ijl},e_{ijk})$ as the off diagonal elements for $l\neq k$. In the following example, the over dispersion parameter σ_{el}^2 is constrained to equal one. As in the continuous case, by setting all the off diagonal elements to zero, the model is analogous to fitting L univariate models.

4.5.2 Reporting of patient symptoms on the RSCL in the CRC NSCLC study

The objectives of this analysis were to investigate the prevalence of reporting each of the

six chosen symptoms of the RSCL between the two treatment groups and over time. It will also yield an overall global estimate for difference in reporting of symptoms in the two treatment groups if appropriate, as well as information about the association across different symptoms. For example, if a patient is reporting chest pain at a particular time, what other symptoms are they likely to report (within subject across symptom correlation)? Similarly, if a patient has a higher than average odds of reporting symptoms of chest pain over the period, is their odds of reporting a second symptom similarly high (between subject across symptom correlation)? A sequence of four models were used. Marginal profiles giving the odds of reported symptoms at each time are shown in figure 4.7. They show a large degree of fluctuation over time with a slight downward trend across all dimensions and very little difference between the two treatment groups.

Model one - constant odds

The initial model assumed a constant dimension specific log odds over time which was allowed to vary across subjects, written

$$\text{logit}(\theta_{ijl}) = \sum_{l=1}^{6} \{\alpha_l + u_{il}\} z_{ijl}^{(l)} \qquad (4.28)$$

Model two - adjustment for baseline

An adjustment for baseline symptoms at baseline was then added to the model in the form of six dummy variables, $base_{il}=1$ if a patient had symptom l at baseline, 0 otherwise, for $l=1,\ldots,6$.

$$\text{logit}(\theta_{ijl}) = \sum_{l=1}^{6} \{\alpha_l + \gamma base_{il} + u_{il}\} z_{ijl}^{(l)} \qquad (4.29)$$

Model three - constant treatment difference in odds

A constant treatment difference in log odds was then added. Initially six separate treatment covariates for each symptom were fitted where $rt_{il}=1$ for continuous, 0 for the split course.

$$\text{logit}(\theta_{ijl}) = \sum_{l=1}^{6} \{\alpha_l + \gamma base_{il} + \delta_l rt_{il} + u_{il}\} z_{ijl}^{(l)} \qquad (4.30)$$

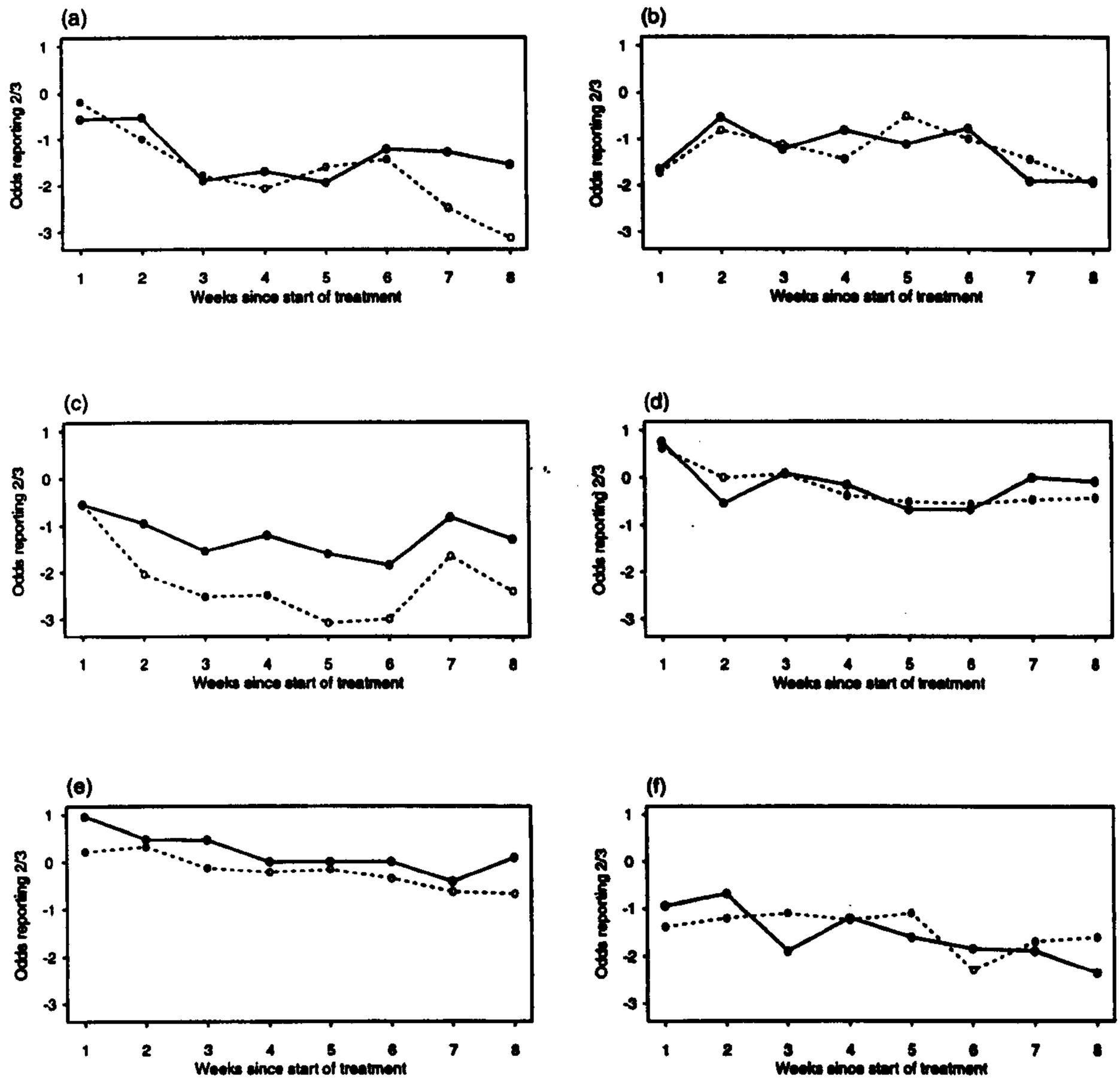

Figure 4.7: Marginal profiles for the odds of reporting symptoms of (a) chest pain; (b) heartburn; (c) anxiety; (d) cough; (e) shortness of breath; (f) dry mouth recorded on the RSCL in the CRC NSCLC study.

A Chi-squared test for heterogeneity of dimension specific treatment difference was constructed. Given no evidence of heterogeneity, a single treatment effect was used in the model.

Model four - linear trend over time

Finally a linear trend over time was added to the model. Again six separate parameters were used, occ_{il}, coded 0 to 7 for weeks 1 to 8, for the lth symptom, 0 otherwise. Consistent with the heterogeneity test of the previous model, a single treatment covariate across dimensions was

used.

$$\text{logit}(\theta_{ijl}) = \sum_{l=1}^{6} \{\alpha_l + \gamma base_{il} + \beta_l occ_{ijl} + u_{il}\} z_{ijl}^{(l)} + \delta rt_i \qquad (4.31)$$

Again a Chi-squared test for heterogeneity of dimension specific linear trends was constructed. In this case, there was strong evidence of a difference in the effects and hence a single covariate was not fitted.

The results from the analysis are given in table 4.6.

The overall odds of five of the six symptoms was low, signified by estimated odds less than one for all but *shortness of breath*. The estimated between subject variance for each symptom was, however, very high in all cases, illustrating large differences in the odds of symptoms between subjects. For example, the estimated variance of 2.03 for the log odds of -1.82 of grade 2/3 chest pain gave an estimated 95% reference range for the subject specific odds of reporting such symptoms of [0.01, 2.63]. This translated to subject specific range for the probability of [0.01, 0.72]. On addition of the symptoms at baseline, a reduction in the between subject variance was seen for four of the six symptoms: *anxiety*, *cough*, *shortness of breath* and *dry mouth*. Similarly for these symptoms, the fixed parameter estimates for this model showed strong relationships between reporting symptoms at baseline and subsequently. Although such strong relationships with baseline response were also seen for the remaining symptoms of *chest pain* and *heart burn*, their respective variances increased in model two.

The addition of a separate treatment effect in each symptom dimension showed a marginally higher odds of anxiety in subjects on the split course therapy with the reverse in each of the remaining symptoms. With the exception of *dry mouth*, which was marginally significant, these

Table 4.6: Fixed parameter and between subject variance estimates (SE) for four sequential multivariate binary models for individual items on the RSCL in the CRC NSCLC study.

	1. Constant odds	2. Baseline adjustment	3. Treatment difference	4. Rate of change over time
	Log odds (SE)	**Odds [95% CI]**	**Odds ratio [95% CI]**	**Odds ratio [95% CI]**	**Odds ratio [95% CI]**
Fixed parameter estimates					
Chest pain	-1.82 (0.30)	0.16 [0.09, 0.30]	0.11 [0.05, 0.24]	1.08 [0.292, 3.97]	0.74 [0.63, 0.88]
Heart burn	-1.45 (0.30)	0.24 [0.13, 0.43]	0.18 [0.09, 0.35]	1.02 [0.27, 3.81]	0.86 [0.74, 1.00]
Anxiety	-1.78 (0.34)	0.17 [0.09, 0.33]	0.06 [0.02,0.15]	0.60 [0.14, 2.50]	0.92 [0.79, 1.08]
Cough	-0.29 (0.31)	0.75 [0.41, 1.37]	0.13 [0.05, 0.35]	1.09 [0.33,3.68]	0.93 [0.81, 1.06]
Shortness of breath	0.11 (0.45)	1.12 [0.46, 2.73]	0.08 [0.02, 0.28]	1.24 [0.23, 6.82]	0.67 [0.56, 0.80]
Dry mouth	-1.61 (0.37)	0.20 [0.10, 0.42]	0.10 [0.05, 0.23]	4.52 [0.94, 21.7]	0.85 [0.73, 0.99]
			Chi-squared test	7.34 (p=0.02)	18.6 (p=0.002)
Single parameter estimate over all dimensions			**Estimate**	0.94 [0.38, 2.32]	Not applicable
Between subject (level three) variance estimates					
Chest pain	2.03 (0.78)		2.21 (0.82)	2.27 (0.84)	2.87 (0.99)
Heart burn	2.33 (0.81)		2.60 (0.88)	2.66 (0.89)	2.37 (0.81)
Anxiety	2.83 (1.00)		2.55 (0.96)	2.61 (0.97)	3.01 (1.08)
Cough	2.70 (0.82)		2.25 (0.73)	2.32 (0.75)	2.34 (0.75)
Shortness of breath	6.24 (1.77)		4.35 (1.39)	4.43 (1.42)	5.93 (1.84)
Dry mouth	3.74 (1.22)		3.42 (1.18)	3.58 (1.22)	4.00 (1.32)

The odds ratios for models 2, 3 and 4 are given for: symptoms versus no symptoms at baseline; continuous versus split course radiotherapy; and a one week increase in time. Estimated from a random effect model, they reflect odds ratios for given underlying subject specific odds.

estimates were all small in comparison with their standard error and gave no evidence of a difference in the odds of all symptoms between the two treatment groups. A test for homogeneity of the treatment difference across all symptoms gave no evidence against the null hypothesis (Chi-squared=7.34 on 5 df, p=0.20). Fitting a single treatment parameter gave a combined odds ratio estimate of 0.94, 95% CI=[0.38, 2.63], interpretable as a 6% reduction in a subject's odds of symptoms if they were to receive the continuous course as opposed to the split course.

Fitting a linear trend in log odds over time suggested a fall in the odds of symptoms in all dimensions over follow-up. This trend was particularly evident for *shortness of breath* with an estimated 33% reduction per week in the subject specific odds of reporting such symptoms, 95% CI=[18%, 46%] reduction. Testing for a homogeneity of linear trend across all symptoms gave a Chi squared statistic of 18.6 on 5 df (p=0.002). A single occasion effect was therefore not fitted. The estimated overall treatment effect in this model was relatively unchanged from that in model three (0.91, 95% CI=[0.36, 2.27]).

The estimated covariance structure of the data is of particular interest in this analysis and is its advantage over six separate univariate analyses. The estimates given from model four with a single treatment effect are presented in table 4.7. The first half of the table gives the correlations between subjects which measure the degree of association between the level three residuals, u_{il}, across symptoms. The estimated covariances with standard errors are given below the diagonal, translated to correlations above the diagonal. These give an indication whether a subject with a higher than 'average' odds of one symptom has a higher than average odds for a second symptom. Those in the bottom half of the table were estimated within subject. Since σ^2_{el} was constrained to be equal to one for l=1,...,6, the covariance and correlation are equal and indicate whether subjects who have higher than expected odds of one

symptom at a particular occasion have a similarly higher than expected odds in a second.

Between subjects, a high degree of positive association was seen between most symptoms.

Table 4.7: Covariance (SE) and correlation estimates between and within subjects for the reporting of symptoms on the RSCL for the CRC NSCLC study.

	Chest pain	**Heartburn**	**Anxiety**	**Cough**	**Shortness of breath**	**Dry mouth**
Between subjects						
Chest pain	**2.87**	0.52	0.51	0.64	0.44	0.56
Heartburn	1.34 (0.71)	**2.37**	0.24	0.44	-0.22	0.35
Anxiety	1.50 (0.83)	0.64 (0.70)	**3.01**	0.55	0.48	0.34
Cough	1.65 (0.71)	1.04 (0.61)	1.46 (0.72)	**2.34**	0.49	0.73
Shortness of breath	1.80 (1.07)	-0.82 (0.88)	2.00 (1.11)	1.81 (0.93)	**5.93**	0.57
Dry mouth	1.89 (0.93)	1.07 (0.79)	1.17 (0.91)	2.21 (0.83)	2.78 (1.28)	**4.00**
Within subjects						
Chest pain	**1.00**					
Heart burn	0.38 (0.05)	**1.00**				
Anxiety	0.57 (0.40)	0.33 (0.06)	**1.00**			
Cough	0.38 (0.05)	0.34 (0.06)	0.35 (0.05)	**1.00**		
Shortness of breath	0.49 (0.05)	0.40 (0.05)	0.45 (0.05)	0.43 (0.05)	**1.00**	
Dry mouth	0.62 (0.03)	0.51 (0.04)	0.59 (0.04)	0.42 (0.05)	0.61 (0.04)	**1.00**

Within the (6x6) block between subjects, the covariance estimates are given below the diagonal, the variance estimates on the diagonal, and the correlation estimates above the diagonal. Given $\sigma^2_{el}=1$, for $l=1,\ldots,6$, within subjects, the covariance and the correlation are equal and are given below the diagonal.

However, the estimated correlation observed between *shortness of breath* and *heartburn* was negative. The corresponding covariance was however, very small in relation to its standard error. The scatter plot matrix in figure 4.8 of the level three residuals from model 4 illustrates these associations. Contrasted with the very crude representation of these correlations in figure 2.17, figure 4.8 demonstrates a major advantage of using the three level model to analyse these data giving a very clear impression of the association across dimensions within the data.

Similarly the multivariate model also gives estimates of covariances (correlations) between dimensions within subjects. These estimates (also shown in table 4.7) are adjusted for an individual's expected response and show a degree of positive association between symptoms which was not at all visible from the crude representation in figure 2.18. Unfortunately the

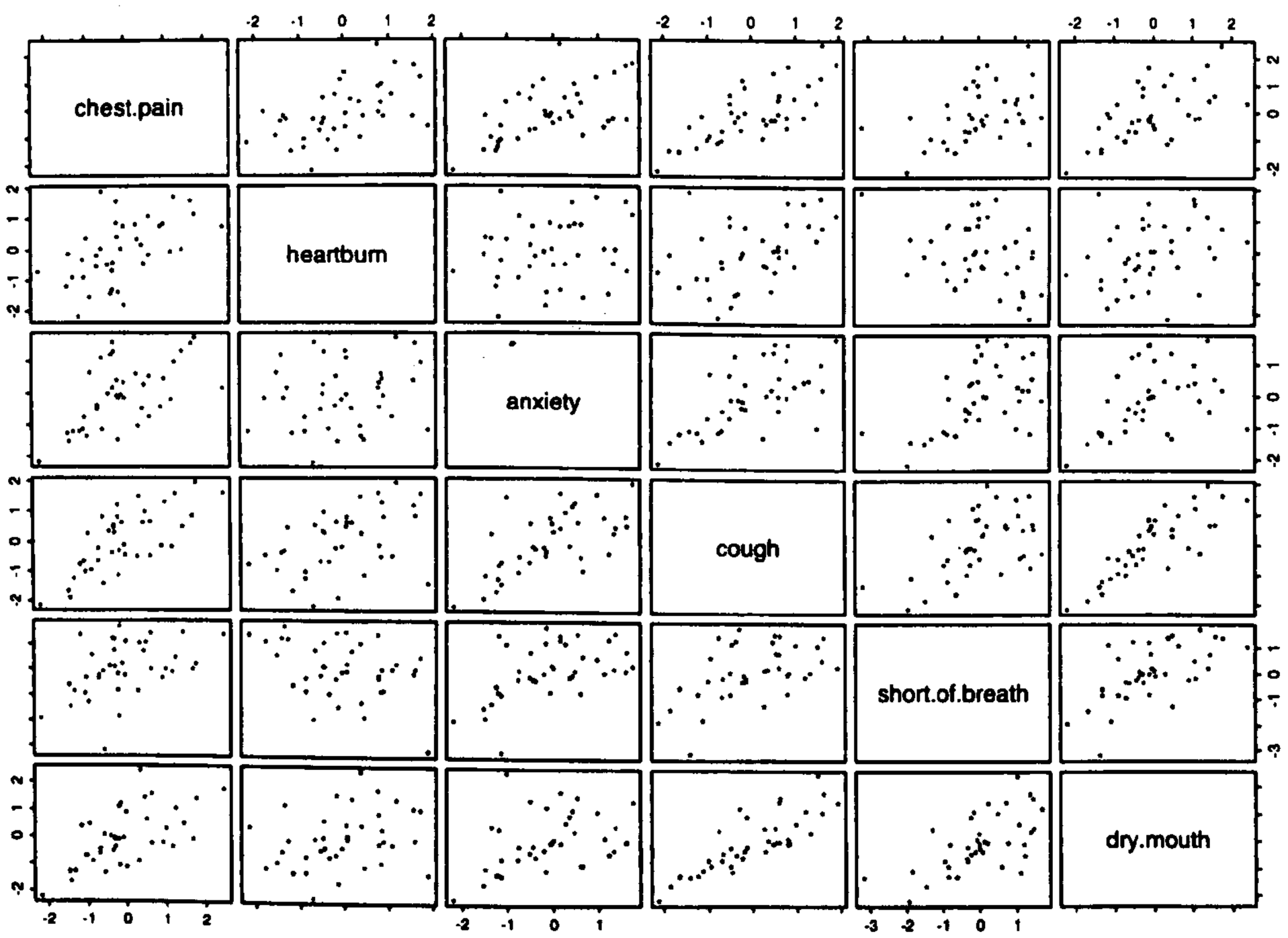

Figure 4.8: Scatter plot matrix of level three residuals of model four showing the between subject across dimension correlations for six symptoms measured on the RSCL in the CRC NSCLC study.

Binomial nature of the residuals at this level does not allow a similar graphical representation of these estimates.

As demonstrated in Section 4.3, residual diagnostics for these random effect models are extremely important given the dependency of the results to the assumed distribution of random effects. Univariate Normal plots and a multivariate Gamma plot are shown in figure 4.9. In contrast to those of the earlier example, the univariate distributions showed reasonable Normality. The Gamma plot in figure 4.9(g) shows the squared distances of the estimated residuals calculated using the fitted level three covariance matrix and shows their distribution to be much less dispersed than expected. That shown in figure 4.9(h) shows the squared distances calculated using the empirical covariance matrix of the estimated residuals. This closely resembles the expected Chi-squared distribution and highlights an important further area of work to determine the distribution of these estimated residuals in order to aid future model checking.

4.5.3 Conclusion

This analysis showed evidence of a fall in the level of reported symptoms over the follow-up period after the start of treatment and no evidence of differences in the odds of the six symptoms according to treatment received. For subjects who reported symptoms at baseline, their odds of reporting symptoms following the start of treatment was increased. A high degree of correlation was seen between symptoms both between and within subjects.

In general the multivariate model allows a greater insight into the subject response on the RSCL than is gained from analysing the summary scores as a continuous measure. Further it gives an intuitive outcome measure, allowing the discussion of the odds of symptoms and how this changes over time rather than, say, a change in the total RSCL score. However, the model

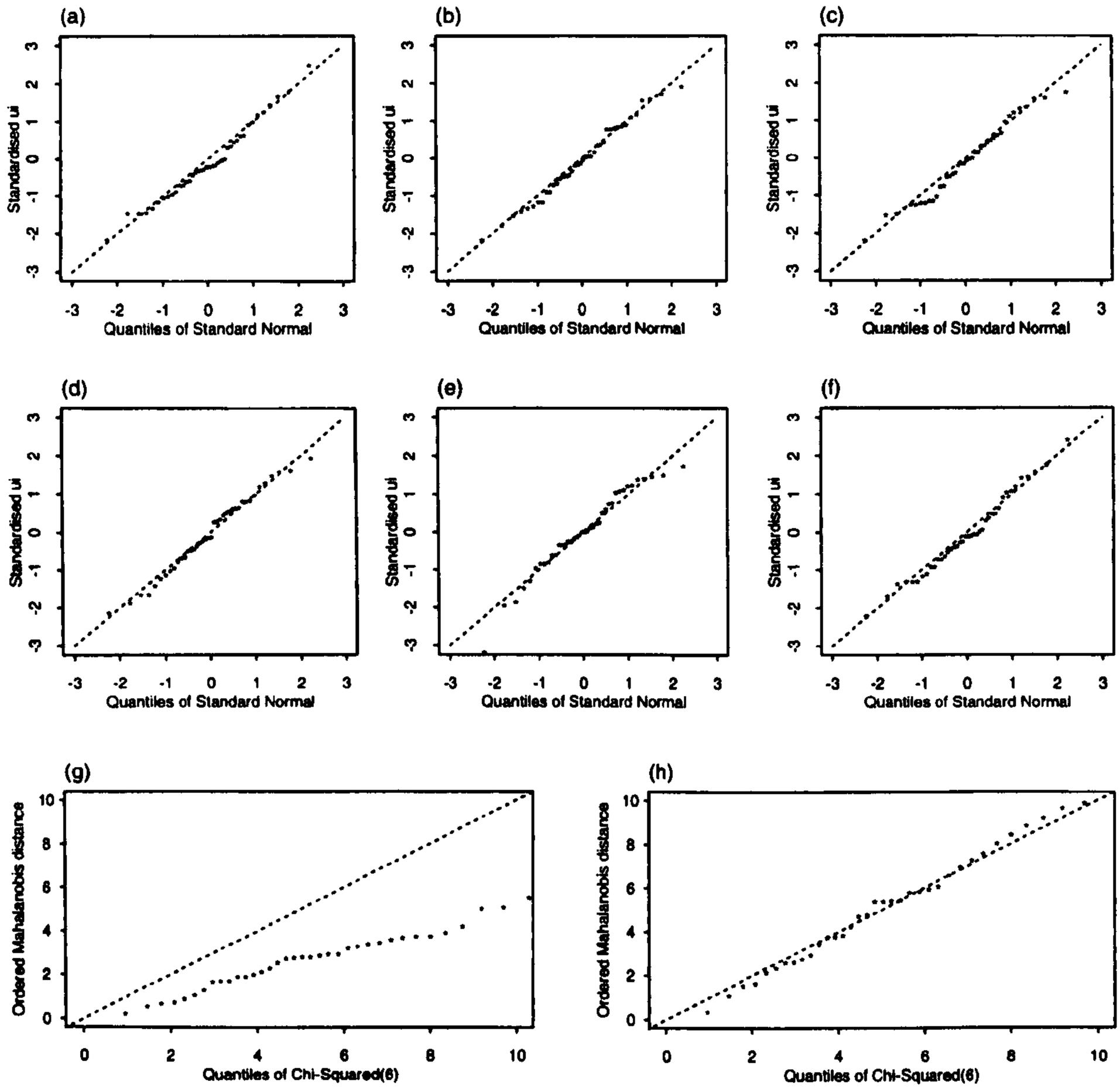

Figure 4.9: Residual diagnostics for the multivariate binary model for the CRC NSCLC study showing univariate Normal plots for (a) chest pain; (b) heartburn; (c) anxiety; (d) cough; (e) shortness of breath; (f) dry mouth; (g) a Gamma plot for multivariate Normality based on the estimated level three covariance matrix; and (h) a Gamma plot for multivariate Normality based on the empirical covariance matrix of the estimated level three residuals.

will be sensitive to the Normality assumptions of the level three residuals and so model checking is necessary.

4.5 Summary and discussion

This work has reviewed the use of marginal and random effect models (or more specifically, hierarchical models), both of which can usefully help the analysis of repeated binary outcomes in quality of life data. Given the difference in their interpretation they should not be thought of as alternative competing ways of answering the same question.

Marginal models give overall population covariate effects and treat the associations between repeated observations on the same subject as nuisance parameters. Such models are very simple to use. In particular, they can be very easily extended to model complex patterns of binary response over time which are often seen in quality of life data. The interpretation of their parameter estimates is also familiar as they can be considered in the same way as those of a cross-sectional analysis. However, if information about the variance structure of the data, that is, the dependency between observations on the same subject, is required, more complex specifications of these model are needed which may be impractical when the number of measurement occasions is high. The models presented here represented a standpoint in middle ground, where some attempt to model the correct correlation structure was attempted in order to maximise the efficiency of the parameter estimates. Most simply, it is possible to ignore the correlation structure altogether in parameter estimation, treating the data as independent and using analysis techniques for independent binary outcomes. Using jack-knifing techniques (Efron and Tibshirani, 1993), whereby the analysis is repeated n times with one subject removed each time, robust standard errors can then be obtained from the standard deviation of the n jack-knife estimates. These results will be very close to those of a GEE1 with an assumed independence 'working' correlation matrix.

Random effect models give subject specific covariate effects. That is, the estimated effect of the covariate of interest is on a patient's underlying log odds of positive response which is

treated as a random effect from some underlying distribution. In the work presented here, and within much of the available software for these models, the random effects are assumed Normally distributed. An advantage of the random effect model is that by integrating out over these random effects, the random parameters can be transformed to give marginal inferences. However, the ability to do this relies on knowing the correct distribution of random effects which is the main problem of the random effect model - its results are not robust to failure in the assumed distribution of the random effects. For longitudinal binary data, this can be a major problem when large numbers of patients give complete positive or complete negative responses over the whole follow-up period, as was demonstrated in Section 4.3. Further work is therefore needed to determine suitable methods to accommodate this feature of the data for more satisfactory modelling of such data. It was shown in Section 4.3.2 that adjusting for baseline may sometimes be a simple way to handle the problem. Alternatively, a different choice of distribution for the random effects may be used. Such models have been discussed by Lee and Nelder (1996), although their formulation is not accessible when the number of measurement occasions is high. Alternatively, Monte-Carlo simulation methods may be used for full likelihood estimation under more general distributions for the random effects (Gilks *et al.*, 1993).

Despite the problem of specification of the distribution of random effects, the possibility of modelling several dimensions or symptoms within a multivariate repeated measurement model could be very valuable for the analysis of quality of life data. Despite its complexity, this analysis not only allows examination of the behaviour of response for different symptoms of interest, by using a single treatment covariate over all symptoms, it also gives an overall estimated effect of treatment which is perhaps more informative and intuitive than a difference in the summary scores. Given that the response to the same symptoms is under consideration, it would present an estimate which may be compared across studies.

Other methods of analysis are also possible. In particular, Zeger and Liang (1992), and Diggle *et al.* (1994) both consider the use of transitional models in which the dependency between observations is incorporated into the analysis by explicitly conditioning on previous observations by including them as covariates in the model. In addition, summary statistic analyses could be used. For example, the proportion of days spent with symptoms calculated for each subject could be analysed within a logistic regression analysis. As discussed in Chapter 3, with unbalanced data, the analysis would have to suitably weight each subject specific estimate in terms of its precision. Again, the parameter estimates from these models will have a specific interpretation highlighting that, with binary data, it is important that the required focus of research is first determined in order that the most suitable analysis can be completed. It should be noted however, that although unfamiliar, within a clinical trial it is often the subject specific treatment effects (obtainable from a random effects analysis) that are of scientific interest.

The main restriction with the work demonstrated here is that quality of life data is perhaps more commonly measured on an ordered categorical scale rather than as a binary response. Although, as done in the examples here, such scales may be dichotomised to form a binary outcome, such an approach is not the most satisfactory. This problem is addressed in the next chapter which concentrates on the analysis of repeated ordered categorical outcomes.

5 The Analysis of Repeated Ordered Categorical Data

5.1 Introduction

Many quality of life items are measured on an ordered categorical scale which may, in some instances, be combined to give an overall summary score as analysed in Chapter 3. In cases when this is not appropriate (as with the daily diary card) or when it is desirable to examine particular items on a questionnaire in more detail, a common approach may be to dichotomise the score into a simple binary response and apply analyses of the type described in Chapter 4. However, both the combination of items to a summary score and the dichotomisation of the score, may result in a loss of information as well as power. A solution is to consider analyses of all possible dichotomies, and then summarise these to give an overall estimate of the effect of interest. A more parsimonious model is to assume the effect of covariates is the same regardless of the position of the cut off defining the dichotomy, reducing the question to a binary data problem with correlated errors. Such analyses are well developed for cross-sectional data. For example, the *cumulative odds* model, proposed by McCullagh (1980), considers the probability of being in category k or below, for k=1,...,K-1, and the *continuation ratio* model, a generalization of the Cox proportional hazard survival model, summarises the conditional probability of being in category k given a response in category k or below for k=2,...,K (Armstrong and Sloan, 1989). The critical assumption of both of these models is the lack of an interaction between the choice of cut point and covariates with the differences between cut points are treated as nuisance parameters.

The extension of these models for repeated measurements has been discussed by several

authors. Grizzle *et al.* (1969) proposed a weighted least squares (WLS) approach which considered the full multinominal distribution generated from repeated ordinal data thus giving equivalent maximum likelihood estimates for the parameters of interest in the models above. Alternatively, Kenward *et al.* (1994) have used a Dale model (Dale, 1986) to specify fully the multinominal distribution of the data for direct maximum likelihood estimation of the model parameters. A problem with the WLS analysis is that the data need to be stratified by all possible combinations of the covariates, thus requiring continuous covariates to be categorised. In addition, as the number of covariates, measurement occasions or response categories increases, the observed cell counts fall which can lead to estimation difficulties. The Dale model used by Kenward *et al.* (1994) is similarly restricted by the number of repeated measurements that can be easily handled (Lessaffre *et al.*, 1996) and has the additional problem of being computationally non-trivial to apply in practice. A further approach which has been suggested is to consider the model as a generalised linear model with a correlated error structure. Estimation is then a simple extension of that discussed in Chapter 4 and can easily cope with large numbers of repeated measurements (Ware *et al.*, 1988, Zeger, 1988). Given their relative ease in practical application and their ability to cope with large numbers of repeated measurements, it is these models which are applied here for both the cumulative odds and continuation ratio parameterisations. Reviews of these and other methods are given by Landis *et al.* (1988), Agresti (1989) and Ware *et al.* (1988). Ashby *et al.* (1992) give a general annotated bibliography of other proposed methods.

Section 5.2 describes the parameterisation and the interpretation of the cumulative odds and continuation ratio models for cross-sectional data. The extensions of these models for repeated measurements in terms of marginal and random effects models are then given in Section 5.3 . Their different interpretations are highlighted. The practical application of both models is then demonstrated using data from the shortness of breath item on the RSCL measured in the CRC

NSCLC study that was analysed as a binary response in Section 4.3.2.

5.2 Regression models for ordered categorical data

Within this section, the parameterisation of the cumulative odds and continuation ratio models for cross-sectional data is described. The notation assumes that y_{ij} is a $(K-1)\times 1$ response vector for subject i, at fixed time j, where $y_{ijk}=1$ for a response in category k for ordered response categories $k=1,...,K-1$, 0 otherwise. The probability that subject i, is in category k at the fixed time j, is $E(y_{ijk})=\pi_{ijk}$. Within the notation used in this section, the subscript j, that denotes the time of measurement, is redundant as it is assumed fixed. Its incorporation throughout, however, is to allow simple generalisation of the models for repeated measurement data. Each model is described in terms of a single covariate, x_i. Their extension to p covariates is trivial.

5.2.1 Cumulative odds model

The cumulative odds model was introduced by McCullagh (1980). The motivation for the model was that the ordinal response represents an underlying continuous unobservable latent variable. The model is parameterised in terms of the cumulative probability of response in category k or below

$$\mu_{ijk}=\sum_{r=1}^{k}\pi_{ijr} \qquad (5.1)$$

By defining $z_{ij}=Ly_{ij}$, where L is a $(K-1)\times(K-1)$ lower triangular matrix of ones, the cumulative odds model for the effect of a covariate x_i can be represented in a logistic regression model for $\mu_{ijk}=E(z_{ijk})$ with linear predictor

$$\log\frac{\mu_{ijk}}{1-\mu_{ijk}}=\alpha_k+\beta x_i, \quad (k=1,...,K-1) \qquad (5.2)$$

Within this model, β is assumed independent of the 'cut-point' k, and can be interpreted as the log odds ratio for a one unit change in the value of x_i of being in category k or below at time j. The $\text{var}(z_{ij})=\sigma_e^2 \mathbf{L} V_i \mathbf{L}^T$ where $V_i=\text{var}(y_{ij})$ and σ_e^2 is an over dispersion parameter.

This transformation by $\mathbf{L}$ introduces an extra complication into the usual logistic regression model by inducing a dependency between $(z_{ijk}, z_{ijk'})$ so that, for $k \neq k'$, $\text{cov}(z_{ijk}, z_{ijk'})=\mu_{ijk}(1-\mu_{ijk'})$. Ignoring this dependence will lead to an underestimation of the off-diagonal elements $\text{var}(z_{ij})$ and through this the variance of the model parameters. However, as in this case it has a known form, it can be modelled directly (McCullagh and Nelder, 1989). Alternatively, it can be left unspecified, and robust variance estimates can be obtained using generalised estimating equations (GEE) introduced in Chapter 4 with an appropriate working correlation structure.

5.2.2 Continuation ratio model

The continuation ratio is an expression of the Cox proportional hazards model for discrete data and models the conditional probability of being in category k given a response is in category k or below (Armstrong and Sloan, 1989). Again this is done within a logistic regression model. Following the same notation as above, this is written

$$\log \frac{\pi_{ijk}}{\mu_{ijk}} = \alpha_k^* + \beta^* x_i, \quad (k=2,\ldots,K) \tag{5.3}$$

Although this provides a different interpretation to the parameters of the cumulative odds model - log odds ratios of a response in category k, conditional upon the response being in category k or below for each unit increase in x_i - again the parameter β^* is assumed constant over all cut-points. In practice, this is most simply derived by defining a new response variables z_{ijk}^*, $k=2,\ldots,K$, where each z_{ijk}^* is defined only for the n_{jk} subjects in category k or below at time j and takes the value 1 for a response in category k, 0 otherwise. For $\mu_{ijk}^*=E(z_{ijk}^*)$,

equation (5.3) can be re-written

$$\log\frac{\mu^{*}_{ijk}}{1-\mu^{*}_{ijk}}=\alpha^{*}_{k}+\beta^{*}x_{i}, \quad (k=2,\ldots,K) \tag{5.4}$$

An advantage of the continuation ratio logit over the cumulative odds model derives from this definition of z^{*}_{ijk}, $k=2,\ldots K$, which means that $\mathrm{cov}(z^{*}_{ijk}, z^{*}_{ijk'})=0$ for $k\neq k'$ (Cox, 1972, Armstrong and Sloan, 1989). Thus for the cross-sectional analysis, the usual logistic regression estimation procedures may be used to obtain correct inferences about the model parameters.

5.3 Two model extensions for repeated measurements

The extension of these cross-sectional analyses for longitudinal data involves the incorporation of the additional dependence derived from repeated observations taken on the same subject. As an extension of the work of Chapter 4, this is demonstrated here for marginal and random effect (or more specifically, hierarchical) models for the cumulative odds and continuation ratio in turn.

5.3.1 Marginal models using generalised estimating equations

Since they yield consistent parameter estimates with robust standard errors without the variance structure of the data needing to be correctly specified, the extension of the marginal model of Chapter 4 for both the cumulative odds and continuation ratio is trivial. The only drawback is that estimates obtained using a poorly specified working correlation structure, although consistent, can be inefficient. However, for an improvement over an independence working correlation structure, care is needed to realise the correct structure between cut-points within occasions as well as that within subjects over time.

For the cumulative odds model in particular, various authors have approached this problem. Kenward *et al.* (1994) noted that $\rho_{ijkk'}=\text{corr}(z_{ijk},z_{ijk'})$ for $k<k'$ can be derived from the cut point parameters α_k in equation (5.2). They then show that more efficient parameter estimates may be obtained by using the $m_i(K-1)$ x $m_i(K-1)$ block diagonal matrix working correlation matrix $\boldsymbol{R}_i$, where the element of kth row and k'th column of the jth diagonal block, $\{\boldsymbol{R}_{ij}\}_{kk'}$, is $\rho_{ijkk'}=\sqrt{\exp(\alpha_k-\alpha_{k'})}$ for $k<k'$, with $\rho_{ijkk'}=\rho_{ijk'k}$. This matrix is assumed the same for all i and is updated at each iteration. Such a structure, although improving on the specification of the dependence between $(z_{ijk},z_{ijk'})$, still assumes an independence structure for observations over time on the same subject. Alternatively, Clayton (1992) used an empirical estimate for $\boldsymbol{R}_i$ based on the observed proportions in the different categories at different times. A further possible estimate for $\boldsymbol{R}_i$ can be obtained using the Pearson residuals given from an independence fit.

For the continuation ratio model, the conditional independence between observed data for each cut point at each time makes specification of an efficient working correlation matrix more straightforward. For instance, in a simple case with j=1,2,3 and k=0,1,2,3, exchangeable or auto-regressive working correlation matrices for the $3(K-1)$ variables z_{ij}^{*} with lag one correlation ρ can be written

$$\boldsymbol{R}_i^{EX}(\rho)=\begin{pmatrix} \mathbf{I}_3 & \rho\mathbf{I}_3 & \rho\mathbf{I}_3 \\ \rho\mathbf{I}_3 & \mathbf{I}_3 & \rho\mathbf{I}_3 \\ \rho\mathbf{I}_3 & \rho\mathbf{I}_3 & \mathbf{I}_3 \end{pmatrix}, \quad \boldsymbol{R}_i^{ARI}(\rho)=\begin{pmatrix} \mathbf{I}_3 & \rho\mathbf{I}_3 & \rho^2\mathbf{I}_3 \\ \rho\mathbf{I}_3 & \mathbf{I}_3 & \rho\mathbf{I}_3 \\ \rho^2\mathbf{I}_3 & \rho\mathbf{I}_3 & \mathbf{I}_3 \end{pmatrix} \tag{5.5}$$

where $\mathbf{I}_3$ is a (3x3) identity matrix. Estimation of ρ for this problem is, however, not trivial.

5.3.2 Random effect models using a hierarchical structure

A multilevel (hierarchical) model for repeated ordered categorical data treats the dependence between category cut points as an additional level in the hierarchy, thus extending

the two level repeated binary response model to a three level multivariate model (Hedeker and Gibbons, 1994). In the same notation as before, the most simple random effects cumulative odds model is written

$$z_{ijk}=\mu_{ijk}+e_{ijk} \tag{5.6}$$

where

$$\text{logit}(\mu_{ijk})=\alpha_k+\beta x_i+u_i \tag{5.7}$$

At level two, the residuals $e_{ij}=(e_{ij1},...,e_{ijK-1})$ are constrained to have a multinominal distribution, that is, $\text{cov}(e_{ijk},e_{ijk'})=\sigma_e^2\mu_{ijk}(1-\mu_{ijk'})$ for all combinations of $\{k,k'\}$, where σ_e^2 is an overdisperson parameter. At the level three, u_i is assumed Normally distributed with mean zero and variance σ_u^2. It therefore follows that a positive u_i signifies a subject specific higher than average odds of being in category k or below (on the log scale) and therefore typically a lower than average symptom score, whereas a negative u_i indicates a lower than average odds of being in category k or below and thus a typically higher than average symptom scores.

Although alternative estimation procedures are feasible (Hedeker and Gibbons, 1994), IGLS (RIGLS) with a Taylor series expansion to linearise the link function will be used here to give penalised quasi-likelihood estimates (Breslow and Clayton, 1993, Goldstein, 1995).

Similarly, the random effect (or multilevel) model for the continuation ratio is also a three level multivariate binary model written

$$z^*_{ijk}=\mu^*_{ijk}+v^*_{ijk}e^*_{ijk} \tag{5.8}$$

where

$$\text{logit}(\mu^*_{ijk})=\alpha^*_k+\beta^* x_i+u_i^* \tag{5.9}$$

However, because of the definition of z^*_{ijk}, the variance structure of the continuation ratio model

has a much simpler form than that of the cumulative odds model as $\text{cov}(e^*_{ijk}, e^*_{ijk'})=0$ for $k \neq k'$. In this case a positive u_i^* suggests a higher than average odds of a response in category k given a response in category k or below and therefore a typically higher than average symptom scores.

5.3.3 Model interpretations

As with the repeated binary case, the interpretation of the marginal and random effects models are very different. The marginal model gives population average effects which have the same interpretation as those of a cross-sectional analysis whereas the random effect or multilevel analysis gives subject specific covariate effects. These will be larger in absolute size than their population average counterparts by an amount directly related to the extent of between subject variation.

There is also a very important distinction to be made between the parameters of the cumulative odds and continuation ratio models. The former represent covariate effects as log odds ratios of being in category k or below for each unit change in the covariate of interest, whereas the continuation ratio gives the log odds ratio of a response in category k given the response is at least in that category. In each case, the fundamental model assumption is that of equivalent covariate effects over all categories. A summary of the different interpretations of a parameter β for symptoms graded an ordered category scale for an arbitrary treatment covariate, x_i, taking the value 0 or 1 are summarised in table 5.1. These practical differences are demonstrated with an example from the CRC NSCLC study.

5.4 Example - CRC NSCLC shortness of breath

Shortness of breath was measured in the CRC NSCLC study using the RSCL on a four point scale: not at all (0); a little (1); somewhat (2); very much (3). Ignoring responses at baseline,

Table 5.1: Summary of the interpretations of a parameter β for repeated ordered categorical data where x_i is a treatment covariate taking the values 0 or 1.

	Model	Interpretation
1	Marginal cumulative odds	Log odds ratio of symptoms in category *k* or below for treatment 1 versus treatment 0
2	Marginal continuation ratio	Log odds ratio of symptoms in category *k* given symptoms are in that category or below for treatment 1 versus treatment 0
3	Random effect cumulative odds	The additional effect of treatment 1 versus treatment 0 on each subject specific log odds of being in category *k* or below
4	Random effect continuation ratio	The additional effect of treatment 1 versus treatment 0 on each subject specific log odds of symptoms in the highest category given symptoms are in that category or below

information was available on 62 patients. Over the entire period, there was little difference between the two groups in the proportions reporting symptoms in categories 0 or 2 (table 5.2). There were however slight differences for categories 1 and 3, with a larger proportion of patients in the continuous course reporting in category 1 and therefore the converse for category 3. Overall this indicates slightly worse symptoms of shortness of breath in the intensive split course. Plotting these proportions on a weekly basis (figure 2.15) showed a slight increase over time in the reporting of symptoms graded 1, 2, 3 in both treatment arms. In the light of these observations, this example uses each of the four models described above to address both the question of whether there is any evidence of a difference in the reporting of shortness of breath in the two treatment arms, or that the reporting of symptoms changed with time.

5.4.1 Model parameterisations

To demonstrate their differences, the results of the different models were compared in terms of their interpretations, as well as the validation of the model assumptions. For the marginal models, independence working correlation matrices were initially used. These results were

Table 5.2: Proportion of subjects in each response category over the entire 8 week follow-up.

	Category			
	0	**1**	**2**	**3**
Split course	0.12	0.37	0.25	0.26
Continuous course	0.13	0.45	0.24	0.18

then compared with those from models with specification of the working correlation matrix more closely approximating that expected, as detailed in Section 5.3.1. For the multilevel model analyses, RIGLS PQL estimation was used. The four models used are defined below.

Marginal cumulative odds

$$\log\frac{\mu_{ijk}}{1-\mu_{ijk}} = \alpha_0^M + \alpha_1^M + \alpha_2^M + \beta^M occ_{ij} + \delta^M rt_i$$

Random effect cumulative odds

$$\log\frac{\mu_{ijk}}{1-\mu_{ijk}} = \alpha_0^{RE} + \alpha_1^{RE} + \alpha_2^{RE} + \beta^{RE} occ_{ij} + \delta^{RE} rt_i + u_i$$

Marginal continuation ratio

$$\log\frac{\mu^*_{ijk}}{1-\mu^*_{ijk}} = \alpha_1^{*M} + \alpha_2^{*M} + \alpha_3^{*M} + \beta^{*M} occ_{ij} + \delta^{*M} rt_i$$

Random effect continuation ratio

$$\log\frac{\mu^*_{ijk}}{1-\mu^*_{ijk}} = \alpha_1^{*RE} + \alpha_2^{*RE} + \alpha_3^{*RE} + \beta^{*RE} occ_{ij} + \delta^{*RE} rt_i + u_i^*$$

(5.10)

Within the marginal models, the cut point parameters, α_k^M and α_k^{*M}, k=1, 2 for the cumulative odds model and k=2, 3 for the continuation ratio model were parameterised in terms of differences from α_0^M and α_1^{*M}, respectively. Similarly for the random effect models. In all four models, a linear occasion effect was parameterised as occ_{ij}=0,...,7 for weeks 1,...,8, and constant treatment effect, rt_i, modelled with the split course as baseline. In each case, the proportional odds assumption was tested by fitting a covariate by cut point interaction and constructing the combined Chi-squared statistic for simultaneous contrasts on 2 df.

5.4.2 Results - marginal model

The results for the two marginal models are given in table 5.3. These should both reflect

Table 5.3: Results for marginal cumulative odds and continuation ratio models results for shortness of breath in the CRC NSCLC study as an ordinal response.

		Estimate (SE)	Odds ratio [95% CI]	Proportional odds assumption (χ^2 on 2 df)
Cumulative odds model				
α_0^M	(*Category 0 or below*)	-2.62 (0.54)	-	-
α_1^M	(*Category 1 or below*)	2.16 (0.34)	-	-
α_2^M	(*Category 2 or below*)	3.34 (0.39)	-	-
β^M	(*occ*)	0.09 (0.04)	1.09 [1.01, 1.18]	8.64 p=0.01
δ^M	(*rt*)	0.39 (0.44)	1.48 [0.62, 3.50]	0.35 p=0.84
Continuation ratio model				
α_1^{*M}	(*Category 1 given 0/1*)	1.66 (0.54)	-	-
α_2^{*M}	(*Category 2 given 0/1/2*)	-2.02 (0.43)	-	-
α_3^{*M}	(*Category 3 given 0/1/2/3*)	-2.59 (0.45)	-	-
β^{*M}	(*occ*)	-0.06 (0.04)	0.94 [0.87, 1.02]	7.91 p=0.02
δ^{*M}	(*rt*)	-0.27 (0.37)	0.76 [0.37, 1.58]	0.62 p=0.73

The standard errors quoted are robust standard errors from a GEE with an independence working correlation matrix.
The treatment effect is given with the split course group as baseline and the occasion effect modelled as week 1 to 8 by a linear effect from 0 to 7.

the patterns in the proportions over time shown in figure 2.15. This is particularly relevant for the cumulative odds model in which the category cut point parameters can be transformed from log odds (and log odds ratios) to give an estimated intercept for each of the lines displayed in figure 2.15(a) for the split course. The estimated treatment effect in this model for the log odds ratio of being in category *k* or below (for all *k*) for the continuous course versus the split course was 0.39 (SE=0.44). Translated in terms of an odds ratio this implies that patients on the continuous radiotherapy course were 48% (95% CI=[-38%, 250%]) more likely to be in category *k* or below than those in the split course. There was no evidence to suggest this effect

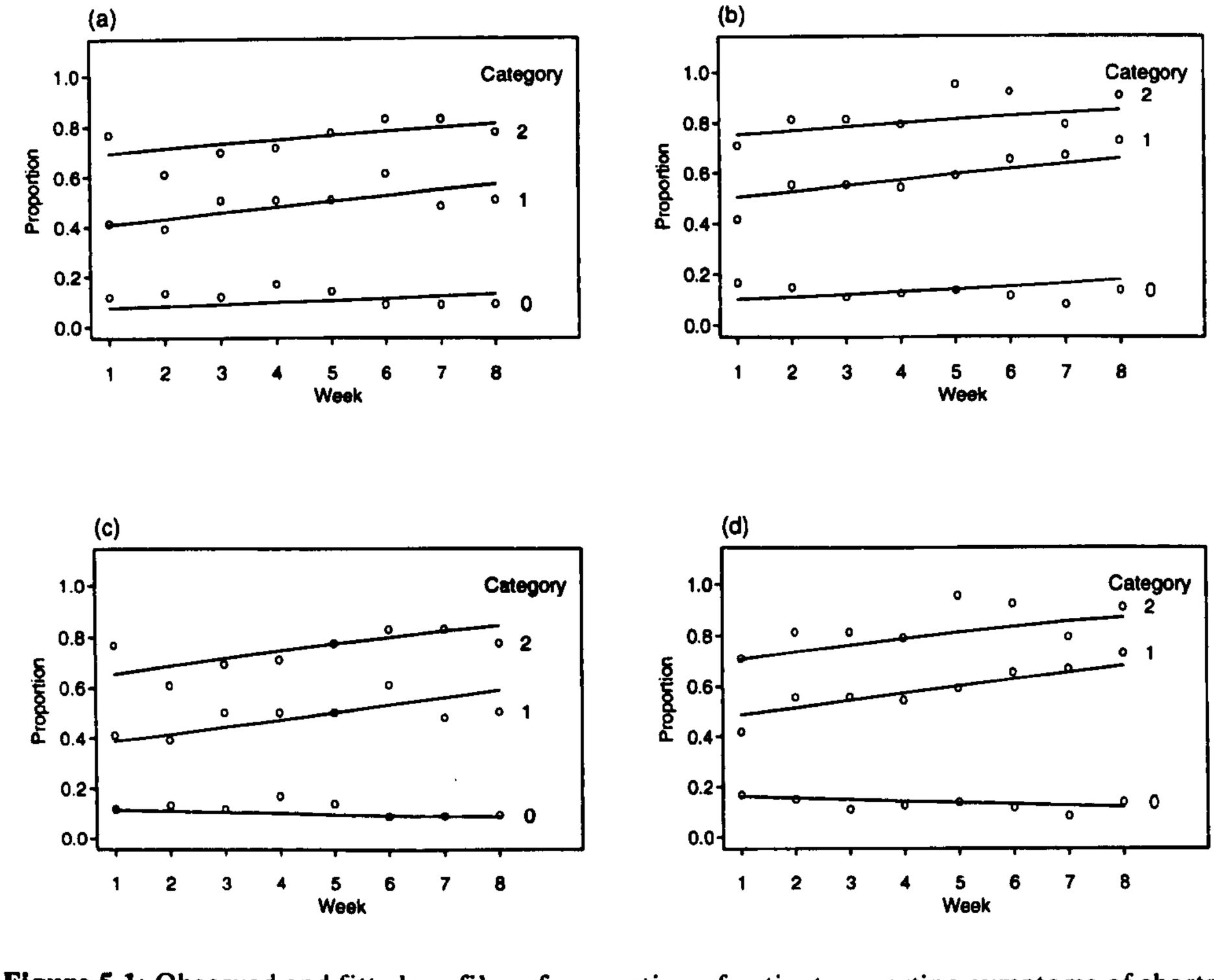

Figure 5.1: Observed and fitted profiles of proportion of patients reporting symptoms of shortness of breath in category *k* or below. The fitted profiles have proportional occasion effects for (a) intensive split course; (b) continuous course; and non-proportional occasion effect for (c) split course; and (d) continuous course in the CRC NSCLC study.

was not due to chance (p=0.37) or of a violation of the proportional odds assumption (p=0.84). This was not so for the linear occasion effect which gave an estimated 9% [1%, 18%] increase in odds per week of being in category *k* or below. The test of the proportional odds assumption gave some evidence of an interaction between occasion and cut point (p=0.01). This is shown in figure 5.1 by the fitted profiles for the models with proportional and non-proportional occasion effect. They demonstrate that the non-proportionality derived as a result of a fall in the proportion of patients in category zero in both treatment arms.

The interpretation of the parameters of the continuation ratio are very different, although

they do reflect the raw data plotted in figure 2.15. The individual category effects give the log odds of being in category *k* conditional on being in category *k* or below (for week 0, split course). For category 1, the estimated odds was high suggesting that of the responses in category 0 or 1, a larger proportion were in category 1. As expected, the corresponding estimated odds fall for the higher categories. Their was no evidence that the estimated treatment effect varied across categories (p=0.73) and, consistent with the cumulative odds model, there was no evidence that the estimated overall treatment difference was not due to chance (p=0.77). The direction of this estimated effect was different to that of the cumulative odds model highlighting the very different interpretation of the parameters of the two models. For the continuation ratio parameterisation, the estimated treatment odds ratio gave the relative difference in odds of being in the highest category, given the response was in that category or below for the continuous versus split course radiotherapy. The estimate of 0.76 (95% CI=[0.37, 1.58]) suggested that patients on the continuous course of radiotherapy had lower odds of responding in the higher of a set of categories than those on the split course. This was therefore consistent with the estimated effect of the cumulative odds model. A similar comparison occurs with the linear occasion effect which was estimated to fall over time. As with the cumulative odds model, there was evidence to suggest that this effect was not consistent over categories (p=0.02).

The results from the marginal models reported above, were based on an independence working correlation matrix. Although inferences based on the robust standard errors from this model will be consistent, they may be inefficient because of the poor approximation of the working correlation structure to the true covariance structure of the data (Zeger *et al.*, 1988). For improved efficiency a number of alternative working correlation structures were assumed.

For the cumulative odds model, although the dependency between cut points has a known

form, the added complexity of the dependence between repeated measurements makes the correlation structure more difficult to specify. The matrices used here attempted to improve the approximation of the dependence between cut points, although they all incorrectly assumed independence between repeated observations on the same subject. They were therefore block diagonal with diagonal blocks $\boldsymbol{R}_{ij}$, j=1,...,8, that is,

$$\boldsymbol{R}_i=\begin{pmatrix} \boldsymbol{R}_{i1} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \boldsymbol{R}_{i7} & 0 \\ 0 & \dots & 0 & \boldsymbol{R}_{i8} \end{pmatrix} \qquad \textbf{(5.11)}$$

Four matrices were used. Each has been previously described in Section 5.3.1. The first, an *empirical* matrix, was based on the observed correlation between the Pearson residuals calculated from an independence model. For computational simplicity, this was based only on subjects who responded at all occasions and was assumed the same for all j. The second matrix was suggested by Kenward *et al.* (1994). It used a two stage process with the working correlation matrix based on the estimated cut point parameters of the marginal model. Since the occasion effect was assumed constant over k (k=0, 1, 2), $\boldsymbol{R}_{ij}$ was therefore the same for all j=1,...,8. The two remaining matrices used the method of Clayton (1992) and were based on the observed proportions within each category: *Clayton (1)* combined data over all occasions and assumed the same structure for all j; *Clayton (2)* used the observed proportions at each occasion separately, and thus allowed $\boldsymbol{R}_{ij}$ to vary across j. A summary of these matrices and the results of the analyses are given in table 5.4.

For the cumulative odds model these illustrated that the exact choice of working correlation matrix although making very slight changes to the parameter estimates of the model, did not affect the model conclusions. For this example, the *Kenward* and *Clayton (1)* working correlation matrices were very similar and so it was unsurprising that their results varied very

Table 5.4: Estimates (robust SE) {estimate/SE} for marginal cumulative odds with different working correlation matrices.

	Working correlation structure			
	Empirical	**Kenward**	**Clayton (1)**	**Clayton (2)**
Category 0 or below	-2.56 (0.52) {-4.90}	-2.63 (0.54) {-4.87}	-2.64 (0.54) {-4.89}	-2.63 (0.53) {-4.96}
Category 1 or below	2.16 (0.32) {6.75}	2.16 (0.34) {6.35}	2.15 (0.34) {6.32}	2.13 (0.34) {6.26}
Category 2 or below	3.34 (0.38) {8.79}	3.34 (0.39) {8.56}	3.34 (0.44) {7.59}	3.30 (0.39) {8.46}
Occasion	0.08 (0.04) {2.00}	0.09 (0.04) {2.25}	0.09 (0.04) {2.25}	0.09 (0.04) {2.25}
Treatment	0.36 (0.45) {0.80}	0.39 (0.44) {0.89}	0.39 (0.44) {0.89}	0.39 (0.43) {0.91}

For each case the (24x24) working correlation matrix, $\boldsymbol{R}_i$, was assumed block diagonal and the same for all i. For the *Empirical, Kenward and Clayton(1)* examples the diagonal blocks, $\boldsymbol{R}_{ij}$ were assumed the same for all j, and took the values

$$\begin{pmatrix} 1.00 & 0.317 & 0.173 \\ 0.317 & 1.00 & 0.243 \\ 0.173 & 0.243 & 1.00 \end{pmatrix}, \begin{pmatrix} 1.00 & 0.341 & 0.189 \\ 0.341 & 1.00 & 0.556 \\ 0.189 & 0.556 & 1.00 \end{pmatrix}, \begin{pmatrix} 1.00 & 0.345 & 0.193 \\ 0.345 & 1.00 & 0.560 \\ 0.193 & 0.560 & 1.00 \end{pmatrix}$$

respectively. For *Clayton(2)*, $\boldsymbol{R}_i$ has the same block diagonal structure, but $\boldsymbol{R}_{ij}$ was varied across occasions j=1,...,8.

little. Allowing the diagonal blocks to vary over occasions (*Clayton (2)*) also made little difference to the results although it did exhibit a very slight gain in efficiency over the independence and the other three working correlation matrices (illustrated by slight reductions in the standard errors of the model parameters of interest). Given the simplicity of each of these structures, particularly the *Kenward* and *Clayton (1)* working correlation matrices, it is recommended that such matrices be used for practical situations.

For the continuation ratio logit, three different matrices were used. The first two had the exchangeable structure given in equation (5.5) with lag one correlation ρ chosen arbitrarily to equal 0.64 and 0.20 respectively. The third matrix had the autoregressive structure described

in equation (5.5), with lag one correlation, again arbitrarily chosen at ρ=0.6. The results are given in table 5.5.

In contrast to the results of the cumulative odds model, in this case the choice of working correlation matrix over an independence matrix had some impact on the estimated coefficients. This was particularly striking for the cut point parameters but did not alter the conclusions of the analysis. The most marked change was seen with the exchangeable matrix with lag one correlation, ρ=0.64. The parameter estimates from this model were also more inefficient than those of the independence matrix given in table 5.3 suggesting that this matrix was less appropriate than the independence matrix. The results of the model with the autoregressive working correlation matrix with ρ=0.64, were much more efficient, suggesting that the poor performance of this first exchangeable matrix, derived from an over specification of the correlation between distant occasions. The exchangeable working correlation matrix with lag one correlation ρ=0.2 illustrated the highest gain in efficiency over the independence model.

Table 5.5: Estimates (robust SE) {estimate/SE) for continuation ratio with different working correlation structures.

	Working correlation matrix		
	Exchangeable (ρ=0.64)	**Exchangeable (ρ=0.2)**	**AR1 (ρ=0.6)**
Category 1 given 0/1	2.73 (0.64) {4.27}	1.95 (0.52) {3.75}	1.86 (0.53) {3.51}
Category 2 given 0/1/2	-1.69 (0.49) {-3.45}	-1.66 (0.40) {-4.15}	-1.12 (0.40) {-2.80}
Category 3 given 0/1/2/3	-3.76 (0.50) {-7.52}	-2.78 (0.45) {-6.18}	-3.00 (0.49) {-6.12}
Occasion	-0.09 (0.04) {-2.25}	-0.08 (0.03) {-2.67}	-0.08 (0.04) {-2.00}
Treatment	-0.25 (0.54) {-0.46}	-0.31 (0.41) {-0.76}	-0.33 (0.45) {-0.73}

5.4.3 Results - random effect models

The results for the random effect models are given in table 5.6. Deriving from random effect models, the parameter estimates represent the effect of a particular covariate on a subject's underlying odds of being in category *k* or below for the cumulative odds model, or of being in category *k* given a response in category *k* or below for the continuation ratio model. Their results can therefore not be compared directly with the plots of proportions over time given in figure 2.15. As discussed in Section 4.2, the parameter estimates will be larger in absolute size than those of the marginal models with the relative difference proportional to the variance of subject specific (log) odds, σ_u^2 or $\sigma_{u..}^2$.

For the cumulative odds model, the variance in subject specific log odds of being in category *k* or below, was estimated as 14.0, indicating a large amount of variation between subjects. For example, this gave a 95% reference range for the subject specific probability of being in category 1 or below (for patients in the split course) of [0.0003, 0.999]. The effect of treatment was consistent in sign with that of the respective marginal model and estimated a 108% increase (95% CI=[-71%, +1370%]) in the subject specific odds of being in category *k* or below if treated with continuous versus a split course of radiotherapy. There was no evidence to suggest this was not due to chance or that it was dependent on the category *k*. Also consistent with the results of the marginal cumulative odds model, there was some evidence of an increase in the odds of being in category *k* or below over time. The estimated change in odds over one week was +23% (95% CI=[+10%, +39%]). There was also some evidence of non-proportionality of this effect over *k*. Given the estimated between subject variability and assumed Normality of the random effects, the expected ratio of the marginal (population average) parameter estimates to these random effect (subject specific) estimates was 0.41. Those observed (relative to the marginal model estimates of *Clayton (2)* in table 5.4) 0.33 and 0.45 for the occasion and treatment effects respectively.

Table 5.6: Results for random effect cumulative odds and continuation ratio models for shortness of breath in the CRC NSCLC study as an ordinal response.

		Estimate (SE)	Odds ratio [95% CI]	Proportional odds assumption (χ^2 on 2 df)
Cumulative odds model				
Fixed parameters				
α_0^{RE}	(*Category 0 or below*)	-6.00 (0.85)	-	-
α_1^{RE}	(*Category 1 or below*)	5.27 (0.46)	-	-
α_2^{RE}	(*Category 2 or below*)	8.23 (0.54)	-	-
β^{RE}	(*occ*)	0.21 (0.06)	1.23 [1.10, 1.39]	8.51 p=0.014
δ^{RE}	(*rt*)	0.73 (1.00)	2.08 [0.29, 14.7]	1.32 p=0.52
Random parameters				
σ_u^2		14.0 (2.74)		
Continuation ratio models				
Fixed parameters				
α_1^{*RE}	(*Category 1 given 0/1*)	5.65 (0.81)	-	-
α_2^{*RE}	(*Category 2 given 0/1/2*)	-5.12 (0.46)	-	-
α_3^{*RE}	(*Category 3 given 0/1/2/3*)	-7.70 (0.54)	-	-
β^{*RE}	(*occ*)	-0.20 (0.05)	0.82 [0.74, 0.91]	8.55 p=0.014
δ^{*RE}	(*rt*)	-0.68 (0.94)	0.51 [0.08, 3.20]	1.37 p=0.50
Random parameters				
$\sigma_{u.}^2$		12.5 (2.43)		

The treatment effect is given with the split course group as baseline and the occasion effect modelled as week 1 to 8 by a linear effect from 0 to 7.

All estimates were obtained using RIGLS with 1st order PQL.

For the continuation ratio model, there was similarly a high degree of variation between subjects in the underlying odds of responding category *k* given *k* or below. For category 2 given 0, 1, 2, the estimated 95% reference range in terms of probability was [0.002, 0.999]. In

terms of the estimated treatment effect, the estimated difference in subject specific log odds of a response in category *k* given a response in category *k* or below, for continuous radiotherapy versus a split course of radiotherapy, was -0.68 (SE=0.94). This translated to a 49% lower odds of being in the higher category if the patient was given continuous therapy versus a split dose (95% CI=[92% lower, +220% higher]). As for the population average effect estimated from the marginal continuation ratio model, there was no evidence to suggest this was not due to chance (p=0.77) or that the effect varied according to the category cut point (p=0.50). Based on the estimated between subject variance, and Normality of subject specific residuals, the expected ratio of the estimated parameters of the marginal model to its random effect counterpart was 0.43. Those observed (relative to the marginal model with exchangeable working correlation with ρ=0.20 given in table 5.5) were 0.46 and 0.40 for the occasion and treatment parameters respectively. Residual diagnostics for each of these random effect models are shown in figure 5.2. In both cases they showed no substantial deviation from Normality.

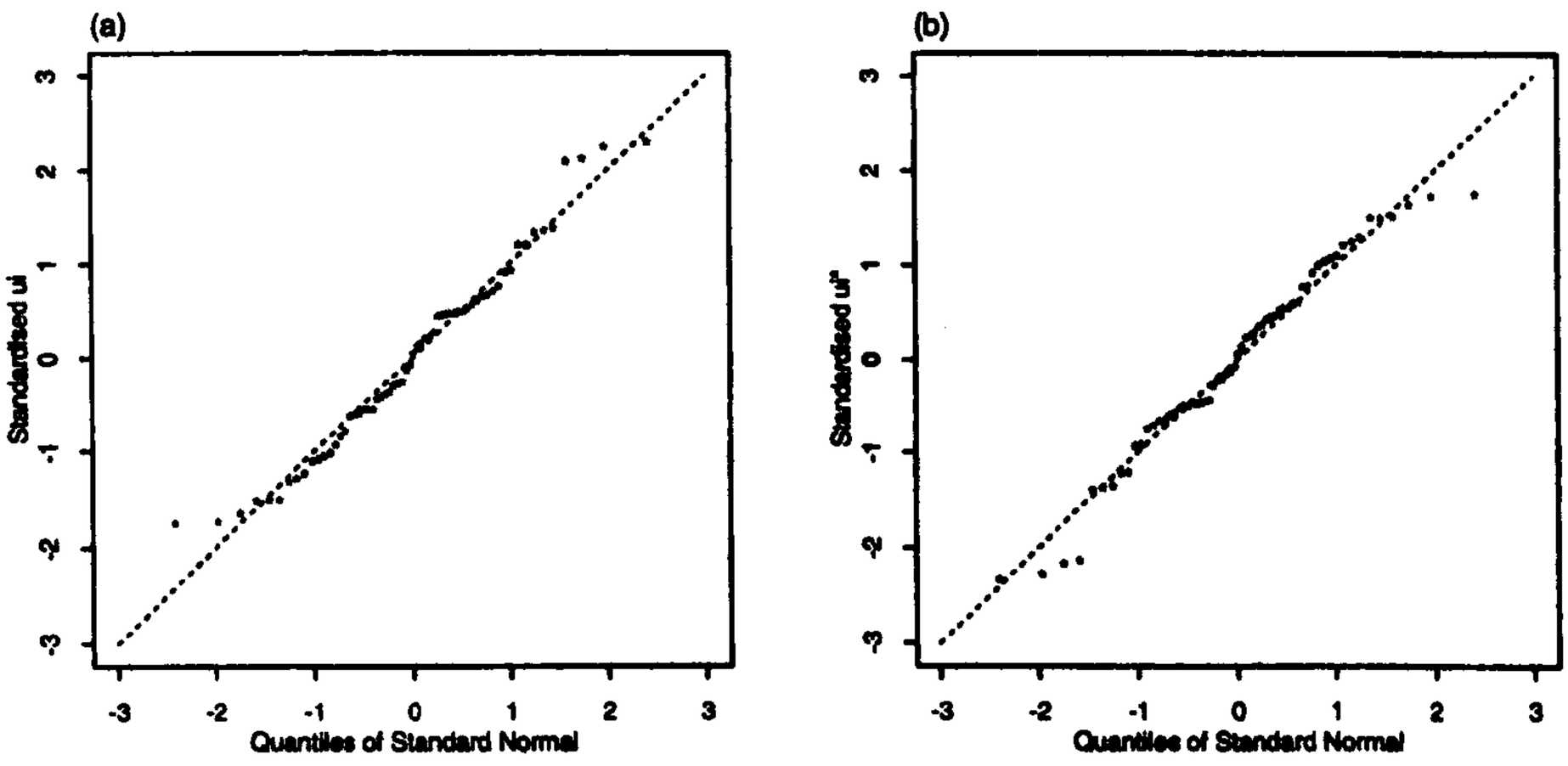

Figure 5.2: Standardised level three (between subject) residuals for the (a) cumulative odds; (b) continuation ratio random effect models for shortness of breath in the CRC NSCLC study analysed as an ordinal response.

5.4.4 Analysis conclusions

In conclusion, each of these models gave evidence of a change in the odds of reporting symptoms over time, either as a population average effect, or in terms of a change in odds for a subject. The direction of this change was shown in all cases to be dependent on the category, *k*. There was no evidence of a difference in the odds in the reporting of the severity of shortness of breath between the two treatment groups, although patients on the continuous course were shown to have less severe symptoms than those on the split course.

5.5 Summary and discussion

As most quality of life measurement scales can, to some degree, be reduced to an ordinal scale, this work has assessed the use of two proposed models for such data. Their extensions for repeated measurement data using random effect (hierarchical) models and marginal models using GEEs for estimation have been explained and applied to the RSCL data in the CRC NSCLC study. Other approaches to both the model parameterisation and its extension for repeated measurement are possible, in particular the use of transitional models that allow for the dependence induced by the repeated measurements by conditioning explicitly on previous observations. Such models give a different interpretation again to both the marginal and random effect models and have been discussed by several authors, for example, Lindsey *et al.*, 1995, Follmann, 1994. In particular, Lindsey *et al.* (1995) use a transitional model in conjunction with a continuation ratio for a model that can be fitted using any conventional logistic regression software. In terms of parameter estimation, full likelihood parameter estimation approaches have been suggested by Grizzle *et al.* (1969) and Kenward *et al.* (1994). Unfortunately, these are restricted to cope with few repeated measurements and would therefore have been difficult to apply to the example data. As these data are typical of those generated within quality of life studies, the use of such models is perhaps restricted for this application.

An exception to this is, for example, in some breast cancer trials where patient life expectancy is relatively long and quality of life measured at a small number of infrequent occasions during this time (Fallowfield *et al.*, 1987).

Once again, within the example presented here, it has been assumed that missing data is unrelated to the response process and it has therefore been ignored. A more detailed discussion into the implications of this are provided in a Chapter 7. However, Mark and Gail (1994) demonstrated with a simple example, that a marginal model for the cumulative odds parameterisation with a generalised estimating equation and an empirical working correlation matrix gave relatively unbiased estimates even in cases when the missing data observed was informative of the underlying response.

The motivation for analysing the data on its original ordinal scale rather dichotomising it to a binary response was to avoid the loss of information which may result from such a dichotomisation, the arbitrariness of choosing where to dichotomise and to increase the power of the analysis. The information gain of the analysis is unfortunately offset by the increased complexity in interpretation of these models. In terms of the gain in power, although this is apparent with the random effect analyses (seen by comparing tables 5.6 and 4.2), it was not the case for the marginal models (tables 5.4, 5.5, and 4.2). This was almost certainly due to the loss of efficiency induced by a poorly specified working correlation structure in that they allow for the dependence between cut points rather than between repeated measurements. Introducing a second set of estimating equations to fully model this structure may help in this respect, but the added efficiency is again offset by the additional complexity (Kenward *et al.*, 1994).

As with the analyses of Chapters 3 and 4, alternative methods of estimation are available for

the marginal or random effect analyses of ordinal repeated measurement data. Those presented here were chosen specifically for their application as limited programming is required in order for them to be applied, therefore making them accessible in practice. For random effect analyses, an equally accessible method is available using Gauss-Hermite quadrature which has been implemented by Hedeker and Gibbons (1994). The main problem with the models presented here is the unfamiliar interpretation of the covariate effects of both cumulative odds and continuation ratio models. Although the equally unfamiliar subject specific interpretation of the random effect model may deter application of these models, as for the binary case, it should be noted that, although unfamiliar, it is often these subject specific effects which are of interest in a clinical trial.

6 The Analysis of Quality of Life Censored by Death

6.1 Background

In many cancer studies treatment is given in order to improve the quality of a limited survival prognosis. In such cases it can be expected that patient death during follow-up will occur, leaving a truncated or censored quality of life profile for that individual. The consequences of such patient 'dropout' when attempting to draw inferences about treatment efficacy may be great. This becomes a particular problem when there is a trade-off between quantity and quality of survival. Cox *et al.* (1992) suggest that in such cases, in order to aid clinical decision making, it may be best to report quality of life and survival outcomes separately, thus allowing the clinician and patient to weigh up the trade-off. Although initially appearing to be straightforward, considering quality of life and survival as two distinct outcomes is not altogether without problems. The potential for bias introduced as a result of early subject dropout has been well documented (Diggle and Kenward, 1994, Little, 1995) and a particular, perhaps philosophical, problem is whether the quality of life of patients who subsequently die should influence inferences about response beyond their time of death, or whether it is the quality of life of only surviving individuals at each time which is relevant. If it is the former, ignoring subjects who die and considering solely the remaining subjects may give misleading conclusions.

As an alternative, much work has been done with methods of combining the two endpoints to give a *quality adjusted survival* analysis (Glasziou *et al.* 1990, Gelber *et al.* 1986, Korn 1993). It has been suggested that these methods allow more insight into the quality of life and

survival trade-off. Until recently such quality adjusted techniques focused on defining a number of *health states* through which patients progress and summarising the weighted survival time which patients spend in each state for some arbitrary weights assigned to each state. This work has been used most extensively to analyse patient toxicity and disease progression data where health states are naturally defined, whereas their application to self assessed quality of life data, where state definitions are more arbitrary, has been extremely limited. Recent work by Glasziou (1995) has attempted to address this by considering the use of patient responses to self assessed questionaries over time as a weight for their survival, thus avoiding the definition of arbitrary health states, but as yet the method has not been tested in practical situations.

The work within this chapter examines the use of both these approaches to quality of life data censored by death, contrasting a number of analyses of the quality of life and survival data from the CRC HAP trial restricted to that available at June 1st 1993 (described in Chapter 1). As full quality of life and survival are also available for all but three subjects in the full version of this data set, for some of the analyses presented, these data will also be analysed as a comparison. Section 6.2 considers treating death in quality of life studies as a dropout problem and discusses types of dropout model which are most appropriate for application to quality of life data. Quality adjusted survival techniques based on both health state and continuous quality of life responses are presented in Section 6.3.

6.2 Modelling dropout mechanisms for informatively censored data

The term *dropout* in repeated measurements is used to refer to the cessation of a subject's response for reasons which may or may not be related to that response. For example, when blood pressure is measured on patients being treated for hypertension, patients whose blood

pressure reaches a certain level may by design be excluded from the study as they undergo extra treatment (Murray and Findlay, 1988). Diggle and Kenward (1994) reported an alternative example where dropouts were not due to study design. In this example, which concerned measuring the amount of protein in the milk of cows receiving different feeds, dropout was due to cows ceasing to lactate before the end of the study. For self assessed quality of life data, the censoring or death of a patient are generally the reasons why a patient is regarded to have dropped out.

If it is planned that subjects are to be observed over a predetermined time period, and interest lies in the nature of response over this entire period, inferences should naturally be based on the hypothetical complete data - that is the sequence of measurement which would have been observed in the absence of dropout. It is on such *complete data inferences* that interest in the literature has focused. This work has demonstrated that the complexity of the analysis required will depend on the relationship between this sequence of measurements (the measurement process) and the probability of patient dropout (the dropout process) (Little, 1995).

When the probability of a subject dropping out is unrelated to the observed responses of interest, the problem with the data analysis is simply one of being able to cope with unbalanced data, for which any of the techniques discussed in the Chapters 3, 4 and 5 are appropriate. As it is usually reasonable to assume that such an assumption is valid for dropout which arises due to censoring because of staggered entry, subsequent discussion for quality of life data is focused purely on the problem of dropout due to patient death.

When there are no *a priori* ties between the parameters which define the dropout process and those of interest for the measurement process, the parameters of the two processes are said

to be distinct. For example, supposing the measurement process can be parameterised in terms of a linear trend over time defined by $\beta=(\alpha,\beta)$, and the dropout process is defined by some underlying probability function with parameters ϕ. The two processes are said to be distinct if ϕ and β are *a priori* independent. When this assumption holds, and in addition the probability of dropout is related only to observed measurements, it has been well documented that, because it is possible to factorize the joint likelihood for the complete data and dropout process into two distinct parts, consistent estimates for the parameters of the measurement process may be obtained using likelihood analyses of the observed data ignoring the dropout process (Little and Rubin, 1977, Zwinderman, 1992, Diggle and Kenward, 1994, Little, 1995). This was the case in the example of Murray and Findlay (1988), where their study design determined that subjects were withdrawn and placed on an open program of treatment if their blood pressure at any time *t* was greater than 110mmHg. This defined a dropout process such that a patient is considered to have dropped out at time *t* if their measurement at time *t*-1 was greater than 110mmHg.

When the dropout process does depend on the unobserved data or the parameters of the two processes are not distinct, this factorisation of the likelihood is not possible and explicit modelling of the dropout process is then required. The possible modelling strategies suggested can be separated into 3 classes: *selection*; *pattern mixture*; and *informatively right censored* models (Little, 1995).

Selection and pattern mixture models evolve from different factorisations of the joint likelihood for the measurement and the dropout process. Selection models use a model for the hypothetical complete data along with one for the dropout process conditional upon the hypothetical data (Diggle and Kenward, 1994), whereas pattern mixture models stratify the population by the pattern of dropout (Little, 1993, Little, 1995). This is summarized in box 6.1.

Box 6.1

Derivation of selection and pattern mixture models

Dropping the subscript i, for $y=(y_1,...,y_m)$, $j=1,...,m$, the complete data, made up of that observed, (y_{obs}), and that missing due to dropout, (y_{mis}), and $r=(r_1,...,r_m)$, a corresponding indicator such that $r_j=1$, if y_j is observed, 0 otherwise, the **selection** model factorises the joint distribution, $f(y,r)$ as

$$f(y,r)=f(y)f(r|y)=f(y_{obs},y_{mis})f(r|y_{obs}y_{mis});$$

whereas, for the **pattern mixture** model it is factorised as

$$f(y,r)=f(r)f(y|r)=f(r)f(y_{obs},y_{mis}|r)$$

Both models rely heavily on assumptions about the relationship between the observed and missing responses and the dropout process which cannot be validated. For selection models this is done implicitly in modelling the dropout process and its relationship with the complete data. Most commonly used examples are probit or logistic models which relate the probability of dropout to the unknown observation at the time of dropout (Diggle and Kenward, 1994). For a pattern mixture model the assumptions are more explicit and concern how the observed data from complete 'patterns' relate to the observed and unobserved data from different incomplete patterns. Under certain assumptions for the pattern mixture model, the two approaches will give equivalent results (Molenburghs *et al.*, 1995).

Unfortunately, since these models condition the dropout process explicitly on past, present or future responses, they require responses to be measured at the same times for all subjects and more inconveniently, that dropout also occurs at a small numbers of distinct occasions. This makes them impractical for the problem of dropout due to death in quality of life studies when measurement occasions may not be consistent across subjects, and dropout occurs at many different times. In addition, they attempt to make inferences about the complete data (that is,

in the absence of dropout). Philosophically, when the reason for dropout is itself is an outcome of interest, it is unclear whether such inferences are relevant.

The third class of models, informatively right censored models, attempt to lessen these restrictions by assuming that dropout is a function of some latent variable for each subject. Examples have been restricted to the case where the outcome of interest is the rate of change in response over time, with dropout, in each case a result of patient death, assumed dependent on subject specific intercept and slope (Wu and Carroll, 1988, Wu and Bailey, 1988, Wu and Bailey, 1989, Schlucter, 1992).

Wu and Carroll (1988) used a two stage iterative algorithm with a *probit censoring model* to combine subject specific intercept and slope estimates taken from a linear random effects model for the response over time to give a *pseudo maximum likelihood* estimate of an 'average' rate of change over time. Wu and Bailey (1989) showed that, under this probit censoring model, the rate of change of the subject specific slope is a monotonic increasing or decreasing function of the dropout time (with the direction dependent on the sign of the dropout parameter of the probit model). By defining a model for the subject specific intercept and slope as a function of the dropout time, they derived a *linear minimum variance unbiased* and a *linear minimum mean squared error* summary estimates for the rate of change. Each use the conditional linear model to give an estimated slope for a subjects with an 'average' survival time which can be compared across different patient groups.

Although an improvement over pattern mixture and selection models, these models do still require that dropout occurs at a limited number of time points during the follow-up. They are therefore of restricted use when dropout is indeed due to death and therefore occurs at different times across all subjects. Further, subjects with fewer than two observations have to be

excluded for tractability.

These restrictions are overcome by Schlucter (1992) who generalised the conditional linear models of Wu and Bailey (1989) to allow the analysis of more unbalanced data generally seen in clinical trials and as a result of dropout due to death. Assuming that individual intercept, slope and logarithm of survival time (which may also be censored) follow a trivariate Normal distribution, an E-M algorithm was used to estimate the covariance structure of the three components and, in turn, and give appropriately adjusted estimates for the average response pattern for a randomly chosen individual with mean survival. Further, by modelling the joint distribution of survival and response, the estimated covariance between response and survival allow the conditional expectation of these parameters for different survival times to be evaluated allowing inferences which may be particularly relevant for quality of life studies.

Although Schlucter (1992) expressed his model in terms of its full likelihood and used an E-M algorithm to estimate its parameters, it may also be formulated as a multilevel model and fitted using the RIGLS algorithm described in Chapter 3 (Touloumi, 1996). This is demonstrated here using the restricted CRC HAP trial data. The transformation from the joint to the conditional model as a function of survival is also presented. These conditional inferences are then compared against those of a *conditional linear model* which simply includes known patient survival times in the multiple regression analysis. As this latter analysis is only possible when full survival data is available (that is, the time of death is known for all subjects) the CRC HAP trial full data are used for the second example.

Given the nature of dropout due to death in quality of life studies, none of the other dropout models discussed here are considered appropriate and they will therefore not be discussed further.

6.2.1 Trivariate Normal model

The multilevel version of Schlucter's log-Normal survival model treats the quality of life response and the survival as a bivariate problem. This gives a two level model for the quality of life response as described in Chapter 3, modelled alongside a *log duration* model for the survival outcome (Goldstein, 1995, Touloumi, 1996).

A response y_{ijk} where y_{ij1} is the jth quality of life response for the ith subject and y_{ij2} (j=1 for all i) is the single survival response for that subject defined as above. Two dummy variables, z_{ij1} and z_{ij2}, define these responses with z_{ij1}=1 for a quality of life response, 0 otherwise, and $z_{ij2}=1-z_{ij1}$. Given these definitions, a simple trivariate Normal model can be expressed as

$$y_{ijk}=f_1(\alpha_1+\beta_1 t_{ij1}+u_{i1}+v_{i1}t_{ij1}+e_{ij1}).z_{ij1} + f_2(\alpha_2+s_{i2}).z_{ij2} \quad \textbf{(6.1)}$$

where f_1 is an identity link function giving the simple variance components model discussed in Chapter 3 for a trend over time. The timing of the jth measurement for the ith subject, $j=1,\ldots,m_i$ is denoted t_{ij1}. As before, the subject level residuals, (u_{i1},v_{i1}), for the intercept and slope for subject i for this part of the models are assumed Normally distributed with mean **0**, and variance, $\Sigma_1=\begin{pmatrix}\sigma_u^2 & \sigma_{uv}\\ \sigma_{uv} & \sigma_v^2\end{pmatrix}$. The level one residuals, denoted e_{ij1} for the jth measurement for the ith subject are assumed to be Normally distributed with mean zero and variance σ_e^2.

Patient survival is modelled using a log duration model denoted f_2 which allows incorporation of censored individuals (Goldstein, 1995). Although the residual error for this part of the model may take a number of distributional forms (Normal, extreme value, Gamma), in this example a Normal distribution with zero mean and variance σ_s^2 is assumed. As there is only one survival observation for each subject this varies only at the patient level with residual, denoted s_{i2} for subject i. Although the survival part of the model in equation (6.1) does not include covariates, this is a simplification for clarity and not a model restriction.

Corresponding with the model used by Schlucter, these two parts of the model are combined and the level two residuals (u_{i1}, v_{i1}, s_{i2}) assumed to follow a trivariate Normal distribution with variance

$$\Omega_2 = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{us} \\ \sigma_{uv} & \sigma_v^2 & \sigma_{vs} \\ \sigma_{us} & \sigma_{vs} & \sigma_s^2 \end{pmatrix} \qquad (6.2)$$

Although all survival data (censored or observed) is used in estimation of the fixed components of this model, only the uncensored individuals contribute information in estimation of the variance components relating to the survival outcome.

Defining $\beta_1 = (\alpha_1, \beta_1)^T$ and $\sigma_{bs} = (\sigma_{us}, \sigma_{vs})^T$ from multivariate Normal theory, the conditional distribution of $\beta_1 | s$, for some known survival time s will also be Normally distributed with

$$E(\beta_1 | s) = \beta_1 + \sigma_s^{-2}\sigma_{bs}(s - \alpha_2) \text{ and } \text{var}(\beta_1 | s) = \Sigma_1 - \sigma_s^{-2}\sigma_{bs}\sigma_{bs}^T \qquad (6.3)$$

giving explicit conditional estimates for describing quality of life conditional on a given survival time. These are given explicitly in box 6.2.

As an application of this model, the restricted RSCL physical data from the CRC HAP trial

Box 6.2

Mean quality of life profiles conditional on survival

$$E(\alpha_1 | s) = \alpha_1 + \frac{\sigma_{us}(s - \alpha_2)}{\sigma_s^2} \qquad \text{var}(\alpha_1 | s) = \sigma_u^2 - \frac{\sigma_{us}^2}{\sigma_s^2}$$

$$E(\beta_1 | s) = \beta_1 + \frac{\sigma_{vs}(s - \alpha_2)}{\sigma_s^2} \qquad \text{var}(\beta_1 | s) = \sigma_v^2 - \frac{\sigma_{vs}^2}{\sigma_s^2}$$

$$\text{cov}(\alpha_1, \beta_1) = \sigma_{uv} - \frac{\sigma_{us}\sigma_{vs}}{\sigma_s^2}$$

were analysed with the basic model of equation (6.1) extended to include a treatment covariate for both the quality of life (rt_{i1}) and the survival (rt_{i2}) outcomes . Each of these variables were defined to take the value 1 for patients in the control arm, 0 for those receiving an HAI. Mean profiles of these quality of life data were given in figures 2.8 and 2.9 and suggested a slight downward trend over time, with an apparent constant treatment difference. To investigate this behaviour in the light of patient death, three progressive models were used. The first model (model one) assumed the survival and quality of life endpoints were independent, or that dropout is uninformative. In the second model (model two) a dependence between survival and intercept residuals was allowed, whereas the final model (model three) allowed a full covariance structure between survival and both subject intercept and slope. In terms of the covariance parameters of equation (6.2), models one and two correspond respectively to assuming that $\sigma_{us}=\sigma_{vs}=0$ and $\sigma_{vs}=0$ respectively. The results of each model are given in table 6.1.

Model one (the independence model) gave some evidence of a small increase in physical quality of life score over time which translated to an 0.34 (95% CI=[0.13, 0.56]) unit increase over a month. There was no evidence of a difference in level of quality of life between the two treatment arms (p=0.20). The survival coefficient gave an estimated mean log survival of 5.97 for the HAI group, translating to a geometric mean survival of 393 (95% CI=[311, 496]) days. The estimated relative survival difference was 0.64 (95% CI=[0.46, 0.89]). These results corresponded well with those presented in Appendix 1.2 which gave an estimated median survival of 404 and 274 days in the HAI and control groups respectively. The random parameter estimates showed a large degree of variation both in patient quality of life intercept and slope, but there was no convincing evidence of an association between the two.

The extension of the model to incorporate dependency between the subject intercept and

Table 6.1: Results of a trivariate Normal model for the RSCL physical quality of life scores in the CRC HAP trial.

		Model one	Model two	Model three
		Estimate (SE)	Estimate (SE)	Estimate (SE)
Fixed parameters				
α_1	($cons_1$)	10.7 (1.10)	9.88 (1.10)	9.83 (1.10)
β_1	($time_1$)	0.011 (0.004)	0.012 (0.004)	0.013 (0.004)
δ_1	(rt_1)	-1.33 (1.54)	-0.89 (1.55)	-0.94 (1.55)
α_2	($cons_2$)	5.97 (0.12)	5.95 (0.11)	5.95 (0.11)
δ_2	(rt_2)	-0.45 (0.17)	-0.38 (0.16)	-0.37 (0.16)
Variance parameters				
Level two	σ_u^2	43.0 (8.50)	44.5 (8.51)	43.4 (8.57)
	σ_v^2	0.0004 (0.0001)	0.0004 (0.0001)	0.0005 (0.0001)
	σ_s^2	0.50 (0.09)	0.48 (0.08)	0.48 (0.08)
	σ_{uv}	-0.024 (0.026)	-0.025 (0.024)	-0.017 (0.027)
	σ_{us}	-	-2.66 (0.66)	-2.52 (0.68)
	σ_{vs}	-	-	-0.002 (0.003)
Level two	σ_e^2	15.72 (1.09)	15.64 (1.01)	15.60 (1.01)
-2 log lh		3870.2	3849.4	3847.7

Within these models, *time* was recorded in days since randomisation, and treatments (rt_{i1} and rt_{i2}) were both coded with the HAI group as baseline (rt_{i1}=0 for HAI group, 1 for control).

their survival times made very little difference to the estimated survival parameter for the HAI group despite the slight change in interpretation of the coefficient now representing mean survival for subjects with an 'average' quality of life score at the start of the study. Since the direct comparison being made is conditional on average patient quality of life, there was some reduction in the estimated relative difference between the two treatment groups (0.68, 95% CI=[0.50, 0.94]). Similarly the quality of life intercept parameter had a modified interpretation being the mean intercept for a randomly selected individual with average survival. This was reflected by a reduction in the parameter estimate from 10.7 (SE=1.10) units to 9.88 (SE=1.10)

units. There was little change in the estimated average slope which retained the interpretation of the previous model. Again there was no evidence of an absolute difference in the level of physical quality of life scores between the two groups, although that of the control group was again lower than in the HAI group. The random parameter of most interest in model two was that for the covariance between intercept and survival. The change in -2 log lh following the introduction of this additional parameter was 20.9 gave strong evidence of a negative association between the level of a subject's initial quality of life and their survival. This indicated that subjects with a lower than average initial quality of life score (better quality of life) tended to have a higher than average survival. The estimated correlation coefficient was 0.58. There was very little change in all parameter estimates from model two to model three. This was not surprising given the size of the estimated covariance between survival and slope and the lack of evidence to support this as a real effect (change in -2 log lh of 1.65).

The basic assumption of this model is trivariate Normality of the level two residuals. Residual diagnostics to assess this assumption are shown in figure 6.1. Figures 6.1(a)-(c) show Normal plots of each of the standardised level two residuals in turn showing little evidence of a deviation from Normality for each univariate distribution. A Gamma plot to test for trivariate Normality is shown in figure 6.1(g). This plots the Mahalanobis distances estimated for each subject using their level two residuals and the estimated correlation matrix from model 3 against the quantiles of a Chi-squared distribution on 3 df. Like the univariate Normal plots, this gave little evidence of a deviation away from trivariate Normality. Bivariate plots of the possible pairs of residuals are given in figures 6.1(d)-(f). The strong negative association between patient survival and intercept is highlighted in figure 6.1(e).

An advantage of this model is that it allows estimation of the conditional distribution of response given survival time using equation (6.3). This was done using the results given in

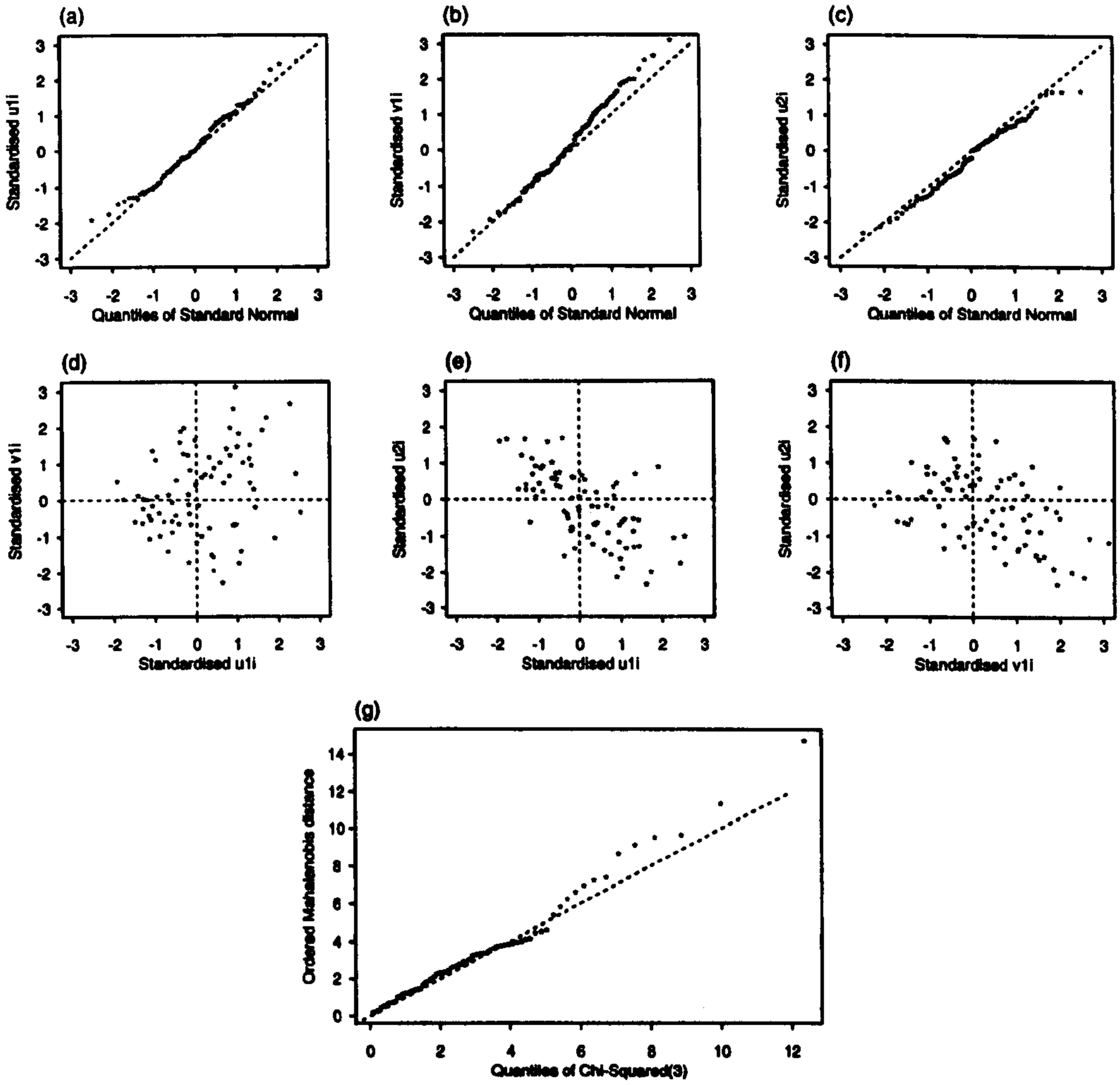

Figure 6.1: Residual diagnostics for the level two residuals of the trivariate Normal model for the restricted data of the CRC HAP trial (a)-(c) give univariate Normal plots; (d)-(f) give bivariate scatter plots; and (g) a Gamma plot.

model three for particular values of s corresponding to 6 to 30 months in 6 month intervals. The resulting expected conditional intercepts and slopes for each of these values are given in figure 6.2 for the HAI and control groups. The intercept and slope are plotted separately as a function of observed survival time in figures 6.2(a) and 6.2(b). Realisations of these average profiles for the HAI and control group respectively are shown in figures 6.2(c) and 6.2(d).

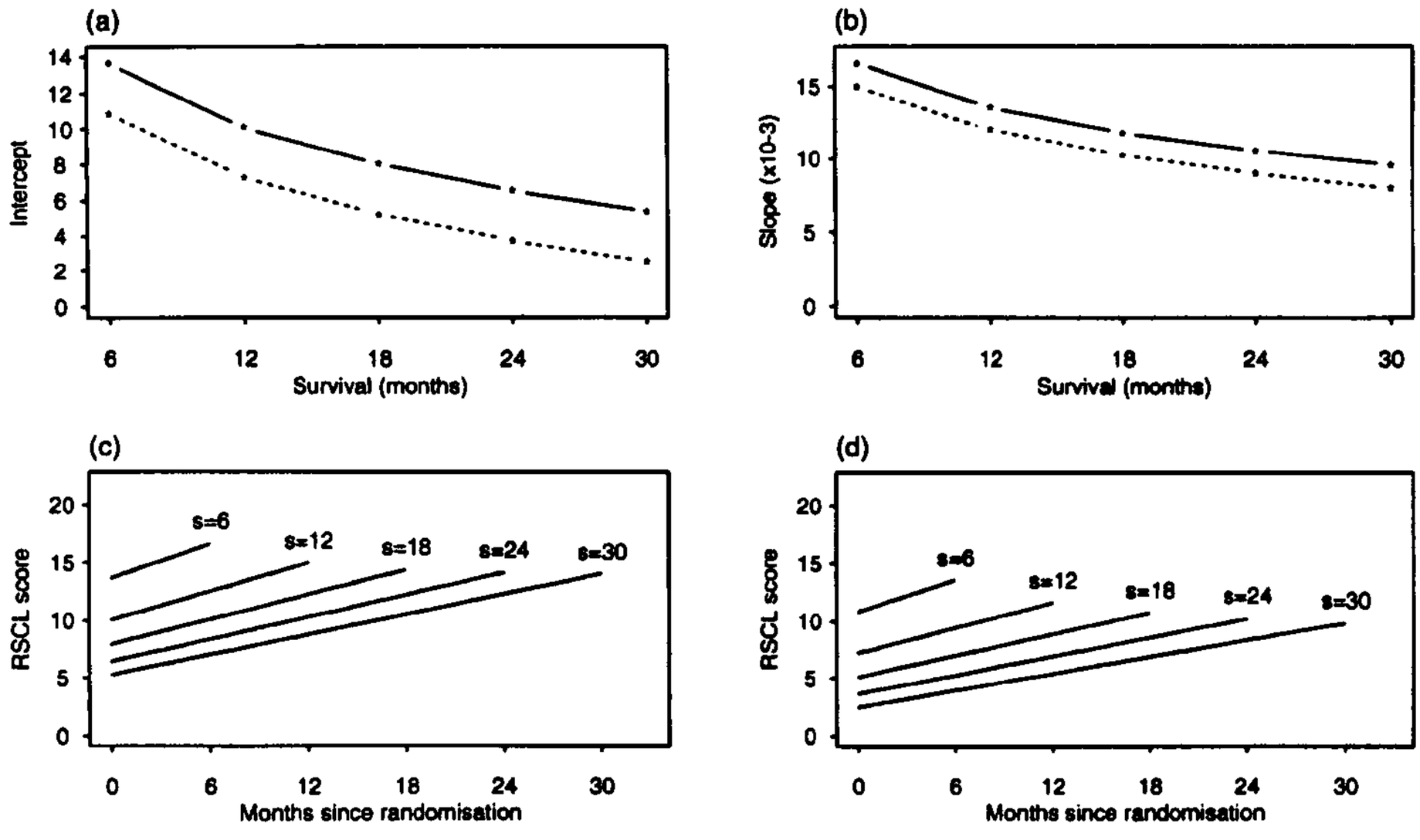

Figure 6.2: Quality of life profiles conditional on survival as estimated for the restricted physical quality of life from the CRC HAP trial from the trivariate Normal model summarised by (a) average intercept and (b) average slope as a function of survival time. HAI:————; control: --------, and average profiles for survival, s=6, 12, 18, 24, 30 months (c) HAI and (d) control.

These figures clarified the relationships indicated by the covariance structure of the joint distribution showing lower initial physical scores (better quality of life) for patients with increased length of survival. Although there was not the same convincing evidence of a relationship between rate of change and survival, the same pattern of a lower rate of change for increasing survival was also illustrated. The realisations of these average profiles when plotted over time showed that, although starting from different points on the physical quality of life scale, by death the profiles had all tended to reach a similar point in the level of quality of life. This suggested that differences in quality of life for different survival times may be due simply to patients being at different states of their disease progression rather than an overall difference in their experience. An important feature highlighted by this analysis was the very different conclusions of the descriptive analyses of figures 2.8 and 2.9 which indicated slight downward trends in quality of life scores in contrast to the positive trends shown here. As such positive

trends were also seen from model one which did not condition on patient survival, this emphasises the need not only for analyses that adjust for patient survival, but more importantly, the problems associated with these descriptive analyses that ignore repeated measurement data structure.

6.2.2 Conditional linear model

If survival time was known for all subjects, the problem of obtaining the conditional inferences shown in figure 6.2 would be simple using a multilevel or random effect structure as discussed in Chapter 3. For example, a model to examine the rate of change in response y_{ij} for subjects i, i=1,...,n over time, t_{ij} for j=1,...,m_i, dependent on survival s_i could simply be written

$$y_{ij}=\alpha+\beta t_{ij}+\zeta s_i+\xi s_i t_{ij}+u_i^c+e_{ij}^c \qquad (6.4)$$

where ζ is the difference in the intercept and ξ is the difference in rate of change of response for each additional day of survival. This model can then be fitted as demonstrated in Chapter 3.

The fundamental difference between this model and the conditional transformation of the trivariate Normal model is that within the conditional linear model, the distribution of the survival times is left unspecified whereas in the trivariate Normal model they are assumed to follow a log Normal distribution. To assess the implications of this difference, both were applied to the full data from the CRC HAP trial. In these full data, survival times were known for all but three patients who were therefore necessarily excluded. The trivariate Normal model used corresponded to model three of the previous section and fitted a full covariance structure. For the conditional linear model, the model of equation (6.4) was extended to include a treatment coefficient. For consistency with the trivariate Normal model, survival was modelled

Table 6.2: Results for a trivariate Normal versus a conditional linear model for the RSCL physical quality of life data in the CRC HAP trial full data.

		Trivariate Normal model			Conditional linear model
		Estimate (SE)			Estimate (SE)
Fixed parameters					
α_1	($cons_1$)	10.52 (1.09)	α	(*cons*)	10.83 (0.97)
β_1	($time_1$)	0.018 (0.004)	β	(*time*)	0.015 (0.003)
δ_1	(rt_1)	-1.57 (1.57)	δ	(*rt*)	-1.55 (1.54)
α_2	($cons_2$)	5.89 (0.10)	ζ	(*surv*)	-3.99 (1.04)
δ_2	(rt_2)	-0.34 (0.15)	ξ	(*time.surv*)	-0.019 (0.006)
Variance parameters					
Level two	σ_u^2	43.8 (8.47)		$\sigma_{u^c}^2$	34.7 (7.00)
	σ_v^2	0.0005 (0.0001)		$\sigma_{v^c}^2$	0.0003 (0.0001)
	σ_s^2	0.50 (0.08)			
	σ_{uv}	-0.002 (0.003)		$\sigma_{u^c v^c}$	-0.02 (0.02)
	σ_{us}	-2.09 (0.62)			
	σ_{vs}	-0.007 (0.003)			
Level one	σ_e^2	16.13 (1.05)		σ_e^2	16.18 (1.05)
-2 log lh		3849.6			3663.6

on the log scale as a deviation from the estimated group mean from the trivariate Normal model for these data (5.89 and 5.55 for the HAI and control groups respectively (table 6.2)). The parameterisation of the model in this way enables a direct comparison of the parameters $\{\zeta,\xi\}$ of the conditional linear model and $\{\sigma_s^{-2}\sigma_{us},\sigma_s^{-2}\sigma_{vs}\}$ used in the transformation from joint to conditional inference for the trivariate Normal model.

Comparing the results for the trivariate Normal model using the full data and those for the restricted data given in the final column of table 6.1 showed a number changes although none

were substantial. These may be attributed both to the difference in samples, and more importantly, because the fixed parameters (α_1, β_1) represent estimated effects for subjects with mean survival (and vice versa) some deviation can be expected.

Given the parameterisation of survival times which has been used, it is possible to compare the parameter estimates of the two models. For instance, the estimates of β_1 and β are directly comparable and give the average rate of change for survival times equal to the respective group means. Similarly, the intercept terms α_1 and α correspond. In both cases, good agreement between the estimates was seen (10.5 versus 10.8 for the intercepts, and 0.018 versus 0.015 for the slopes). Estimates of $\{\sigma_s^{-2}\sigma_{us}, \sigma_s^{-2}\sigma_{ws}\}$ from the results in table 6.2 were $\{-4.18, -0.014\}$. Again these were in good agreement with those of the conditional linear model for $\{\zeta, \xi\}$ of $\{-3.99, -0.019\}$. The estimates of the covariance structure of the conditional model calculated from the trivariate Normal model (using the equations given in box 6.2) corresponded well with those of the conditional linear model with $\delta^2_{u_c}$=35.1, $\delta^2_{v_c}$=0.0004 and $\delta_{u_c v_c}$=-0.04. from the trivariate Normal model versus 34.7, 0.0003 and -0.02 from the conditional linear model.

As in the previous section, realisations of the conditional model for a number of survival times have been plotted. Figure 6.3(a) and (b) give the results for the trivariate Normal conditional transformation, (c) and (d) give the equivalent figures for the conditional linear model. These highlight a substantial difference between the inferences drawn form the two models with figures 6.3(a) and (b) showing less variability with changing *s* than those of the conditional linear model in figures 6.3(c) and (d). This is thought to be due to the assumption of log Normality of survival times restricting the behaviour of the estimates of the trivariate Normal model. In terms of the conclusions of the analyses for the data, these analyses gave very different quality of life profiles for different survival times to those of the restricted data. In contrast to the respective plots for the restricted data shown in figure 6.2, they do not suggest

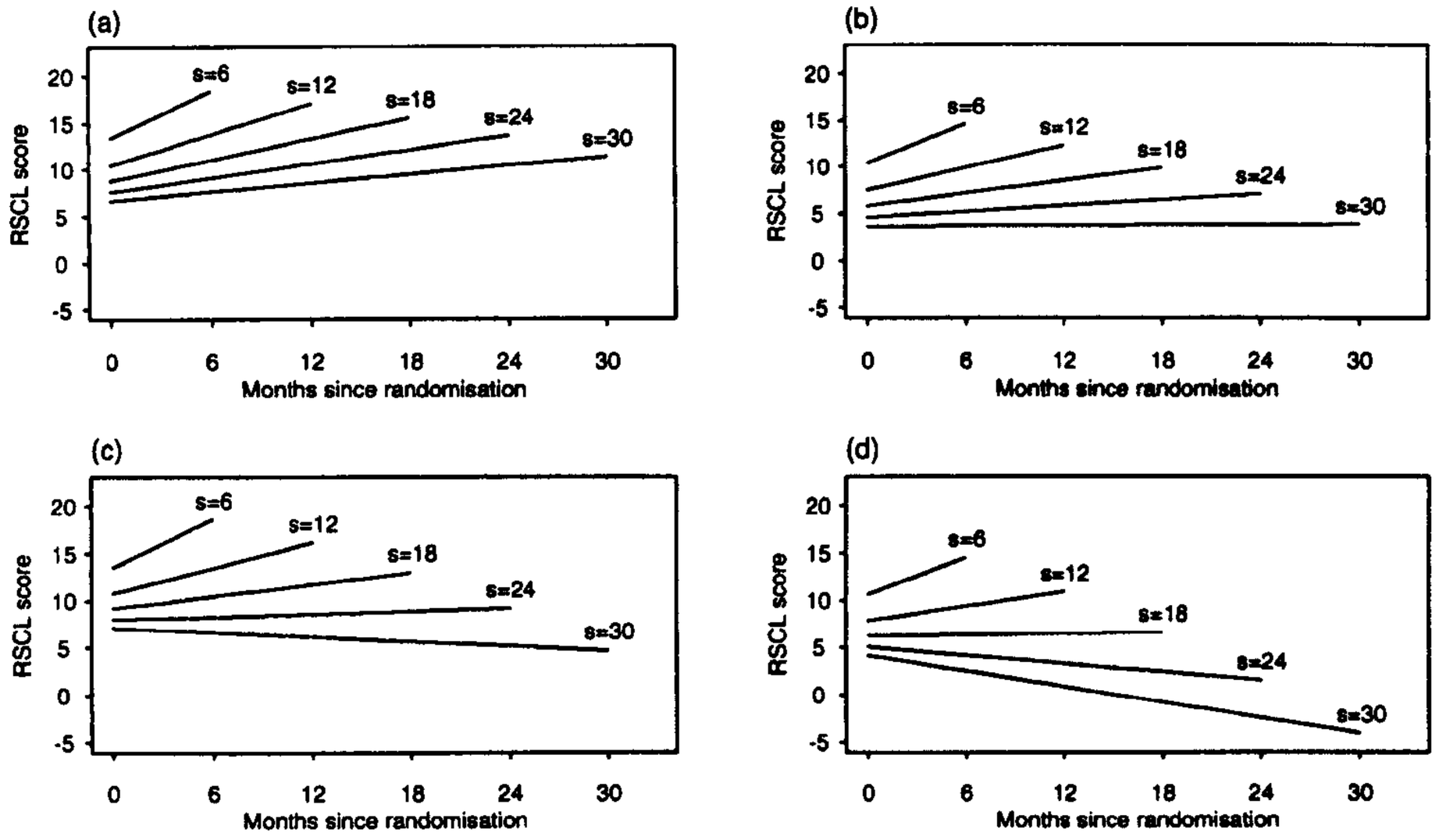

Figure 6.3: Average quality of life profiles for survival of 6, 12 , 18, 24, 30 months, (a)-(b) as estimated from the trivariate Normal model for the HAI and control groups and (c)-(d) as estimated from the conditional linear model for RSCL physical quality of life in the CRC HAP trial.

a similar profile of quality of life for different survival times. Rather, they show the longer term survivors to have a more gradual decline in quality of life than the short term survivors (from the conditional linear model, p=0.002).

6.2.3 Conclusions

These analyses have demonstrated the use of a trivariate Normal model to analyse quality of life and survival data together to obtain inferences about the joint distribution of the two outcomes, or inferences about patient quality of life conditional upon survival times. The analysis performed well and gave clear and comprehensive representation of the data which could be fairly easily explained to a clinician or patient.

In comparison with the conditional linear model which may be seen as a simple solution to

analysing quality of life data alongside survival, the trivariate model performed well although more work is needed to determine the properties of this model particularly when a substantial amount of data are censored. As the analysis can be easily extended to include more covariates than have been considered here, can be fitted easily even when some subjects are censored, and may be applied to continuous, binary or ordinal data, it may serve as a useful tool in future practical data analysis of quality of life data.

6.3 Quality adjusted survival analysis

Rather than considering quality of life and survival as separate endpoints, quality adjusted survival analyses have also been proposed (Schumacher *et al.*, 1991, Korn, 1993, Glasziou, 1995). The combination of quality of life and survival outcomes to give a *quality adjusted survival* was first, and still is commonly, used in health economics for decision making (Weinstein and Stason, 1977). In its most simple form, it is assumed that patient quality of life may be divided into distinct health states (s=1,...,S) which are assigned weights, $w=(w_1,...,w_S)$, reflecting the value of survival spent in each state. Applying these weights to the time spent in each state, denoted t_{is} for subject i in state s, a patient's *quality adjusted life years* (*QALYs*) is then defined as the sum of the weighted times spent in each state

$$QALY_i=\sum_{s=1}^{S} w_s t_{is} \qquad (6.5)$$

These weighted times are then summarised and compared across different patient groups of interest. In recent years, such methods have been adapted and used for the analysis of patient toxicity data in cancer and AIDS research under the name of TWiST, that is *Time WIthout Symptoms and Toxicity* (Gelber and Goldhirsch, 1986, Gelber *et al.*, 1991, Gelber *et al.*, 1992). In its original form, the TWiST metric combined patient survival and toxicity data by subtracting periods of time during which the patient experiences toxic effects of treatment or

symptoms of disease recurrence from overall survival (Gelber and Goldhirsch, 1986). In addition, for some side effects or symptoms additional survival time was subtracted to allow for recovery. It was therefore equivalent to *QALY*s where the specific symptoms represented the health states with the weights assigned to these health states being less than or equal to zero.

When survival times are complete for all subjects, a comparison of *QALY*s across patient groups is straightforward using usual methods for continuous outcome data to compare the mean *QALY*s across patient groups. When some survival times are censored, as is often the case in clinical trials, survival analyses techniques have been used. It has been shown however, that such techniques can lead to biased estimation of the true distribution of *QALY*s in the form of overestimation of the survival function. This is caused by the weighting of individual survival times because patients who have poorer quality of life accumulate quality adjusted time very slowly, and are therefore more prone to early censoring leading to an underestimation of the hazard function, and an overestimation of the survival function (Gelber *et al.*, 1989). The problem can be reduced to some degree by restricting the follow-up period by some upper limit, L, and focusing the analysis on *QALY*(L), the amount of *QALY* accumulated within L time units. This has the effect of reducing the extent of censoring, and in turn the extent of bias. Unfortunately simulation studies have shown that the degree of bias still remains fairly high even when the amount of censoring is low (Gelber *et al.*, 1989). An alternative solution which has been shown to perform better is to impute *QALY*s for censored individuals and then to perform an analysis as for uncensored data. This is discussed in more detail in Section 6.3.1.

In the light of these problems with survival analyses for censored *QALYs*, the TWiST metric was redefined using a *partitioned quality adjusted survival analysis* (PQAS). Rather than weighting individual patient TWiST, a PQAS analysis weights the estimated group means for time spent in each health state. Its limitation is that it requires that the health states are

progressive, with the final state defined as patient death. For each other state, the 'survival' event of interest is defined as an individual progressing from that state to the next. The analysis is then performed by estimating the survival for each state in turn. By restricting survival by some upper limit L, the restricted mean survival for each state is estimated by the finite area under each survival curve. The mean time spent in each state is then calculated by subtracting the observed mean survival in state s from that observed in state $s+1$. Q-TWiST is then defined as the weighted sums of these restricted mean survival times for some weights $w=(w_1,...,w_S)$. Although a simple formula for the variance of the estimates is not available, variances can be obtained by bootstrapping, thus enabling formal comparisons of Q-TWiST across patient groups (Glasziou *et al.*, 1990).

Since the quality adjusted techniques outlined so far have relied on the existence of definable health states they may be inappropriate for the analysis of continuous self assessed quality of life. As an alternative, Glasziou *et al.* (1995) note that a possible definition of the rate of gain of *QALYs* at time t is simply the product of the proportion of people still alive at time t multiplied by the average quality of life of the survivors at time t. The *integrated survival-quality product* or mean *QALYs* over some time period bounded by an upper limit L, is then estimable by calculating the area underneath this profile. A similar analysis has also been suggested by Korn (1993) in which, rather than summarizing the quality of life at each time point for the group as a whole, individual patient profiles are first summarized by calculating the area underneath the profile, where the profile is bounded by either the patient's survival or censoring time. These potentially censored quality of life summaries for each individual are analysed for the group as a whole. Unfortunately, in the same way as for the quality adjusted times from health state models, conventional survival estimates for the distribution of these individual summaries will be biased because of induced informative censoring on the quality adjusted time scale. Although Korn (1993) gives an adjusted survival

estimation algorithm which reduces this bias to some degree, the method can be computationally intensive and can cope only with quality of life measured at a limited number of occasions. Its practical application to the much quality of life data in cancer trials is therefore limited and its use will not be pursued further here.

Although the use of these quality adjusted survival techniques have been much discussed in the quality of life literature (Cox *et al.*, 1992, Fayers and Jones, 1983, Schumacher *et al*, 1991) very little practical application of use with self assessed quality of life data in clinical trials has been reported. Indeed only two practical applications have been found (Allen-Mersh *et al*, 1994, Korn, 1993). The aim of this section of work is therefore to assess the methods outlined above to determine whether they are appropriate for future reporting of self assessed quality of life data where large numbers of individuals have died during the measurement process. To do this the RSCL physical scores from the CRC HAP trial are analysed using modifications to TWiST, PQAS and the integrated quality-survival product in turn. Along with each example, the particular analysis is discussed in more practical detail than given so far. Results for both the restricted and full data sets will be given. For the two health state methods, states were defined on the basis of the recommended RSCL 'normal' physical score classification, that is a score of less than 20 units. For both the restricted and full data, approximately 13% of total follow-up time was classed as abnormal. A discussion of the problems and advantages of each analysis is given in Section 6.3.4.

The lack of practical application of all these methods has meant that software is not available. S-Plus (Becker *et al.*, 1988) functions have therefore been written for each case. These are listed in Appendix 3.

6.3.1 Time with normal quality of life

Time with normal quality of life (TNQOL) is here defined as that period of a patient's survival spent with 'normal' RSCL scores. Its definition is based on that of the TWiST metric but is more specific to patient quality of life as evaluated by self assessed questionnaires rather than patient symptoms and toxicity. In similar notation to that of Gelber *et al.* (1989) it was defined as $TNQOL_i=TR_i-AQOL_i$ where TR_i is the time from the start of treatment to death and $AQOL_i$ is the amount of time of abnormal quality of life patient i experiences during this time. Although these quantities may or may not be observed because of censoring, given U_i, the follow-up time for patient i, their observed values can be obtained by $OTR_i=U_i$, and $OTNQOL_i=OTR_i-OAQOL_i$ where $OAQOL_i$ is the observed amount time spent with abnormal quality of life. The censoring variable for OTR_i and $OTNQOL_i$ is given by $\delta_i=1$ if a death is observed, 0 otherwise.

In their published report of the CRC HAP data, Allen-Mersh *et al.* (1994) analysed such $OTNQOL_i$ for $i=1,...,n$, using Kaplan-Meier estimation of the survival function, and comparing the two treatment arms with a log rank test. They concluded that there was a "*significant prolongation in normal (...) survival for physical symptoms (p=0.04)*" in HAI treated patients compared with controls. However, as discussed earlier, the discounting of patient time with abnormal quality of life results in informative censoring. This requires some refinement to these simple survival analyses to assess the robustness of conclusions to varying severity of plausible bias. Based on the work of Gelber *et al.* (1989) this was done by defining a bounded TNQOL, $OTNQOL(L)_i$, defined as the amount of TNQOL accumulated within L days from the start of treatment. This was given by $OTNQOL(L)_i=OTR(L)_i-OAQOL(L)_i$ where $OTR(L)_i=\min(L,U_i)$ and $OAQOL(L)_i$ is the length of time spent with abnormal quality of life within L time units. The censoring variable for $OTR(L)_i$ and $OTNQOL(L)_i$ is then $\delta(L)_i=1$ if $OTR(L)_i=L$ or the patient has died, 0 otherwise. Thus a patient is only considered censored if they are still

alive, but have not been observed for L time units.

With the exception of the first and last recorded measurements, when 0 and the time of death or censoring were used, it was assumed that the transition between quality of life states took place midway between the current and the previously observed measurement time. Based on these assumptions, the definitions above are illustrated by a hypothetical example in table 6.3.

As the choice of L influences the results of the analysis, a range of values were used with their results plotted against the value of L thus giving an indication whether the analysis conclusions change with increasing L, as well as assessing the possible extent of bias. For the illustrated example the bounds were chosen at 6, 12, 18 and 24 months, as well as for the maximum follow-up within the sample which was 1273 days (42 months). These bounded times were then analysed either by Kaplan-Meier estimation of the censored $OTNQOL(L)_i$ or by a two sample *t*-test of an uncensored data set with maximum, minimum or mean values imputed for censored individuals. These were again based on the definitions of Gelber *et al.* (1989) and are summarised in box 6.3. These authors noted that for values of L no larger than median follow-

Table 6.3: Hypothetical example illustrating accumulation of TNQOL.

	Time of measurement										
	Quality of life score					*OAQOL*	*OTNQOL*	δ	*OAQOL*(L)	*OTNQOL*(L)	δ(L)
A	**23**	**51**	**79**	**107**	**135**	65	85	0	65	75	1
	21	24	12	14	16						
B	**42**	**70**	**98**	**116**	**154**	28	143	1	28	112	1
	15	14	18	21	16						

Within the body of the table the numbers given in bold type refer to the time of quality of life measurement in days since randomisation with the quality of life scores given in normal font. Patient A was censored at 150 days, and patient B died at 171 days. In the case of the definitions requiring an upper bound for survival, L=140 days was assumed.

Box 6.3

Definition of imputed $OTNQOL(\mathrm{L})_i$ for censored individuals

$$\max(OTNQOL(\mathrm{L})_i) = \mathrm{L} - OAQOL_i$$

$$\min(OTNQOL(\mathrm{L})_i) = U_i - OAQOL_i$$

$$\mathrm{mean}(OTNQOL(\mathrm{L})_i) = U_i - OAQOL_i + \frac{1}{2}(\mathrm{L} - U_i).$$

up, all four methods performed very well. As median survival in the restricted CRC HAP trial data were 404 and 274 days for the HAI and control groups respectively, it was expected that estimation beyond 18 months may not perform as well as that at 6 and 12 months.

The Kaplan-Meier estimated survival curves for censored TNQOL for the restricted data are shown in figure 6.4 and shows a very slight early advantage for the control group which was reversed later in follow-up. Within these data, 22% and 15% of the HAI and control groups

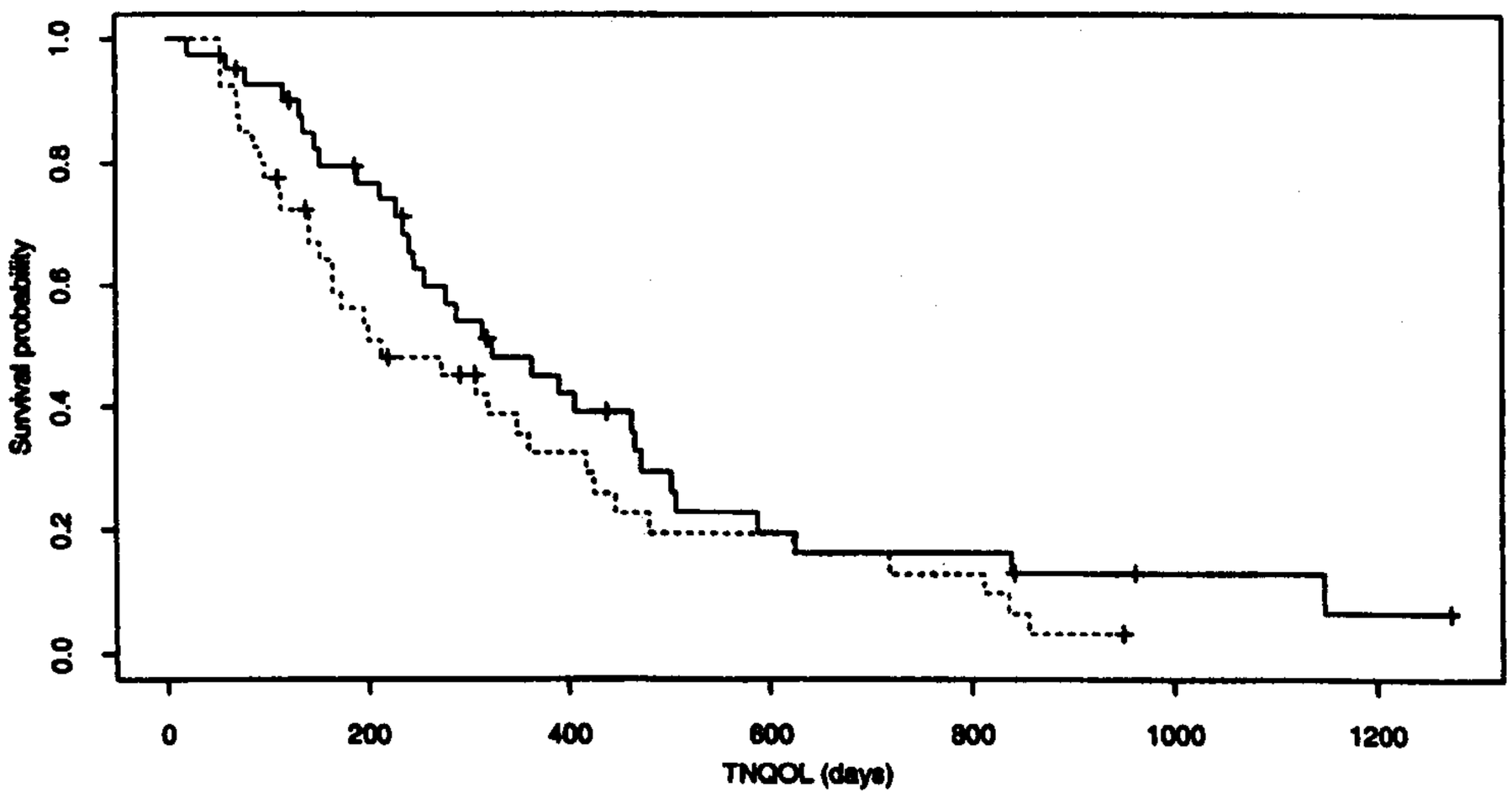

Figure 6.4: Kaplan-Meier TNQOL based on the RSCL physical scores for the restricted data of the CRC HAP trial. HAI:————; control: --------. Censored observations are marked +.

respectively were censored. Although these proportions were not high, it was observed by Gelber *et al.* (1989) that even when the degree of censoring was only 10%-20%, the Kaplan-Meier derived estimates may still be subject to an unacceptably large bias. The truncated and imputed methods to approximate for the bias are given in table 6.4 for L=1273 days (42 months). Mean survival for the former was estimated by calculating the area under the respective truncated Kaplan-Meier curves. Standard errors for these estimates were obtained by bootstrapping. A simple bootstrap was used which repeatedly re-sampled from the data with replacement and estimated the area under the curve for the resulting sample. This was done separately for each treatment group. The number of repeated samples was arbitrarily chosen to be 2000. The confidence intervals for the estimated mean difference were calculated assuming unequal variances with the degrees of freedom based on Welch's test. The final column of the table gives a *p*-value for an alternative test to Welch's test in each case. For the

Table 6.4: Mean TNQOL(L) in days over a 42 month period (1273 days) estimated using Kaplan-Meier (K-M) and three different imputation methods (detailed in box 6.3) for the RSCL physical scores for the CRC HAP trial.

	Mean TNQOL(L) (SE)		Mean difference [95% CI]	*p* value (Welch's test)	*p* value (alternative test)
	HAI	Control			
Restricted data					
K-M area under curve	441.7 (62.7)	326.6 (45.1)	115.1 [-38.9,269.1]	0.141	0.120
Maximum imputation	475.7 (56.7)	380.8 (54.3)	95.0 [-63.3, 251.3]	0.230	0.230
Minimum imputation	366.1 (45.5)	290.8 (39.6)	75.3 [-44.8, 195.4]	0.215	0.216
Mean imputation	461.2 (49.8)	361.0 (47.1)	100.2 [-36.2, 236.6]	0.148	0.148
Full data					
K-M area under curve	448.1 (47.8)	310.7 (40.5)	137.4 [12.8, 262.0]	0.031	0.033

Kaplan-Meier analysis, a log rank test comparing the survival curves was used, and for the imputation methods, an unpaired two sample *t*-test assuming equal variances. At 1273 days (42 months) 9 subjects in the HAI group and 6 in the control group were censored. Also given in the table are the results from an analysis of the full data over the same 42 month period. These results are also presented over time in figure 6.5 which shows the estimated mean difference from each of the analyses of the restricted data for all values of L plotted against L.

Naturally there was some difference in the point estimates for each of the four estimation procedures, but all the analyses of the restricted data in table 6.4 gave consistent results with an estimated treatment difference in mean TNQOL accumulated within 42 months in favour of the HAI group. Although these were results were based on a slightly smaller data set, they were consistent with the conclusions of Allen-Mersh *et al.* (1994). The strength of evidence for all

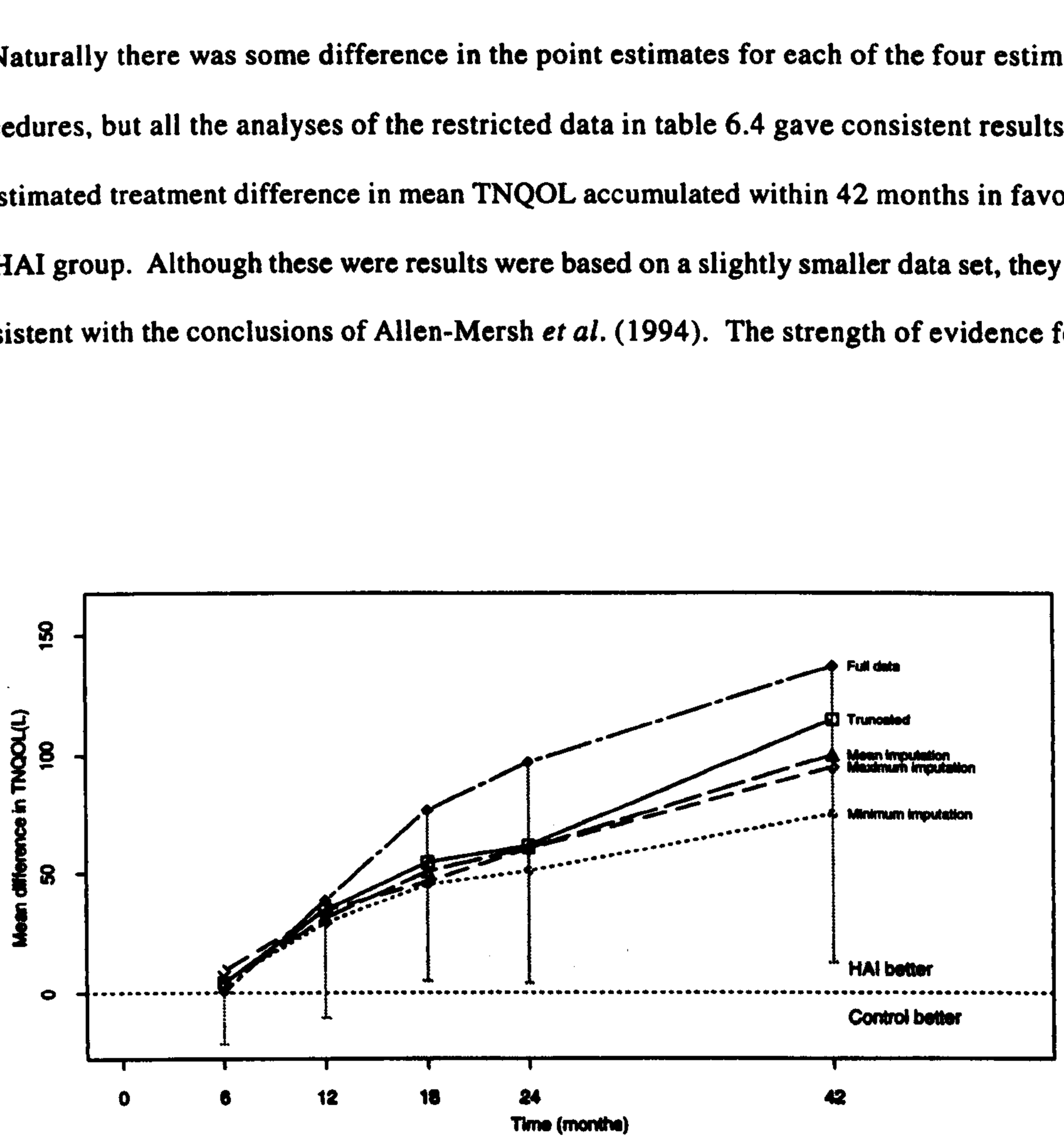

Figure 6.5: Mean difference in TNQOL(L) (HAI-Control) estimated for restricted data using K-M following truncation (□), maximum (◇), minimum (○) and mean (△) imputation with the observed TNQOL(L) for the full data (♦). Vertical bars show the lower bounds of a 95% CI for the full data.

of the analyses using the restricted data was, however, not convincing in the restricted data. At 42 months the difference between the results of each of the three imputation methods and that of the truncated Kaplan-Meier for the restricted data was at its greatest and may indicate bias in the latter analysis as a result of censoring. At earlier time points (when censoring was naturally lower) the differences between the four methods were much reduced. This replicates the findings of Gelber *et al.* (1989). Over the 42 month period, although never reaching statistical significance, all of the results for the restricted data showed better quality of life for patients receiving HAI over the controls. Such behaviour was also seen with the analyses of the full data.

A possible criticism of this analysis, is the choice of definition for 'normal' quality of life. Although the chosen cut-off in score was that recommended by the RSCL developers, the sensitivity of the results to the cut-off were assessed by lowering the cut off to 16. These results are presented in table 6.5 for each of the 4 approximate methods for the smaller data as well as for the full data. Although obviously reducing the estimated mean TNQOL, in the HAP trial example, changing the cut off in this way did not change the conclusions of the analysis that the HAI patients had a longer TNQOL(L) than the control patients although without the full data, which is now available, there was no convincing evidence to conclude that this was due to a real treatment effect rather than due to chance.

6.3.2 Partitioned quality adjusted survival analysis

A partitioned quality adjusted survival (PQAS) analysis assumes that patient quality of life moves progressively through a number of defined health states and thus requires the definition of appropriate progressive health states. Given the invasive nature of the HAI treatment and thus the early expected poor quality of life, improving later in follow-up, for a PQAS analysis of the CRC HAP trial data three progressive states (s=1,...,3) were defined as '*treatment related*

Table 6.5: Mean TNQOL(L) in days over a 42 month period (1273 days) estimated using Kaplan-Meier (K-M) and three different imputation methods for a 'normal' quality of life cutoff of 16 units.

	Mean TNQOL(L) (SE)		Mean difference [95% CI]	*p* value (Welch's test)	*p* value (alternative test)
	HAI	Control			
Restricted data					
K-M area under curve	399.4 (62.6)	314.6 (45.9)	84.9 [-68.1, 237.9]	0.273	0.316
Maximum imputation	44.3 (58.9)	368.5 (55.4)	74.8 [-86.2, 235.8]	0.358	0.359
Minimum imputation	322.1 (44.5)	277.8 (40.1)	44.3 [-74.9, 163.5]	0.461	0.462
Mean imputation	417.2 (49.5)	348.0 (47.9)	69.1 [-68.0, 206.3]	0.319	0.319
Full data					
K-M area under curve	393.1 (46.1)	291.3 (40.8)	101.8 [-20.7, 224.2]	0.102	0.100

abnormal quality of life', '*normal quality of life*' and '*quality of life deterioration*'. These definitions are described in table 6.6 and give a conservative estimate of the amount of time spent with normal quality of life scores. This is because once a patient had recorded a quality of life deterioration, all remaining survival time was classed as abnormal regardless of whether

Table 6.6: Partitioned quality adjusted survival state definitions for PQAS analysis of the CRC HAP RSCL physical data.

State name	Definition
Treatment related abnormal quality of life	Abnormal quality of life immediately following treatment. A patient is deemed to have left this state on the first occurrence of a normal quality of life score.
Normal quality of life	Periods of normal quality of life scores. The time of leaving this state is defined by the first occurrence of an abnormal quality of life score following at least one normal score.
Quality of life deterioration	Periods of abnormal quality of life following periods with normal scores. A patient leaves this state at death regardless of their quality of life score.

Table 6.7: Hypothetical example illustrating transition times for a PQAS.

	Time of measurement					Treatment related abnormal quality of life		Normal quality of life		Quality of life deterioration	
	Quality of life score					Survival	Status	Survival	Status	Survival	Status
A	**23**	**51**	**79**	**107**	**135**	**65**	1	150	0	150	0
	21	24	12	14	16						
B	**42**	**70**	**98**	**116**	**154**	0	1	107	1	171	1
	15	14	18	21	16						

Within the body of the table the numbers given in bold type refer to the time of quality of life measurement in days since randomisation with the quality of life scores given in normal font. Patient A was censored at 150 days, and patient B died at 171 days.

normal scores were subsequently recorded. This is illustrated by patient B in the hypothetical example given in table 6.7. As with the TNQOL(L) example, the cut-off for a normal score was taken according to the RSCL guidelines at 20 units and the time of state to state transitions, if they occurred, were taken as midway between the time of current and the previously observed measurement.

Having determined the survival and censoring indicator for each patient and each state, the mean quality adjusted survival time was estimated as described by Glasziou *et al.* (1990). Survival curves, $S_s(t)$, for the survival in state s or worse, were estimated for each state separately. Given $\hat{S}_s(t)$ (s=1,...,3) and an upper bound, L, the restricted mean survival, $\mu_s(L)$, was then estimated by calculating the area under $\hat{S}_s(t)$ bounded by L. The restricted mean time spent in each state, $T_s(L)$, was then estimated by

$$\hat{T}_s(L)=\hat{\mu}_s(L)-\hat{\mu}_{s-1}(L) \qquad (6.6)$$

Assuming that $\hat{\mu}_0(L)=0$, this difference gives the estimated mean time spent in each adjacent states. Finally, for some vector of weights, $w=(w_1,...,w_3)$, the restricted mean quality adjusted survival, $QAS(L)$, was estimated by the weighted sum of these differences,

$$\hat{QAS}(\mathrm{L})=\sum_{s=1}^{3} w_s \hat{T}_s(\mathrm{L}) \qquad (6.7)$$

The variance of this quantity, $\mathrm{var}(\hat{QAS}(\mathrm{L}))=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{w}$, where $\boldsymbol{V}=\mathrm{var}(\hat{\boldsymbol{T}}(\mathrm{L}))$ for $\hat{\boldsymbol{T}}(\mathrm{L})=(\hat{T}_1(\mathrm{L}),\ldots,\hat{T}_3(\mathrm{L}))$ and was obtained by bootstrapping. Again this was a simple bootstrap of size 2000 in which subjects within each treated group were repeatedly re-sampled with replacement. The empirical estimate of the variance of $\hat{\boldsymbol{T}}(\mathrm{L})$ estimated for each sample was then used as an estimate for $\boldsymbol{V}$.

For the analyses presented, a range of upper bounds for survival corresponding to 6, 12, 18, 24 and 42 months were used in order to check for consistency in the analysis conclusions for increasing L. The estimated restricted mean times spent in each state for the restricted and the full data are shown in table 6.8 for L=1273. The partitioned estimated survival curves for the restricted data are shown in figure 6.6.

As expected, the analysis showed that patients who received the HAI treatment spent on

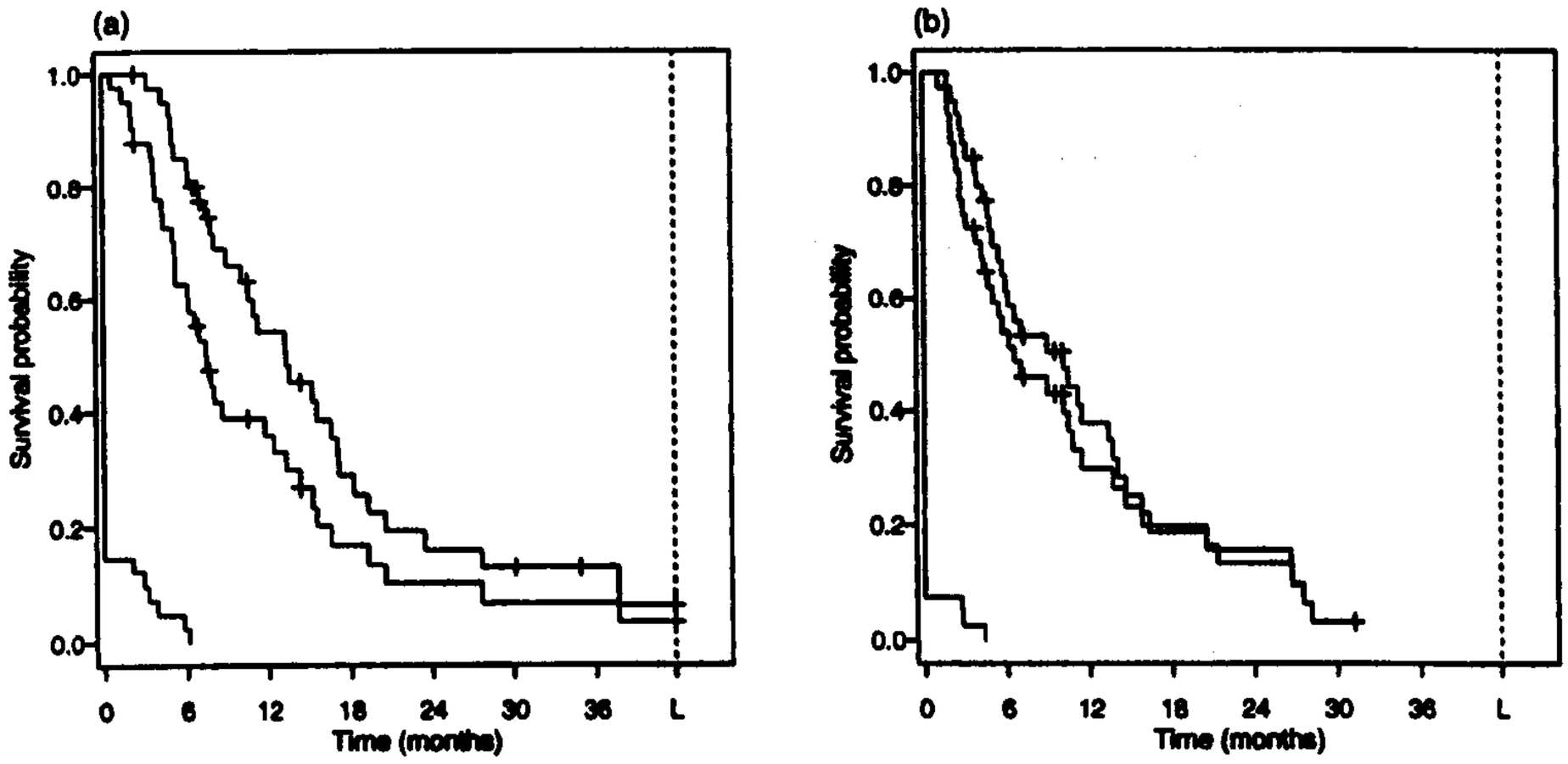

Figure 6.6: Partitioned quality adjusted survival analysis for the restricted CRC HAP trial RSCL physical quality of life data: (a) HAI; (b) control. Censored observations are marked +.

Table 6.8: Restricted mean (SE) survival time (days) in each progressive health state based on 42 months (1273 days) follow-up for both the full and the restricted CRC HAP trial RSCL physical quality of life data.

State	Restricted data		Full data	
	HAI	**Control**	**HAI**	**Control**
Treatment related abnormal quality of life	17.8 (7.29)	7.36 (4.24)	22.0 (7.98)	7.01 (4.02)
Normal quality of life	322.2 (52.2)	306.4 (46.6)	349.1 (47.9)	290.7 (40.9)
Quality of life deterioration	142.9 (46.8)	36.6 (12.7)	123.7 (38.7)	36.2 (12..2)

In order that the area underneath each survival curve is defined, an upper bound for survival was required. For the data presented in the table this was 42 months (1273 days) for both the restricted and full data.

average a greater length of time with abnormal quality of life following treatment with 17.8 days versus 7.36 in the control group. The mean time spent with normal quality of life scores was also greater in the HAI group, although the relative difference between the groups was much less than that seen for the first state. Similarly, the mean survival following a deterioration in quality of life was greater in the HAI group. Over the entire follow-up period, this implies that for any given positive weights less than one assigned to each state, the overall

Table 6.9: Estimated mean *QAS* for HAI and control groups for different choice of weights for a PQAS analysis of the restricted RSCL physical data of the CRC HAP trial.

Weights			QAS (SE)			
w_1	w_2	w_3	**HAI**	**Control**	**Difference [95% CI]**	***p* value**
0	1	0	322.2 (52.2)	306.4 (46.6)	15.8 [-123.5, 155.1]	0.822
0.5	1	0.5	402.5 (50.0)	328.4 (44.5)	74.1 [-59.2, 207.4]	0.272
0.25	1	0.75	366.8 (49.3)	319.2 (45.2)	47.6 [-85.6, 180.7]	0.479
0.75	1	0.25	371.2 (49.0)	321.0 (44.9)	50.2 [-82.1, 182.5]	0.755
0.9	1	0.9	466.8 (56.2)	345.9 (43.5)	120.0 [-20.7, 262.5]	0.093
1	1	1	482.9 (58.6)	350.3 (43.4)	132.6 [-12.7, 277.9]	0.073

weighted restricted survival time will always favour the HAI treated arm of the study.

Although the direction of effect was not altered by the choice of weights given to each state, obviously the size of the effect is still dependent on the choice and hence also the level of statistical significance. This is illustrated in table 6.9 for different choices of weight. The extremes of reasonable positive weights are shown in the first and sixth row of this table and corresponding to counting only time with normal quality of life and discarding the quality of life states and recognising all accumulated survival time. From these limits it is demonstrated that any reasonable choice of positive weight will not only favour the HAI group in terms of the estimated difference in mean QAS(L) as was already known, but that there will be no convincing evidence of a treatment difference for any given weights as shown by the lack of

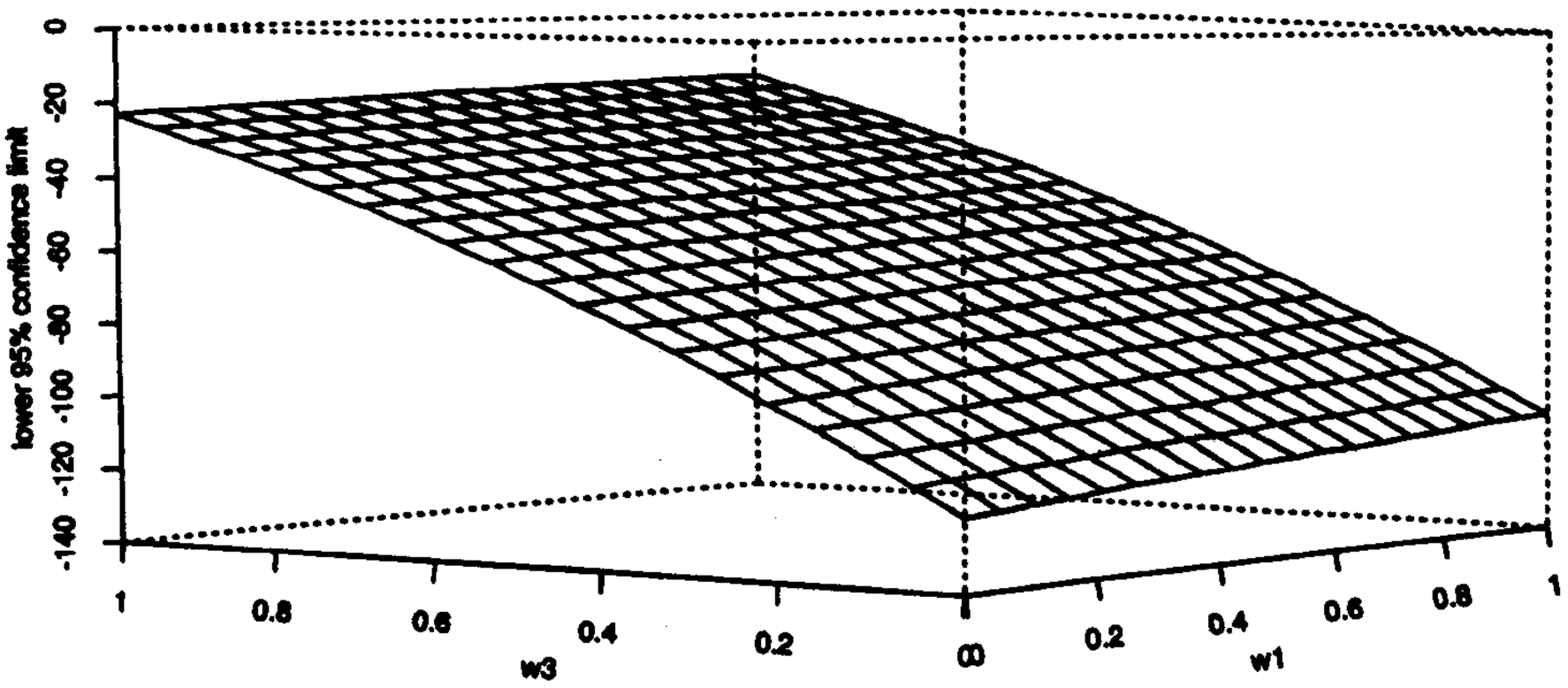

Figure 6.7: Surface of the realised values of the lower limit of 95% confidence interval for mean QAS(L) for all possible combinations of w_1 and w_3 in the interval [0, 1].

evidence from the two extreme cases. This is also shown in figure 6.7 which shows the surface of 95% lower bounds for mean QAS(L) for all combinations of w_1 and w_3, with w_2 always equal to one. The figure shows that for all possible combination of weights, the lower confidence bound was always less than 0, implying that the 95% confidence interval included the null value of no difference. The maximum lower bound shown on the figure corresponds to the final row of table 6.9 when total survival in each group is compared.

As with the TNQOL analysis, a similar picture was also seen at the shorter restricted time points as shown in figure 6.8. One exception to this was at 24 months at which time the control group had a very slightly higher estimated restricted mean time spent with normal quality of life.

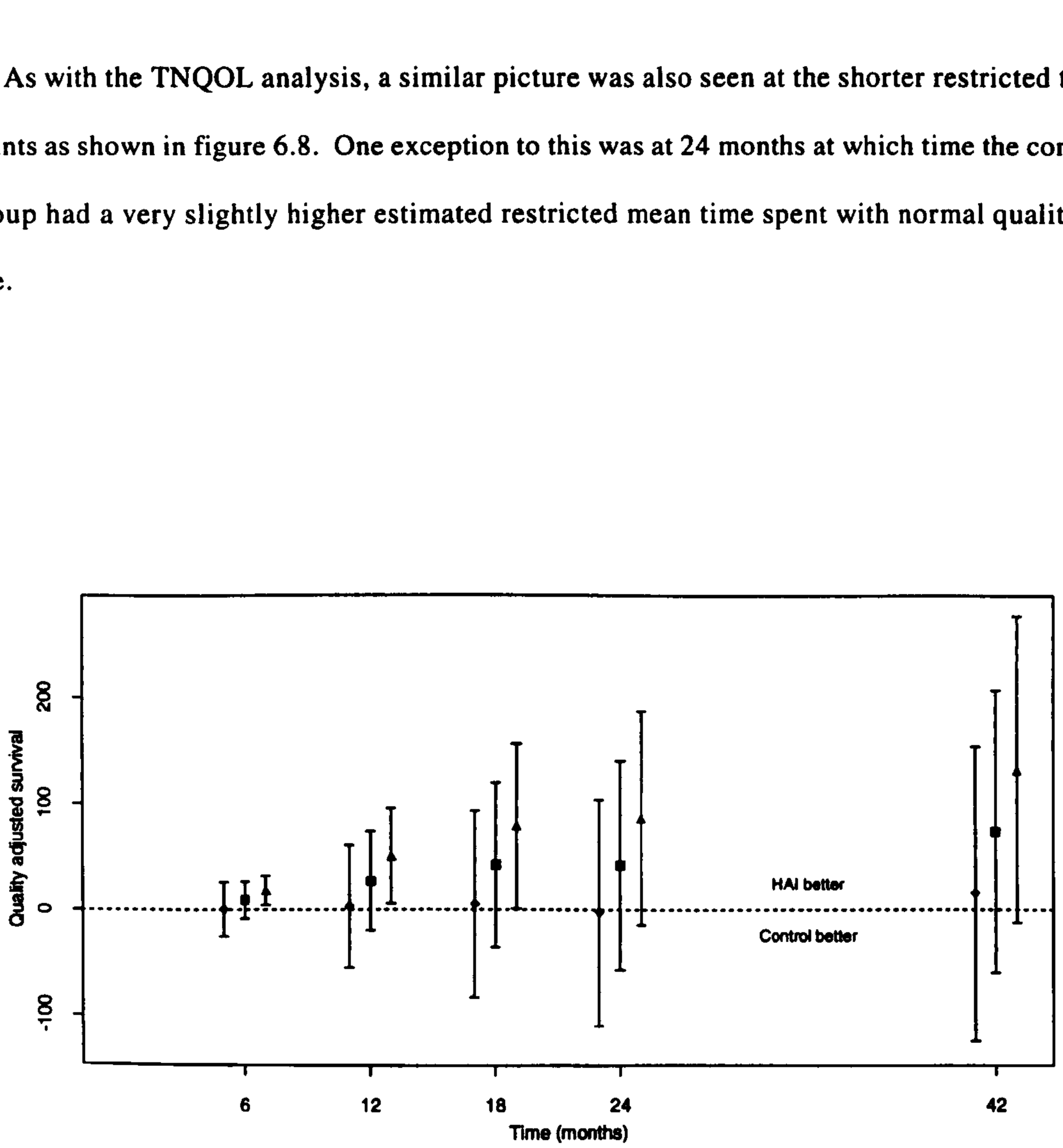

Figure 6.8: Estimated QAS(L) and 95% CI for L=6, 12, 18, 24, 42 months with weights given by: $(w_1, w_2, w_3) = (0, 1, 0)$: ♦; (0.5, 1, 0.5): ■; and (1, 1, 1): ▲ for the restricted RSCL physical quality of life scores of the CRC HAP trial.

6.3.3 Integrated quality-survival product

The integrated quality-survival product was outlined by Glasziou (1995) as an alternative to health state based models. The basis of the analysis is the observation that the rate of gain of quality adjusted life years at times t, denoted $qaly(t)$, may be expressed as the product of the survival function and the average quality of life of survivors at that time,

$$qaly(t)=Q(t)S(t) \tag{6.8}$$

This then allows the mean quality adjusted life years gained up to some time limit L, denoted $QALY(\mathrm{L})$, to be estimated as the area under this quality-survival product

$$QALY(\mathrm{L})=\int_0^{\mathrm{L}} Q(t)S(t)\,\mathrm{d}t \tag{6.9}$$

A Kaplan-Meier estimate is the obvious estimator for the survival function, but estimation of $Q(t)$ has more options. Glasziou (1995) suggest this could be done by simply estimating group means at fixed time points, t, and then interpolating between distinct times using either a step function assuming quality of life changes at these fixed times or by linear interpolation. Alternatively they suggest that the order of estimation could be reversed, such that quality of life measured at distinct times for individuals is interpolated to continuous time, with $Q(t)$ estimated as an average of these continuous functions. Such an approach is particularly helpful when measurement of quality of life takes place at very different times for each subject.

In the subsequent application, three alternative methods of extrapolation were considered. The first used the most simple formulation possible, and assumed a simple group specific linear regression model through all the available data. This was then evaluated at all observed measurement occasions across subjects. The second analysis used a lowess smoother for an average quality of life which was not necessarily of a linear form. The third analysis first obtained subject specific quality of life profiles in continuous time. These were obtained by

first evaluating quality of life at the planned 30 day measurement times for each subject using the mean of the two closest measurements at days 0, 30, 60, ...etc, and then interpolating between these points using a step function.

Although the idea of integrated quality-survival product was to solve the problem of subjectively chosen weights by using patients' quality of life scores to weight survival, the scores resulting from typical quality of life measurement instruments still may not give ideal weights. For example, the RSCL is scored such that low scores indicate a better physical quality of life. In order to obtain a reasonable weighting some transformation of this score is obviously needed. This gives a similar arbitrariness problem for choice of weights. For example, for the RSCL physical score, the percentage of the total score calculated as $w_1 = 1 - \left(\frac{score}{40}\right)$ could be used. Alternatively, to give a weighting which recognises the small clinical differences between low scores in contrast to the large clinical differences between

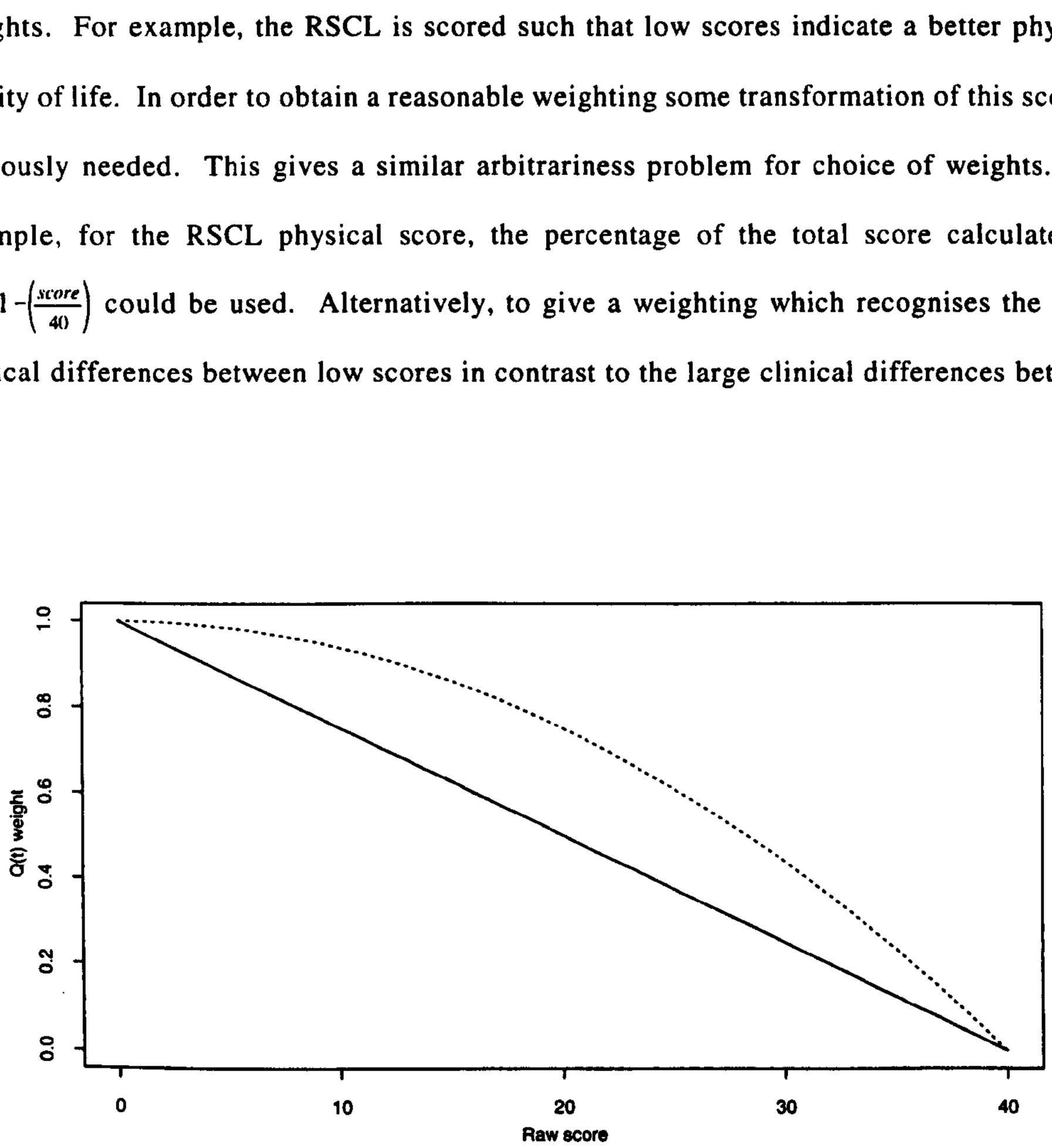

Figure 6.9: Profile of weight versus quality of life response as used for the continuous quality adjusted survival. 1-(score/40):————; 1-(score/40)^2: --------.

Table 6.10: Estimation of $Q(t)$ for an integrated quality-survival product analysis of the restricted CRC HAP trial data.

Analysis	Weighting	Method of estimation
1	w_1	A linear function of time evaluated at all observed measurement times
2	w_1	A lowess smoother evaluated at all observed measurement times
3	w_1	Mean of the subject specific quality of life evaluated at the planned 30 day measurement occasions, t=(0, 30, 60, ..., 1260)
4	w_2	A linear function of time evaluated at all observed measurement times
5	w_2	A lowess smoother evaluated at all observed measurement times
6	w_2	Mean of the subject specific quality of life evaluated at the planned 30 day measurement occasions, t=(0, 30, 60, ..., 1260)

high scores, the transformation $w_2=1-\left(\frac{score}{40}\right)^2$ could be used. Both of these weighting functions are shown in figure 6.9 and were used for the analysis of the restricted RSCL physical scores of the CRC HAP trial giving six analyses in total which are outlined in table 6.10. All analyses were restricted to L=42 months follow-up.

Following estimation of $Q(t)$ and $S(t)$, the quality-survival product, $Q(t)S(t)$, was simply the product of $Q(t)$ and $S(t)$ evaluated over t. This is shown in figure 6.10 for the CRC HAP trial RSCL physical data for analyses 1,...,3. Figure 6.10(a) shows $S(t)$, 6.10(b), (d) and (f) give the estimated $Q(t)$ for analyses 1, 2 and 3 respectively. The product of the survival curve and these respective estimated quality of life functions is given in figure 6.10(c), (e) and (g).

These figures highlight the different degrees of smoothing used in each estimation of $Q(t)$, ranging from the linear function in figure 6.10(b), to that based on the subject specific means at each planned 30 day interval in figure 6.10(f) which was very noisy. The interesting thing

to notice from these figures, is that the degree of smoothing had little impact on the final results of the analysis in this example with figures 6.10(c), (e) and (g) all showing very similar

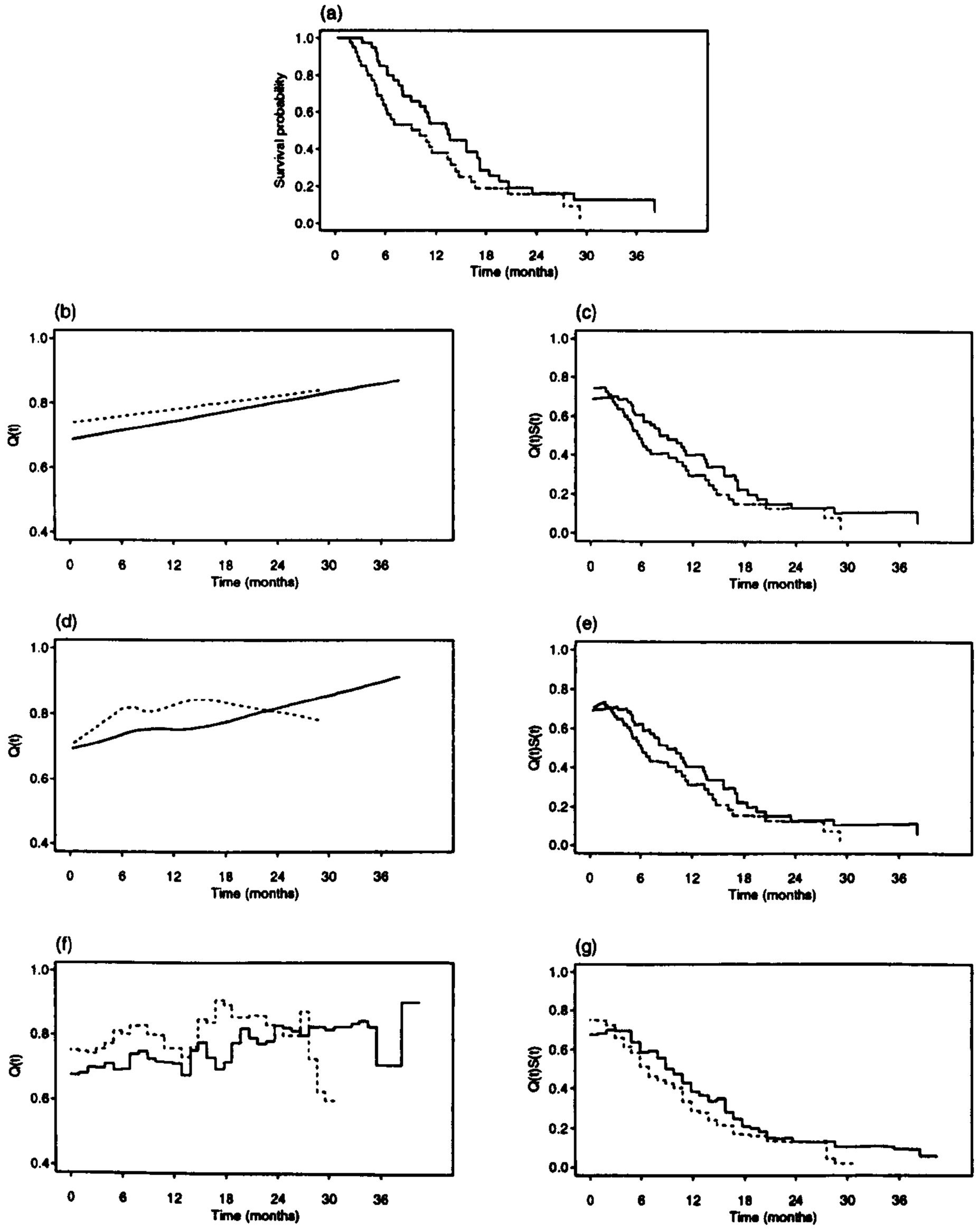

Figure 6.10: Estimated (a) survival $S(t)$; and quality of life function $Q(t)$ and quality-survival product for (b) and (c) analysis 1; (d) and (e) analysis 2; (f) and (g) analysis 3 . HAI: ————; control: --------.

Table 6.11: Integrated quality-survival product for the restricted RSCL physical data from the CRC HAP over a 42 month period.

	Mean (SE)			
Analysis	**HAI**	**Control**	**Difference [95% CI]**	***p*-value**
1	346.7 (54.4)	258.2 (32.7)	88.5 [-38.3, 215.3]	0.168
2	353.4 (58.6)	265.9 (32.5)	87.5 [-46.5, 221.4]	0.196
3	353.2 (47.6)	276.5 (40.1)	76.7 [-47.2, 200.6]	0.222
4	420.1 (58.9)	302.9 (40.4)	118.0 [-24.5, 260.5]	0.103
5	433.8 (65.6)	319.5 (38.3)	114.3 [-37.5, 266.1]	0.137
6	431.6 (55.0)	322.2 (42.2)	109.4 [-28.7, 247.5]	0.119

behavioural patterns. This is further demonstrated by the estimated integrated quality-survival products for the three example over the 42 month period which are given in the first three rows of table 6.11. Although the smoothed estimates of $Q(t)$ gave slightly increased points estimates of the difference between the two groups, there was little to suggest a difference between these estimates. Similar results were seen for the alternative transformation of quality of life scores.

6.3.4 Conclusion

The aim of this section of work was to assess a number of quality adjusted survival methods proposed for the analysis of self assessed quality of life data. Three alternative analyses were used. The first two analyses were based on health state models, and although they have been extensively used in toxicity studies, their application for self assessed quality of life data has been limited. Practical application of the third analysis, although specifically designed for such data, has yet to be reported in the quality of life literature. The aim of this work was therefore to assess the practical use of these analyses for the analysis of self assessed quality of life data and to highlight the issues which they generate.

The main problem with the TNQOL analysis (based on TWiST as defined by Gelber *et al.*, (1989)) is that of bias induced by the informative censoring of discounted survival times. Because of this issue, which lead to the method being abandoned for the analysis of toxicity data, it is concluded that, if used, results obtained using Kaplan-Meier estimation of the survival curve in analyses in the light of a moderate or large degree of censoring, should be presented alongside those of the imputation methods suggested by Gelber *et al.* (1989). This enables the consistency of conclusions over a range of approximate methods to be assessed. As demonstrated in these analyses however, although these will give a range of possible values for patient group estimated mean TNQOL, differing degrees of bias in these estimates may result from the different quality of life behaviour in the two groups. A range containing all possible differences may be achieved by comparing the minimum imputed mean in one group with the maximum imputed mean in the second group and vice versa. However, unless there is very strong evidence of a difference it is unlikely that these estimates will give results on which strong inferences can be made.

A further problem with this analysis is that it is impossible to determine whether any apparent advantage in TNQOL(L) for the one group over another is due to the previously established survival benefits or indeed a superior quality of life. This has been one of the major criticisms of the use of quality adjusted survival techniques in the literature and it is difficult to see how it may be overcome.

The main restriction of the partitioned quality adjusted survival (PQAS) analysis is that it requires that the health states defined for the analysis are progressive. Such are inherent for the analysis of toxicity data where states of toxicity, TWiST and disease recurrence occur naturally. Self assessed quality of life data however can fluctuate throughout treatment and follow-up, particularly when periods of treatment are repeated several times during follow-up

as is often the case with treatments of chemo- and radio- therapy. This can be overcome to some degree by careful definition of states, and incorporation of additional states which allow a little more movement in the quality of life profile to occur. For example, '*poor quality of life after treatment*' and '*poor quality of life after good*'. However, this requires some knowledge about the expected behaviour of quality of life within a study, and will not always be practicable. In such cases, a partitioned quality adjusted survival analysis will not be feasible. The method does have an advantages over the TNQOL (TWiST) analysis in that it overcomes the main criticism of such methods in the literature by having a clear interpretation of the results in terms of it being possible to determine how the estimated weighted quality adjusted survival is made up in terms of time spent in each state. Perhaps more importantly, it is also not subject to bias as a result of informative censoring.

The integrated quality-survival product was suggested specifically as a quality adjusted survival analysis for self assessed quality of life data, (Glasziou, 1995). Like TWiST however, it suffers from an interpretational problem as it is difficult to understand how the quality adjusted survival time has been accumulated and the presentation of the individual components of the quality-survival product ($Q(t)$ and $S(t)$) is vital to facilitate interpretation. Although the main aim of this analysis is to eliminate the need for a researcher to subjectively assign weights for survival, it is unclear whether the scores obtained from quality of life measuring instruments define a reasonable weighting system for the analysis. Transformations of the scores to achieve more realistic weights from the scores may be applied, but these suffer from the same subjectivity as choosing weights in the first place.

An unfortunate feature that all these analyses share, is the large number of assumptions or subjective decisions that have to be made by the researcher. These include: the definition of health states (TNQOL and PQAS); the choice of weighting for each state (PQAS and weighted

TNQOL); the choice of transformation of quality of life scale (integrated quality-survival product); and the length of time over which quality adjusted survival is evaluated (TNQOL, PQAS and integrated quality-survival product), and demand that a number of sensitivity analyses are presented along with the study results in order to study the robustness of conclusions drawn. These need to include presenting the results over a range of restricted times, L and varying the choice of weights. Within an analysis of the integrated quality-survival product, the impact of the choice the transformation of the quality of life scores needs also to be investigated.

If feasible, it is concluded that the PQAS based analysis is the most favourable as a quality adjusted survival analysis of self assessed quality of life data. The basis of this conclusion is its ability to overcome the interpretational criticisms these analyses have faced in the literature. When extensive censoring occurs within the data TWiST based analyses like TNQOL, although simple to apply, are best avoided. Although the integrated quality-survival product has some potential, its lack of an intuitive interpretation of its results may limit its application as a reasonable alternative to PQAS. Whatever analysis is finally chosen though, possibly the most important part of a quality adjusted survival analysis is an extensive sensitivity analysis to accompany any results to support the conclusions in the light of the many arbitrary assumptions which all of the analyses require.

6.4 Summary and discussion

This work has examined alternative analyses for self assessed quality of life data which are censored as a result of patient death. Two very contrasting approaches were considered. The first included models for the analysis of data which are incomplete as a result of patient dropout. It was concluded that, of the three general classes of such models which have been

discussed in the literature, those termed *informatively right censored* models are the most appropriate in relation to the problem at hand. This conclusion was made on the grounds that, both *pattern mixture* models (Little, 1993) and *selection* models (Diggle and Kenward, 1994) rely on measurement (and therefore dropout) occurring at unique time points which is not guaranteed in quality of life studies. In addition they attempt to make full data inferences - that is, inferences based on expectation in the absence of dropout. This was not regarded as appropriate in the context of quality of life data and patient death, since it is the expectation of quality of life conditional on a patient being alive which is of interest. For such inferences, two different models were presented. Both gave very similar results. The trivariate Normal model, attributable to Schlucter, (1992), modelled the joint distribution of patient quality of life and survival, which could then be transformed to give inferences for the conditional distribution of quality of life given a particular survival time. In the second model, this distribution was modelled directly. The latter was unfortunately restricted to circumstances when survival times are known for all subjects, whereas the former incorporates censored survival times within the analysis. Although both analyses were presented with quality of life modelled in terms of a linear trend over time, more complex patterns of response over time may be incorporated. In addition, the analyses are not restricted to continuous outcomes, and may be used as extensions of the models for the random effect binary or ordinal outcomes discussed in Chapters 4 and 5. Further work is however needed in order to fully determine the properties of these models and their robustness to differing degrees of censoring.

The second analysis options considered were different methods for quality adjusted survival analyses. Such analyses have been used successfully for the analysis of toxicity data, but have been criticised in the quality of life literature as they make it impossible to determine how quality adjusted survival has been accumulated, for instance whether patients have a short survival with good quality of life or a long survival with poor quality of life. Of the three

different quality adjusted survival analyses evaluated, it was concluded that the partitioned quality adjusted survival (PQAS) (Glasziou *et al*, 1990) was the most useful. Unfortunately, the analysis requires that progressive quality of life health states are definable, which may make it infeasible in many practical situations. In addition, as bootstrapping is required in order to obtain measures of precision for estimated effects, they are also computer intensive.

A third analysis option which has not been considered here is the use of multi-state models. These models have been used to study disease progression (Kay, 1986, Andersen *et al.*, 1991) where survival and disease development are modelled in the setting of a Markov chain, estimating transition intensities (or instantaneous hazard rates) between health states. It has been suggested (Olschewski and Schumacher, 1990, Abrams, 1992) that by defining appropriate quality of life states with death as a final absorbing state, such analyses could be useful for the analysis of self assessed quality of life data. Since there has been no application of such models in the quality of life literature, it is unclear how useful they may prove to be, and it is an interesting area of research in which further work is needed.

It should be noted that all of the work included in this chapter once again ignored the problem of intermittent missing data. This is a very different problem to that of censored quality of life responses due to death that discussed in the following chapter.

7 Missing Data in Quality of Life Studies

7.1 Introduction

Missing data is a common problem in studies measuring quality of life. In all of the work covered so far in this thesis it has been assumed that the implications of some of the data being missing can be ignored and it has been considered simply as a problem which generates unbalanced data rather than a source of potential bias. The implications of this assumption will depend on the reasons underlying the missing data, or more formally the *missing value process*. If responses are missing simply because subjects forget to complete or return questionnaires and is not related to the underlying level of quality of life at that time then, although reducing the precision of estimates, it is reasonable to assume that the occurrence of missing data will not introduce a bias. On the other hand if the reason underlying the incidence of missing responses is in some way related to the underlying level of response, for instance if subjects tend not to respond when they are depressed or when they are feeling particularly well, inferences from analyses that ignore the missing data may be subject to bias.

Even within a randomised controlled trial, when it is believed that the underlying reasons for missing responses are the same across the two treatment groups, a bias in the estimate of a treatment comparison may occur if indeed there is a difference in the underlying level of response in the two groups. For example, if the probability of subject non-response is related to an underlying high level of response, and if one group of patients experience higher underlying levels of response than the other, then this group will be more susceptible to missing data and the observed crude mean difference in response between the groups will be an

underestimation of the true underlying difference.

Although the implications of missing data have been well documented in terms of missing data due to dropout or attrition, very little has been written relating to intermittent missing data. In one of the few references to address the problem, Diggle *et al.* (1994) have suggested that if data is intermittently missing, "*it may be reasonable to assume that they arise from mechanisms unrelated to the measurement process and are therefore missing completely at random*". Intuitively, this does not seem reasonable for quality of life data. The work in this chapter investigates this explicitly and attempts to determine how missing data typical of that seen in quality of life studies may affect the conclusions of an analysis. This is done by developing two models to investigate how the observed missing data relate to the observed quality of life response. These are exemplified using the data from the CRC NSCLC study. Before these models are developed however, Section 7.2 describes the notation and well known classification of missing response processes as defined by Little and Rubin (1987). This is followed by a simulation exercise to exemplify these definitions and their implications for particular intermittent missing value processes. The current literature on examining the nature of the missing value process within a data set is then discussed and two models that extend this work are developed. Using the results of both the simulation exercise and these two models, some conclusions as to the effect of missing data on the inferences already drawn from the CRC NSCLC study are then discussed. A more general discussion is given in Section 7.4.

7.2 Missing value processes

7.2.1 Notation

The classifications of missing value processes defined by Little and Rubin (1987) will be used throughout using the following notation.

The vector, $y_i=(y_{i1},...,y_{im})$, is the vector of underlying responses for subject i over occasions, $j=1,...,m$. The components of this vector which are observed are denoted y_{io} with those that are missing denoted y_{im}. As such these two sub-vectors form a partition of $y_i=(y_{io},y_{im})$.

x_i is a design matrix of known covariates which relates y_i to θ, a vector of unknown parameters describing the relationship of x_i and y_i about which inferences are to be made. For example, x_i may be a $(m \times 2)$ matrix to model $\theta=(\alpha,\beta)$, an intercept and slope to examine how the quality of life response changes over time.

A second $(m \times 1)$ vector, $r_i=(r_{i1},...,r_{im})$ denotes the missing value process, where $r_{ij}=1$ if y_{ij} is observed, 0 if it is missing. The design matrix z_i relates r_i to ϕ, a vector of unknown parameters for the missing value process where x_i and z_i may or may not be distinct. For example, both the missing data process r_i and response y_i may change with respect to time.

The density of y_i given θ and x_i is written $f_1(y_i|\theta,x_i)$. Similarly the density of r_i given ϕ, z_i and y_i is given by $f_2(r_i|\phi,z_i,y_i)$. The joint density of the observed data (y_{io},r_i) can then be obtained by integrating the joint distribution of (y_i,r_i) over the sample space of the missing data y_{im}.

$$f(y_{io},r_i|\theta,\phi,x_i,z_i) = \int_{y_{im}} f_1(y_i|\theta,x_i) f_2(r_i|\phi,z_i,y_i)\, dy_{im} \qquad (7.1)$$

It is the partitioning of the joint density in equation (7.1) that is critical in defining the implications of particular missing data processes.

7.2.2 Missing completely at random

An observation is said to be *missing completely at random* (MCAR) if missingness is completely independent of the underlying measurement process. That is

$$f_2(r_i|\phi,z_i,y_i) = f(r_i|\phi,z_i) \tag{7.2}$$

Since y_i and r_i are independent, y_{io} can be seen as a random sample from y_i, and thus inferences based only on the observed data (y_{io},r_i) are valid. Further, substituting equation (7.2) into equation (7.1) illustrates that ignoring a missing value process that is MCAR will introduce no bias although there will be a loss of precision in the analysis because of the reduced data size. Treating the missing data as simply a problem of unbalanced data is therefore valid under such a missing value process.

7.2.3 Missing at random

An observation is defined as *missing at random* (MAR) if missingness is related to the observed data, y_{io}. Formally

$$f_2(r_i|\phi,z_i,y_i) = f(r_i|\phi,z_i,y_{io}) \tag{7.3}$$

Combining this with equation (7.1), the joint density of the observed data (y_{io},r_i) reduces to

$$f_2(r_i|\phi,Z_i,y_{io}) \int_{y_{im}} f_1(y_i|\theta,x_i)\, dy_{im} = f_2(r_i|\phi,Z_i,y_{io}) f_3(y_{io}|\theta,x_i) \tag{7.4}$$

Following this partitioning, Rubin (1976) showed that if θ and ϕ are distinct, likelihood based inferences about θ can be made based only on y_{io} and x_i. Hence, for continuous outcomes, variance component models which use the full available data, such as those used in Chapter 3, will be valid. Alternative models are also available. In particular, Zwinderman (1992) discussed alternative models for continuous outcomes subject to missing data classed as 'dropout'. Also, the Dale model proposed by Kenward *et al.* (1994) for the analysis of ordinal data will also be valid under MAR.

It should be noted however, that unlike the MCAR process, equation (7.4) shows that y_i and r_i are not independent. This implies that the sampling properties of maximum likelihood estimates will depend on the missing value process and thus precision estimates based on the expected information matrix will be incorrect. It has been suggested therefore, that under MAR, precision estimates and test statistics should be based on the observed rather than expected information (Laird, 1988). A further problem with all likelihood based analyses is that, they implicitly impute missing data, and can therefore be sensitive to model misspecification.

As an alternative, a *weighted* GEE for binary and continuous outcomes have also been shown to perform well for the MAR case. The basis for this proposed model is a note identifying the source of bias in the GEE analysis under MAR and adjusting for this accordingly (Rotnitzky and Wypij, 1994, Robins *et al.*, 1995, Robins and Rotnitzky, 1995).

7.2.4 Not missing at random

Finally, an observation is said to be *not missing at random* (NMAR) if given observed measurements, there is some residual association of the missingness to the realised value of the missing observations. That is

$$f_2(r_i|\phi, z_i, y_i) = f(r_i|\phi, z_i, y_{io}, y_{im}) \qquad (7.5)$$

In situations where data are NMAR, inferences drawn from an analysis ignoring the missing data will be biassed. Simulation exercises by many authors have attempted to address the extent of the problem. Wang-Clow *et al.* (1995) saw that when the missing data process was non-ignorable but the same in both treatment groups, non-likelihood based approaches tended to do generally better than those based on likelihood. However, in practical situations, in the

event of non-ignorable non-response it is impossible to determine the extent of possible bias. Given this, the most satisfactory way to determine the implications of the missing data on parameter estimation has been suggested to be a combination of pattern mixture models using plausible missing data processes and multiple imputation (Laird, 1988, Glynn *et al.*, 1993) to give a sensitivity analysis for the different missing data assumptions.

7.2.5 Two simulated illustrations

To illustrate the problem of intermittent missing data for analyses of typical quality of life data, two simulated examples were performed. Although a number of simulated examples to illustrate the problems of the different missing value processes have been recently presented in the literature (Wang-Clow *et al*, 1995, Wu and Carroll, 1988) these have concentrated on the problem of patient dropout rather than intermittent missing data which is an additional problem in quality of life studies. It was therefore hoped that this exercise with simulation parameters typical of those seen in quality of life data, may give more information about the problems that intermittent missing data may cause in quality of life studies.

In each scenario, full data were simulated for eighty-two subjects on eight occasions with the subjects split equally between two groups. In the first example, it was assumed that both groups had the same underlying response, constant through time. In the second example, the mean response, again assumed constant through time, was different in the two groups. In both cases, responses for each subject i were generated from a Normal distribution with mean μ_i and variance σ_e^2. These subject specific means were then in turn assumed to be Normally distributed with mean μ and variance σ_μ^2. The values taken for these parameters are defined in table 7.1. The choice of parameters was based on previous experience with the results of the NSCLC CRC study reported in Chapter 3.

Table 7.1: Simulation parameters for two intermittent missing value simulations.

	Marginal parameters		Variance parameters		Missing data process		
					MCAR	MAR	NMAR
	μ_1	μ_2	$\sigma^2_{u1}=\sigma^2_{u2}$	$\sigma^2_{e1}=\sigma^2_{e2}$	$Pr(r_{ij}=0)$	$Pr(r_{ij}=0 \mid y_{ij}.>18.33)$	$Pr(r_{ij}=0 \mid y_{ij}>18.33)$
1	16	16	10	4	0.20	0.55	0.75
2	14	17.5	10	4			

For the MAR process missing data is assigned with a probability of 0.55 on the basis of the last observed measurement, y_{ij}..

The impact of missing data for three methods of data analysis were was assessed: an unweighted mean of subject specific parameters obtained by least squares (UWLS) (Wu and Carroll, 1988); a generalised estimating equation with an exchangeable correlation structure, (GEE) (Liang and Zeger, 1986); and full random effect analysis using multilevel models to give REML estimates (MLn) (Goldstein, 1995). For the UWLS analysis, subjects who following the assignment of missing data had less than three observations overall were excluded. Consistent with the simulated data structure, the parameter of interest, the mean response in each group was assumed constant over time. Separate variance parameters were estimated for each group using MLn thus giving an estimate of the intraclass correlation in each group. A single estimate of the intraclass correlation was used for the GEE analysis.

The results using each of these methods for the 300 simulated full data sets are given in tables 7.2 and 7.3 for the marginal and variance parameters respectively. These very closely reflected the simulation parameters used and were the same regardless of estimation method. Responses within this full data set were then designated to be missing according to each of the missing data processes defined above. For the MCAR case a value was designated missing with a probability of 0.2. Under MAR, missingness was determined on the basis of the previously observed measurement such that a response was assigned missing with a probability of 0.55 if the value of the last available observation was greater than 18.33. For example, if the second

Table 7.2: Estimated marginal parameters (SD) of the full data simulated for missing data simulations.

	Method of analysis		
	UWLS	**GEE**	**MLn**
Simulation one : μ_1=16, μ_2=16, δ=0			
μ_1	16.03 (0.51)	16.03 (0.51)	16.03 (0.51)
μ_2	16.01 (0.51)	16.01 (0.51)	16.01 (0.51)
δ	-0.018 (0.72)	-0.018 (0.72)	-0.018 (0.72)
Simulation two : μ_1=14, μ_2=17.5, δ=3.5			
μ_1	14.03 (0.51)	14.03 (0.51)	14.03 (0.51)
μ_2	17.51 (0.51)	17.51 (0.51)	17.51 (0.51)
δ	3.48 (0.72)	3.48 (0.72)	3.48 (0.72)

The mean and standard deviation of the estimates from the 300 simulations are given as the estimate in each case.

Table 7.3: Estimated variance parameters (SD) of the full data simulated for the missing data simulations.

	MLn		GEE
	Variance estimates		ρ
	Group 1	**Group 2**	**Overall**
Simulation one : $\sigma^2_{u1}=\sigma^2_{u2}=10$; $\sigma^2_{e1}=\sigma^2_{e2}=4$; $\rho=0.71$			
σ^2_u	9.97 (2.36)	9.80 (2.35)	
σ^2_e	3.99 (0.34)	3.99 (0.32)	0.70 (0.04)
Intraclass correlation	0.71 (0.05)	0.70 (0.05)	
Simulation two : $\sigma^2_{u1}=\sigma^2_{u2}=10$; $\sigma^2_{e1}=\sigma^2_{e2}=4$; $\rho=0.71$			
σ^2_u	9.97 (2.36)	9.80 (2.35)	
σ^2_e	3.99 (0.34)	3.99 (0.32)	0.70 (0.04)
Intraclass correlation	0.71 (0.05)	0.70 (0.05)	

The mean and standard deviation of the estimates from the 300 simulations are given as the estimate in each case.

observation was designated missing on the basis of the first, missingness of the third observation was also based on the value of the first. Being impossible to determine the value of the previous measurement, no datawere deleted at the first measurement occasion. For the NMAR process, a value was designated missing on the basis its own value. If the value exceeded 18.33 it was assigned missing again with a probability of 0.75. The observed mean proportion of missing data for each simulation are shown in table 7.4. Given the simulation parameters, this was approximately 20% overall for the MCAR and NMAR analyses. Under MAR the overall proportion of missing data in simulation one was slightly lower at 16%, and for simulation two slightly higher at 22%. The distribution of proportion of missing data by treatment group and overall for each individual simulation within simulations one and two are shown in figure 7.1 along with a summary of the numbers of subjects omitted from the UWLS analysis in each case .

For the MCAR cases, all the distributions were the same. This was because the assignment of missing data in both simulations was set to start from the same random seed. For simulation one where both groups had the same underlying parameters the distributions of missing data

Table 7.4: Observed mean proportion of missing data in missing data simulations one and two by group and overall.

	Group 1	Group 2	Overall
Simulation one			
MCAR	0.20	0.20	0.20
MAR	0.16	0.16	0.16
NMAR	0.20	0.20	0.20
Simulation two			
MCAR	0.20	0.20	0.20
MAR	0.18	0.25	0.22
NMAR	0.09	0.31	0.20

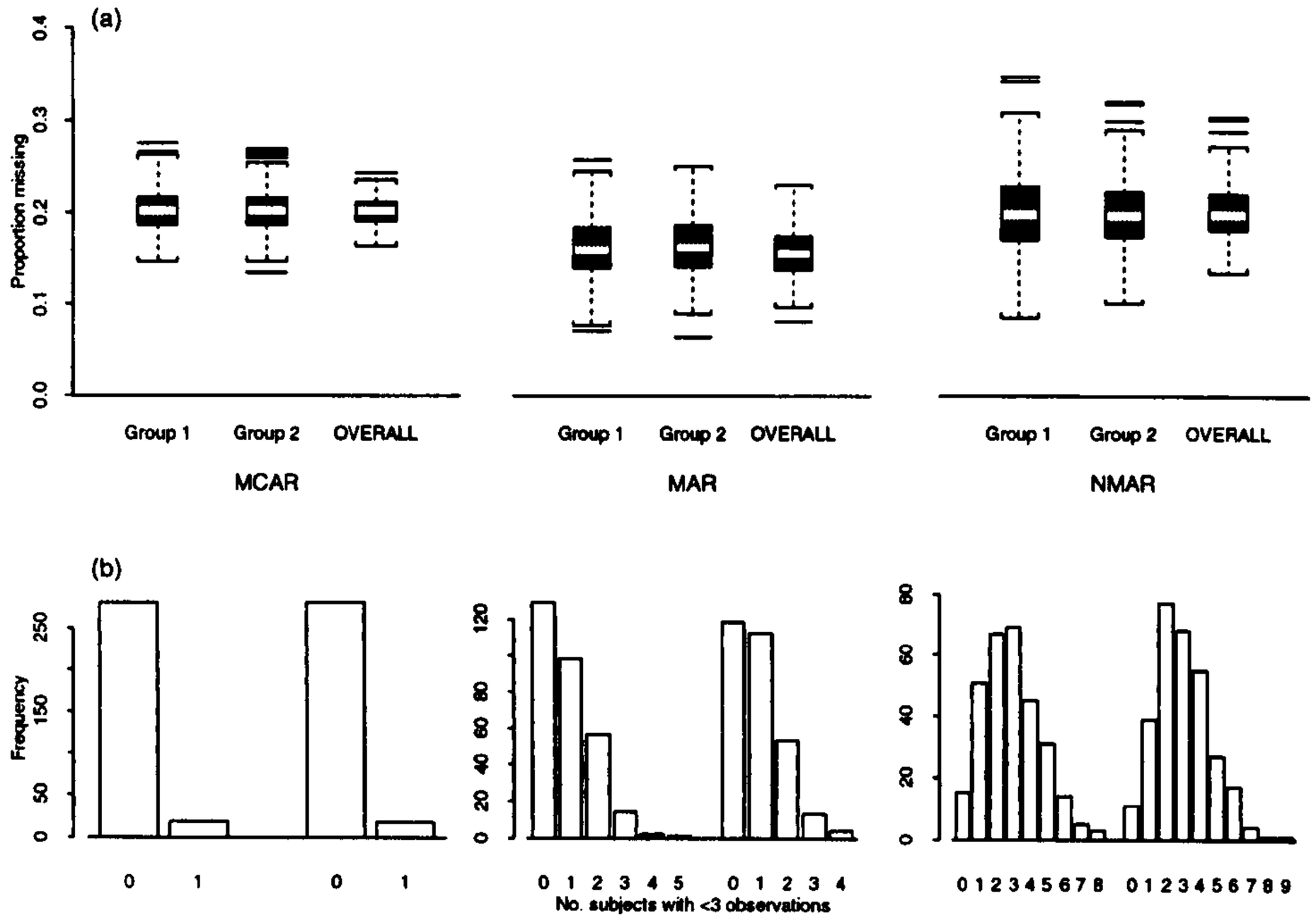

Figure 7.1: Missing data summary box plots and histograms for simulations one and two showing the distribution of the proportion of missing data ((a) & (c)) and the number of subjects who were omitted from the UWLS analysis because they had less than three observations ((b) & (d)) for each of the 300 simulations.

were very similar in terms of the proportions missing as well as the number of subjects with less than three observations who were omitted from the UWLS analysis. For simulation two where the different group means were used, in line with their expectations, the distribution of missing data between the two groups was different for the MAR and NMAR cases where the missing data process was related to the underlying response. This was also reflected in the distribution of the number of subjects with less than three observations who were omitted from the UWLS analysis.

The results of the three different analyses of these data following the different missing data simulations are given in table 7.5 for the fixed (marginal) effects and table 7.6 for the random (variance) components. A series of one sample *t*-tests were used to test for bias in these estimates following data deletion under the various missing value procedures with the original simulation parameters. All estimates have been rounded to 2 decimal places which accounts for some apparent discrepancies of *p*-values in the tables.

All three analysis methods performed well for the MCAR case, and gave no evidence of any bias in the estimation for both simulations. For simulation one, in the MAR case the UWLS analysis produced estimates lower than the true mean effects in each group, although the evidence for this as a real effect was not convincing. These were both consistently estimated in the GEE and MLn analyses. For all three analyses, the estimated group difference was still consistently estimated. However, for simulation two the UWLS analysis overestimated the true mean effects in each group, although it did consistently estimate the group difference. Although not to the same degree, the estimated group means from the GEE and MLn analyses were greater than the true effects. Again the estimated group difference was still consistently estimated. In the NMAR case, the mean parameter estimates were underestimated in both simulation examples.

Table 7.5: Estimated marginal parameters (SE) {two sided *p*-value} for simulations one and two following the assignment of missing data.

	Method of analysis		
	UWLS	**GEE**	**MLn**
Simulation one : μ_1=16, μ_2=16, δ=0			
Missing completely at random (MCAR)			
μ_1	16.03 (0.03) {0.31}	16.03 (0.03) {0.31}	16.03 (0.03) {0.31}
μ_2	16.01 (0.03) {0.74}	16.01 (0.03) {0.74}	16.01 (0.03) {0.74}
δ	-0.020 (0.04) {0.64}	0.020 (0.04) {0.62}	-0.020 (0.04) {0.62}
Missing at random (MAR)			
μ_1	15.97 (0.03) {0.28}	16.01 (0.03) {0.83}	16.01 (0.03) {0.76}
μ_2	15.96 (0.03) {0.19}	16.00 (0.03) {0.99}	16.00 (0.03) {0.93}
δ	-0.020 (0.70) {0.98}	-0.019 (0.71) {0.98}	-0.007 (0.04) {0.87}
Not missing at random (NMAR)			
μ_1	15.25 (0.02) {<0.001}	15.45 (0.03) {<0.001}	15.48 (0.03) {<0.001}
μ_2	15.22 (0.03) {0.08}	15.43 (0.03) {<0.001}	15.46 (0.03) {<0.001}
δ	-0.026 (0.04) {0.45}	-0.019 (0.04) {0.61}	-0.023 (0.04) {0.59}
Simulation two: μ_1=14, μ_1=17.5, δ=3.5			
Missing completely at random (MCAR)			
μ_1	14.03 (0.03) {0.74}	14.03 (0.03) {0.74}	14.03 (0.03) {0.74}
μ_2	17.51 (0.03) {0.74}	17.51 (0.03) {0.74}	17.51 (0.03) {0.74}
δ	3.48 (0.04) {0.48}	3.482 (0.04) {0.48}	3.482 (0.04) {0.48}
Missing at random (MAR)			
μ_1	14.24 (0.03) {<0.001}	14.08 (0.03) {0.01}	14.07 (0.03) {0.02}
μ_2	17.74 (0.03) {<0.001}	17.58 (0.03) {0.01}	17.57 (0.03) {0.02}
δ	3.50 (0.04) {0.97}	3.49 (0.04) {0.89}	3.50 (0.04) {0.90}
Not missing at random (NMAR)			
μ_1	13.66 (0.03) {<0.001}	13.72 (0.03) {<0.001}	13.73 (0.03) {<0.001}
μ_2	16.31 (0.03) {<0.001}	16.74 (0.03) {<0.001}	16.77 (0.03) {<0.001}
δ	2.65 (0.04) {<0.001}	3.01 (0.04) {<0.001}	3.04 (0.04) {<0.001}

The standard error of the estimates from the 300 simulations is given for each case and used to construct a one sample *t*-test comparing each estimate with the simulation parameter.

For the random parameters, the full random effects analysis using MLn performed reasonably well in the both MCAR and MAR cases in simulation one. For the latter, the estimated between subject variances were slightly lower than the full data example, but there

Table 7.6: Estimated variance parameters (SE) {two sided *p*-value} for simulations one and two following the assignment of missing data.

	MLn		GEE
	Variance estimate		ρ
	Group 1	Group 2	Overall
Simulation one : $\sigma^2_{u1}=\sigma^2_{u2}=10$; $\sigma^2_{e1}=\sigma^2_{e2}=4$; $\rho=0.71$			
Missing completely at random (MCAR)			
σ^2_u	9.98 (0.14) {0.71}	9.79 (0.14) {0.13}	0.70 (0.002) {0.33}
σ^2_e	4.01 (0.02) {0.81}	4.00 (0.02) {1.00}	
Missing at random (MAR)			
σ^2_u	9.81 (0.14) {0.18}	9.78 (0.13) {0.08}	0.69 (0.002) {0.03}
σ^2_e	4.00 (0.02) {0.88}	4.01 (0.02) {0.60}	
Not missing at random (NMAR)			
σ^2_u	7.66 (0.13) {<0.001}	7.44 (0.14) {<0.001}	0.62 (0.004) {<0.001}
σ^2_e	3.55 (0.02) {<0.001}	3.54 (0.36) {0.20}	
Simulation two : $\sigma^2_{u1}=\sigma^2_{u2}=10$; $\sigma^2_{e1}=\sigma^2_{e2}=4$; $\rho=0.71$			
Missing completely at random (MCAR)			
σ^2_u	9.98 (0.14) {0.86}	9.79 (0.14) {0.87}	0.70 (0.002) {0.37}
σ^2_e	4.01 (0.02) {0.83}	4.00 (0.02) {1.00}	
Missing at random (MAR)			
σ^2_u	10.13 (0.14) {0.37}	9.72 (0.14) {0.04}	0.67 (0.003) {<0.001}
σ^2_e	3.99 (0.02) {0.51}	4.00 (0.02) {0.91}	
Not missing at random (NMAR)			
σ^2_u	8.02 (0.12) {<0.001}	8.02 (0.15) {<0.001}	0.63 (0.003) {<0.001}
σ^2_e	3.65 (0.02) {<0.001}	3.52 (0.02) {<0.001}	

The standard error of the estimates from the 300 simulations is given for each case and used to construct a one sample *t*-test comparing each estimate with its simulation parameter.

was no evidence to suggest this was not due to chance. For simulation two, although the parameters were consistently estimated, under the MAR process estimates for the between subject variance were slightly greater than their true values for group 1 and slightly lower for group 2. The evidence was, however, not convincing. For the NMAR case, all the parameters were under estimated. This is consistent with what was expected, as removal of the larger observations will have the effect of drawing the available data for all subjects closer together, thus lowering the between subject variance, as well as reducing the variance around a subject's fitted parameters.

Overall these analyses have shown that under an MCAR process, the method of analysis is not important in order to avoid biased estimation. They also highlighted that for an NMAR process, bias is unavoidable. The most interesting scenario for which some information is retrievable is under an MAR process. As expected, the UWLS analysis showed most bias, however, although to a smaller degree, some was also seen in estimates using GEE and MLn. When compared to the parameter estimates of the full data (tables 7.2 and 7.3) however, the extent of this bias was reduced somewhat.

7.2.6 A third simulated example

Since in practical situations it may be expected that the level of response will change over time a third simulation which allowed a fall in the level of response of time was also performed. Its simulation parameters are given in table 7.7. In this example, a fall in the level of response over time has been assumed. This has been allowed to vary across subjects. The variation between individuals in the underlying level of response has also been increased in this example. The missing data processes assumed in each case were the same as in the two previous examples. The observed proportions of missing data for each missing data process are shown in table 7.8. The distribution of observed proportions for each 300 simulations, as well as the

Table 7.7: Simulation parameters (2).

Mean response			Variance (random effects)		Missing data process		
					MCAR	MAR	NMAR
μ_1	μ_2	β	$\Sigma_1=\Sigma_2$	$\sigma^2_{e1}=\sigma^2_{e2}$	$Pr(r_{ij}=0)$	$Pr(r_{ij}=0 \mid y_{ij}>18.33)$	$Pr(r_{ij}=0 \mid y_{ij}>18.33)$
17	14	-0.7	$\begin{pmatrix} 35 & -0.9 \\ -0.9 & 0.8 \end{pmatrix}$	10	0.20	0.55	0.75

For the MAR simulation, data was designated missing with a probability of 0.55 on the basis of the last observed value denoted y_{ij}•.

number of subjects which were omitted from the UWLS analyses due to an excess of missing data are shown in figure 7.2. As expected, these show a higher proportion of missing observations in group 1 for both the MAR and NMAR cases. Similarly, for the UWLS analysis, the distribution of number of subjects removed because of lack of data was different in the two groups.

Table 7.8: Observed proportions of missing data in a third missing data simulation exercise.

	Mean proportion of missing values observed		
	Group 1	Group 2	Overall
MCAR	0.20	0.20	0.20
MAR	0.19	0.12	0.16
NMAR	0.25	0.14	0.20

The results for the full data and those following the application of each of the missing data processes are shown in table 7.9 for the marginal parameters and table 7.10 for the variance parameters. For the marginal parameters, as for the previous two examples, parameter estimation with MCAR missing data was consistent for all analysis methods. When data was MAR, the UWLS analysis consistently estimated the group difference, but its performance in estimating the group intercepts and slopes was poor. For GEE, although the group means were

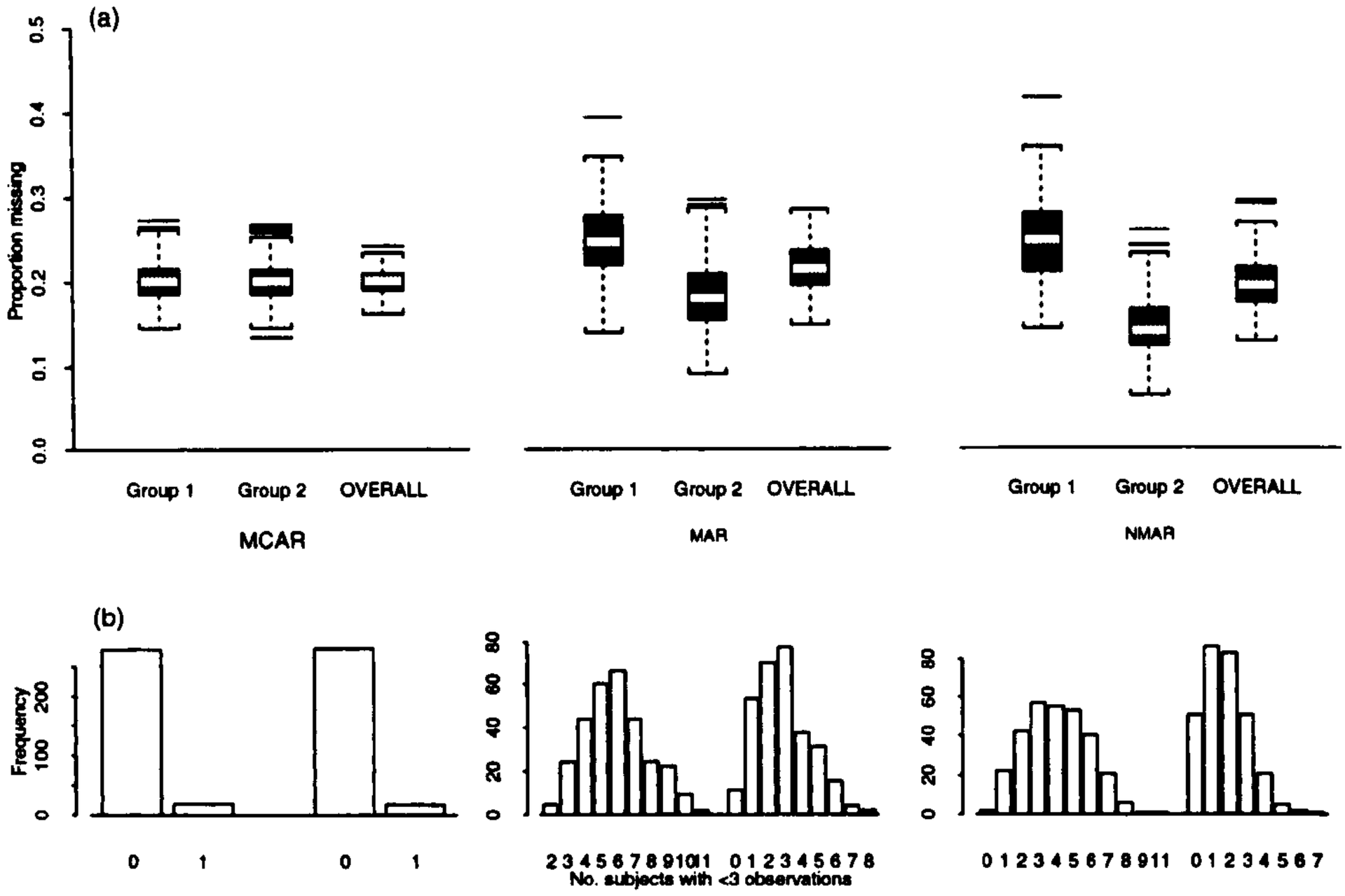

Figure 7.2: Missing data summaries for simulation three showing (a) the distribution of the proportion of missing data; and (b) the number of subjects who were omitted from the UWLS analysis because they had less than three observations for each of the 300 simulations.

again consistently estimated, there was some evidence of a bias in the estimate for the slope which was greater in absolute size than the true simulation parameter. Estimation of the fixed parameters using MLn was much better with no evidence of a bias for any of the parameters. When the data were NMAR all analyses performed badly. With the exception of the slope estimated with the GEE analysis, all parameter estimates were significantly lower than their respective underlying parameters.

Estimation of the variance parameters using MLn indicated a bias in estimation of the variance between subject intercepts and the covariance between slope and intercept under all three missing value processes. However, when compared to the estimates of the full data the

Table 7.9: Estimated marginal parameters (SE) {two sided *p* value} following missing data simulation.

	Method of analysis		
	UWLS	**GEE**	**MLn**
Simulation parameters : μ_1=17, μ_2=14, δ=-3, β=-0.7			
Full data (Estimate (SD))			
μ_1	17.06 (1.00)	17.07 (1.01)	17.07 (0.99)
μ_2	14.04 (0.96)	14.03 (0.98)	14.04 (0.95)
δ	-3.03 (1.38)	-3.03 (1.42)	-3.03 (1.34)
β	-0.70 (0.12)	-0.70 (0.12)	-0.70 (0.12)
Missing completely at random (MCAR)			
μ_1	17.05 (0.06) {0.40}	17.05 (0.06) {0.40}	17.05 (0.06) {0.41}
μ_2	14.02 (0.06) {0.72}	14.03 (0.06) {0.60}	14.02 (0.06) {0.73}
δ	-3.04 (0.08) {0.64}	-3.02 (0.08) {0.81}	-3.03 (0.08) {0.81}
β	-0.69 (0.007) {0.30}	-0.69 (0.007) {0.37}	-0.69 (0.007) {0.35}
Missing at random (MAR)			
μ_1	16.87 (0.06) {0.03}	17.01 (0.06) {0.89}	17.02 (0.06) {0.78}
μ_2	13.90 (0.06) {0.08}	14.02 (0.06) {0.77}	13.99 (0.06) {0.80}
δ	-2.97 (0.08) {0.72}	-2.99 (0.08) {0.92}	-3.02 (0.08) {0.78}
β	-0.66 (0.007) {<0.001}	-0.74 (0.007) {<0.001}	-0.69 (0.007) {0.18}
Not missing at random (NMAR)			
μ_1	15.41 (0.05) {<0.001}	15.89 (0.05) {<0.001}	15.79 (0.05) {<0.001}
μ_2	13.03 (0.05) {<0.001}	13.26 (0.05) {<0.001}	13.13 (0.05) {<0.001}
δ	-2.38 (0.08) {<0.001}	-2.64 (0.07) {<0.001}	-2.65 (0.07) {<0.001}
β	-0.68 (0.007) {0.01}	-0.70 (0.007) {0.81}	-0.64 (0.007) {<0.001}

The standard error of the estimates from the 300 simulations is given for each case and used to construct a one sample *t*-test comparing each estimate with their simulation parameter.

rating of its performance under MCAR and MAR was improved. Again in the NMAR case, all the estimated parameters were further away from their true values than in the MCAR and MAR examples with substantial evidence of bias.

Table 7.10: Estimated variance parameters (SE) {two sided *p* value} following missing data simulation.

	Simulation parameters : σ^2_u=35, σ^2_v=0.8, σ_{uv}=-0.9 σ^2_e=10			
	Full data Estimate (SD)	**MCAR**	**MAR**	**NMAR**
σ^2_u	37.02 (6.91)	37.03 (0.42) {<0.001}	36.90 (0.42) {<0.001}	29.16 (0.004} {<0.001}
σ^2_v	0.80 (0.16)	0.79 (0.01) {0.60}	0.79 (0.009) {0.48}	0.69 (0.43) {0.001}
σ_{uv}	-1.68 (0.87)	-1.68 (0.05) {<0.001}	-1.72 (0.05) {<0.001}	-1.39 (0.009) (<0.001}
σ^2_e	9.98 (0.63)	10.02 (0.04) {0.60}	10.05 (0.04) {0.22}	9.24 (0.04) {<0.001}

The standard error of the estimates from the 300 simulations is given for each case and used to construct a one sample *t*-test comparing each estimate with the respective simulation parameter.

7.2.7 Conclusion

These simulations have illustrated some elements of the discussion of the previous section. Primarily, they have shown that under MCAR missing data processes the choice of estimation procedure was not crucial in avoiding bias in the parameter estimation. As expected, under MAR the MLn estimated parameters showed little evidence of bias. Surprisingly however, GEE, a non-likelihood approach, also performed well. This was not so for the UWLS analysis which is some degree is due to the need to exclude completely, all subjects with less that three responses. As expected, under NMAR, without taking into account the missing data process explicitly in the analysis, all analyses performed badly in all but the most simple case of estimating a group difference of zero.

7.3 Determining the nature of a missing value process

As is evident from the previous section, the type of analysis required to make valid inferences about parameters of interest will always depend on the missing value process. For instance, an ignorable process which is MCAR rather than MAR will allow full flexibility in

the analysis procedure, whereas a non ignorable (NMAR) process as distinct from an ignorable one (MCAR or MAR) will give rise to potentially biased parameter estimation. Methods to assess the nature of the missing data process are therefore an important part of any applied analysis of data which are subject to missing data.

Some work has been reported in the literature discussing possible approaches in this area. In particular, Diggle (1989) proposed a non parametric test assessing whether missing data due to dropout occurred at random within distinct groups of the sample. Rideout (1991), in response to Diggle's paper, showed how using a logistic regression model for the odds of dropout for a given mean level of response gives an equivalent test for MCAR versus MAR processes. Other more recent work in the area has involved stratifying the population according to their observed pattern of missing data and then testing for the equivalence of response in each stratum (Park *et al.* 1993, Dawson 1994).

Unfortunately, the basic nature of missing data means that a robust test to highlight an ignorable (MCAR or MAR) process from a non ignorable one (NMAR) is impossible. Diggle and Kenward (1994) presented a model which related the odds of dropout to the observed measurement history and the conditional expectation of the unobserved response at the time of dropout. Essentially the method involves evaluating the expectation of the missing response at the time of dropout conditional on the observed responses and an underlying missing value process (based on the logistic regression model) in a full data likelihood model. Using these conditional estimates, the missing value process is then updated. The procedure continues iteratively until convergence. Naturally the results are very sensitive to the model specification for which no possibility of validation is available.

All of this work to be found in the literature refers to the missing data due to dropout as

opposed to intermittent missing values. Particularly when the proposed methods involve stratification by the observed pattern of missing data the extension of the methods for the intermittent case is not feasible. The remaining work in this chapter discusses an extension of the logistic regression model to help distinguish an MCAR from an MAR process for intermittent missing values in a repeated measurement problem. This is then extended and modelled alongside the quality of life measurement process to model jointly missing data and quality of life.

The first model is a multilevel logistic regression that is an extension of Rideout's logistic regression model for the intermittent missing values in a repeated measurement analysis. The model is a first order transition model with a random effect which relates the subject specific odds of missing responses to the underlying level of response observed at the previous measurement occasion (if available). Before applying this model to the missing data problem in the CRC NSCLC study, an exploratory data analysis of the missing data in this study is performed in more detail than given previously in Chapter 2.

7.3.1 The data

A description of the extent of the missing data within the CRC NSCLC study was given in tables 2.1 and 2.2. They showed that on a weekly basis, a little over 50% of data was available each month. Ignoring the complete non-responders and those subjects who responded only at pre-treatment, as they can offer nothing to any data analysis of post treatment response, 57 subjects remained. The amount of missing data among these subjects is summarised in table 7.11. Missing data that was due to patient death during follow-up was ignored in the analysis the reasoning for this is discussed in the subsequent sections. A 75% to 80% response rate was seen overall which was very evenly distributed between treatment groups and over time.

Table 7.11: Number of missing data each week in the CRC NSCLC study taken as a proportion of those subjects giving at least one post baseline response.

	n	Week							
		1	2	3	4	5	6	7	8
Overall	**57**	**0.28**	**0.21**	**0.19**	**0.25**	**0.26**	**0.19**	**0.18**	**0.23**
Continuous course	29	0.21	0.17	0.17	0.28	0.24	0.14	0.14	0.24
Split course	28	0.36	0.25	0.21	0.21	0.29	0.25	0.21	0.21

Further, in a logistic regression analysis of the proportion of responses given by each subject against several baseline covariates (table 7.12) it was seen that women tended to give a larger proportion of possible responses. There was also evidence of relationships with subject weight and Karnofsky performance score at baseline. The relationship with weight suggested that heavier subjects tended to complete a smaller proportion of responses. It was felt that this was likely to be confounded with the sex relationship. However, in a multiple regression analysis it was the relationship with weight that retained significance rather than

Table 7.12: Examination of patterns of missing data in the CRC NSCLC study: a logistic regression of number of responses recorded on baseline variables.

	Univariate analyses		Partial regression coefficients
Baseline variable	**Estimate (SE)**	**95% CI**	**Estimate (SE)**
Age (years) *	-0.08 (0.10)	[-0.28,0.118]	-
Weight (kgs) *	-0.21 (0.09)	[-0.38,-0.04]	-0.22 (0.09)
Sex (M=0, F=1)	0.35 (0.17)	[0.009,0.68]	-0.09 (0.22)
Karnofsky *	0.35 (0.08)	[0.20, 0.50]	0.42 (0.11)
FEV1 (l) *	3.31 (1.85)	[-0.30,6.94]	-
Treatment (short=0, long=1)	0.026 (0.15)	[-0.26,0.32]	-

* Results given per 10 unit increase.

that of sex. The relationship with Karnofsky performance indicator was as expected: subjects with the lower scores at baseline tended to complete a smaller proportion of their responses. This was relatively unchanged in a multiple regression analysis. No evidence of a relationship was seen with age, FEV1 (as a marker for disease severity) or treatment.

7.3.2 Logistic regression model for MCAR versus MAR

In Chapter 4 a random effects logistic regression analysis for repeated measurements was introduced which assumed that each subject has an underlying propensity of a positive response which was modelled as an odds on the log scale. Within this model, the coefficients of covariates in the model represent the absolute changes in this subject specific log odds for unit changes in the covariate value.

The first model presented here to investigate the nature of the missing data process investigates the dependency of missing data on the value of the previously observed measurement, if available, using this two level logistic regression model with binomial error e_{ij} at level one for a response $r_{ij}=1$ if data is available, 0 if missing for $j=2,\ldots,m$:

$$r_{ij}=p_{ij}+e_{ij} \tag{7.6}$$

with

$$\log\frac{p_{ij}}{1-p_{ij}}=\alpha+\gamma\, prevna_{ij}+\delta\, prevobs_{ij}+u_i \tag{7.7}$$

where $prevna_{ij}=1$ if $r_{ij-1}=1$, 0 otherwise (ie $prevna_{ij}=1$ if the previous observation is available, 0 otherwise), $prevobs_{ij}=y_{ij-1}$ if $r_{ij-1}=1$, 0 otherwise (ie $prevobs_{ij}$ is the value of the previous observation when available). The random effect u_i is assumed Normally distributed with zero mean.

Under this parameterisation $\exp(\alpha + u_i)$ is the subject specific log odds of having a response available at occasion j when the response at the previous occasion j-1 is missing. Within subject this is assumed constant over all j. The covariate effect γ has little interpretable value representing the modification on the log scale in subject specific odds of response when the previous observation is available and takes the value zero. The parameter of particular interest is δ which give the change in the subject specific log odds for each unit change in the previous response when observed. Although not a full test (as the missing value process may be related to more observations than simply that observed at the previous measurement occasion), a simple test for H_0: MCAR versus H_1: MAR is H_0: $\delta=0$.

It is worth noting that these covariate effects, although giving the same relative changes in the odds between subjects, absolute changes in odds for equivalent levels of response at occasion j-1 will differ in accordance with each subject's underlying propensity for missing data given by u_i. As discussed Chapter 4, these subject specific parameters u_i are not estimated explicitly. Instead we estimate their variance and thus the extent of variation between subjects in terms of this underlying propensity.

An assumption of the model is that within subject, the incidence of missing data is constant over time. A relaxation of this assumption involves incorporating an occasion covariate into the model. It may also be possible that variation between subjects may be explained by measured baseline covariates. By introducing these as additional covariates into the model of equation (7.7) and examining their effect on the extent of variation in the random effect u_i this may be examined. Similarly, this may help explain any dependency of the missing value process on available responses.

Within their analyses of dropout data, Diggle and Kenward (1994) observed in many of their

examples the incidence of dropout appeared related to changes in the level of response. Their model involved investigating changes between the conditional expectation of the unobserved response and the observed response at the previous measurement occasion. The role of changes in level can also be easily examined by a simple extension of the model of equation (7.7). This is done simply by changing the definition of the indicator $prevna_{ij}$ to take the value 1 only when the previous two observations are available (and hence a change in response can be calculated) and 0 at all other times with this modelled alongside the difference between these two measurements. For clarity this model is represented in equation (7.8) with the redefined variables denoted $chok_{ij}$ and $change_{ij}$ respectively. In order that it is possible to measure the change in the previous two responses, with this parameterisation, measurement occasions are limited to, j=3,...,m.

$$\log\frac{p_{ij}}{1-p_{ij}} = \alpha + \gamma\, chok_{ij} + \delta\, change_{ij} + u_i \qquad (7.8)$$

Estimation of the model parameters can be done as an iterative process using RIGLS (Goldstein, 1995) and a first (or second) order Taylor series expansion to linearise the logistic model. As discussed in chapter 4, if only the fixed parameter estimates are used to formulate the Taylor series expansion (marginal quasi-likelihood (MQL) estimation) all the parameters in the model will be subject to a downward bias which is proportional in size to var(u_i). Therefore penalised quasi-likelihood (PQL) estimation, which uses the estimated residuals from the model in addition to the fixed effects in evaluating evaluate the Taylor series expansion is used here.

To demonstrate the use of this model and its extensions, an analysis was carried out using quality of life responses from all 57 patients in the CRC study who had at least one post

treatment quality of life response available. Baseline responses were included in the analysis in the same way as those taken post baseline, giving j=1,...,8 for the model of equation (7.7) and j=2,...,8 for that of equation (7.8). Of these 57 patients, two died during the eight-week follow-up and hence the responses for these individuals were truncated at time of death.

The final data set analysed therefore contained 452 observations consisting of 55 patients each contributing eight observations and 2 patients contributing seven and five observations respectively. Out of these 452 observations, 102 were classified as missing on the HAD scale. Of the 102 missing observations, 47 responses were available at the previous measurement occasion. The mean anxiety score at this previous occasion was estimated at 5.66 (SD=4.10). The corresponding mean for available observations was 5.14 (SD=4.68). This information is summarised in table 7.13 along with the same information for the RSCL. A comparison of the two questionnaires shows that overall there were slightly fewer missing responses on the RSCL than on the HAD scale. There was some indication that the mean level of previous physical

Table 7.13: Crude summary of the missing data for the HAD scale.

		Overall	By previous value		Mean (SD) of available previous scores	
			Previous available	Previous missing		
Hospital anxiety and depression scale					Anxiety	Depression
n	Available	350	298	52	5.14 (4.68)	5.94 (4.31)
	Missing	102	47	55	5.66 (4.10)	6.70 (4.87)
Odds missing		0.29	0.16	1.06		
Rotterdam Symptom Checklist					Physical	Psychological
n	Available	369	324	45	13.56 (7.91)	12.13 (7.24)
	Missing	83	39	44	17.26 (9.72)	14.69 (7.22)
Odds missing		0.23	0.12	0.98		

scores was higher when subsequent responses were missing than when available.

Initially the relationship of missing data and anxiety response was examined using the basic model given in equations (7.6) and (7.7). The results for this model are given in the first column of table 7.14. Within the text they have been transformed to refer to the odds of missing data by taking the reciprocal of the estimated odds. For a more natural interpretation of the parameter of the $prevobs_{ij}$ variable, it was modelled as a deviation from the observed mean response for the whole data set. In the case of anxiety responses this was 7.03 units.

This model gave no evidence to suggest that the incidence of missing responses was related to the value of the previous anxiety response when available was given. The estimated within subject odds ratio and confidence interval for each unit in the value of the previous response was 1.03 [0.95, 1.10]. This was consistent with the summary results in table 7.13 where the mean anxiety response was slightly higher on occasions when the subsequent response was missing than when it was available.

The estimated variance of the random component u_i expresses how the underlying (log) odds of missing data varied across subjects. From this, a 95% reference range for the subject specific odds of missing data was calculated to give ranges on the odds scale of [0.14, 3.45] where previous values were missing and [0.03, 0.62] when previous values were available. On a probability scale these ranges correspond to [0.12, 0.72] and [0.03, 0.26] respectively. These intervals signified a large degree of variation between subjects and suggested that the incidence of missing data is to a large degree a subject specific phenomena.

The results for the remaining quality of life dimensions in the CRC NSCLC study are also given in table 7.14. Again, the $prevobs_{ij}$ variable was modelled in terms of a deviation from the

Table 7.14: Estimated coefficients (SE) for the basic model for each quality of life dimension estimated by PQL.

	HAD scale		RSCL	
	Anxiety	**Depression**	**Physical**	**Psychological**
Fixed parameters				
α (*cons*)	0.35 (0.24)	0.40 (0.25)	0.54 (0.27)	0.55 (0.27)
γ (*prevna*)	1.57 (0.28)	1.53 (0.28)	1.78 (0.32)	1.74 (0.31)
δ (*prevobs*)	-0.028 (0.040)	-0.056 (0.039)	-0.059 (0.024)	-0.051 (0.027)
Random parameters				
σ_u^2	0.65 (0.30)	0.77 (0.33)	0.81 (0.38)	0.81 (0.38)
-2 log lh	312.2	322.3	172.1	175.9

sample mean. For the depression, physical and psychological scores these were 7.90, 15.45 and 14.0 units respectively. Again, estimation used PQL with a second order Taylor series expansion.

There was little evidence of a relationship between missing data and the level of previous depression scores also obtained from the HAD scale, ($\exp(-\hat{\delta})$=1.06., 95% CI=[0.98, 1.18] for each unit score increase in the depression response). However, some evidence of a relationship of subsequent missing data with previous quality of life scores on the Rotterdam Symptom Checklist was seen. The estimated relative changes in subject specific odds were 1.06 (95% CI=[1.01, 1.11] and 1.05 (95% CI=[1.00, 1.11]) for each unit increase in the score for physical and psychological dimensions respectively. Such relationships corresponded to an increase in the incidence of missing data for higher levels of response (lower levels of quality of life). The lower limits of the respective confidence intervals for the confidence intervals were both very close to one demonstrating that this evidence was not strong. Fitting both dimensional responses simultaneously showed the information of the responses on the psychological

dimension could be explained by that in the physical dimension with a resulting change in deviance on adding the previous psychological scores of 0.4 (on 1 df). Again, these results were consistent with those seen in table 7.11.

Normal plots of level two residuals for each of the four models in table 7.14 are given in figure 7.3. These were all very skewed at the upper tail of the distribution signifying an upper constraint on the odds of missing data due to the short follow-up period and restricted number of possible responses. Given a longer follow-up or more measurement occasions, these distributions could be expected to improve.

The basic model for the relationship of missing data and previous physical quality of life responses was then extended. These results are given in table 7.15 to model changes in previous responses. Also presented is a repeated analysis for the level of previous observation.

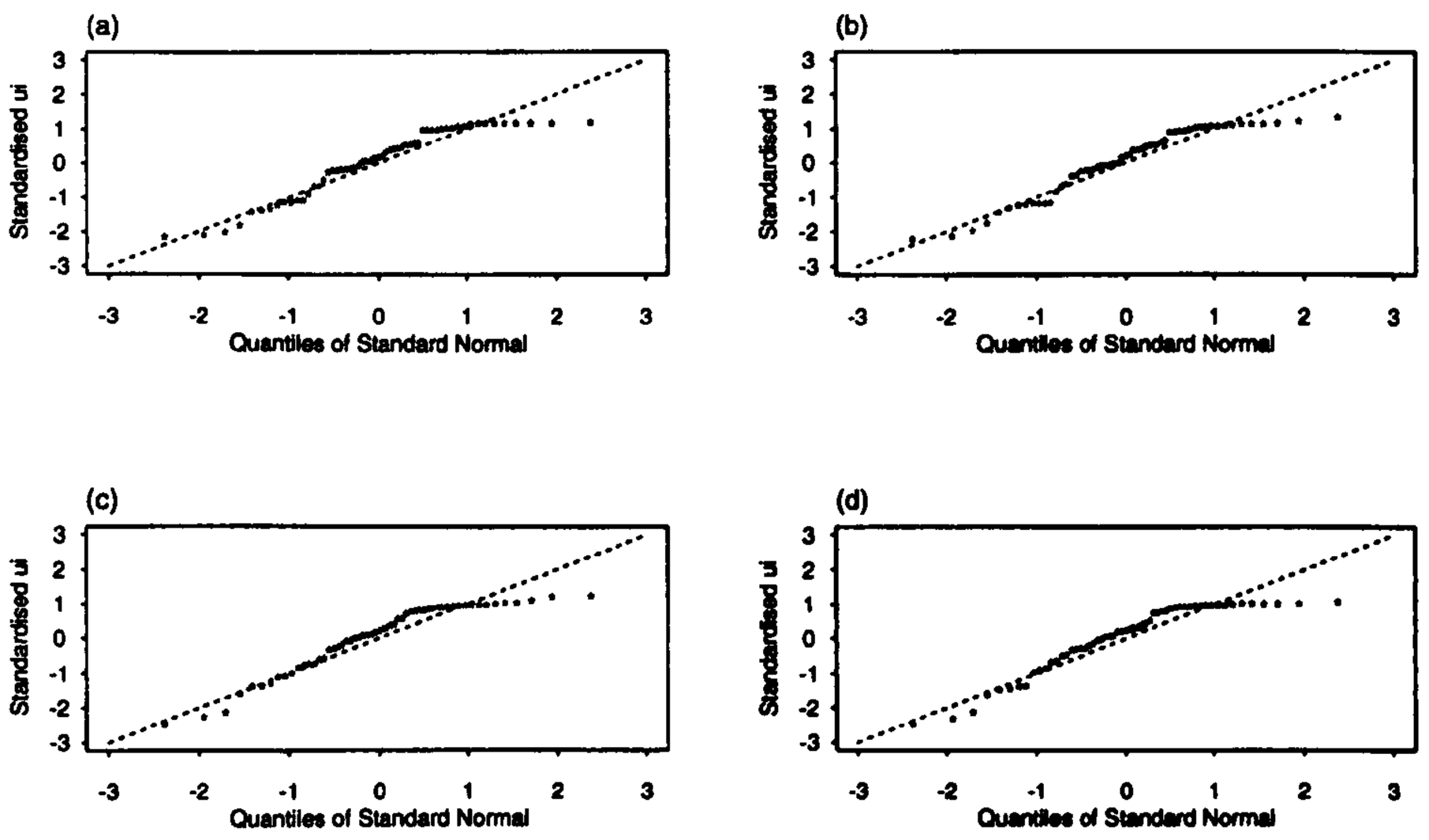

Figure 7.3: Normal plots of standardised level two residuals for the (a) anxiety; (b) depression; (c) physical; and (d) psychological quality of life responses.

Table 7.15: Parameter estimate (SE) for modelling changes in the previous two responses for physical quality of life estimate by PQL.

	Estimate (SE)	
	Previous response	**Change in previous two response**
Fixed parameters		
α (*cons*)	0.54 (0.37)	1.65 (0.40)
γ (*prevna/chok*)	2.85 (0.66)	1.36 (0.46)
δ (*prevphy/change*)	-0.049 (0.032)	-0.015 (0.022)
Random parameters		
σ_u^2	1.91 (0.76)	3.04 (1.07)
-2 log lh	-12.76	-13.00

These results differ slightly from those given in table 7.14 as the analysis was based on a slightly smaller data set with j=2,...,8 (as opposed to j=1,...,8 in the previous model).

There was evidence of an obvious reduction in power for this analysis shown by the ratio of parameter estimates to their standard errors being reduced in all cases. In particular the analysis failed to give as much evidence for the relationship between the incidence of missing data and previous physical response seen previously (p=0.13). Similarly, the lack of evidence to support a relationship between the incidence of missing data and changes in the previous two responses may have been be due to a lack of statistical power as only 28 cases of missing data with a change in previous level of quality of life responses were available. Indeed lack of power is a problem in all of the models in this section.

In the first analyses, a relationship was seen between the incidence of missing data and the level of physical quality of life observed on the previous measurement occasion. This gave evidence that the underlying missing data process was not missing completely at random.

(MCAR). However, if it were possible to explain the observed association by other known covariates at baseline, this conclusion may be revised. In the following analysis, the Karnofsky performance indicator, a measure of a patient's overall physical condition as judged by a clinician, was included in the basic model to assess this. The Karnofsky scale is scored in units of 10 between 0 (dead) and 100. Within the CRC NSCLC study data the lowest Karnofsky score reported was 50. For modelling purposes, the data were considered as (observed Karnofsky score - 50). Treatment and measurement occasion were also considered as potentially important covariates of interest.

Due to missing covariate measurements, 17 patients for whom the Karnofsky score was missing were excluded from the analysis. Within this smaller data set there were 319 observations that consisted of 71 missing values. Of these 71, 33 (46%) had a previous observation available versus 213 (86%) of the 248 available responses. This was in contrast to 46% and 85% in the larger data set. There was very little difference in the summary of previous observations in this reduced data set compared to that given in table 7.13.

The impact of the inclusion of each of the three variables was considered separately (table 7.16). In each case inclusion of the extra covariate changed the interpretation, and therefore estimates, of the intercept parameter. For each model the parameter represents the log odds of an available response when the previous response is missing and the covariate of interest takes the value zero for a subject with a random effect of zero. For the model parameterisation used here, this represents a pre-treatment Karnofsky=50 and split course radiotherapy group. The occasion effect was modelled as a linear term with j=2,...,8 modelled as j=0,...,6.

The relationship seen with baseline Karnofsky performance indicator was similar to that seen in the earlier model (table 7.12) and gave some evidence that subjects with a lower

Table 7.16: Fixed parameter estimates (SE) for a logistic MCAR versus MAR model for previous physical quality of life scores extended for baseline covariates (PQL estimation).

		Estimate (SE)			
		α (*cons*)	γ (*prevna*)	δ (*prevobs*)	σ_u^2
Baseline model		**0.61 (0.43)**	**2.88 (0.77)**	**-0.053 (0.036)**	**1.77**
Univariate extensions to the baseline model	**Karnofsky**	0.062 (0.026)	-0.87 (0.69)	2.80 (0.74)	-0.047 (0.034)	1.24
	Treatment	-0.27 (0.68)	0.78 (0.53)	2.89 (0.79)	-0.055 (0.037)	1.96
	Occasion	-0.034 (0.11)	0.69 (0.50)	2.92 (0.78)	-0.055 (0.036)	1.77

performance score at baseline tended to have a lower proportion of responses. There was a little evidence of confounding between this relationship and that observed with previous physical quality of life scores with only a small reduction in the estimated effect of the previous scores. The main effect of adding the Karnofsky indicator to the model was the noticeable reduction in the between subject variance from 1.77 to 1.24. Consistent with the results of the preliminary analyses there was no evidence to suggest a relationship of treatment group and the missing data. Similarly there was little evidence of a linear trend over time.

From these series of analyses it may be concluded that the prevalence of intermittent missing data in the NSCLC study was related to the patients underlying condition at the start of the study (given by their Karnofsky score) as well as their physical recorded physical quality of life during the study. It would therefore not be realistic that the missing data in study were missing completely at random.

7.3.3 Joint modelling of the quality of life and missing data process

The model in the previous section explicitly conditioned the missing data process on the observed data in an attempt to determine the nature of the missing data process. Such a model stems from the background of selection and pattern mixture models discussed in Chapter 6. As discussed in that chapter, an alternative way of approaching the problem of missing data is to condition not on the actual measurements, but some latent variables determining the underlying patient response (Wu and Carroll, 1989, Schlucter, 1992). This was demonstrated in Chapter 6 when patient survival was modelled alongside their quality of life in a multivariate model which enabled estimation of the joint distribution of survival and quality of life. A similar model may be applied to the intermittent missing data problem. Such a model assumes that missing data is not due to the level of observed quality of life at a particular time, but on a patient's underlying quality of life response as well as their own propensity for non response. The type of inferences this allows are as to whether a subject who has a higher than average intercept and slope tends also to have a higher than average propensity for missing data. The model therefore does not concentrate on the particular instances of missing data, it focuses more on the overall level of missing data for each subject and addresses the question as to whether the parameters underlying the quality of life (response) and missing data processes may be assumed distinct.

Unlike the survival model of Chapter 6, the model is a three level model. This difference arises because in the case of missing data, there exists random variation both within and between subjects for both outcomes of interest. The variation at level three is between subject and at level two between occasions. No variation is modelled between dimensions (quality of life and missingness) at level one. The response y_{ijk} is made up of the quality of life responses y_{ij} and missing data process r_{ij} such that $y_{ijk}=y_{ij}$ for $k=1$, r_{ij} for $k=2$. Given this, the most simple model is then expressed as

$$y_{ijk}=z_{ij1}\{\alpha_1+\beta_1 x_{ij1}+u_{i1}+v_{i1}x_{ij1}+e_{ij1}\} \\ +z_{ij2}\{[1+\exp(-\alpha_2-u_{i2})]^{-1}+e_{ij2}\} \quad (7.9)$$

where $z_{ij1}=1$ for all y_{ij1}, 0 otherwise and $z_{ij2}=1-z_{ij1}=1$ for all y_{ij2}, 0 otherwise. This gives a model for the quality of life responses corresponding exactly to the simple variance components model of equation (3.1), where x_{ij1} denotes the time of measurement, and one for the missing data process that is analogous to the multilevel binary response model of equations (4.14) and (4.15). In the latter, the subject residual u_{i2} denotes the deviation from the average log odds of missing data for the ith subject assumed to be Normally distributed with zero mean and variance σ^2_{u2}. The residual component of this model, e_{ij2}, is assumed to come from a binomial distribution and has with mean θ_{ij} and variance $\theta_{ij}(1-\theta_{ij})$, where $\theta_{ij}=E(r_{ij})$ and may or may not vary with j.

The extra variance components of this model over and above those of two separate models represent the associations at level three between (u_{i1},u_{i2}) and (u_{i2},v_{i1}) denoted σ_{u12} and σ_{uv12} and at level two between (e_{ij1},e_{ij2}) denoted σ_{e12}. When all of these covariance terms are fitted, the model coefficients for the quality of life response give the mean intercept and slope for a subject with an underlying mean propensity for missing data (as given by the missing data coefficient α_2). On the other hand, constraining these covariances to be equal to zero, will give results exactly comparable with fitting two univariate models.

An extension to the model of equation (7.9) including additional covariates was fitted to the CRC NSCLC physical quality of life data. The extra covariates were fitted within the quality of life response part of the model included a treatment covariate denoted $rt_{ij}=1$ for the continuous course, 0 for the intensive split course, and a baseline response, $base_{ij}$ which was centred around the overall mean of 15.45 units. The data set used therefore consisted only of the 42 subjects who responded at baseline and at least one post treatment follow-up. Out of

these 42 subjects, the proportion of missing data in each treatment arm was 15% overall, and 14% and 16% for the long and short courses respectively.

Three successive models were used. The first assumed independence between the two outcomes, with all the covariances between them constrained to be zero. The second model allowed a dependency between the two intercept terms, and the final model had a full covariance structure. The results are given in table 7.17.

The results from model one are analogous to those of two separate models: for the quality of life response there was some evidence of a slight fall in the level of physical quality of life over the period, but no evidence of any difference between the two treatment groups. There was a large degree of variability between the subject specific profiles. The missing data part of the model fit only an intercept term therefore assuming the same constant level of missing data over the period in the two treatment arms. The fitted intercept term of 1.77 (95% CI=[1.30, 2.24]) gave an estimate of the overall proportion of missing data of 0.15 (95% CI=[0.10, 0.21]). Again there was a large amount of variation between subjects with a 95% reference range for the subject specific proportion of missing data of [0.02, 0.55].

On extending model one to allow covariance between the level of quality of life and the log odds of missing data, the intercept parameter α_2 may be interpreted as the log odds of available data for a subject with average (baseline adjusted) intercept. Similarly the quality of life intercept term α_1 reflects the average baseline adjusted intercept for a subject with 15% missing observations. Very little change in any of the parameter estimates was seen. This was indicative of the small estimated correlation between the two intercept terms with very little evidence of a real effect with a change in -2 log lh of 0.5 on 2 df for the addition of two extra parameters. This was also true for the estimated covariance between the odds of missing data

Table 7.17: Parameter estimate (SE) for modelling changes in the previous two responses for physical quality of life.

		Estimate (SE)		
		Model one	**Model two**	**Model three**
Fixed parameters				
α_1 (*cons*$_1$)		15.15 (1.41)	15.01(1.40)	15.03 (1.40)
β_1 (*occ*)		-0.61 (0.18)	-0.61 (0.18)	-0.62 (0.18)
γ_1 (*base*)		0.68 (0.13)	0.67 (0.13)	0.68 (0.13)
δ_1 (*rt*)		1.64 (2.09)	1.56 (2.08)	1.56 (2.08)
α_2 (*cons*$_2$)		1.77 (0.24)	1.77(0.24)	1.78 (0.24)
Random parameters				
Level three	σ^2_{u1}	36.7 (9.92)	36.4 (9.86)	36.5 (9.90)
	σ_{uv1}	-1.02 (1.23) {-0.19}	-1.02 (1.23) {-0.19}	-1.04 (1.23) {-0.19}
	σ^2_{v1}	0.81 (0.27)	0.81 (0.27)	0.81 (0.27)
	σ^2_{u2}	1.01 (0.47)	1.00 (0.47)	1.01 (0.47)
	σ_{u12}	-	-0.93 (1.47) {-0.15}	-1.05 (1.56) {-0.17}
	σ_{uv12}	-	-	0.06 (0.27) {0.07}
Level two	σ^2_{e1}	10.1 (1.07)	10.3 (1.10)	10.3 (1.10)
	σ_{e12}	-	0.49 (0.23) {0.15}	0.48 (0.23) {0.15}
-2 log lh		1667.0	1666.5	1666.5

Standard errors are given on the variance parameters only as an indication of their variability. The estimated correlations are given within the curly brackets{ } for the covariance estimates. No estimate is given for σ^2_{e2}=var(e_{ij2}) as in all cases this has been constrained to be equal to one for binomial variation.

and the rate of change in quality of life - the extension of model two to model three.

Residual diagnostics for model one are shown in figure 7.4. These show a very poor validity of the main Normality assumptions of the model, particularly in respect to the missing data part of the model. This is the result of the restricted possible available responses per subject and indicates that the model is perhaps more suited to data with a longer follow-up period with more

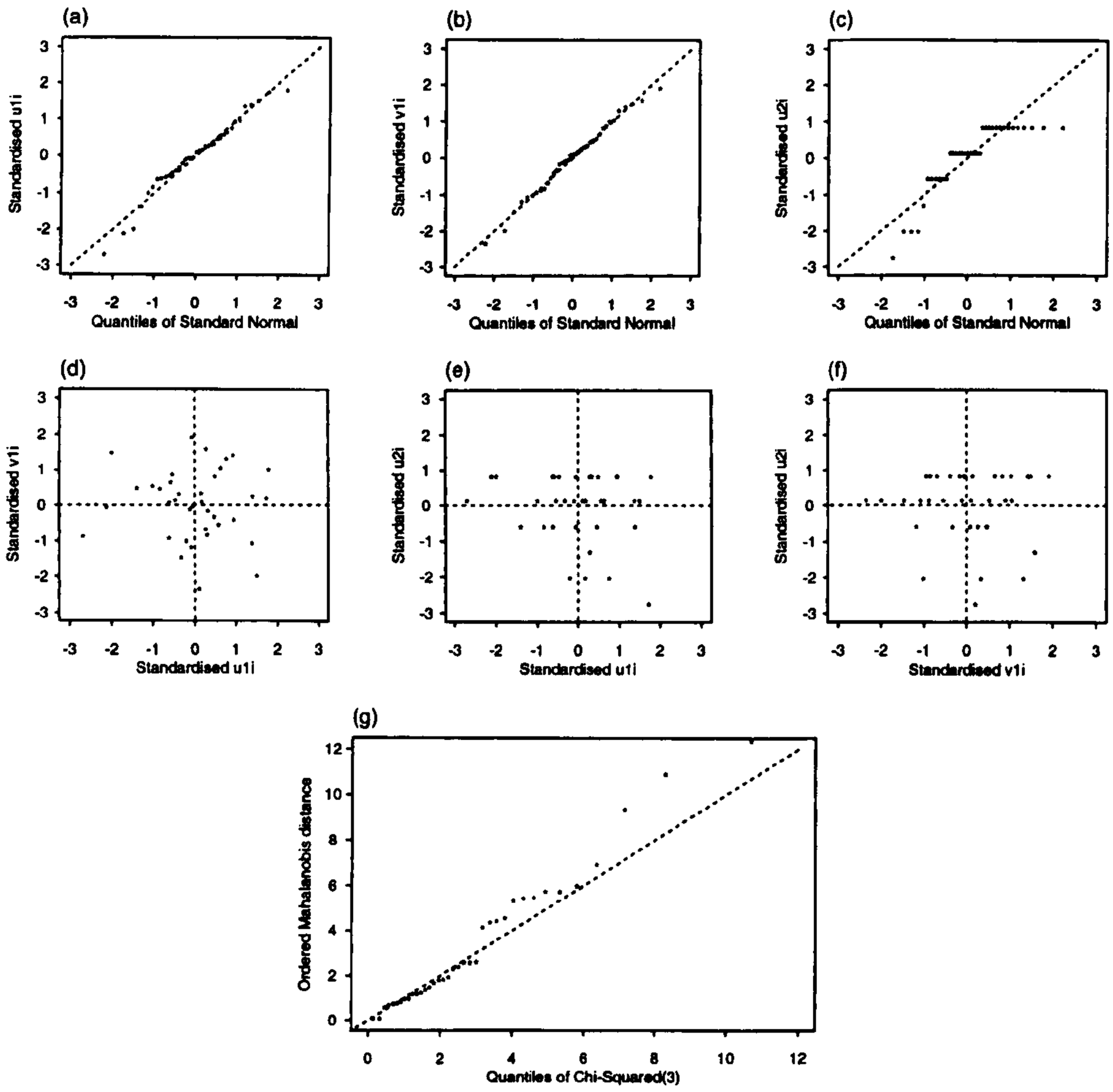

Figure 7.4: Residual diagnostics for joint quality of life and missing data model showing univariate Normal plots for (a) u_{i1}; (b) v_{i1}; and (c) u_{i2}; bivariate plots of (d) (u_{i1},v_{i1}); (e) (u_{i1},u_{i2}); f) (v_{i1},u_{i2}); and (g) a Chi-squared plot of the Mahalanobis distances.

measurement occasions.

7.3.4 Conclusions

Two models have been used to help determine how the intermittent missing data relates to the response process which is observed in a study. The first model examined the relationship between the incidence of missing data and previous responses either observed or missing using a two level repeated measurement model which considers each subject to have an underlying

propensity to miss responses. This subject specific odds is assumed to be acted on in the same relative way by covariates in the model for all subjects. For the CRC NSCLC study, this model gave some evidence to suggest that missing data was related to the previously observed score from either dimension of the Rotterdam Symptom Checklist. This relationship was slightly more pronounced for the physical quality of life score, suggesting that subjects with missing responses tended to have recorded higher scores (poorer quality of life) in their previous response. No similar evidence was seen for the anxiety and depression scores of the HAD scale although the observed relationships in the data were in the same direction as those for the physical and psychological dimensions. The lack of evidence to support these relationships may have been due to a lack of statistical power. In addition, although a relationship between missing data and baseline Karnofsky performance was also seen, this did not seem to detract substantially from the quality of life association. Therefore, assuming the missing data was ignorable (MCAR or MAR) these analyses have given some evidence against the assumption of MCAR particularly in terms of patient physical well being.

In the second model, rather than conditioning on the level of the observed response, the model considered allowed the underlying propensity for missing data to be related to the underlying level of response characterised by subject specific random effects rather than the actual level of observed responses. It therefore addresses the question whether the parameters underlying the quality of life (response) process may be assumed distinct from those of the missing data process. For the CRC NSCLC RSCL physical data, little evidence of a dependence both in terms of underlying level and rate of change of response over time was seen. However, residual diagnostics of the model put its conclusions into doubt. This poor performance was thought to be due to the small number of measurement occasions in the data set.

From these results it must be concluded the missing data in the study could not be considered as missing completely at random with missingness showing at least some relationship to observed physical well being.

7.4 Summary and discussion

Intermittent missing data has been reported as a problem in most cancer clinical trials with quality of life assessment. Although clear definition of how theoretically the relationships between missingness and response may affect analysis options, little practical experience in terms of the performance of different methods of estimation has been reported for intermittent missing data (as opposed to dropout). In the first half of this chapter, a simulation exercise was therefore performed to investigate the performance of three methods of analysis suitable for the analysis of continuous outcomes: subject specific analyses with unweighted combination (UWLS); generalised estimating equations with an exchangeable correlation structure (GEE); and random effect (or hierarchical) models using MLn. Data were simulated under three different underlying response processes, and missing data assigned under MCAR, MAR and NMAR missing value processes. The results showed that, with approximately 20% of data missing, all three analyses performed well under MCAR. When data were MAR, estimates from the UWLS analyses were always biassed, with MLn performing best of all. Under NMAR, when the underlying response process and missing data processes were the same across patient groups, although the estimated patient group means were biased, the estimated group differences were unbiased. This was not the case when group differences did exist and evidence of bias was seen in the parameter estimates. The GEE analysis performed best of all in this situation. The main conclusion to be drawn from these simulations is that, missing data (and more particularly the type of missing value process) has a large impact on the choice of analysis highlighting the importance of investigating the nature of missing value process.

Although some work has been presented in the literature in this area, much of it has dealt with the special case of missing data due dropout (or 'monotonic' missing data) rather than that which is missing intermittently. The second half of this chapter concentrated on developing such models that may help in this process when data is missing intermittently. The first model looked at determining whether subsequent missing data was related to previously observed quality of life. The model is similar to that suggested by Rideout (1991) for investigating the nature of a dropout process. The major difference is that for intermittent missing data, repeated missing responses for subjects required the problem to be set in a repeated measurement framework. There are a number of drawbacks with this model however, in particular its potential lack of power which will be inversely related to the extent of the missing data problem, such that the greater the missing data problem, the lower the power of the analysis. This in turn means that relationships between missing data and underlying response in small data sets should not merely be ruled out on the basis of a lack of statistical evidence. In addition, the results are all very model dependent. For instance, in the example given, there may have been a relationship with missing data and subsequent response which would never have been determined from the series of models fitted here.

Under the assumption that missing data is MAR, Rubin (1976) showed that inferences based on likelihood analyses would be valid provided the parameters which define the missing data process and measurement process are distinct. Although a fundamental assumption for Rubin's assertion, little work has been done to verify the assumption in practice. The second model presented here attempted to do this. Unfortunately in the example here the data set was too small to justify the assumptions made. However, with a larger data set the model may be useful although more work is needed to truly assess the properties of the model and determine in what circumstances its performance is maximised.

The type of missing data process which presents the greatest analysis problem is a non-ignorable or NMAR process which the work in this chapter has not directly addressed. The fundamental problem with such a process is that the validity of any models which are developed are impossible to verify due to the very nature of the process one is trying to determine. The best current solution to the problem, suggested by Glynn *et al.* (1993) is that a combination of pattern mixture models (Little, 1994) and multiple imputation (Rubin, 1987) be used under reasonably assumed non-ignorable processes. This produces a series of sensitivity analyses such that the robustness of estimates made ignoring the process can be explicitly assessed. Their examples related to survey data and it was suggested that follow-up surveys of non-respondents could be used to help decide on reasonable missing data processes. This will not be feasible in quality of life assessment in clinical trials studies where the number of missed assessments per individual is great and often patients have died by the time of analysis. In many studies quality of life assessment, however, often takes place at a similar time to clinical follow-up so that some information about a patient's general condition may be available to aid in the definition of appropriate missing data processes.

One of the main limitations of the work in this chapter is that it is focused on the analysis of continuous outcomes. Given the amount of data in quality of life studies which is measured on an ordinal scale (which may then be dichotomised), further work is needed to asses the ways in which intermittent missing data will affect the types of models discussed in Chapters 4 and 5 for the analysis of such data. A particular example is the case of missing item response for which the multivariate binary model of Section 4.5 may be of great help when missingness is MCAR. It is unclear however how this model will perform under MAR or NMAR missingness processes.

8 Discussion and Recommendations

The assessment of quality of life as a primary outcome in cancer clinical trials is now almost universal. To date, its reporting in the applied literature has generally used simple descriptive summaries. This means that the statistical inferences that can be drawn about differences in quality of life between patient groups, as well as changes in quality of life over time, can be limited. Further, as these analyses often ignore many natural characteristics of the data, their conclusions could be misleading. There is therefore a definite need to improve the analysis of such data. For this end, the aim of this thesis was to assess the practical application of recent developments in statistical methodology to handle the problems faced in the analysis of self assessed quality of life data from cancer clinical trials. Those of particular concern were highlighted in a review paper by Cox *et al.* (1992) relating to the analysis of typically unbalanced repeated measurement data, multiple dimensional outcomes and missing data, where the latter may be in the form of censoring of the quality of life response due to patient death, or simply because patients have failed to complete questionnaires.

Descriptive analyses that have been typically used to present quality of life data in the applied literature were reviewed in Chapter 2, and it was concluded that, given the quantity of data that a quality of life study generates, these are essential in the analysis process. However, they should concentrate not only on examination of average behaviour over time, as has generally been done to date, but also consider individual patient data, as well as distributions of summary statistics calculated for each subject. Additionally, in the light of missing data, patterns of missingness both in relation to observed quality of life and disease characteristics need to be examined. Given the frequency of missing data as well as the repeated

measurements and multiple dimensionality typical of quality of life data, these analyses should not constitute the whole data analysis and the use of statistical models that account for the data structure and allow solid conclusions to be drawn need to be encouraged. It was on such analyses that the remainder of this thesis was focused.

Essentially two classes of model that appropriately allow for the dependence between repeated measurements within subjects were considered: random coefficient (or hierarchical) and marginal models. The fundamental difference between the two models is the way in which they incorporate this dependency into the modelling framework. In Chapter 3 it was seen that the dependency within subjects can be accommodated by incorporating a subject specific random effect into the model. In turn, such *random coefficient* models then allow examination of the components of variance of the data. In the most simple case, this is the variance within and between subjects. These models were shown also to offer the ability to examine this variance structure further in terms of its relationship to other patient characteristics, as well as treatment. Further, they can be extended to incorporate the assessment of multiple dimensions of quality of life which not only allows estimation of inter-dimensional correlations within and between subject, but permits overall covariate effects (for example the effect of treatment) to be estimated if appropriate.

Although the work of Chapter 3 concentrated on the analysis of continuous outcomes, it was shown in Chapter 4 that these random coefficient models can also be used for the analysis of repeated binary outcomes that arise from arbitrary dichotomisation of the ordinal scales of individual items on a questionnaire. This not only allows the pattern of behaviour of specific aspects of patient quality of life to be examined using the most basic model, but also a multivariate model can be used to obtain an overall analysis of patient quality of life which has more intuitive appeal than the more conventional summary scores. Again the associations

between the individual symptoms are also examinable in this model. Further, it also offers a simple solution to the problem of missing item responses (Fayers, 1996) - that is, when patients fail to answer individual questions on a questionnaire as opposed to missing the whole questionnaire. Missing items usually result in the loss of the occasion response for a subject as they mean that a summary score cannot be evaluated for the questionnaire. However, since the random coefficient model can easily handle unbalanced data, if it can be assumed that the reasons underlying the missing item are ignorable such missingness is a trivial matter in the analysis.

Marginal models, the second class of models used, treat the dependency between repeated observations as nuisance parameters. The resulting parameter estimates have a very different interpretation to those of the random coefficient model - they give the estimated effect of covariates of interest on the response of population as a whole, whereas those of the random coefficient model are estimated in terms of their impact on a subject specific response. It was demonstrated in Chapter 4 that, although this distinction is not important when a identity link function is used (as is generally done with continuous outcomes), it can be vital when a logit link function is used with binary outcomes and can lead to large differences in the resulting parameter estimates. In this case, it is therefore important to determine whether it is the population average or subject specific covariate effects that are of interest prior to deciding on the analysis strategy. In terms of a treatment covariate, this relates to whether it is a general public health perspective that is of interest (population average) or simply the effect of treatment on an individual patient's quality of life (subject specific).

Because of the way both of these models incorporate the variance structure into the analysis procedure, they can easily be extended to incorporate dependency arising from alternative sources. One such additional source variance arises when logistic regression models are used

for the analysis of repeated ordinal data following transformations from the ordinal scale to a series of correlated binary responses (McCullagh and Nelder, 1983). The application of both the marginal and the random coefficient model for two such transformations - *cumulative probability* and *continuation ratio* - were discussed in Chapter 5. The models presented have the same properties as those of Chapter 4 with an important distinction needed between the way in which they incorporate the dependency between repeated measurements (as random effect or nuisance parameters). In addition, the two transformations presented also result in very different parameter interpretations between which a clear distinction is needed. Unfortunately, this can make the models difficult to interpret and their practical use may therefore be more limited than that of the more simple binary models for chosen dichotomies, despite their often arbitrary choice.

The application of many of the models in Chapter 3 to 5 was demonstrated assuming the response to be a simple linear function of time. The exception to this was the use of natural cubic splines for the analysis of complex patterns of response due to transient symptoms following treatment (Chapter 4). Such behaviour is common in cancer clinical trials, particularly involving chemotherapeutic treatment (MRC lung cancer working party, 1991a). Although the application demonstrated the cubic spline within a marginal model, when the timing of treatment varies across patients it may also be possible to incorporate such complex patterns at a subject level within a random coefficient model. Since this is a feature which can frequently occur within cancer clinical trials, it is a particularly relevant, and very interesting, area of further research.

Another important consideration in the analysis of quality of life data is the problem of informative censoring of patient quality of life as a direct result of patient death. Two different classes of models to address this issue were reviewed in Chapter 6. *Dropout models* (Little,

1995) attempt to obtain correct inferences about the response in the light of such patient dropout whereas *quality adjusted survival* techniques combine the quality of life and survival endpoints to give a quality adjusted survival about which inferences are then made. Although the latter have been applied extensively and successfully for toxicity data, their use in self assessed quality of life data was shown to be problematic and it was concluded that only a *partitioned quality adjusted survival* (PQAS) (Glasziou *et al.*, 1990) analysis is currently useful. Unfortunately, this analysis is restricted to instances when progressive quality of life states can be defined which somewhat limits their use for self assessed quality of life data that is often seen to fluctuate between states. In terms of the three classes of dropout models that have been discussed in the literature (Little, 1995), it was concluded the *informatively right censored* dropout models (Wu and Bailey, 1988) and in particular, the *trivariate Normal model* (Schlucter, 1992) is the most suitable for application with quality of life data. Although this latter model considers quality of life and survival together, it treats them as a multivariate problem estimating their joint distribution, from which their conditional distribution can also be determined. This is an area in which more work is needed. In particular, with the application of such models to binary outcomes, and their behaviour when many survival times are censored. A further class of model, that has not been addressed here, is *multi state* models (Kay, 1985). These are transitional (Markov) models that incorporate death as an absorbing state and present a further area of interesting research in this area.

The value of all of these models for analysing self assessed quality of life data depends on their ability to cope with the presence of intermittent missing data during follow-up. This is a particular problem in quality of life studies and was discussed in Chapter 7. In the notation of Little and Rubin (1987), since the marginal model analyses presented in Chapter 4 and 5 for binary and ordinal outcomes used quasi-likelihood estimation procedures, these assume data are *missing completely at random* (MCAR), as do the quality adjusted survival analyses of

Chapter 6. Although the random coefficient models presented in Chapters 3 and 6 make the assumption that the data are *missing at random* (MAR), it is unclear under what conditions the random coefficient binary models, used in Chapters 4 and 5, will be valid (whether just MCAR or MAR). This needs to be assessed with a simulation exercise similar to that presented for continuous outcomes in Chapter 7. For such data, these confirmed that under MCAR the choice of analysis procedure makes little difference to the analysis conclusions. Under MAR, the random coefficient models which used RIGLS (Goldstein, 1986) estimation or the marginal models using GEE (Liang and Zeger, 1986) performed adequately, with an unweighted least squares analysis leading to estimation bias. When data were *not missing at random* (NMAR) all analyses performed badly, although those using GEE still gave unbiased estimation of rate of change and group differences within one example.

Because of the differential bias of different analysis strategies in the light of missing data, investigation of the nature of a missing value process is important. Unfortunately, by definition it is impossible to identify a missingness process that is (NMAR) purely on the basis of the observed data. Although a number of models to distinguish between MAR and MCAR processes have been applied in determining the nature of dropout processes (Rideout, 1991, Park *et al.*, 1993, Dawson, 1994), these are generally unsuitable for application with intermittent missing data. Two alternative models were suggested in Chapter 7 and, when applied to data of a recent cancer clinical trial, gave some evidence of an MAR process as opposed to MCAR. However, caution is needed in interpreting these models as they may lack power and so be prone to false negative results. Further, as it is impossible to determine whether a process may be NMAR, the best current solution seems to be that offered by Glynn *et al.* (1993). This work, however, is little developed in terms of practical application, and is an area which offers much scope for interesting and challenging further research.

All of the analyses of the work presented in the thesis were performed using MLn (Rasbash and Woodhouse, 1995) or S-Plus (Becker *et al.*, 1988) software. MLn has been specifically written to estimate the parameters of random coefficient models using (restricted) IGLS. Other packages can also fit such models (Kreft *et al*, 1994), in particular, all the analyses presented could be performed using HLM (Bryk *et al.*, 1988), VARCL (Longford, 1988) and BUGS (Thomas *et al.*, 1992). In addition, two level models for continuous outcomes can be fitted using SAS Proc Mixed (SAS Institute Inc., 1992), S-Plus (version 3.3) and BMDP-5V (Zwinderman, 1990). For each case, different estimation procedures are used: for instance, HLM uses the EM algorithm (Dempster *et al.*, 1977), and BUGS uses Markov Chain Monte Carlo methods (Gilks *et al.*, 1993). Marginal models can be fitted using most software packages, although the choice of package will determine the ease in which robust standard errors are obtained. The models that were presented here used a first order generalised estimating equation (GEE1) which can be fitted using OSWALD (Smith and Diggle, 1994), a library of S-Plus function. These functions also allow the alternating logistic regression (ALR) models, that were outlined in Chapter 4, to be fitted. Macros which fit models using GEE1 are also available for SAS software. Alternatively, robust standard errors can be obtained for the simple repeated measurement model by bootstrapping or jack-knifing (Efron and Tibshirani, 1993) where the individual forms the experimental unit. STATA (StataCorp., 1995) also has facilities to obtain Huber (1967) estimates of the standard errors for both continuous and binary outcomes that will be very close to those of a GEE1 with an independence working correlation matrix. As software to perform the quality adjusted survival analyses is not generally available, a number of S-Plus functions were specifically written for the analyses performed here. These are listed in Appendix 3.

This thesis has shown that there exist a number of recently developed statistical methods that can be very successfully used to tackle the major issues that have been raised concerning

the analysis of quality of life data. Essentially, the central feature required of such models is their ability to handle unbalanced repeated measurement data and appropriately adjust for the dependence between observations measured on the same subject. Since all the examples in the thesis gave evidence of a large degree of between subject variation, the impact of ignoring the repeated measurement structure for these data would have been great. As it is believed that these data are typical of quality of life studies in general, it is concluded that the use of statistical models, such as those presented here, are not only essential for the analysis of quality of life data in general, but that they are generally accessible and should therefore be encouraged within the applied quality of life literature. This requires not only that the methods themselves are communicated within the quality of life research field, but also that the limitations of the simple descriptive analyses are highlighted. There are however a number of areas which do require further attention, the most challenging of which relate to coping with non-ignorable intermittent missing data and the informative censoring of response as a result of patient death.

References

Aaronsen, N. K. and Beckman, J. *The quality of life of cancer patients*, Raven Press, **17**, (1987).

Aaronsen, N. K., Bullinger, M. and Ahmedzar, S. 'A modular approach to quality of life assessment in cancer clinical trials', *Recent Results in Cancer Clinical Trials*, **111**, 231-249, (1988).

Aaronsen, N. K. 'Quality of life assessment in clinical trials: methodological issues', *Controlled Clinical Trials*, **10**, 195S-207S, (1989).

Abrams, K. A. 'Discussion of the paper by Cox *et al.*' *Journal of the Royal Statistical Society, Series A*, **155**, 353-393, (1992).

Agresti, A. *Categorical data analysis*, John Wiley, (1990).

Agresti, A. 'A survey of models for repeated ordered categorical response data', *Statistics in Medicine*, **8**, 1209-1224, (1989).

Agresti, A. and Lang, J. 'A proportional odds model with subject specific effects for ordered categorical response', *Biometrika*, **80**, 527-534, (1993).

Allen-Mersh, T. G., Earlham, S., Fordy, C., Abrams, K. and Houghton, J. 'Quality of life and survival with continuous hepatic-artery floxuridine infusion for colorectal liver metastases', *The Lancet*, **344**, 1255-1260, (1994).

Altman, D. *Practical Statistics for Medical Research*, Chapman and Hall, first edition, (1991).

Andersen, H., Hopwood, P., Prendiville, J., Radford, J. A., Thatcher, N. and Ashcroft, L. ' A randomised study of bolus *vs* continuous pump infusion of ifosfamide and doxorubicin with oral etoposide for small cell lung cancer', *British Journal of Cancer*, **67**, 1385-1390, (1993).

Andersen, P. K. Hense, L. S., and Keiding, N. 'Assessing the influence of reversible disease indicators on survival', *Statistics in Medicine*, **10**, 1061-1067, (1991).

Armitage, P and Berry. G. *Statistical Methods in Medical Research*, Blackwell Scientific Publications, second edition, (1987).

Armstrong, B. G. and Sloan, M. 'Ordinal regression for epidemiological data', *American Journal of Epidemiology*, **129**, 191-204, (1989).

Ashby, M., Neuhaus, J. M., Hauck, W. W., Bacchetti, P., Heilbron, D. C., Jewell, N. P., Segal,.

References

M. R. and Fusaro, E. E. 'An annotated bibliography of methods for analysing correlated categorical data', *Statistics in Medicine*, **11**, 67-69, (1992).

Becker, R. A., Chambers, J. M. and Wilks, A. R. *The new S Language*, Wadsworth and Broks-Cole, Pacific Grove, (1988).

Benjamin, B. and Pollard, J. H. *The analysis of mortality and other actuarial statistics*, Heinemann: London, (1980).

Bergner, M., Bobbit, R. A., Carter, W. B. and Gilson, B. S. 'The sickness impact profile: development and final revision of a health status measure', *Medical Care*, **19**, 787-806 (1981).

Bland, J. M. and Altman, D. G, 'Correlation, regression and repeated data.', *British Medical Journal*, **308**, 896, (1994).

Bland, J. M. and Altman, D. G, 'Calculating correlation coefficients with repeated observations: Part 1, correlation within subjects.', *British Medical Journal*, **310**, 446, (1995a).

Bland, J. M. and Altman, D. G, 'Calculating correlation coefficients with repeated observations: Part 2, correlation between subjects.', *British Medical Journal*,**310**, 633, (1995b).

Bowling, A. *Measuring Health: A review of quality of life measurement scales*, Open University Press, (1983).

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

Carey, V., Zeger, S. L. and Diggle, P. 'Modelling multivariate binary data with alternating logistic regressions', *Biometrika*, **80**, 517-526, (1993).

Chinn, S. and Burney, P. G. J. 'On measuring repeatability of data from self administered questionnaires', *International Journal of Epidemiology*, **16**, 121-127, 1987).

Clayton, D. 'Repeated ordinal measurements: A generalised estimating equation approach', Technical report. Medical Research Council Biostatistics Unit, Cambridge, UK, (1992).

Cleveland, W. S. 'Robust locally weighted regression and smoothing scatter plots', *Journal of the American Statistical Association*, **74**, 829-836, (1979).

Coates, A., Gebski, V., Bishop, J. F., Jeal, P. N., Woods, R. L., Snyder, R., Tattersall, H. N., Byrne, M., Harvey, V., Gill,G., Simpson, J., Drummond, R., Browne, J., van Cooten, R. and Forbes, J. F. 'Improving the quality of life during chemotherapy for advanced breast cancer', *The New England Journal of Medicine*, **317**, 1490-1495, (1987).

Cole, B. F., Gelber, R. D. and Goldhirsch, A. 'Cox regression models for quality adjusted survival analysis', *Statistics in Medicine*, **12**, 975-987, (1993).

Cox, D. R. 'The analysis of multivariate binary data', *Applied Statistics*, **21**, 113-120, (1972).

Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Speigelhalter, D. J. and Jones D. R.

'Quality of life assessment: Can we keep it simple ?', *Journal of the Royal Statistical Society, Series A*, **155**, 353-393, (1992).

Crawford, S. L., Tennstedt, S. L. and McKinlay, J. B. ' A comparison of analytical methods for non-random missingness of outcome data', *Journal of Clinical Epidemiology*, **48**, 209-219, (1995).

Crowder, M. J. and Hand, D. J. *Analysis of Repeated Measures*, Chapman and Hall, (1993).

CRC CTC Study Protocol. Palliative treatment of symptomatic inoperable non small cell lung cancer.

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

Dawson, J. D. 'Stratification of summary statistic tests according to missing data patterns', *Statistics in Medicine*, **13**,1853-1863, (1994).

de Haes, J. C. J. M., Raatgever, J. M., van der Burg, M. E. L., Hamersma, E. and Neijt, J. P. 'Evaluation of the quality of life of patients with advanced ovarian cancer treated with combination chemotherapy', In *the Quality of life of Cancer patients*, Aaronsen, N. K. and Beckman, J. (eds), p215. Raven Press: New York, (1989).

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*, **62**, 1034-1038, (1990).

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

Devlen, J. 'Anxiety and depression in migraine', *Journal of the Royal Statistical Society*, **87**, 338-341, (1994).

Diggle, P. J. 'Testing for random dropouts in repeated measures data', *Biometrics*, **45**, 1255-1258, (1989).

Diggle, P. 'Discussion of paper by K-Y Liang, S. L. Zeger and B. Qaqish', *Journal of the Royal Statistical Society, Series B*, **45**, 28-29, (1992).

Diggle, P. and Kenward, M. G. 'Informative dropout in longitudinal data analysis', *Applied Statistics*, **43**, (1994).

Diggle, P. J., Liang, K-Y. and Zeger, S. L., *Analysis of Longitudinal Data*, Oxford Science Publications, (1994).

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, (1993).

Everitt, B. S. 'The analysis of repeated measures: a practical review with examples', *The Statistician*, **44**, 113-135, (1995).

Fallowfield, L. J., Baum, M. and Maguire, G. P. 'Effects of breast conservation on

psychological morbidity associated with diagnosis and treatment of early breast cancer', *British Medical Journal*, **293**, 1331-1334, (1986).

Fallowfield, L. J., Baum, M. and Maguire, G. P. 'Addressing the psychological needs of the conservatively treated breast cancer patient: discussion paper', *Journal of the Royal Society of Medicine*, **80**, 696-700, (1987).

Fallowfield, L. *Quality of Life*, Souvenir Press, (1990).

Fayers, P. M. and Jones, D. R. 'Measuring and analysing quality of life in cancer clinical trials: A review', *Statistics in Medicine*, **2**, 429-446, (1983).

Fayers, P. 'MRC quality of life studies using a daily diary card - practical lessons learned from cancer trials', *Quality of life research*, **4**, 343-352, (1995).

Fayers, P. 'Aspects of incomplete quality of life data in randomised trials: II missing items', Paper presented at the *'Missing quality of life research on cancer clinical trials: practical and methodological issues'*, (1996).

Feldstein, M. L. 'Quality of life adjusted survival for comparing cancer treatments', *Cancer*, **67**, 851-854, (1991).

Firth, D. 'Discussion of paper by K-Y Liang, S. L. Zeger and B. Qaqish', *Journal of the Royal Statistical Society, Series B*, **45**, 28-29, (1992).

Fitzmaurice, G. M. and Laird, N. M. ' A likelihood based method for analysing longitudinal binary responses', *Biometrika*, **80**, 141-151, (1993).

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

Fitzpatrick R., Fletcher, A. E., Gore, S. M., Jones, D. R., Speigelhalter, D. and Cox, D. R. 'Quality of life measures in health care: applications and issues of assessment', *British Medical Journal*, **305**, 1074-1077, (1992).

Fletcher, A. E., Gore, S. M., Jones, D. R., Fitzpatrick, R. and Cox, D. R. 'Quality of life measures in measures in health care: design, analysis and interpretation', *British Medical Journal*, **305**, 1145-1148, (1992).

Follmann, D. 'Modelling transitional and joint marginal distributions in repeated categorical data', *Statistics in Medicine*, **13**, 467-477, (1994).

Ford, I., Norrie, J. and Ahmadi, S. 'Model inconsistency illustrated by the Cox proportional hazards model', *Statistics in Medicine*, **14**, 735-746, (1995).

Frison, L. and Pocock, S. J. 'Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design', *Statistics in Medicine*, **11**, 1685-1704 (1992).

Gail, M. H., Wiehand, S. and Piantadosi, S. 'Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates', *Biometrika*, **71**, 431-444, (1984).

Garret, M., Fitzmaurice, G. M. and Laird, N. M. 'A likelihood based method for analysing longitudinal binary responses', *Biometrika*, **80**, 141-151 (1993).

Gelber, R. D. and Goldhirsch, A. 'A new endpoint for the assessment of adjuvant therapy in postmenopausal women with operable breast cancer', *Journal of Clinical Oncology*, **4**, 1772-1779 (1986).

Gelber, R. D., Gelman, R. S. and Goldhirsch, A. 'A quality of life orientated output for comparing therapies', *Biometrics*, **45**, 781-795, (1989).

Gelber, R. D., Goldhirsch, A. and Cavalli, F. 'Quality of life adjusted evaluation of adjuvant therapies for operable breast cancer', *Annals of Internal Medicine*, **114**, 621-628, (1991).

Gelber, R. D. and Goldhirsch, A. 'Discussion of the paper by Cox *et al.*', *Journal of the Royal Statistical Society, Series A*, **155**, 353-393, (1992).

Gelber, R. D., Lenderking, W. R., Cotton, D. J., Cole, B. J., Fischl, M. A., Goldhirsch, A. and Testa, M. A. 'Quality of life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection', *Annals of Internal Medicine*, **12**, 961-966 (1992).

Gentleman, R. C., Lawless, J. F., Lindsey, J. C. and Yan, P. 'Multi-state models for analysing incomplete disease history data with illustrations for HIV disease', *Statistics in Medicine*, **11**, 805-821, (1994).

Gilks, W. R., Clayton, D. G., Spiegelhalter, D, J, Best, N. G., Sharples, L. D. and Kirby, A. J. 'Modelling complexity: Application of Gibbs sampling in medicine', *Journal of the Royal Statistical Society, Series B*, **55**, 39-102, 1993).

Girling, D. J., Hopwood, P. and Ahmedzai, S. 'Assessing quality of life in palliative oncology', *Progress in Palliative care*, **2**,80-86, (1994).

Glasziou, P. P., Simes, R. J. and Gelber, R. D. 'Quality adjusted survival analysis', *Statistics in Medicine*, **9**, 1259-1276, (1990).

Glasziou, P. P. 'Quality adjusted survival analysis with repeated quality of life measurements', *Contributed paper read at the 16th meeting of the International Society for Clinical Biostatisticians*, (1995).

Glimelius, B., Hillner, B. E. and Smith, T. J. 'Efficacy and cost effectiveness of adjuvant chemotherapy in women with node negative breast cancer. a decision analysis model', *New England Journal of Medicine*, **324**, 160-168 (1991).

Glimelius, B., Hoffman, K., Wilhelm, R. N., Pahlman, L. and Sjoden, P-O, 'Quality of life during chemotherapy in patients with symptomatic advanced colorectal cancer', *Cancer*, **73**, 556-562, (1994).

Glynn, R. J., Laird, N. M. and Rubin, D. B. 'Multiple imputation in mixture models for non-ignorable non-response with follow-ups', *Journal of the American Statistical Association*, **88**, 984-993, (1993).

Godambe, V. P. 'An optimal property of regular maximum likelihood estimation', *Annals of*

Mathematical Statistics, **31**, 1208-1212, (1960).

Goldstein, H. 'Multilevel mixed linear model analysis using iteratively generalised least squares', *Biometrika*, **73**, 43-56, (1986).

Goldstein, H., *Multilevel models in educational and social research*, Griffen, (1987).

Goldstein, H. and McDonald, R. P. 'A general model for the analysis of multilevel data', *Psychometrika*, **53**, 455-467 (1988).

Goldstein, H. 'Restricted unbiased iterative generalised least squares estimation', *Biometrika*, **76**, 622-623 (1989).

Goldstein, H. 'Nonlinear multilevel models with an application to discrete response data', *Biometrika*, **78**, 45-51, (1991).

Goldstein, H. and Rasbash, J. 'Efficient computational procedures for the estimation of parameters in multilevel models based on iteratively generalised least squares', *Computational Statistics and Data Analysis*, **13**, 63-71, (1992).

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nutall, D. and Thomas, S. 'A multilevel analysis of school examination results', *Oxford Review of Eduction*, **19**, 425-433, (1993).

Goldstein, H., Healy, M. J. R. and Rasbash, J. 'Multilevel time series models with applications to repeated measures data', *Statistics in Medicine*, **13**, 1643-1655, (1994).

Goldstein, H., *Multilevel statistical models*, Second edition, Kendall's Library of statistics 3, (1995).

Goldstein, H. and Rasbash, J. 'Improved approximations for multilevel models with binary responses', *Journal of the Royal Statistical Society, Series A*, **159**, 505-513, (1996).

Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489-504, (1969).

Guyatt, G., Patrick, D. and Feeny, D. 'Postscript to the proceedings of the international conference on the measurement of quality of life as an outcome in clinical trials', *Controlled Clinical Trials*, **4S**, 266S-269S, (1991).

Hastie, T. J. and Tibshirani, R. J. *Generalized additive models*, Chapman and Hall, (1990).

Hedeker, D. and Gibbons, R. D. 'A random effects ordinal regression model for multilevel analysis', *Biometrics*, **50**, 933-944, (1994).

Hillner, B. E. and Smith, 'Efficacy and cost effectiveness of adjuvant chemotherapy in women with node negative breast cancer. A decision analysis model', *New England Journal of Medicine*, **324**, 160-168, (1991).

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 palliative

chemotherapy trial', *Quality of life research*, **3**, 339-352, (1994).

Huber, P. J. 'The behaviour of maximum likelihood estimates under non standard conditions', *Proceedings of fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 221-233, (1967).

Hurny, C., Bernard, J., Joss, R., Willems, Y., Cavalli, F., Kiser, J., Brunner, K., Favre. S., Alberto, P., Glaus, A., Senn, H., Schatzmann, E., Ganz, P. A. and Metzger, U. 'Feasibility of quality of life assessment in a randomised phase III trial of small cell lung cancer - a lesson from the real world', *Annals of Oncology*, **3**, 825-831, (1992).

Ibbotson, T., Maguire, P., Selby, P., Priestman, T. and Wallace, L. 'Screening for anxiety and depression in cancer patients: the effects of disease and treatment', *European Journal of Cancer*, **30A**, 37-40, (1994).

Johnson, R. A., *Applied multivariate statistical analysis*, Wichern, (1988).

Jones, R. H. 'Serial correlation of random subject effects ?', *Communications in Statistics - simulation and computation*, **19**, 1105-1123, (1990).

Kalbfleisch, J. D. and Prentice, R. L., *The Statistical analysis of failure time data*, John Wiley, (1980).

Karnofsky, F. A., Abelmann, W. H., Craver, L. F. and Burchenal, J. H. 'The use of nitrogen mustards in palliative treatment of carcinoma', *Cancer*, **20**, 634-656, (1949).

Kay, R. 'The analysis of transition times in multi-state stochastic processes using proportional hazards regression models', *Communications in statistics Part A - Theory and Methods*, **11**, 1743-1756, (1982).

Kay, R. 'A Markov model for analysing cancer markers and disease states in survival studies', *Biometrics*, **42**, 855-865, (1986).

Kenward, M. G. 'A methods for comparing profiles of repeated measurements', *Applied Statistics*, **36**, 296-308, (1987).

Kenward, M. G., Lesaffre, E. and Molenburghs, '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*, **50**, 945-953, (1994).

Kenward, M. G. and Welham, S. 'Use of splines in extending random coefficient models for the analysis of repeated measurements', *Personal communication*, (1996).

Korn, E. L. 'On estimating the distribution function for quality of life in cancer clinical trials', *Biometrika*, **80**, 535-542, (1993).

Kreft, I. G. G., de Leeuw, J. and van der Leeden, R. 'Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL', *The American Statistician*, **48**, 324-335, (1994).

Laird, N. M. and Ware, J. H. 'Random effects models for longitudinal data', *Biometrics*, **38**,

963-974, (1982).

Laird, N. M. 'Missing data in longitudinal studies', *Statistics in Medicine*, **7**, 305-315, (1988).

Landis, J. R., Miller, M. E., Davis, C. S., and Koch, G. G. 'Some general methods for the analysis of categorical data in longitudinal studies', *Statistics in Medicine*, **7**, 109-137, (1988).

Lee, Y. and Nelder, J. A. 'Hierarchical generalized linear models', *Journal of the Royal Statistical Society, Series B*, **58**, 000-000, (1996).

Lesaffre, E., Molenburghs, G. and Dewulf, L. 'Effect of dropoouts in a longitudinal study: an application of a repeated ordinal model', *Statistics in Medicine*, **15**, 1123-1141, (1996).

Liang, K-Y and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13-22, (1986).

Liang, K-Y, Zeger, S. L. and Qaqish, B. 'Multivariate regression analyses for categorical data', *Journal of the Royal Statistics Society. Series B*, **1**, 3-40, (1992).

Lindsey, J. K., Jones, B. and Ebbutt, A. F. 'Simple models for repeated ordinal responses, with an application to a seasonal rhinitis clinical trial', *Personal communication*, (1995).

Lipsitz S. R., Kim K. and Zhao, L. 'Analysis of repeated categorical data using generalized estimating equations', *Statistics in Medicine*, **13**, 1149-1163, (1994).

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. 'Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association', *Biometrika*, **78**, 153-160, (1991).

Little, R. J. A. and Rubin, D. B., *Statistical Analysis with Missing data*, John Wiley, (1987).

Little, R. J. A. 'Commentary on models for missing data', *Statistics in Medicine*, 347-355, (1988).

Little, R. J. A. 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association*, **88**, 125-134, (1993).

Little, R. J. A. ' A class of pattern mixture models for normal incomplete data, *Biometrika*, **81**, 471-483, (1994).

Little, R. J. A. 'Modelling the drop-out mechanism in repeated measures studies', *Journal of the American Statistical Association*, **90**, 1112-1121, (1995).

Longford, N. T. *VARCL - software for variance component analysis of data with hierarchical nested random effects (maximum likelihood)*, Educational Testing Service, Princeton, NJ. (1988).

Longford, N. T. *Random Coefficient Models*, Oxford Science Publications, (1995).

McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society, Series B*, **42**, 109-142, (1980).

McCullagh, P., and Nelder, J. A., *Generalised Linear Models*, Chapman and Hall, (1989).

Maguire, P. and Selby, P. 'Assessing quality of life in cancer patients', *British Journal of Cancer*, **60**, 437-440, (1989).

Map Workshop, 'Markov models', *Statistics in Medicine*, **12**, 2127-2130, (1993).

Mark, S. D., and Gail, M. H. 'A comparison of likelihood based and marginal estimating equation methods for analysing repeated ordered categorical responses with missing data: application to an intervention trial of vitamin prophylaxis for oesophageal dysplasia', *Statistics in Medicine*, **13**, 479-493, (1994).

Matthews, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P. 'Analysis of serial measurements in medical research', *British Medical Journal*, **300**, 230-235, (1990).

Matthews, J. N. S. 'A refinement to he analysis of serial data using summary statistics', *Statistics in Medicine*, **12**, 27-37, (1993).

Medical Research Council lung cancer working party 'Controlled trial 12 versus 6 courses of chemotherapy in the treatment of small cell lung cancer', *British Journal of Cancer*, **59**, 584-590, (1989).

Medical Research Council lung cancer working party 'Assessment of quality of life in small cell lung cancer using daily diary card developed by the MRC Lung Cancer Working Party', *British Journal of Cancer*, **64**, 229-306, (1991a).

Medical Research Council lung cancer working party 'Inoperable non-small cell lung cancer (NSCLC): a Medical Research Council randomised trial of palliative radiotherapy with two fractions or ten fractions', *British Journal of Cancer*, **63**, 265-270, (1991b).

Medical Research Council lung cancer working party 'A Medical Research Council randomised trial of palliative radiotherapy with two or a single fraction in patients with inoperable non-small cell lung cancer and poor performance status', *British Journal of Cancer*, **65**, 934-941, (1992).

Medical Research Council lung cancer working party 'A randomised trial of 3 or 6 courses of etoposide cyclosphophamide methotrexate and vincristine or 6 courses of etoposide and ifosfamide in small cell lung cancer (SCLC) II: survival and prognostic factors', *British Journal of Cancer*, **68**, 1150-1156, (1993a).

Medical Research Council lung cancer working party 'A randomised trial of 3 or 6 courses of etoposide cyclosphophamide methotrexate and vincristine or 6 courses of etoposide and ifosfamide in small cell lung cancer (SCLC) II: quality of life', *British Journal of Cancer*, **68**, 1157-1166, (1993b).

Molenburghs, G, Kenward, M. G. and Lessafre, E. 'The analysis of longitudinal ordinal data with informative dropout', *submitted*, (1994).

Morris, J. N., Suissa, S., Sherwood, S., Wright, S. M. and Greer, D. 'Last days: a study of the quality of life of terminally ill cancer patients', *Journal of Chronic Diseases*, **39**, 47-62, (1986).

Murray, G. D. and Findlay, J. G. 'Correcting for bias caused by dropouts in hypertension trials', *Statistics in Medicine*, **7**, 941-946, (1988).

Nayfield, S. G., Ganz, P. A., Moinpour, C. M., Cella, D. F. and Hailey, B. J. 'Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials', *Quality of life research*, **1**, 203-210, (1992).

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. 'A comparison of cluster-specific and population averaged approaches for analysing correlated binary data', *International Statistical Review*, **59**, 25-35, (1991).

O'Brien, P. C. 'Procedures for comparing samples with multiple endpoints', *Biometrics*, **40**, 1079-1087, (1984).

Olschewski, M and Schumacher, M. 'Statistical analysis of quality of life data in cancer clinical trials', *Statistics in Medicine*, **9**, 749-763, (1990).

Olschewski, M., Schulgen, G. and Schumacher, M. 'Discussion of the paper by Cox *et al.*', *Journal of the Royal Statistical Society, Series A*, **155**, 387, (1992).

Olschewski, M., Schulgen, G., Schumacher, M. and Altman, D. G. 'Quality of life assessment in clinical cancer research', *British Journal of Cancer*, **70**, 1-5, (1994).

Olweny, C. 'Quality of life in cancer care', *The Medical Journal of Australia*, **158**, 429-432, (1993).

Park, T. and Davis, C. S. 'A test of the missing data mechanism for repeated categorical data', *Biometrics*, **49**, 631-638, (1993).

Park, T., Lee, S. and Woolsen, R. F. 'A test of the missing data mechanism for repeated measures data', *Communications in Statistics. Theory and methods*, **22**, 2813-2829, (1993).

Patterson, H. D. and Thompson, R. 'Recovery of inter-block information when block sizes are unequal', *Biometrika*, **58**, 545-554, (1971).

Pocock, S. J., Geller, N. J. and Tsiatis, A. A. 'The analysis of multiple endpoints in clinical trials', *Biometrics*, **43**, 487-498, (1987).

Pocock, S. J. 'A perspective on the role of quality of life assessment in clinical trials', *Controlled Clinical Trials*, **12**, 257S-265S (1991).

Porcelli, P., Simona, Z., Centonze, S. and Sisto, G. 'Psychological distress and levels of disease activity in inflammatory bowel disease', *Italian Journal of Gastroenterology*, **26**, 111-115, (1994).

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

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

Rasbash, J. and Woodhouse, G. *MLn command reference version 1.0*, (1995).

Rideout, M. S. 'Testing for random dropouts in repeated measurement data', *Biometrics*, **47**, 1617-1621 (1991).

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

Robins, J. M. and Rotnitzky, A. 'Semi-parametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association*, **90**, 122-129, (1995).

Rodriguez, G. and Goldman, N. 'An assessment of estimation procedures for multilevel models with binary responses', *Journal of the Royal Statistical Society, Series A.*, **158**, 73-89, (1995).

Rotnitzky, A. and Wypij, ' A note on the bias of estimators with missing data', *Biometrics*, **50**, 1163-1170, (1994).

Royston, P. and Altman, D. G. 'Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling', *Applied Statistics*, **43**, 429-467, (1994).

Rubin, D. B. 'Inference and missing data', *Biometrika*, **63**, 581-599, (1976).

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

SAS Institute Inc. SAS Technical report, p.229, SAS/STAT Software changes and enhancements, Release 6.07, Cary NC, USA: SAS Institute 620pp. (1992).

Schumacher, M., Olschewski, M. and Schulgen, G. 'Assessment of quality of life in clinical trials', *Statistics in Medicine*, **10**, 1915-1930 (1991).

Schlucter, M. D. 'Methods for the analysis of informatively censored longitudinal data', *Statistics in Medicine*, **11**, 1861-1870, (1992).

Shih, W. J. 'On informative and random dropouts in longitudinal studies', Letter to the Editor, *Biometrics*, **48**, 970-971, (1992).

Shih, W. J., Quon, H. and Chang, M. N. 'Estimation of the mean when data contain non-ignorable missing values from a random effects model', *Statistics and Probability Letters*, **19**, 249-257, (1994).

Slevin, M. L. 'Quality of life: philosophical question or clinical reality?', *British Medical Journal*, **305**, 466-469, (1992).

Smith, D. M. and Diggle, P. J. 'OSWALD: Object oriented software for the analysis of longitudinal data in *S*', *Personal correspondence*, (1994).

Speigelhalter, D. J., Gore, S. S., Fitzpatrick, R., Fletcher, A. E., Jones, D. R. and Cox, D. R. 'Quality of life measures in health care: resource allocation', *British Medical Journal*, **305**, 1205-1209, (1992).

References

StataCorp. *Stata Statistical Software: Release 4.0*, College Station, TX: Stata Corporation, (1995).

Stephens, R. J., Stenning, S. P., Parmar, M. K. B. and Machin, D. 'Discussion of the paper by Cox *et al.*' *Journal of the Royal Statistical Society, Series A*, **155**, 387, (1992).

Stiratelli, R., Laird, N. and Ware, J. H. 'Random effects models for serial observations with binary response', *Biometrics*, **40**, 961-971, (1984).

Tandon, P. K. 'Applications of global statistics in analysing quality of life data', *Statistics in Medicine*, **9**, 819-827, (1990).

Thomas, A., Spiegelhalter, D. J. andGilks, W. R. 'BUGS: A program to perform Bayesian inference using Gibbs sampling', *Bernado, Bayesian Statistics 4*, Clarendon Press Oxford, 837-842, (1992).

Touloumi, G., 'Repeated measures of CD_4 counts over time and relation to mortality and disease progression in HIV clinical trials', *Unpublished PhD upgrading report*, London School of Hygiene and Tropical Medicine, (1996).

Trew, M. and Maguire, P. 'Further comparison of two instruments for measuring quality of life in cancer patients', *Quality of life*, Beckman, J. (ed.) p111 Proc Third Workshop of the EORTC study group on quality of life, Paris, (1982).

UK Hepatic Artery Pump Trial study protocol, *Cancer Research Campaign Clinical Trials Centre, Kings College School of Medicine and Dentistry, London.*

Wang-Clow, F., Lange, M., Laird, N. M. and Ware, J. H. 'A simulation study of estimators for rates of change in longitudinal studies with attrition', *Statistics in Medicine*, **14**, 283-297 (1995).

Ware, J. H., Lipsitz, S. and Speizer, F. E. 'Issues in the analysis of repeated categorical outcomes', *Statistics in Medicine*, **7**, 95-107, (1988).

Wei, L. J., and Stram, D. O. 'Analysing repeated measurements with possibility of missing observations by modelling marginal distributions', *Statistics in Medicine*, **7**, 139-148, (1988).

Weinstein, M. C. and Stason, W. B. 'Foundations of cost-effective analysis for health and medical practices', *New England Journal of Medicine*, **296**, 716-721 (1977).

Welch, B. L. 'On the comparison of several mean values: an alternative approach', *Biometrika*, **38**, 330-336, (1951).

Woodhouse, G. *ML3 Software for Three-level Analysis Users' Guide for V2*, (1991).

Wu, M. C. and Carroll, R. J. 'Estimation and comparison of rates of changes in the response of informative right censoring by modelling the censoring process', *Biometrics*, **44**, 175-188, (1988).

Wu, M. C. and Bailey, K. R. 'Analysing changes in the presence of informative right censoring caused by death and withdrawal', *Statistics in Medicine*, **7**, 337-346, (1988).

Wu, M. C. and Bailey, K. R. 'Estimation and comparison of changes in the presence of informative right censoring: conditional linear model', *Biometrics*, **45**, 939-955, (1989).

Zeger, S. L. 'Commentary', *Statistics in Medicine*, **7**, 161-168, (1988).

Zeger, S. L., Liang, K-Y and Albert, P. S. 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics*, **44**, 1049-1060, (1988).

Zeger, S. L. and Karim, M. R. 'Generalized linear models with random effects; a Gibbs sampling approach' *Journal of the American Statistical Association*, **86**, 79-86, (1991).

Zeger, S. L. and Liang, K-Y, 'An overview for methods of analysis of longitudinal data', *Statistics in Medicine*, **11**, 1825-1839, (1992).

Zigmond, A. S. and Snaith, R. P. 'The hospital anxiety and depression scale', *Acta Psychiatrica Scandinavica*, **67**, 361-370, (1983).

Zhao, L. P. and Prentice, R. L., Correlated binary regression using a quadratic exponential model', *Biometrika*, **77**, 642-648, (1990).

Zwinderman, A. H. 'The measurement of change in quality of life in clinical trials', *Statistics in Medicine*, **9**, 931-942, (1990).

Zwinderman, A. H. 'Statistical analysis of longitudinal quality of life data with missing measurements', *Quality of Life Research*, **1**, 219-224, (1992).

Appendices

List of Appendices

A1 Study Description

A1.1 CRC non-small cell lung cancer study (CRC NSCLC)

The Cancer Research Campaign, Clinical Trials Centre (CRC CTC) non small cell lung cancer trial (NSCLC) was designed to '*compare the results of palliative radiotherapy to the mediastinum in patients with previously untreated bronchial carcinoma using either an intensive short course (a 2-week split course, with the second week of treatment given only to well patients randomised to receive it) or a continuous 4-week treatment*'. The major outcomes of interest in the study were patient survival, measured in days from randomisation; local control, as shown by serial chest X-rays; symptomatic relief; and self assessed quality of life.

Quality of life was measured pre-treatment and then weekly for the 8 weeks following the start of treatment. The measuring instruments used were the Hospital Anxiety and Depression (HAD) scale and, with slight amendments to incorporate some disease specific items of interest, the Rotterdam Symptom Checklist (RSCL). Questionnaires were sent to patients through the post to be returned to the trials office using pre-paid envelopes.

82 patients were randomised into the study, 42 to receive the continuous 4 week treatment and 40 to receive the split course. At the one month assessment of patients on the split course, 15 were subsequently randomised to the one week course, 13 to the two week course. 12 patients were not re-randomised. Table A1.1 shows baseline patient characteristics in each arm of the study in terms of age, sex and Karnofsky performance measure (Karnofsky *et al.* 1949).

Table A1.1: Baseline patient characteristics for the CRC NSCLC study.

		Continuous course	Split course
Number of patients		42	40
Sex	n (%) male	33 (81)	27 (68)
Age (yrs)	mean (SD)	65.0 (7.11)	67.18 (7.29)
Weight (kgs)	mean (SD)	65.6 (9.99)	62.6 (11.3)
Karnofsky	median (IQR)	70 (10)	70 (10)

Survival data for the two treatment arms are given in table A1.2 and figure A1.2. These give no evidence of a survival difference between the two groups.

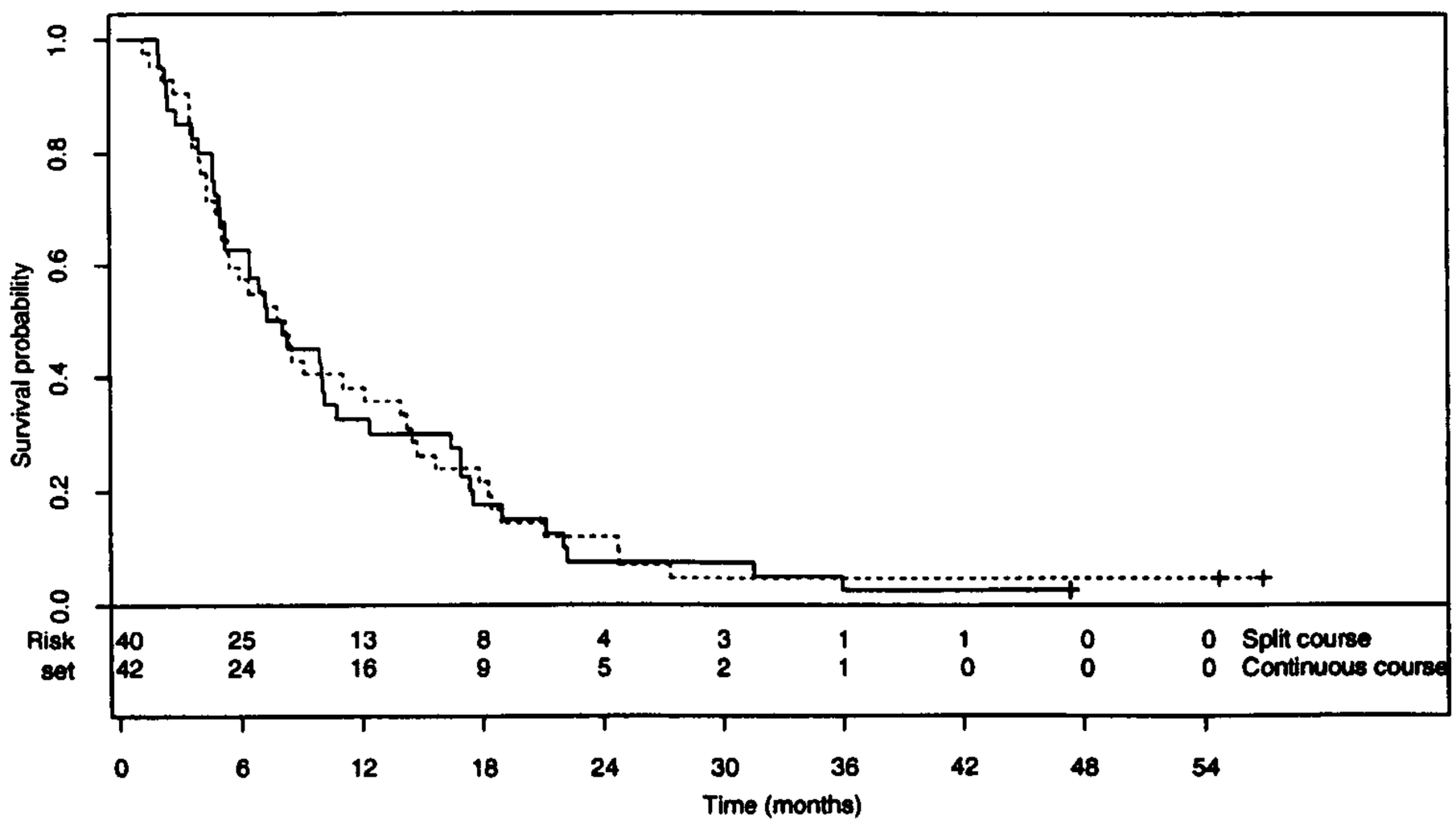

Figure A1.1: Kaplan-Meier survival curves for the CRC NSCLC study. Split course radiotherapy:————; continuous course radiotherapy ---------. Censored individuals are marked +.

Table A1.2: CRC NSCLC study: patient survival.

	Median survival (days)	Survival rate [95% CI]		Log rank test
		6 month	12 month	χ^2_1 (p value)
Split course	223	0.58 [0.44, 0.75]	0.30 [0.19, 0.48]	0.0 (0.922)
Continuous course	239	0.57 [0.44, 0.74]	0.36 [0.24, 0.54]	

A1.2 CRC Hepatic Artery Pump Trial (CRC HAP)

This was a randomised multi-centre trial based in four geographical areas: London, the Midlands, the South West and the North. The objectives of the study were to '*assess the quality of life and tumour response in patients with colorectal hepatic metastases treated by intra-hepatic arterial fluro-deoxyuridine (FUDR) infusion with that in patients with comparable metastases who received the conventional symptomatic treatment*' (HAI versus control).

The trial had three principal outcomes of interest: survival, measured in days from the time of randomisation into the trial until death; tumour response, as given by the percentage tumour involvement measured by a CT scan at 4 monthly intervals; and quality of life, measured by patient self assessment using the Sickness Impact Profile (SIP), the Rotterdam Symptom Checklist (RSCL) and the Hospital Anxiety and Depression (HAD) scale. These questionnaires were completed prior to randomisation and then monthly at the time of the regular monthly clinical follow-ups.

100 patients were recruited to the study, 51 to the HAI group and 49 to the control group. The results of the study, published elsewhere (Allen-Mersh *et al*, 1994) showed some evidence of a survival advantage for patients undergoing pump implantation. Since publication of these results, more complete data has become available. These data are shown in figure A1.2 and strengthened the evidence for a survival difference. For the purpose of the examples in the thesis, the complete data set was restricted to include only information available as of June 1st, 1993. This was done to reproduce the situation seen for the initial analysis of the quality of life data in the study where differential death rates in the two treatment groups meant that standard methods of analysis could not be used to analyse the quality of life data. At this time only 79 patients had been randomised into the study, 40 to the HAI group, 39 to the control. Survival data for this 'restricted' data are shown in figure A1.3. Survival rates for both the full and restricted data are given in table A1.3.

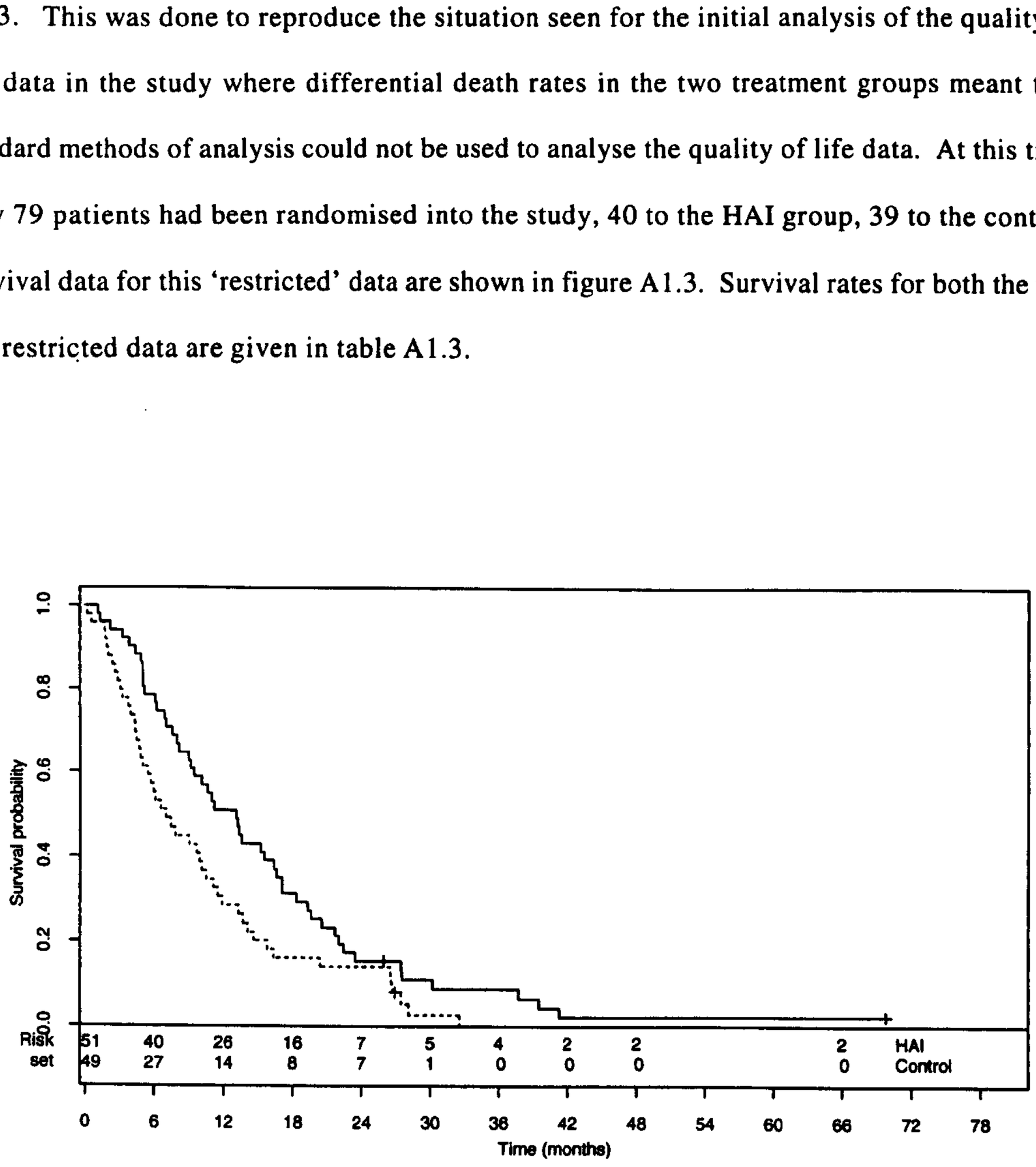

Figure A1.2: Kaplan-Meier survival curves for the full data set from the HAP trial. HAI:————; control: ---------. Censored individuals are marked +.

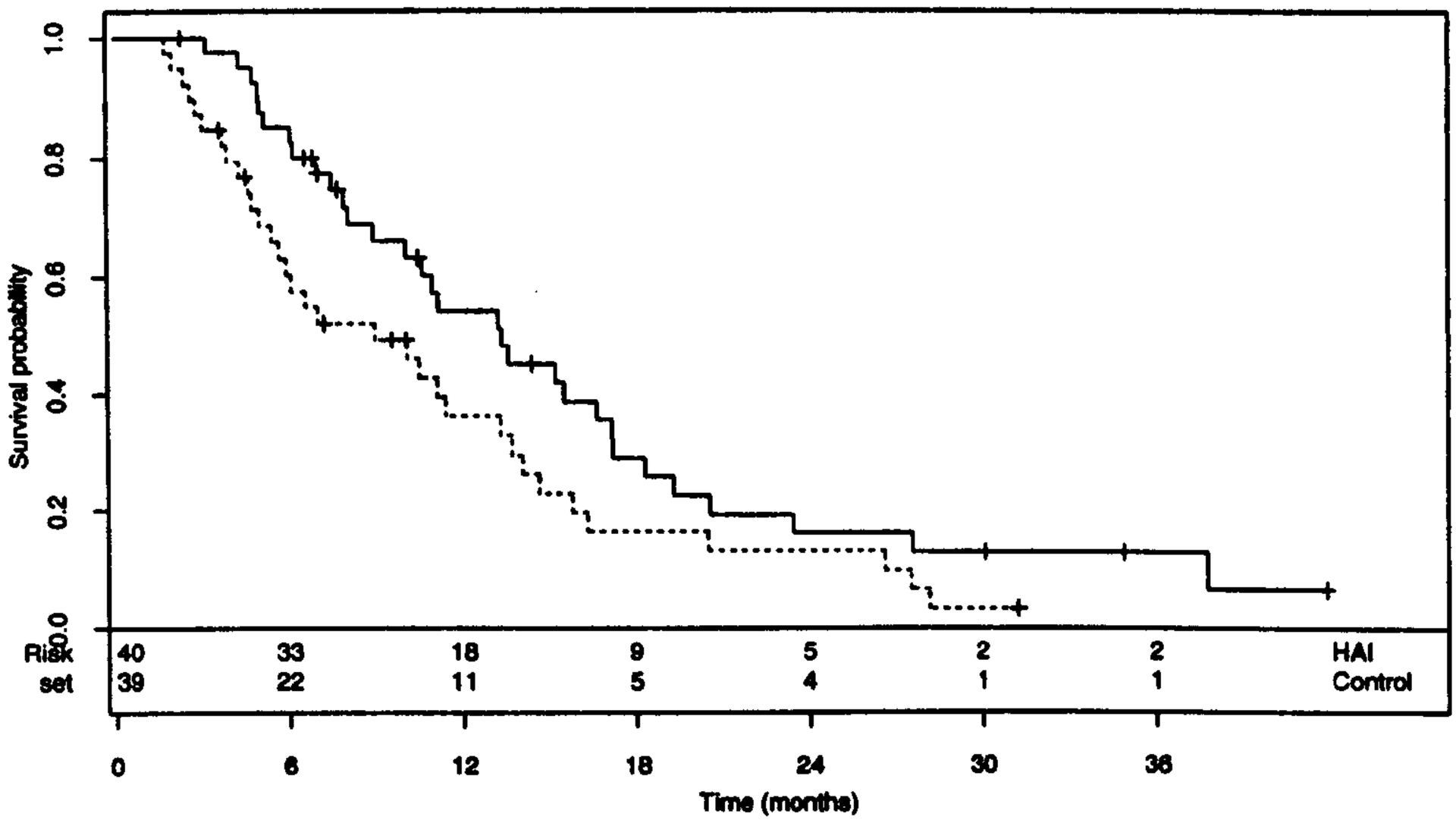

Figure A1.3: Kaplan-Meier survival curves for the restricted data set of the HAP trial. HAI:————; control: ---------. Censored individuals are marked +.

Table A1.3: Survival comparisons for the full and restricted data from the CRC HAP trial.

	Median survival (days)	Survival rate [95% CI]		Log rank test
		6 months	12 months	χ^2_1 (p value)
Full data				
HAI	334	0.83 [0.72, 0.95]	0.51 [0.37, 0.70]	4.0 (0.05)
Control	214	0.58 [0.44, 0.76]	0.33 [0.20, 0.53]	
Restricted data				
HAI	404	0.76 [0.66, 0.89]	0.49 [0.37, 0.65]	5.5 (0.02)
Control	274	0.53 [0.41, 0.69]	0.27 [0.17, 0.42]	

A1.3 MRC Lung Cancer working party (MRC LU07)

This was a randomised multi-centre trial comparing two policies of palliative thoracic radiotherapy. The aim of the trial was to assess whether a shorter dose of two fractions each of 8.5 Gy given one week apart was as effective as the conventional longer dose of a typically 30 Gy in total given in ten fractions over two weeks. Palliation and disease response were assessed monthly by clinicians for the first twelve months and every three months thereafter. Daily assessments of palliation were made by the patients using the daily diary card (Fayers and Jones, 1983).

374 patients were initially recruited into the trial. Following the exclusion of 5 patients later discovered to have been ineligible on entry, 369 patients remained with 184 receiving the shorter dose (F2) and 185 the longer dose (FM).

The results of the trial reported elsewhere (MRC lung cancer working party, 1991b) showed no evidence of a survival differences (table A1.4 and figure A1.4) between the two treatment groups. Performance status on entry was shown to be related to survival in both groups. Descriptive analyses of symptomatic quality of life data highlighted transient dysphagia

Table A1.4: Survival data for the MRC LU07 study.

	Median survival (days)	Survival rate [95% CI]		Log rank test
		6 months	12 months	χ^2_1 (p value)
FM	178	0.47 [0.40, 0.55]	0.21 [0.16, 0.28]	0.1 (0.74)
F2	178	0.46 [0.40, 0.54]	0.20 [0.15, 0.27]	

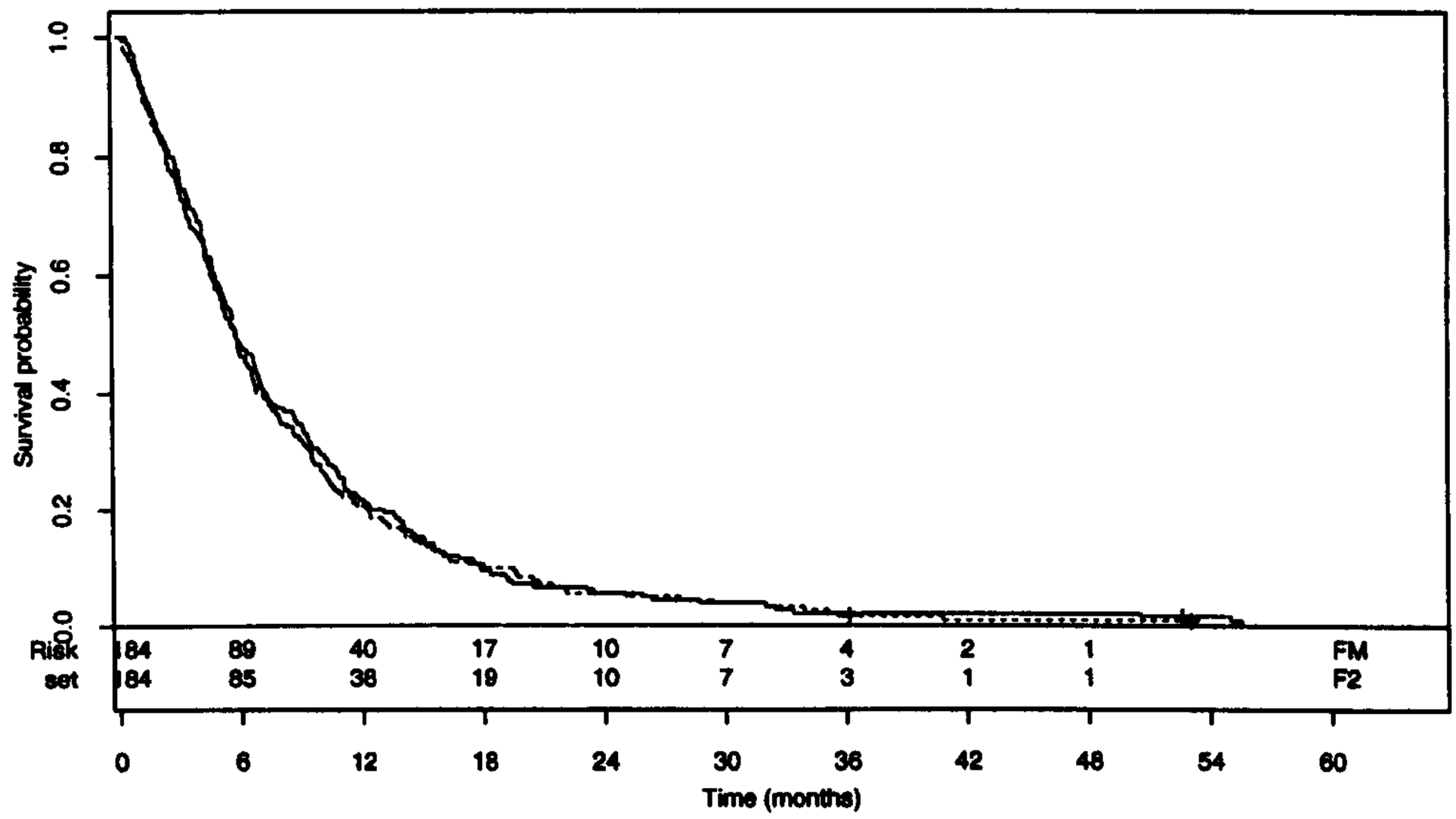

Figure A1.4: Kaplan-Meier survival curves for the MRC LU07 study. Multiple fraction radiotherapy (FM): ————; two fraction radiotherapy (F2:) ---------. Censored individuals are marked +.

following treatment in both groups. No evidence of a palliative gain of the conventional longer dose to the shorter dose was reported.

A2 Description of the Quality of Life Measuring Instruments

A2.1 Hospital Anxiety and Depression (HAD) scale

The Hospital Anxiety and Depression (HAD) scale was originally developed to aid the detection and management of emotional disorders in patients under investigation and treatment in medical and surgical hospital departments. It consists of 14 items measuring patient anxiety and depression. The items on the questionnaire are clearly divided into the two subscales and rated on a 4-point scale.

The items comprising the depression subscale are based largely on the *an hedonic* depressive state - that is, an inability to derive pleasure from day to day activities that the normal person finds pleasurable. This is because it is believed to be the central psychopathological feature of that form of depression that responds well to anti-depressant drug treatment, therefore providing useful information to the clinician. The items for anxiety were chosen following a study of the *Present State Examination (PSE)* (Fallowfield, 1990).

Box A2.1

Example items from the HAD scale

I feel tense or 'wound up'

(3) Most of the time
(2) A lot of the time
(1) From time to time
(0) Occasionally

I get sudden feeling of panic

(3) Very often indeed
(2) Quite often
(1) Not very often
(0) Not at all

Items are scored from 0-3 or 3-0 depending on the direction of the wording which is alternated in order to avoid responder bias. High scores indicate emotional problems. Based on psychiatric diagnosis as a gold standard, the scores can be categorized as: ≤7 - normal; 8-10 - possible case of clinical anxiety or depression; ≥11 - definite cases of clinical anxiety or depression.

It has been suggested that for use in cancer clinical trials, a cut-off of 10 or 11 can distinguish between patients who are coping well with their disease and those who have developed morbid anxiety or depression (Maguire and Selby, 1989).

Validity of the instrument was assessed with over 100 general medical outpatients and hospital staff. Sensitivity tests using PSE as the gold standard showed high correlation between the HAD scores and the interviewers' assessment. The scale scores were not shown to be affected by physical illness.

A2.2 Rotterdam symptom checklist (RSCL)

The Rotterdam Symptom Checklist (RSCL) was developed primarily as a tool to measure symptoms reported by cancer patients in clinical research (de Haes *et al.*, 1990). The original list of items included in the instrument was based on previous analyses done using alternative checklists not specifically designed for cancer patients. A selection of the items from three such lists was taken on the basis of factor loadings of principal components analyses, relevance according to oncology specialists, and the distribution of the response to the questions (very skewed responses were omitted). This gave an initial list of 34 items with an additional eight items referring to activities of daily living (ADL). On the basis of further research, the checklist was amended in terms of the recommended length of time between repeated

administrations and in the inclusion/omission of a number of items. The resulting instrument now has 30 items (plus 8 for ADL) which can be efficiently summarised into a psychological and physical dimension both of which show good reliability.

The items are scored on a four point scale: not at all; a little; quite a bit; very much. Examples are given in Box A2.2. Normal scores for the physical and psychological dimensions have been defined as <20 and <10 respectively.

Box A2.2

Example items from the RSCL

Lack of appetite		**Sore muscles**		**Nausea**	
(1)	not at all	(1)	not at all	(1)	not at all
(2)	a little	(2)	a little	(2)	a little
(3)	somewhat	(3)	somewhat	(3)	somewhat
(4)	very much	(4)	very much	(4)	very much

An advantage of the RSCL is its flexibility in allowing the inclusion of relevant disease specific items. For instance, in studies of ovarian cancer, the inclusion of items which refer to hair loss and sexual interest at the expense of items which were not relevant to the disease, gave great insight into the distinction of treatment regimens at no expense of reliability (Maguire and Selby, 1989).

A2.3 Daily diary card

The daily diary card was originally developed by the Medical Research Council, Tuberculosis and Chest Diseases Unit and has been used in several MRC cancer treatment studies. The cards are completed each day by the patient and require no assistance. They

consist of 5 items to which a patient gives a graded response 1-5. The items which are used on the cards may vary according to the nature of the study. For example, in a study of patients receiving chemotherapy treatment, an item referring to vomiting may be useful, whereas in a study using radiotherapy to treat lung cancer, an item asking about difficulties swallowing are more applicable. Extra information about patient quality of life can also be obtained in an open free format comments box for each week of the card's use. Example questions from the card are given in box A2.3.

Box A2.3

Example questions and responses for the daily diary card

Overall condition		**Activity**	
(1)	very well	(1)	normal work/housework
(2)	well	(2)	normal work but with effort
(3)	fair	(3)	reduced activity but not confined to home
(4)	poor	(4)	confined to home or hospital
(5)	very ill	(5)	confined to bed

The advantage of the cards are the speed with which they can be completed. This allows their use on a daily basis without putting an excessive burden on the patient. This in turn gives a clear picture of the patient quality of life over treatment and follow-up, and can highlight transient side effects which may be missed by less frequent follow-up. Unfortunately, as with many quality of life instruments, their compliance rate can be poor as patients tire of completing the card over too long a follow-up period.

In terms of validity and reliability, very little has been written about the cards, although they have been shown to display the expected features of symptomatic quality of life in the days following traumatic treatment.

A3 S-Plus Functions for Quality Adjusted Survival Analyses

A3.1 Introduction

The following series of S-Plus functions were written in order to perform the quality adjusted survival analyses for the RSCL physical data of the CRC HAP trial. In Section A3.2, the three main functions used to perform each analysis are broadly described in terms of their practical use, arguments and returned value. Listings of the function code are given in Section A3.4. Secondary functions which are called by the functions are described in the same way in the second half of each section.

Each of the main functions require the same basic arguments: a *patient identifier* (Id); *patient survival* (survival); a *censoring indicator* (status); *patient quality of life* (qol); and the *timing of each quality of life measurement* (qol.time). These arguments all have the same length and are defined with one row of data for each unique qol.time of a patient.

Table A3.1: Main and secondary functions used for the quality adjusted survival analyses in Chapter 6.

Analysis	Main function	Calls
TWiST (TNQOL)	TWiST.f	QAS.f
PQAS	PQAS.f	PQASTIMES.f, QAS.f
Integrated quality-survival product	glasziou.f	glas.f, AUC.f

A3.2 Function descriptions and their use

A3.2.1 TWiST.f

The function TWiST.f performs a TWiST based analysis (Gelber *et al.*, 1989) to estimate the time spent with normal quality of life (TNQOL) where normal quality of life is defined by a user specified cutoff for the quality of life dimension of interest. By default, the function performs an analysis which uses the censored quality of life times using a log-rank test to compare patient groups. Alternatively, when many subjects are censored, it is possible to truncate the follow-up time by some time L. Alternatively, TWiST can be imputed for censored individuals, as defined by Gelber *et al.* (1989), and patient groups compared using arithmetic means of the resulting uncensored data. If required, the function provides variance estimates for the means calculated as the area underneath the survival curves for censored or truncated analyses by resampling with replacement from the group data.

```
TWiST.f ( id , survival , status , qol , qol.time , group , cut.off , plotit=F , group.names=NULL
, method = 'censored' , L = max(survival) , what = 'max' , boot=F , R=NULL )
```

where

id, survival, status, qol, qol.time	are as defined in the introduction
group	defines the patient groups and is of the same length as id
cut.off	defines the cut off for normal quality of life
plotit	is an indicator as whether results should be plotted
group.names	patient group names for the plotted data
method	method of analysis required: 'censored', 'truncated', 'imputed'
L	maximum follow-up for truncated or imputed analyses
what	method of imputation: 'max', 'min', 'mean'
boot, R	indicator whether bootstrapped variances are required and number of samples

A3.2.2 PQAS.f

PQAS.f, directs the analysis for a partitioned quality adjusted survival analysis, (Glasziou *et al.*, 1990). Within the function, three health states are determined according to the definitions in table 6.4 and a userspecifiedd cut.off for normal quality of life. The function returns a summary of the time spent in each of the states for each group separately. If a vector of weights for each state are given, a weighted quality adjusted survival time is also returned.

```
PQAS.f ( id , survival , qol, qol.time , status , group=rep( 1,length(id) ) , cutoff=19.5 ,
weight=NULL , L=max(survival) , leg=NULL , boot=F , R=NULL , plotit=F , ... )
```

where

id, survival, qol, qol.time, status	are specified as defined in the introduction
group	defines the patient groups and is of the same length as id
cutoff	defines the cutoff for normal quality of life
weight	a vector of length 3 defining the weights for each health state
L	upper boundary for the length of follow-up
leg	a vector of the same length as the number of patient groups required for plotting the partitioned quality adjusted survival curves if plotit=T
boot, R	indicator whether bootstrapped variances of mean time in each state are required and the number of samples if required
plotit	indicator whether partitioned quality adjusted survival curves should be drawn

A3.2.3 glasziou.f

This function perform an integrated quality-survival product analysis Glasziou (1995). If bootstrapped variances are not required, it returns a list containing a matrix for each group which, in the notation of Section 6.3.3, correspond to t, $Q(t)$, $S(t)$ and $Q(t)*S(t)$. If bootstrapped variances are required, a matrix is returned which has estimates of the integrated quality-survival product in each group on the first row, with their corresponding standard errors as the second. Three different methods of evaluating $Q(t)$ are possible: linear; smooth; and mean. These correspond to the analyses described in Section 6.3.3.

```
glasziou.f ( id , survival , status , qol , qol.time , group=rep(1,length(survival)) , days=seq(
from=0 , to=max(survival) , by=20 ) , method='smooth' , boot=F , R=NULL )
```

where

id, survival, status, qol, qol.time	are as defined in the introduction
group	is a group indicator the same length as id
days	gives the days for which $Q(t)$ is to be evaluated. This needs only to be specified if method='mean'
method	method to be used to evaluate $Q(t)$: 'linear', 'smooth', 'mean'
boot, R	indicator whether a bootstrapped variances are required, and number required

A3.2.4 PQASTIMES.f

This function takes patient details of patient quality of life in terms of the actual value and timing, and calculates the amount of time each patient spends in each state along with a censoring indicator for each state according to whether patients have left that state, or are censored whilst remaining in the state. It returns a matrix with 5 columns containing a patient identifier, time of exiting each state for each subject and each state, the censoring status of each subject on exiting each state, a state identifier and a patient group indicator.

```
PQASTIMES.f ( id , survival , qol , qol.time , status , group=rep( 1,length(id) ) ,
L=max(survival) , cutoff=19.5 )
```

where

id, survival, qol, qol.time, status	specified as defined in the introduction
group	defines the patient groups and is of the same length as id
L	upper boundary for follow-up time
cutoff	defines the cutoff for normal quality of life

A3.2.5 QAS.f

This function calculates the area underneath a survival curve as given by survival and status which have one record for each patient. It returns a vector containing the area underneath the survival curve defined for each state as well as the overall survival curve.

```
QAS.f( survival , status, state=rep( 1 , length(survival) ), plotit=F , L=max(survival) ,
leg=NULL , ...)
```

where

survival	patient survival with one observation per patient per state
status	survival censoring indicator, one observation per patient per state
state	state indicator of the same length as survival
plotit	indicator to determine whether the survival curves are plotted
L	upper bound on follow-up
leg	legend required as a label for the required survival plot
...	any plotting arguments

A3.2.6 glas.f

This function calculates the quality-survival product. It uses a Kaplan-Meier estimate for $S(t)$ and estimates $Q(t)$ according to either a linear function through all the data, a lowess smooth of all the data, or subject specific means at specified days. It returns either a summary of the quality-survival product (a matrix containing t and $Q(t)*S(t)$) or full details (t, $Q(t)$, $S(t)$ and $Q(t)*S(t)$).

glas.f (sam , qt , times , survival , status , comment=F , summary=T , id=NULL , method)

where

sam	identifies which rows of qt are to be used
qt	gives the data to be used to evaluate $Q(t)$. If method='mean', this is a matrix where each row is the mean quality of life for each subject evaluated at each time point. Otherwise, it is simply the observed quality of life listed for each subject in a vector of the same form as qol described in the introduction
times	times at which $Q(t)$ is to be evaluated. It is only required if method='mean'
survival, status	survival time and censoring indicator for each subject. If method='mean', these have one observation per subject. If method='smooth' or 'linear', they are of the same form as survival and status described in the introduction
comment	an indicator whether comments as to the progress of the function should be printed
summary	an indicator whether a summary of the quality-survival product is required or full details
id	patient identifier. Required only if method='linear' or 'smooth'. It should be of the same form as id as defined in the introduction
method	method of analysis: 'linear', 'smooth' or 'mean' as described above

A3.2.7 AUC.f

Calculates the area underneath a curve defined by the two columns of a given matrix, y. It returns either the cumulative area underneath the curve at each value of the second column of y, or a summary giving the total area underneath the curve.

AUC.f (y, cummulative = F)

where

y	is a matrix with two columns. The first corresponds to a response *y*, the second a covariate *x*. It is ordered in ascending *x* and defines a curve under which the area is to be calculated
cummulative	an indicator whether the cumulative sum of the area is required, or just simply the overall area underneath the curve

A3.3 Function listings

A3.3.1 TWiST.f

```
TWiST.f <- function( id , survival , status , qol , qol.time , cut.off , group , plotit=F ,
group.names=NULL , method = 'censored' , L = max(survival) , what = 'max' , boot=F,
R=NULL ){

# Error checks of required information
    if ( boot==T & is.null(R) )
        stop( 'For a bootstrapped sample R is required' )

    if ( method=='imputed' & ( what !='max' | what !='min' | what !='mean' ) )
        stop( 'For imputed analysis specify what=max, min or mean' )

# Check a graphics device is active if plot is required
    if (plotit)
        par( cex=1 )

# Define variables
    aqol <- rep( 0 , length(id) )
    OTR <- numeric( length( unique(id) ) )
    OAQOL <- numeric( length( unique(id) ) )
    OTWiST <- numeric( length( unique(id) ) )
    delta <- numeric( length( unique(id) ) )
    tau <- numeric( length( unique(id) ) )
    newgroup <- numeric( length( unique(id) ) )

# Calculate TNQOL for each patient
    patient <- 1
    for ( i in unique(id) ){
        m <- length( qol[id==i] )

        if ( m > 1){
            if ( qol[id==i][1] > cut.off )
                aqol[id==i][1] <- qol.time[id==i][1] + ( ( qol.time[id==i][2] - qol.time[id==i][1] ) / 2
)

            if ( qol[id==i][m] > cut.off )
                aqol[id==i][m] <- ( ( qol.time[id==i][m] - qol.time[id==i][m-1] ) / 2 ) + (
survival[id==i][m] - qol.time[id==i][m] )

        if ( m > 2){
            for ( j in 2:(m-1) )
                if ( qol[id==i][j] > cut.off )
                    aqol[id==i][j] <- ( qol.time[id==i][j+1] - qol.time[id==i][j-1] ) / 2
        } # END m>2
    } # END m>1

else{
    if ( qol[id==i][1] > cut.off )
```

```
        aqol[id==i][1] <- survival[id==i][1]
      } # END else

if ( method == 'censored' ){
      OTR[patient] <- unique( survival[id==i] )
      delta[patient] <- unique( status[id==i] )
      OAQOL[patient] <- min( OTR[patient] , sum( aqol[id==i] ) )
      OTWiST[patient] <- OTR[patient] - OAQOL[patient]
      } # END if censored

    if ( method == 'truncated'  | method == 'imputed' ){
      OTR[patient] <- min( unique(survival[id==i]) , L)
      delta[patient] <- ifelse( OTR[patient] == L | unique( status[id==i] )==1, 1 , 0 )
      OAQOL[patient] <- min( OTR[patient] , sum( aqol[id==i][dsr[id==i]<L] ) , L )
      OTWiST[patient] <- OTR[patient] - OAQOL[patient]

      if ( method == 'imputed' & delta[patient] == 0 ){
        OTWiST[patient] <- switch( what , min= unique( survival[id==i] ) -
OAQOL[patient],
            max= L - unique( survival[id==i] ),
            mean= unique( survival[id==i] ) - OAQOL[patient] + (0.5*( L - unique(
survival[id==i] ))) )
        } # END imputation
      } # END if truncated or imputed

    newgroup[patient] <- unique( group[id==i] )

    patient <- patient + 1
    } # END for

  if ( method == 'imputed' ){
    test <- t.test( OTWiST[newgroup==0], OTWiST[newgroup==1] )
    statistic <- matrix(c( mean(OTWiST[newgroup==0]) , mean( OTWiST[newgroup==1]
), var( OTWiST[newgroup==0] ) , var( OTWiST[newgroup==1] ) ) , 2 , 2 , T , list( c( 'Mean' ,
'Variance' ) , c( 'Pump' , 'Control' )) )
    print( test )
    } # END imputed

  else {
    test <- surv.diff( OTWiST, delta , newgroup )
    print( test )

    statistic <- c( QAS.f( OTWiST[newgroup==0] , delta[newgroup==0] , plotit=F ,
L=max(OTR) ), QAS.f( OTWiST[newgroup==1] , delta[newgroup==1] , plotit=F ,
L=max(OTR) ) )

#  Bootstrapping variances if required
    if ( boot == T ){
      statistic <- rbind( statistic , rep(NA,2) )
```

```
dimnames(statistic) <- list( c('Mean','SEM') , c('Pump','Control') )

for ( grp in 0:1 ){
   cat('\nBootstrapping variances using ',R,' samples\n')
   samples <- NULL
   n <- length( OTWiST[newgroup==grp] )
   for ( r in 1:R ){
                  cat(r)
                  sam <- sample( x = 1:n , size = n, replace = T )
                  samples <- rbind( samples , QAS.f( OTWiST[newgroup==grp][sam] ,
delta[newgroup==grp][sam] , plotit=F , L=max(survival) ) )
                  cat('\t')
                  } # END resampling
               cat('\n')
               statistic[2,grp+1] <- sqrt( var( samples ) )
               } # END group
            } # END bootstrap


   # Plot the results if required
       if ( plotit ){
            surv.plot( OTWiST , delta , newgroup, lty = unique(group)+1 )
            title(xlab = 'TNQOL (days)' , ylab = "Survival probability" )
            } # END plotit
       } # END else

   if ( method == 'imputed' )
       method <- paste( method , what )

   invisible ( list ( id=unique(id) , method=method , OTR = OTR, OAQOL = OAQOL,
OTNQOL = OTNQOL , delta = delta , group= newgroup , statistic=statistic ))
   } # END function TWiST.f
```

A3.3.2 PQAS.f

```
PQAS.f <- function( id , survival , qol, qol.time , status , group=rep( 1,length(id) ) ,
cutoff=19.5 , weight=NULL , L=max(survival) , leg=NULL , boot=F , R=NULL , plotit=F , ...){

# Error consistency checks
    if ( boot==T & is.null(R) )
        stop('For bootstrapping variances R is required')

# Check the graphics device is active
    if ( plotit )
        par( cex=1 )

    cat('\n Obtaining the cummulative survival times for each state ... \n')

    PQAS <- PQASTIMES.f( id, survival , qol , qol.time , status , group=group , cutoff=cutoff
)
    results <- list()
    cat('\n Calculating the area under the curve ...\n')
    for ( g in 1:length(unique(group)) ){
        grp <- unique(sort(group))[g]
        pqas <- PQAS[ PQAS[,'group'] == grp , ]

# Calculating the mean time spent in each state
        results[[g]] <- QAS.f( survival=pqas[,'survival'] , status=pqas[,'status'] ,
state=pqas[,'state'] , plotit=plotit , L=L , leg=leg[g] , ... )

# Bootstrapping variance if required
        if ( boot ){
            cat('\nBootstrapping variances using ',R,' samples\n')
            samples <- NULL
            n <- length( unique(pqas[,'id']) )
            for (i in 1:R) {
                cat(i)
                sam <- sample(x = 1:n, size = n, replace = T)
                sam <- c( sam,sam+n,sam+(2*n) )
                samples <- rbind( samples , QAS.f( pqas[sam,'survival'] , pqas[sam,'status'] ,
pqas[sam,'state'] , plotit=F , L=L ) )
                cat('\t')
            } # END resampling

# Estimating the weighted QUALITY ADJUSTED SURVIVAL for given weights
            if (!is.null(weight)){
                temp <- matrix( results[[g]][,1:length(weight)] , length(weight) ,1 )
                temps <- samples[,1:length(weight)]
                weight <- matrix( weight,1,length(weight) )
                results[[g]] <- cbind( results[[g]] , 'QAS'=weight%*%temp , 'var(QAS)'=weight
%*% var(temps) %*% t(weight) )
            } # END applying weights
```

```
            results[[g]] <- list ( means = results[[g]] , var=var(samples) )
            }
            cat('\n\n')
        } # END for group g
    names(results) <- paste( 'group' , 1:length(unique(group)) , sep='')

    results
    } # END function PQAS.f
```

A3.3.3 glasziou.f

```
glasziou.f <- function( id , survival , status , qol , qol.time , group=rep(1,length(survival)) ,
days=seq( from=0 , to=max(survival) , by=20 ) , method='smooth' , boot=F , R=NULL  )

# Error consistency checks
# Check the method of estimating qol to be used
    if ( method != 'smooth' & method != 'mean' & method != 'linear' )
        stop( 'For Glasziou analysis method may be smooth, mean or linear')

# For a bootstrapped variance, R is the number of samples
    if ( boot & is.null(R) )
        stop( 'R is required for bootstrapping variances')


# Setting estimation parameters
# The number of people in total
    N <- length( unique( id ) )

# Make sure the vector of days starts at zero. The total number of days is M
    if ( min(days)!=0 )
        days <- c( 0,days )
    M <- length(days)

# Make sure group is > 0
    oldgroup <- group
    group <- group - min(group) + 1

    cat('\nPerforming an Integrated Quality-adjusted survival analysis ...\n')

# For the survival analysis need unique values for each subjects
    ID <- unique(id)
    SURVIVAL <- numeric(N)
    STATUS <- numeric(N)
    GROUP <- numeric(N)

    patient <- 1
    for ( i in unique(id) ){
        GROUP[patient] <- unique( group[id==i] )
        SURVIVAL[patient] <- unique( survival[id==i] )
        STATUS[patient] <- unique( status[id==i] )

        patient <- patient + 1
        }

# If the full follow-up is not used, need to censor individuals who died after last day
    STATUS <- ifelse( STATUS==1 & SURVIVAL>max(days) , 0 , STATUS)
# And also their survival
```

```
    SURVIVAL <- ifelse( SURVIVAL>max(days), max(days) , SURVIVAL )

# qt gives the mean qol for subjects evaluated at each day
    if ( method=='mean' ){
        qt <- matrix( NA , M , N , T, list(as.character(days) , paste('pat',ID,sep='') ))
        cat('\nEstimating qt using subject specific means over time... \n')

        for ( j in 1:M ){
            patient <- 1
            for ( i in ID ){

         # Only if the patient is still alive can their qol be calculated
              if ( SURVIVAL[ID==i] > days[j] ){

                lower <- qol.time[id==i][qol.time[id==i] == max(qol.time[id==i][qol.time[id==i]
<= days[j]  ])][1]
                upper <- qol.time[id==i][qol.time[id==i] == min(qol.time[id==i][qol.time[id==i]
>= days[j]  ])][1]

                qt[j , patient] <- mean( c(qol[id==i][qol.time[id==i]==lower] ,
                    qol[id==i][qol.time[id==i]==upper] ) , na.rm=T )

                } # END is the patient alive

            # Otherwise it is missing
                else
                 qt[j , patient] <- NA

                patient <- patient + 1
                } # END each patient
            } # END each follow-up time
        } # END if mean

    # QSt is the survival and qol product for the Glasziou analysis
    QSt <- list()

    # If a bootstrap is done a summary of the results are given by auc
    auc <- matrix(NA , 2 , length(unique(group)) , T ,
list(c('auc','SE'),paste('group',unique(group),sep='') ) )

    # Analysis done for each group in turn
    for ( g in sort( unique(group) ) ){
        cat('\nGroup',g,'..')
    # How many subjects in group g
        Ng <- sum( GROUP==g )

        if (method=='mean')
            QSt[[g]] <- glas.f( 1:Ng , qt[,GROUP==g] , days , SURVIVAL[GROUP==g],
STATUS[GROUP==g]
```

```
            , comment=T , summary=F , method=method )
        else{
            days <- dsr[group==g]
            QSt[[g]] <- glas.f( 1:Ng , cbind(qol.time,qol)[group==g,] , days ,
SURVIVAL[GROUP==g], STATUS[GROUP==g]
                    , comment=T , summary=F , id=id[group==g] , method=method )
            }

    # Calculate the auc for the data set
        auc[1,g] <- AUC.f(  QSt[[g]][,c('QSt','Time')] )

    # Bootstrapping a variance if required
        if ( boot ){
            cat('Bootstrapping a variance using',R,'samples ... \n')

        # Matrix of generated samples row by row
            samples <- matrix( NA , R , Ng )
            for ( i in 1:R )
               samples[i,] <- sample( 1:Ng , Ng , replace=T )

            if ( method =='mean' )
               bootstrapped.QSt <- apply( samples , 1 , glas.f , qt[,GROUP==g] , days ,
                                          SURVIVAL[GROUP==g], STATUS[GROUP==g] ,
method=method )
                else
                   bootstrapped.QSt <- apply( samples , 1 , glas.f ,
cbind(qol.time,qol)[group==g,] , days ,
                                          SURVIVAL[GROUP==g] , STATUS[GROUP==g] ,
id=id[group==g] , method=method )

                bootstrapped.QSt <-
array(bootstrapped.QSt,c(length(bootstrapped.QSt)/(R*2),2,R),T)

                cat('\n')

            # Calculating the area under the curve for each sample
                bootstrapped.auc <- apply( bootstrapped.QSt , 3 , AUC.f )
                auc[2,g] <- sqrt( var( bootstrapped.auc ) )
                } # END bootstrapping

        cat('\n')
        } # END group

    group <- oldgroup
    names(QSt) <- paste( 'group==',sort(unique(group)) )

    if ( !boot )
        return( QSt )
    else
```

```
        return( auc )

    } # END function glas.f
```

A3.3.4 PQASTIMES.f

```
PQASTIMES.f <- function( id , survival , qol , qol.time , status , group=rep( 1,length(id) ) ,
L=max(survival) , cutoff=19.5 ){

# Set up the variables to be created
n <- length(unique(id))
exit1.time <- numeric(n)
exit1.status <- numeric(n)
exit2.time <- numeric(n)
exit2.status <- numeric(n)
exit3.time <- numeric(n)
exit3.status <- numeric(n)
small.group <- numeric(n)

# IS QOL ABNORMAL
ab.qol <- qol > cutoff
patient <- 1
for (i in unique(id)){
    # START IN POOR
    if ( ab.qol[id==i][1] == T ){
        # STAY IN POOR
        if ( length( rle(ab.qol[id==i])[[1]] ) == 1 ){
            exit1.time[patient]   <- min( L, unique( survival[id==i] ) )
            exit1.status[patient] <- ifelse( exit1.time[patient]==L , 0 , unique( status[id==i]))
            exit2.time[patient]   <- min( L, unique( survival[id==i] ) )
            exit2.status[patient] <- ifelse( exit2.time[patient]==L , 0 , unique( status[id==i]))
            }

        # GO TO GOOD
        else {
            exit1.time[patient]   <- min( L , mean( c( qol.time[id==i][ 1 + rle( ab.qol[id==i]
)[[1]][1] ] , qol.time[id==i][ rle( ab.qol[id==i])[[1]][1] ]) ) )
            exit1.status[patient] <- ifelse( exit1.time[patient]==L , 0 , 1 )

            # STAY IN GOOD
            if ( length( rle(ab.qol[id==i])[[1]] ) == 2 ){
                exit2.time[patient]   <- min( L, unique( survival[id==i] ) )
                exit2.status[patient] <- ifelse( exit2.time[patient]==L , 0 ,unique(
status[id==i] ) )
                }
            # GO BACK TO POOR
            else {
                exit2.time[patient]   <- min( L, mean( c( qol.time[id==i][1 + sum( rle(
ab.qol[id==i])[[1]][1:2])] , qol.time[id==i][ sum( rle( ab.qol[id==i])[[1]][1:2])] )) )
                exit2.status[patient] <- ifelse( exit2.time[patient]==L , 0 , 1 )
                }
            } # END go to good
```

```
} # END start in poor

# START IN GOOD
else {
    exit1.time[patient] <- 0
    exit1.status[patient] <- 1

        # STAY IN GOOD
        if ( length( rle( ab.qol[id==i])[[1]] ) == 1 ){
           exit2.time[patient]   <- min( L , unique( survival[id==i] ) )
           exit2.status[patient] <- ifelse( exit2.time[patient]==L , 0 , unique( status[id==i] )
)
           } # END stay in good

        # GO TO POOR
        else {
           exit2.time[patient]   <- min( L, mean( c( qol.time[id==i][1 + sum( rle(
ab.qol[id==i])[[1]][1])] , qol.time[id==i][ sum( rle( ab.qol[id==i])[[1]][1])] ) )
           exit2.status[patient] <- ifelse(exit2.time[patient]==L , 0 , 1 )
           } # END go to poor
        } # END start in good
     exit3.time[patient]   <- min( L , unique( survival[id==i] ) )
     exit3.status[patient] <- ifelse( exit3.time[patient]==L , 0 , unique( status[id==i] ) )
     small.group[patient] <- unique(group[id==i])
     patient <- patient + 1
     } # END for each individual

  PQAS <- matrix( c( rep(unique(id),3 ) , c( exit1.time , exit2.time , exit3.time ) , c(
exit1.status , exit2.status , exit3.status ) , rep( 1:3 , rep( length(unique(id)),3) ) , rep(
small.group,3 ) ) , 3*length(unique(id)) , 5 , byrow=F , dimnames=list( NULL ,
c('id','survival','status','state','group' ))  )

# Returns the result which is the time spent in each qol state for each subject
  PQAS
  } # END PQASTIME.f
```

A3.3.5 QAS.f

```
QAS.f <- function( survival , status, state=rep( 1 , length(survival) ), plotit=F ,
L=max(survival) , leg=NULL , ...){

# Check the graphics device is active
    if (plotit)
        par( cex=1 )

# Set up the variables to be calculated
    nstate <- length( unique(state) )
    survival <- ifelse( survival>L , L , survival )
    if ( min(state)==0 )
        state <- state+1
    mean.surv <- matrix( 0 , 1 , nstate+1 , F , list( NULL , c( paste( 'state' , 1:nstate , sep='' )
,'survival' ))  )

# Calculate the area under the survival curve for each state
    for( i in 1:nstate ) {
        fit <- surv.fit( survival[state == i] , status[state == i] )
        area <- numeric( length(fit$time) )
        area[1] <- fit$time[1]
        for ( i in 2:length(fit$time) )
            area[i] <- ( fit$time[i] - fit$time[i-1] ) * fit$surv[i-1]

    # Stops missing being generated when no time is spent in a state
        mean.surv[,i] <- ifelse( !is.na( sum(area) ) , sum(area) , 0 )
        } # END estimation for each state


# Plot the estimated survival curves if requested
    if ( plotit ) {
        fit <- surv.fit( survival, status, state, )
        plot( fit, lty = rep(1, nstate) , xlim=c(0,L) , ... )
        abline( v = L, lty = 2)
        } # END plotting

# If only one state, return the area under the survival curve otherwise calculate the areas
within each state
if ( nstate == 1 )
    return(mean.surv[,1])

else {
    mean.surv[,nstate+1] <- mean.surv[,nstate]
    for ( i in nstate:2 )
        mean.surv[,i] <- mean.surv[,i] - mean.surv[,i-1]
    return(mean.surv)
    }
} # END function QAS.f
```

A3.3.6 glas.f

```
glas.f <- function( sam , qt , times , survival, status , comment=F , summary=T , id=NULL ,
method ){
    cat('.')
# Qt gives the mean survival of people alive at each of the specified time intervals given by
the mean qt for each day
    if ( comment )
        cat('\nEstimating Qt ... \n')

    if ( method == 'mean' )
        Qt <- apply( qt[,sam] , 1 , mean , na.rm=T )

    else{
        tempx <- split( qt[,1] , id )
        tempy <- split( qt[,2] , id )
        y <- numeric(0); x <- numeric(0)

        for ( i in 1:length(sam) ){
            y <- c( y , tempy[[ i ]] )
            x <- c( x , tempx[[ i ]] )
            }
        x <- x[!is.na(y)]; y <- y[!is.na(y)]
        y <- y[!is.na(x)]; x <- x[!is.na(x)]
        y <- y[order(x)]; x <- x[order(x)]

        if ( method == 'smooth' ){
            Qt <- lowess( x , y )
            times <- Qt[[1]]
            Qt <- Qt[[2]]
            } # END if smooth
        else {
            beta <- sum( (x-mean(x)) * (y-mean(y)) ) / sum( (x-mean(x))^2 )
            alpha <- mean(y) - (beta*mean(x))
            Qt <- alpha + (beta*x)
            times <- x
            } # END linear
        } # END method != 'mean'
    m <- length( times )

# St gives the probabilities for each group for each day
    if ( comment )
        cat('\nEstimating St ... \n')
    St <-  numeric( m )
    surv <-  surv.fit( survival[sam] , status[sam] )
    for (j in 1:m){
        the.prob <- surv$surv[ surv$time== max(surv$time[ surv$time<times[j] ] ) ]
        St[j] <- ifelse(length(the.prob)==1,the.prob,1 )
```

```
        } # END of survival estimation

# Return all the results
    if ( !summary )
        return( cbind( 'Time'=times , 'Qt'=Qt , 'St'=St , 'QSt'=Qt*St ) )
# Or just a summary
    return( cbind( 'Time'=times , 'QSt'=Qt*St ) )

} # END function glas.f
```

A3.3.7 AUC.f

```
AUC.f <- function( y, cummulative = F){

# y is a nx2 vector
# M is the number of time points
    M <- nrow(y)
    x <- y[,2]; y <- y[,1]

# area is a containing the area in each segment
    area <- numeric(M - 1)
    for( j in 1:(M - 1))
        area[j] <- (x[j + 1] - x[j]) * mean(c(y[j + 1], y[j]), na.rm = T)

# cummulative is an indicator whether the cummulative sum or sum is given
    if(cummulative == T)
        return(cumsum(c(0, area)))
    else
        return(sum(area, na.rm = T))
} # END function AUC.f
```