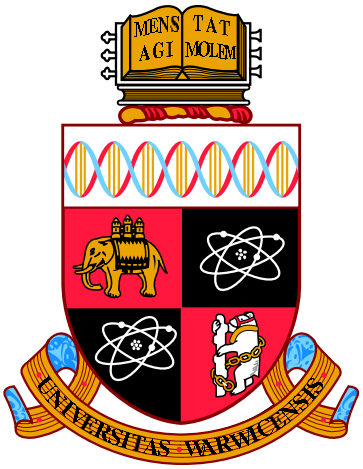


Analysis of Repeated Measurements with Missing Data

by

Mouna Akacha

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Statistics

Department of Statistics

10. March 2011

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgements

This thesis would not have been possible without all those who have supported me one way or another throughout the past three years.

I owe my deepest gratitude to my supervisor, Jane L. Hutton, for her encouragement, guidance and support throughout my PhD. Thank you for the fruitful collaboration, lively discussions and for being more than just a supervisor.

I am also extremely grateful to Norbert Benda without whose constant encouragement I would not have even started a PhD. Many thanks for your involvement and invaluable support.

Throughout the last three years I have also had the chance to discuss my work with various members of the Department of Statistics, the Statistical Methodology Group of Novartis and other institutions. Special thanks go to John Copas, David Firth, Ekkehard Glimm, Dan Jackson, Geert Molenberghs, Jianxin Pan, Heinz Schmidli, Shaun Seaman and Ian White. Furthermore, I would like to thank the CAST team for permitting the use of their data and Sallie Lamb from the Warwick Medical School for very helpful discussions regarding the data.

It is an honour for me to have been part of the Department of Statistics at the University of Warwick. The research environment was very stimulating and the department gave me ample opportunities to interact with researchers and to attend conferences. I am grateful to the Centre for Research in Statistical Methodology for providing financial support. Moreover, I would like to thank the International Biometric Society for providing travel support to attend two international conferences. I would like to extend my sincere

gratitude to the whole administrative staff of the department.

On a personal note I would like to thank all of my family and friends for their encouragement, understanding and extraordinary support. I would not have finished this thesis without you! My special appreciation goes to Giorgia Carta, Maria Costa, Thaís Fonseca, Guy Freeman, Flávio Gonçalves, Bryony Hill, Ben Jacoby, Krzysiek Łatuszyński, Christopher Nam, Murray Pollock, Walter Staiano, Gaidad Tekle, Siren Veflingstad, Peter Windridge and Piotr Zwiernik. I also would like to thank my friends Mareke, Markus, Matthias and Mirjam in Germany, who although far away were always by my side.

Lastly, and most importantly, I am eternally grateful for my parents and would like to dedicate this thesis to them and my siblings. Diese Doktorarbeit möchte ich meinen Eltern und Geschwistern widmen, die mich immer soweit wie moeglich unterstützt haben. Ohne Eure Erziehung und Eure Unterstützung wäre die Erstellung dieser Arbeit sicherlich nicht möglich gewesen. شكرا

Declaration

I hereby declare that this thesis is based on my own research, except when stated otherwise, in accordance with the regulations of the University of Warwick, and has not been submitted elsewhere.

The methods and results contained in Chapters 3 and 6 are joint work with my supervisor Professor Jane L. Hutton and have been accepted for publication in the journal *Statistics in Medicine* under the title: *'Modelling the rate of change in a longitudinal study with missing data, adjusting for contact attempts'*, see the corresponding CRiSM research report [Akacha and Hutton, 2010] for a draft.

The contents of Chapter 7 and 8 are incorporated in a published paper [Akacha and Benda, 2010] under the title *'The impact of dropouts on the analysis of dose-finding studies with recurrent event data'* and are joint work with Dr. Norbert Benda.

In both publications I have taken the leading role both in terms of preparation of the material and conceptual work. In addition, all results presented in Chapters 4 and 5 have been produced by me, except when otherwise indicated by references.

Apart from the work presented in this thesis, a third publication resulting from a competition during my PhD period and joint work with two fellow PhD students, T.C.O. Fonseca and S. Liverani, was accepted for publication under the title: *'First CLADAG Data Mining Prize: Data mining for longitudinal data with different marketing campaigns'*. This work will be available in the book *New Perspectives in Statistical Modeling and Data Analysis* published by Springer in 2011, see the corresponding CRiSM research report [Akacha et al., 2009] for a draft.

Abstract

This thesis discusses issues arising in the analysis of repeated measurement studies with missing data.

The first part of the thesis is motivated by a study where continuous and bounded longitudinal data form the outcome of interest. The aim of this study is to investigate the change over time in the outcome variable and factors that influence this change. The analysis is complicated because some patients withdraw from the study, leading to an incomplete data set.

We propose a non-linear mixed model that specifies the rate of change and the bounds of the outcome as a function of covariates. This mixed model has advantages over transforming the data and is easy to interpret. We discuss different models for the covariance structure of bounded continuous longitudinal data.

To explore the impact of missingness, we perform several sensitivity analyses. Further, we propose a model for informative missingness, taking into account the number and nature of reminders made to contact initial non-responders, and evaluate the impact of missingness on estimates of change. We contrast this model with the traditional selection model, where the missingness process is modelled.

Our investigations suggest that using the richer information of the reminder process enables a more accurate choice of covariates which induce missingness, than modelling the missingness process. Regarding the reminder process, we observe that phone calls are most effective.

The second part of this thesis is motivated by dose-finding studies, where the number of events per subject within a specified study period form the primary outcome. These studies aim to identify a target dose for which the new drug can be shown to be as effective as a competitor medication. Given a pain-related outcome, we expect many patients to drop out before the end of the study. The impact of missingness on the analysis and models for the missingness process must be carefully considered.

The recurrent events are modelled as over-dispersed Poisson process data, with dose as regressor. Additional covariates may be included. Constant and time-varying rate functions are examined. Based on a range of such models, the impact of missingness on the precision of the target dose estimation is evaluated by simulations. Five different analysis methods are assessed: a complete case analysis; two analyses using different single imputa-

tion techniques; a direct likelihood analysis; and an analysis using pattern-mixture models.

The target dose estimation is robust if the same missingness process holds for the target dose group and the active control group. This robustness is lost as soon as the missingness mechanisms for the active control and the target dose differ. Of the methods explored, the direct-likelihood approach performs best, even when a missing not at random mechanism holds.

Abbreviations

AIC	Akaike's Information Criterion
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
BKC	Below Knee Cast
CAST	Collaborative Ankle Support Trial
CC	Complete Case
CD	Complete Data
CI	Confidence Interval
DL	Direct Likelihood
EB	Empirical Bayes
Est.	Estimate
EMEA	European Medicines Agency
FAOS	Foot and Ankle Outcome Score
FAOSS	FAOS Symptoms-Score
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
iid	independent and identically distributed
ind.	independent

LCCF	Last Count Carried Forward
l'HR	l'Hôspital's rule
LOCF	Last Observation Carried Forward
LORCF	Last Observed Rate Carried Forward
MANOVA	Multivariate ANOVA
MAR	Missing At Random
MCAR	Missing Completely At Random
MED	Minimum Effective Dose
MI	Multiple Imputation
MNAR	Missing Not At Random
NLMM-1	Non-Linear Mixed Model (adjusting for gender and age)
NLMM-2	Non-Linear Mixed Model (adjusting for age)
PMM	Pattern-Mixture Model
p-val.	P-value
Q-Q plot	Quantile-Quantile plot
SD	Standard Deviation
SE	Standard Error

Notation

Unless otherwise stated, the following notation is repeatedly used throughout this thesis. In addition to their statement here, they are usually described at their first occurrence. Note that small-case versions of random variables denote realisations thereof.

$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_0^+$	set of non-negative real numbers
$\mathbb{R}^+$	set of positive real numbers
$\{\}$	empty set
N	sample size used in Part I
N^*	reduced sample size used in Chapter 6
m	sample size used in Part II
M_i	number of observations for subject $i \in \{1, ..., N\}$
M	number of observations per subject for balanced data
N_{tot}	number of total observations in Part I, i.e. $M_1 + ... + M_N$

Outcome of interest in Part I

$Y_{i,j}$	random variable, $i \in \{1, ..., N\}$, $j \in \{1, ..., M_i\}$
$\boldsymbol{Y}_i$	M_i-dimensional random vector $(Y_{i,1}, ..., Y_{i,M_i})^\top$
$\boldsymbol{Y}$	N_{tot}-dimensional random vector $(\boldsymbol{Y}_1^\top, ..., \boldsymbol{Y}_N^\top)^\top$
$\boldsymbol{Y}_{i,obs}$	random vector associated with the observed components of $\boldsymbol{y}_i$
$\boldsymbol{Y}_{obs}$	random vector of observed components for all subjects, i.e. $(\boldsymbol{Y}_{1,obs}^\top, ..., \boldsymbol{Y}_{N,obs}^\top)^\top$
$\boldsymbol{Y}_{i,mis}$	random vector associated with the missing components of $\boldsymbol{y}_i$
$\boldsymbol{Y}_{mis}$	random vector of missing components for all subjects, i.e. $(\boldsymbol{Y}_{1,mis}^\top, ..., \boldsymbol{Y}_{N,mis}^\top)^\top$

Missingness Process in Part I

$R_{i,j}$	indicator random variable modelling the missingness process, $i \in \{1, ..., N\}$
$\boldsymbol{R}_i$	M-dimensional random vector $(R_{i,1}, ..., R_{i,M})^\top$

$\boldsymbol{R}$	N_{tot}-dimensional random vector $(\boldsymbol{R}_1^\top, ..., \boldsymbol{R}_N^\top)^\top$

Reminder Process in Part I

K	number of reminder categories
$Z_{i,j}$	random variable modelling the reminder process, $i \in \{1, ..., N^*\}$
$\boldsymbol{Z}_i$	M-dimensional random vector $(Z_{i,1}, ..., Z_{i,M})^\top$
$\boldsymbol{Z}$	N_{tot}-dimensional random vector $(\boldsymbol{Z}_1^\top, ..., \boldsymbol{Z}_{N^*}^\top)^\top$
$\boldsymbol{V}_{i,j}$	indicator random vector associated with $Z_{i,j}$
V_i	indicator random matrix $(\boldsymbol{V}_{i,1}, ..., \boldsymbol{V}_{i,M})^\top$

Outcome of interest in Part II

T	length of the study period in Part II
$N_i(T)$	random variable in Part II, $i \in \{1, ..., m\}$
n_i	realisation of $N_i(T)$, denoting the number of events occurring by the end of the study
$T_{i,j}$	time of occurrence random variables, $i \in \{1, ..., m\}$, $j \in \{1, ..., n_i\}$
$t_{i,j}$	event times and realisation of $T_{i,j}$ with $0 < t_{i1} < ... < t_{in_i} \leq T$
$t_{i,d}$	dropout time of subject $i \in \{1, ..., m\}$
$n_i(t_{i,d})$	number of events of subject i by dropout
$T_{i,d}$	random variable associated with $t_{i,d}$
N_i	complete recurrent event data information for subject i, i.e. $\{N_i(T), T_{i1}, ..., T_{in_i}\}$
$N_{i,obs}$	corresponds to observed part of N_i, i.e. $\{N_i(t_{id}), T_{i1}, ..., T_{id}\}$
$N_{i,mis}$	missing part of N_i, i.e. $\{N_i(T), T_{id+1}, ..., T_{in_i}\}$

Other random variables

U_i	subject-specific random effect vector for subject i
ϵ_i	random vector for within-individual errors
P_i	random variable associated with the dropout pattern of subject i

Explanatory Variables

$\boldsymbol{x}_{i,j}$	vector of explanatory variables for subject i at occasion j
X_i	matrix of explanatory variables for subject i, i.e. $X_i = (\boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,N})^\top$
A_i	matrix of explanatory variables for subject i
$\boldsymbol{w}_{i,j}$	vector of explanatory variables for subject i at occasion j
W_i	matrix of explanatory variables for subject i, i.e. $W_i = (\boldsymbol{w}_{i,1}, ..., \boldsymbol{w}_{i,N})^\top$
$t_{i,j}$	observation times for subject i and occasion j
age_i	age of subject i
a_i	age centered around median age in Part I, i.e. $a_i = age_i - 27$
sex_i	$\in \{f, m\}$, gender of subject i, f female and m male
s_i	randomisation group for subject i
C	comparator drug in Part II

Parameters of interest

θ	parameter associated with non-linear mixed models (Part I) and models for the recurrent event data sequence (Part II)
ϑ	parameter associated with count data models (Part II)
ϕ	parameter associated with missingness process
ψ	parameter associated with reminder process
β	parameter associated with models for the mean
τ	parameter associated with covariance structures
ξ	parameter associated with covariance structures
σ^2	parameter associated with within-individual variances
D^2	parameter associated with variances of random effects
ζ	parameter associated with overdispersion
η	target dose of interest in Part II

Moments

$\mathbb{E}(\cdot)$	expectation
$Cov(\cdot)$	covariance
$Corr(\cdot)$	correlation
$\mu_i, \tilde{\mu}_i$	expectation vector for subject i
$\mu_{i,j}, \tilde{\mu}_{i,j}$	expectation for outcome of subject i at observation time j
Σ	$(M \times M)$-dimensional overall covariance matrix
Σ_i, $\Sigma_i(\tau)$	covariance matrix of observation vector for subject i
$\tilde{\Sigma}_i$, $\tilde{\Sigma}_i(\xi)$	covariance matrix accounting for within-individual heterogeneity
D, $D(\xi)$	covariance matrix of random effect vector U_i
Cov_{emp}	empirical covariance matrix
$Corr_{emp}$	empirical correlation matrix

Distributions

$Bernoulli(\cdot)$	Bernoulli distribution
$Exp(\cdot)$	Exponential distribution
$Gamma(\cdot,\cdot)$	Gamma distribution
$Multinomial(n, \boldsymbol{p})$	Multinomial distribution, where n is the number of trials and $\boldsymbol{p}$ is the vector of event-probabilities
$\mathcal{N}(\cdot,\cdot)$	Normal distribution
$\mathcal{N}_x(\cdot,\cdot)$	Multivariate Normal distribution, where x is the dimension of the random vector
$\mathcal{NB}(\cdot,\cdot)$	Negative-Binomial distribution
$Poisson(\cdot)$	Poisson distribution
$Uniform(\cdot)$	Uniform distribution (continuous)

Functions

$\exp(\cdot)$	exponential function
$\ln(\cdot)$	natural logarithm
$logit(\cdot)$	logit-function, i.e. inverse of logistic function
$\mathbb{P}(\cdot)$	probability mass function

$\mathcal{P}(\cdot)$	a parametric model
$f(\cdot)$, $f_Y(\cdot)$	density function of a random variable Y
$F(\cdot)$, $F_Y(\cdot)$	cumulative distribution function of a random variable Y
$L(\cdot)$, $L_Y(\cdot)$	likelihood function based on Y
$-2\,\ell$	deviance
$\Gamma(\cdot)$	gamma function
$\mathbb{1}(E)$	indicator function, where $\mathbb{1}(E) = 1$ if the event E is true and $\mathbb{1}(E) = 0$ otherwise
$g(\cdot)$	non-linear model function in Part I and (possibly) non-linear function quantifying dose-response relation in Part II
$d(\cdot)$	vector-valued function
$h(\cdot)$	positive-valued function
$\lambda_0(\cdot)$	baseline intensity function
$\lambda_x(\cdot)$	intensity function
$\Lambda_x(\cdot)$	cumulative intensity function

Matrices

I_M	$(M \times M)$-dimensional identity matrix
J_M	$(M \times M)$-dimensional matrix will all elements unity
$rank(\cdot)$	rank of a matrix
$A^\top$	transpose of a matrix A
A^{-1}	inverse of a matrix A

Chapter 1

Introduction

Studies where measurements are collected repeatedly over time have received much attention in recent years. This is mainly due to their ability to characterize the change in the response variable over time and factors that affect the change. They are particularly popular in biomedical and health-care applications, where investigators are interested in identifying the influence of treatment on disease development.

Statisticians involved in the analysis of repeated measurement studies are generally confronted with numerous challenges. For instance, the repeated measurements for each subject are usually correlated. Consequently, the independence assumption that is crucial for many statistical techniques is violated. A misspecification of the correlation structure can lead to incorrect inference for factors that affect the response. Therefore, special statistical techniques have to be adopted.

An additional, commonly observed complication in the analysis of repeated measurement studies arises through missing data due to missed visits, dropouts or non-return of questionnaires. In view of the fact that missingness usually occurs for reasons outside of the control of the investigators and may be related to the outcome measurement of interest, the data analysis is severely complicated. In general there are three potential problems that arise with missing data: loss of efficiency, complication in data handling and analysis, and bias due to differences between the observed and unobserved data [Horton and Lipsitz,

2001].

The present thesis discusses issues arising in the analysis of repeated measurement studies with missing data and is divided in two parts. The first part is devoted to studies that observe bounded, continuous repeated measurements, while the second part focuses on recurrent event data studies. In both parts, regression models and methods to handle missing data are established.

The first part of this thesis is motivated by medical research, where it is very common to measure physical or mental ability repeatedly over time through questionnaires or scales. Based on the answers, summary measures such as scores can be derived for all study points. In many applications, these scores will have finite range, where one bound indicates 'no symptoms' and the other bound 'extreme symptoms'. Examples are the *Barthel* index [Mahony and Barthel, 1965], the *Neck Disability Index* [Vernon and Mior, 1991], the *Foot and Ankle Outcome Score* [Roos et al., 2001] and visual analogue scales. In studies where we expect most patients to recover, we often observe that later measurements are clustered towards one end of the range. In this case, different patients might have the same initial and the same final scores. However, the rate at which they achieve the final score might differ substantially dependent on explanatory variables, e.g. treatment or age. The bounds themselves can also be of scientific interest, e.g. a maximum achievable score can differ substantially for different ages and genders.

For a continuous and bounded score, the classical approach is to transform the data so that a linear regression model fits adequately. For some scores, however, a non-linear dependence of the transformed outcome score on covariates persists due to the bounded nature of the score. In addition, models based on transformations cannot investigate the dependence of bounds on covariates as the bounds need to be specified prior to the transformation. Using transformations can also complicate the interpretation of covariate effects on the original score.

In view of these limitations, we present a valuable alternative to transforming data.

A non-linear mixed model for the mean score on the original scale as a function of covariates is proposed. The model is constructed for scores where the rate of recovery changes over time and has been motivated by the *Collaborative Ankle Support Trial* (CAST), which is the first large randomized controlled trial comparing four types of mechanical support for acute ankle sprains of sufficient severity to prevent weight bearing [Lamb et al., 2005, 2009; Cooke et al., 2009].

Apart from modelling the mean score, we discuss the challenge of modelling the covariance and correlation structure for bounded longitudinal data. With repeated measurements, we expect higher correlation when the measurements are closer in time than when they are further apart. Additionally, with bounded data, correlations increase as measurements reach the bounds regardless of the time interval between measurements. Finally, the variances are rarely constant over time. A data-driven regression approach introduced by Pourahmadi [1999] is adopted and extended to meet these characteristics. In particular, our extension allows for missing values.

Additionally, close attention is paid to the common issue of missing values in studies with repeated measurements. The statistical framework for the analysis of incomplete data introduced by Rubin [1976] is reviewed. Cases where the missingness process can be ignored are distinguished from cases where the missingness and outcome process need be modelled jointly. Popular methods to handle missing data for both cases are reviewed. Their respective merits and drawbacks are depicted. We discuss the necessity of a sensitivity analysis, where the stability of the conclusions is investigated under different assumptions regarding the reasons for missingness. Such a sensitivity analysis is performed for the CAST data set.

The first part of this thesis concludes by adopting the idea of including additional information about the reasons for missingness in the analysis of incomplete data sets [Alho, 1990; Wood et al., 2006; Jackson et al., 2010]. This information usually consists of proxy outcomes [Jackson et al., 2010], follow-up studies on a sample of non-responders [Cooke et al., 2009], collection of the reasons for dropout or extended retrieval efforts. Apart from

questionable model assumptions, all reported attempts to incorporate additional information are limited to cross-sectional studies. We establish a model for longitudinal data that relaxes some of the model assumptions and adjusts for missingness by taking into account the number and nature of reminders made to contact initial non-responders. This model is applied to the CAST data set and contrasted to more traditional approaches to handle missing data.

The second part of this thesis focuses on dose-finding studies that aim to identify the target dose for which a new drug can be shown to be as effective as a competing medication. The selection of a *valid* dose is crucial in clinical drug development. A dose which is too low will hinder the proof of efficacy and a dose which is too high could lead to a poor safety profile.

The dose-finding studies that motivated our work seek to analyze processes which generate events repeatedly over time. Such processes are referred to as recurrent event processes. Examples include seizures in epileptic studies, hot-flushes of postmenopausal women or flares in gout studies. Interest lies in understanding the underlying event occurrence process. This includes the investigation of the rate at which events occur, the inter-individual variation, and most importantly, the relationship between the event occurrence and explanatory variables such as treatment or dose.

Given a pain-related outcome and a repeated measurement study, a considerable number of patients is expected to drop out before the end of the study period. Due to the fact that dropping out may be related to the outcome of interest, the impact of missingness on the target dose selection needs to be carefully considered.

The interplay between recurrent event data modelling, dose selection and different causes for missingness poses the main focus of the second part of this thesis. In this context, the target dose is defined as the dose for which the expected response at the end of the study period is equal to that of the competitor group. Due to dropout the endpoint of interest may be missing. In many situations, however, information about the counting process prior to

the dropout is available, e.g. through patient diaries. This knowledge can be incorporated in a recurrent event data analysis.

The recurrent events are modelled as over-dispersed Poisson process data, with dose as regressor. Additional covariates may be included. Constant and time-varying rate functions are examined. Based on these models the impact of missingness on the precision of the target dose estimation is evaluated. Diverse reasons for dropping out are considered, including dependence on covariates and number of events. The performances of various analysis methods, including pattern-mixture models for recurrent event data, under different scenarios are assessed via simulations. To the best of our knowledge there is no review of several missing data methods for the analysis of recurrent event data or any work on missingness in the case of dose-finding studies.

1.1 Outline of Thesis

This thesis has nine chapters and is organized in two parts. The first part is motivated by the CAST study, which is presented in Chapter 2.

In Chapter 3, characteristics of longitudinal studies and analysis options for normally distributed data are reviewed. Different regression models for the CAST data set are discussed. The regression models considered are based on an exploratory analysis, which reveals that the outcome score evolves non-linearly over time. Also, the scores approach an upper limit as time increases. Different transformations are used in an attempt to reduce the effect of the boundedness and to improve the linearity of the data with respect to covariates. Based on these transformations, linear-mixed models are fitted to the data but lead to a poor fit. As an alternative, a non-linear mixed model accounting for the features of the CAST data set is established and fitted to the data.

Chapter 4 examines the covariance structure of bounded, continuous longitudinal data and adopts the data-driven regression approach introduced by Pourahmadi [1999]. This approach is extended by using the non-linear mixed model proposed in Chapter 3 for the

outcome process and by allowing for missing values. Different model selection tools are used and the results compared.

In Chapter 5, we review issues arising through missing data in repeated measurement studies. The framework of missing data analysis and commonly used methods to handle missing values are presented. Simple approaches such as complete case analysis and single imputation techniques are contrasted with more elaborate techniques, e.g. multiple imputation and pattern-mixture models. The chapter concludes by performing a sensitivity analysis for the CAST data set.

In Chapter 6, we propose a model to account for informative missingness, taking into account the number and nature of reminders made to contact initial non-responders. Using this model for the CAST study, the impact of missingness on the rate of change is evaluated in a sensitivity analysis. We contrast this model with a traditional model, where we adjust for missingness by modelling the missingness process.

In the second part of this thesis we focus on missingness in connection with dose-finding studies, where the number of events per subject within a specified study period form the primary outcome. In Chapter 7, we briefly review dose-finding studies, recurrent event data analysis and missing data issues. The study which motivated our work is presented and different methods to handle missing data in recurrent event data studies, including pattern-mixture models, are established. Regression models enabling a target dose selection are discussed.

In order to compare the performances of the proposed missing data methods a scenario evaluation study will be presented in Chapter 8.

We conclude this thesis with Chapter 9, where we present an overview of the main results and discuss areas for future research.

1.2 Computing Environment and Typeset

For the computational requirements of this thesis we use `SAS`, `Gauss`, `R` and `WinBUGS`. This thesis was typeset with LaTeX 2_ε using the editor TeXnicCenter.

Part I

CAST and Missing Data

Chapter 2

The CAST Study

2.1 Introduction

Acute ankle sprain is one of the most common soft tissue injuries seen in the UK emergency departments. According to Cooke et al. [2009], this type of injury accounts for between 3% and 5% of all emergency department attendances and approximately 5600 injuries each day. The injury is usually painful and incapacitating and makes weight bearing difficult to tolerate. A large variety of different treatments, e.g. immobilisation, physiotherapy or functional treatments, are available. By functional treatment we mean an early mobilisation that may include initial external support of the ankle via elasticated bandages or lace-up boots [Cooke et al., 2009]. Although many different treatments are available, researchers noted a lack of good-quality evidence to aid clinical decision making regarding an 'optimal' treatment.

The *Collaborative Ankle Support Trial* (CAST) is the first large randomized controlled trial comparing different types of mechanical support for acute ankle sprains of sufficient severity to prevent weight bearing [Lamb et al., 2005, 2009; Cooke et al., 2009]. The aim of this longitudinal study was to estimate the clinical and cost effectiveness of three different methods of mechanical support after severe ankle sprain compared to a standard treatment. These treatments include functional treatments but also one treatment that

immobilises the ankle for a certain time period.

We will present the CAST study in this chapter. In Section 2.2 we briefly discuss the study design. We then present the results of the original analysis in Section 2.3. Exploratory findings for CAST are shown in Section 2.4, and in Section 2.5 we discuss the missing data issue.

2.2 Study Design

The data for this trial were obtained from a randomised and multicentre study, which was run in 6 National Health Service trusts (8 hospitals) across the UK. Patients attending the selected emergency departments who had sustained a severe sprain of the lateral ligament complex of the ankle, were unable to weight bear, aged 16 and older, and gave informed consent were randomised into one of four treatment groups –*Tubigrip* (standard treatment), 10-day *below knee cast* (BKC), *Aircast brace* and *Bledsoe boot* [Lamb et al., 2005]. In this context, Tubigrip, Aircast brace and Bledsoe boot belong to the class of functional treatments, whereas the 10-day below knee cast immobilises the ankle during the course of the treatment.

The clinical status of these patients was measured at four points in time (baseline and follow-up at 4 weeks, 12 weeks and 39 weeks) via the *Foot and Ankle Outcome Score* (FAOS), which is a questionnaire containing 42 items and 5 subscales that ascertains functional limitations and the severity of other symptoms after ligament sprains [Roos et al., 2001]. The subscales are: pain subscale (7 items); other symptoms sub-scale (9 items); function in activity of daily living (ADL) sub-scale (17 items); function in sport and recreation sub-scale (5 items) and the foot and ankle-related quality of life (QoL) sub-scale (4 items).

Based on the answers, a continuous score, with 100 indicating no symptoms and 0 indicating extreme symptoms, was calculated for each sub-scale and point in time.

The total sample size was $N = 584$. Due to the fact that some patients did not

receive the FAOS questionnaire but another questionnaire called *Ankle Performance Scale* (APS), see Karlsson and Peterson [1991], during the baseline assessment and because of missing outcome measures for one person in the main study, the data of 559 persons instead of 584 persons will be investigated in this thesis. Moreover, this analysis will concentrate on the *symptoms* sub-scale score which will be referred to as *FAOSS-score* (FAOS-symptoms score). Note that the methodology derived in this thesis could be applied to any of the other four sub-scales, because they are very similar to the FAOSS-score from a qualitative point of view.

2.3 Original Analysis

According to the original analysis, reported in Lamb et al. [2005], the recovery was monitored at each of the time points separately. Linear regression models adjusting for gender, age and baseline scores were used to provide estimates of the recovery, with 95% confidence intervals. The explanatory variable *randomisation group* is used rather than the *treatment group*, because the analysis was performed on an intention-to-treat basis, i.e. all participants were analysed in the groups to which they were randomised, regardless of the treatment that they received. The distinction becomes important when investigators are confronted with non-compliance, which was the case in the CAST study. We note that there were marked differences in the compliance between the randomisation groups: BKC was least popular (16% non-compliance) compared with the other groups having less than 3% non-compliance. Generally, non-compliance may break the randomisation to the different treatment arms and create bias when analyzing treatment effects. An intention-to-treat analysis avoids the effects of non-compliance by investigating the effect of a treatment policy rather than the actual treatment effects.

For the FAOSS-score, the original analysis showed that the BKC offered a small but statistically significant benefit to the tubular bandage at 4 weeks. Neither the Aircast brace nor the Bledsoe boot conferred a siginifant advantage. At 12 weeks and by 39 weeks

there were no significant differences between the three comparator supports and the tubular bandage. The results for the other subscales are presented in Cooke et al. [2009].

From a statistical view point there are limitations to this analysis. For each patient, the outcome of interest was measured repeatedly over time. However, the collected data were analysed at every time point separately. Hence, the comparison of the different treatments was reduced to per time point conclusions and did not enable an overall statement regarding the rate of recovery. We will present alternative analysis techniques that model the four time points jointly, and that are able to account for the correlation between the four measurements of each subject, see Chapter 3. Furthermore, the issue of missing data was not addressed directly. Given the analysis method, all available data per time point were simply incorporated into the investigations for the same time point.

2.4 Exploratory Analysis

For initial exploratory re-analysis, the individual evolution of the FAOSS-score for a small subset of patients was plotted against time, see Figure 2.1. We connect the measurements for each subject to demonstrate the evolution over time.

From this plot we see the scores were usually an increasing function of time. Also, the scores increased much faster at the beginning of the study than towards the end.

The achieved scores at the study points and the rates at which these scores were achieved varied across the subjects. This variation appears to be smaller at the end of the study period than at the beginning. In fact, we observe that later measurements are clustered towards the upper end of the range. In particular, different patients might have the same initial and the same final scores, but the rate at which they achieve the final score can differ substantially, see black and pink lines. This observation suggests that the rate of recovery is more of interest than the actual scores at the beginning or the end of the study.

Although the individual evolutions show these differences, we note that in general the responses exhibited similarly shaped curves.

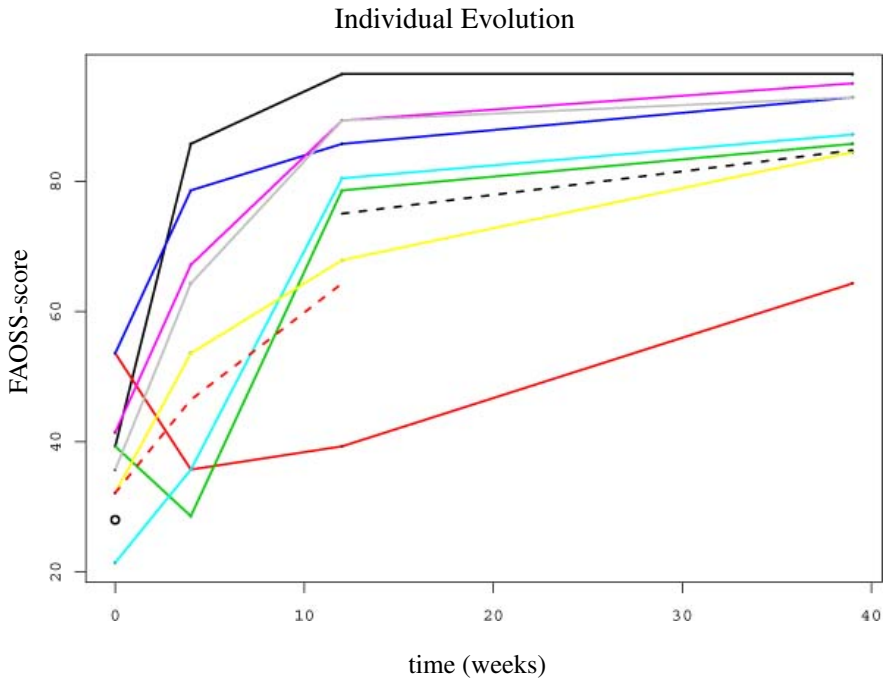


Figure 2.1: Individual evolution of the FAOSS-score for a random subset of 10 patients. The dashed lines correspond to patients with missing outcomes.

The final observation we want to make concerns the dashed lines which belong to patients with missing observations. The red dashed line refers to a patient who dropped out after 12 weeks. In contrast, the black line belongs to a patient who shows a non-monotone missingness pattern. This patient returned the questionnaire at baseline, 12 weeks and 39 weeks, but not at week 4. We will discuss the different missing data patterns in Section 2.5.

In order to show the aforementioned differences in the variation at the different points in time, we show side-by-side boxplots of the observations for each point in time and each randomisation group, see Figure 2.2. We note that due to the bounded nature of the score the distribution of the score data at the later time points is skewed. In particular, we observe that variations are not constant over time. For a detailed discussion of the variance structure we refer to Chapter 4.

These two plots, but also many studies that measure recovery from actue injury, show that the natural time course of recovery of ankle sprains is likely to stabilise within a

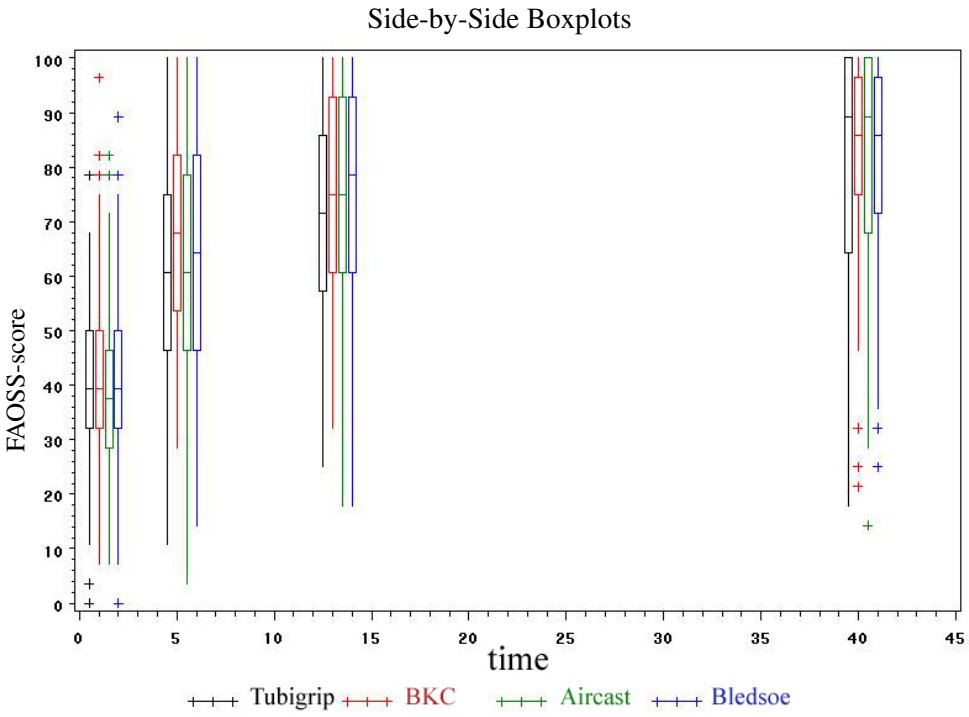


Figure 2.2: Boxplots of the FAOSS-score for different time points and randomisation groups

certain period (here: 3 to 9 months) and it is seen that the difference between the treatments will narrow in the longer term because the majority of people will recover, see Figure 2.2, Linde et al. [1986]; Schapp et al. [1989]; Avci and Sayli [1998] and Lamb et al. [2005].

An important aim of treatment is to accelerate the rate of recovery. Understanding the impact of explanatory covariates on the rate of recovery is important for guiding patients' and clinicians' expectations.

Our aim is to model the recovery rate and the bounds by modelling the responses at the four time points jointly. In this context, we aim to adjust for the explanatory variables of gender, age and randomisation group.

The randomisation groups were generally well matched in terms of gender and age. There was a slightly larger number of males in the BKC group. Overall there was a greater proportion of men (58%) than women (42%). The mean age of participants was 30 years (SD 10.8, median 27, range 16 – 72).

Time		Randomisation Groups			
		Tubigrip	BKC	Aircast	Bledsoe Boot
Baseline	Mean	40.3	41.8	38.8	41.1
	SD	14.1	16.4	15.1	16.9
4 weeks	Mean	60.7	67.4	62.8	61.6
	SD	19.5	19.0	20.5	20.7
12 weeks	Mean	70.0	76.0	73.8	75.1
	SD	20.5	18.4	20.6	20.4
39 weeks	Mean	80.4	82.8	81.0	81.2
	SD	20.4	17.0	20.3	19.0

Table 2.1: Summary Statistics (SD: standard deviation) for the FAOSS-score and the different randomisation groups and time points.

Summary statistics for the FAOSS-score and the different randomisation groups and time points are given in Table 2.1. For further details regarding the data set we refer to Cooke et al. [2009].

2.5 Missing Data and the Reminder Process

As with many clinical trials that collect longitudinal data, we are confronted with *missing data* in the CAST trial. By missing data we mean that intended measurements were not collected, see Chapter 5. Postal questionnaires were used in an attempt to minimise loss to follow-up, and a system of reminder letters and telephone calls was instituted to follow up those who did not return their questionnaire. We distinguish between the following 'reminder categories' $z \in \{0, 1, 2, 3, 4, 5\}$:

$z = 0$**:** questionnaire returned - no chasing;

$z = 1$**:** questionnaire returned after telephone chase;

$z = 2$**:** questionnaire returned after 2nd copy sent with no further telephone chasing;

$z = 3$**:** questionnaire returned after 2nd copy sent with further telephone chasing;

$z = 4$**:** core outcomes obtained over the telephone;

	attempts					
time point	0	1	2	3	4	5
baseline	553 (100%)	0	0	0	0	0
4 weeks	187 (33.8%)	152 (27.5%)	53 (9.6%)	40 (7.2%)	35 (6.3%)	86 (15.6%)
12 weeks	146 (26.4%)	141 (25.5%)	46 (8.3%)	48 (8.7%)	78 (14.1%)	94 (16.7%)
39 weeks	124 (22.4%)	117 (21.3%)	42 (7.6%)	59 (10.7%)	81 (14.7%)	130 (23.5%)
# Total quest. returned	1010	410	141	147	194	310

Table 2.2: Overview of the reminders needed to retrieve a questionnaire. In brackets the percentage of the returned questionnaires per attempt category is given for each time.

$z = 5$**:** non responder.

The frequency for each category and time point is shown in Table 2.2. As this information is only available for $N^* = 553$ out of the $N = 559$ patients, we show this information only for this subsample. Note that even though a questionnaire was received, the FAOS-scores could not be calculated in certain cases due to item-level missingness, e.g. sport subscales – asking for the physical functionality during physical exercises such as running – were completed less well than other subscales. Basic mean imputation was used to substitute likely values in order to be able to calculate a score for these questionnaires [Cooke et al., 2009]. In Chapter 5 we will discuss why the use of such *single imputation methods* is inadvisable.

In spite of the described procedure to gain data, the FAOS-score for all subscales and points in time was measured for only 51% of the randomised patients. The FAOSS-score was completely measured for 67% of the participants. An overview of the extent and the different patterns of missingness for the FAOSS-scale is given in Table 2.3. Note that approximately 10% of the participants show a *non-monotone missingness pattern*, i.e. missing values do not appear in a sequential order.

A first attempt to understand the underlying missing data mechanism is to explore the average scores for the different observation times and missingness patterns. Such a summary is given in Table 2.4. Apparently, participants with a low baseline score tend to not return the questionnaire at the 4 week time point. Furthermore, Table 2.4 suggests that participants with a high 4 week or a high 12 week score tend to not return the question-

Baseline	4 weeks	12weeks	9 months	Frequency	Percent
Completers					
O	O	O	O	376	67.26
Dropouts					
O	O	O	M	50	8.94
O	O	M	M	25	4.47
O	M	M	M	50	8.94
Non-monotone missingness					
O	O	M	O	14	2.50
O	M	O	O	21	3.76
O	M	O	M	12	2.15
O	M	M	O	11	1.97

Table 2.3: Overview of missingness patterns and the frequencies with which they occur in the FAOSS data (where: O, observed; M, missing).

Baseline	4 weeks	12weeks	9 months	Frequency	Percent
Completers					
40.87	62.49	73.46	81.17	376	67.26
Dropouts					
41.79	65.14	75.43	M	50	8.94
41.29	63.43	M	M	25	4.47
38.14	M	M	M	50	8.94
Non-monotone missingness					
40.56	73.72	M	86.48	14	2.50
39.63	M	72.79	80.27	21	3.76
30.95	M	72.27	M	12	2.15
40.58	M	M	84.42	11	1.97

Table 2.4: Overview of the means for different missingness patterns in the FAOSS data (where: M stands for 'missing').

naire for the following point in time. Hence, missingness seems to depend on the previous observed measurements. However, a study to investigate the response issues to the CAST questionnaire through interviews with 14 'responders' and 8 'non-responders', published in Nakash et al. [2008], comes to the conclusion that: 'Almost half of the participants who did not respond to follow-up considered themselves to have made a full recovery by the second time point (i.e. 12 weeks post injury). The term 'full recovery' is used to describe those participants who, on questioning, used terminology such as 'back to normal' or 'perfect' to describe their ankle. The effect of recovery on response is an important consideration in acute injury trials. Full recovery is likely to occur in the majority of participants before the end of the follow-up period. Participants may then feel that their further input is unnecessary and hence fail to respond to follow-up attempts. This appears to be the case with CAST participants.' Seemingly, missingness does also depend on non-observed future outcomes. A statistical analysis investigating these hypotheses will be presented in Chapter 5 and Chapter 6.

The issue of missing data and the implications for the analysis and inference will be discussed in more detail in Chapter 5.

2.6 Summary

In this chapter we introduced the motivating study, CAST, and discussed its characteristics. This included the longitudinal and bounded nature of the outcome of interest, but also the fact that a substantial part of the data is missing.

The presented plots show that the scores change faster at the beginning of the study than towards the end. This suggests that the comparison of the clinical effectiveness of the different treatments should be conducted in terms of the recovery rates. Furthermore, the plots show that the data at the later time points are clustered due to the bounded nature of the score.

We discussed the results of the original analysis and argued why this analysis is not

suitable for the analysis of longitudinal data. Extensions which account for the inter- and intra-individual variation, but also for the bounded nature of the score will be discussed in the next chapter.

Additionally, the original analysis does not adjust for missing data. However, the exploratory analysis (Section 2.5) suggests that missingness is outcome-related. Hence, the impact of missingness on the conclusions should be carefully explored, see Chapter 5.

Chapter 3

Longitudinal Data and CAST

3.1 Introduction

Longitudinal data are a special case of correlated data, which also encompass such structures as multivariate observations, clustered data, repeated measurements, time series data and spatially correlated data [Molenberghs and Verbeke, 2005].

We speak of longitudinal data, when for every subject the same characteristic is measured at a sequence of observation times. The collection of this type of data has become increasingly important in different areas of contemporary quantitative research. Lindsey [1999] lists more than 15 different scientific fields where longitudinal data arise, for example agriculture, biology, business, criminology, economics, geography, medicine, meteorology, politics and sociology. This work focuses mainly on medical applications for longitudinal studies. The interest in longitudinal data for biomedical and health-care applications arose with the need to understand the development and persistence of diseases over time [Fitzmaurice et al., 2004]. Investigators were particularly interested in identifying factors, such as treatment or age, that influence the disease development.

In a longitudinal study, every subject gives rise to a vector of measurements. The components of this vector have a temporal order. Thus, longitudinal data combine elements of multivariate and time series data [Diggle et al., 1996]. They are a special case of multi-

variate data, because the *same* characteristic is measured at different times, as opposed to measuring a number of *different* characteristics. The measurements of a single subject are ordered in time and the corresponding measurement vectors for all subjects can be seen as replicated (short) time series data [Diggle et al., 1996]. However, an important difference of longitudinal and time series data is that a time series usually arises from a study with a single replication and a large number of repeated measures, whereas a longitudinal study usually involves a large number of replications with a relatively small number of repeated measurements [Fitzmaurice et al., 2004].

3.1.1 Objectives and Advantages of Longitudinal Studies

The main objective of longitudinal studies is to characterize the change in the response variable over time and the factors that affect the change [Fitzmaurice et al., 2004]. Longitudinal studies are capable of achieving this goal by separating changes over time within individuals from differences among subjects in their baseline levels [Diggle et al., 1996; Hedeker and Gibbons, 2006]. One cannot distinguish these two sources of heterogeneity in cross-sectional studies, where a single outcome is measured at a defined time for each subject. In fact, Fitzmaurice et al. [2004] claim that: 'it is an inescapable fact that the assessment of within-subject changes in the response over time can be achieved only within a longitudinal study design'.

Longitudinal studies also tend to be statistically more powerful than cross-sectional studies with the same number of subjects, because each person acts as his or her own control [Diggle et al., 1996]. In most studies, measurements vary beyond what can be explained through available covariates. This variability can be due to unmeasured subject-specific factors, for example genetic, environmental, social or behavioral factors [Fitzmaurice et al., 2004]. In general, these factors can be assumed to be stable over the duration of the study. By measuring the response variable repeatedly over time for every subject, longitudinal studies are able to account for this source of variability. The change in the response variable can be estimated with greater precision [Diggle et al., 1996]. In contrast, cross-sectional

studies are not capable of accounting for unmeasurable subject-specific factors.

3.1.2 Challenges in the Analysis of Longitudinal data

The price to pay for the advantages of longitudinal studies are the numerous challenges in the analysis of longitudinal data. The repeated observations for each subject are usually correlated as the *same* characteristic is measured repeatedly over time. This correlation violates the independence assumption that is very important for many statistical techniques. According to Diggle et al. [1996], there are at least three consequences if existing correlation is ignored in longitudinal data analysis:

- inference for regression parameters is incorrect;
- estimates of the regression parameters are inefficient; and
- insufficient protection against bias caused by missing data.

For instance, Fitzmaurice et al. [2004] present a simple example, where ignoring the correlation structure leads to a substantial overestimation of the variability of the estimate of change. This leads to standard errors and p-values for the test of no change over time that are too large.

In general, there are three likely sources of random variation that cause the correlation among repeated measurements on the same subject: between-subject heterogeneity, serial correlation and measurement error [Diggle et al., 1996; Fitzmaurice et al., 2004; Molenberghs and Verbeke, 2005]. The between-subject heterogeneity reflects the natural variation in individuals' response. For example, some subjects can be intrinsically 'high responders' and others are 'low responders'. As mentioned above, this variation can be caused by unmeasured factors such as genetic predisposition.

The serial correlation is caused by *inherent within-individual biological variability* [Fitzmaurice et al., 2004]. Many biomedical or health-related outcomes show random variation within the repeated measurements of one subject which cannot be explained by covariates or circadian rhythms [Fitzmaurice et al., 2004]. Indeed, these outcomes can be

seen as realisations of time-varying stochastic processes. The induced variation results in correlation among the repeated measurements of one subject, where the correlation depends on the time lags between measurements. In general, correlations become weaker as the time lags increase [Diggle et al., 1996].

The final source of variability is measurement error, a component that is closely related to the reliability of the outcome measure and a common problem in many studies. Generally, measurement error induces a shrinkage of the correlation among the repeated measurements [Fitzmaurice et al., 2004].

Measurement error and serial correlation are elements that describe the within-individual heterogeneity. Fitzmaurice et al. [2004] note that many longitudinal studies do not have sufficient data to estimate these two sources of random variation separately. Many longitudinal studies may therefore only distinguish between-individual heterogeneity from within-individual heterogeneity. For a more detailed discussion on the different sources of variation we refer to Diggle et al. [1996] and Fitzmaurice et al. [2004].

As mentioned earlier, the correlation among repeated measures invalidates the crucial independence assumption for many statistical techniques and thus special statistical methods are required. The added correlation could be seen as a negative feature; however, it is this characteristic that makes longitudinal studies more powerful than cross-sectional studies [Fitzmaurice et al., 2004].

An added, commonly observed complication in the analysis of longitudinal studies arises through *unbalanced* data. We speak of unbalanced data when the observation times are not common to all subjects. We distinguish unbalanced data due to study design from unbalanced data due to missing data. Examples of the former case are studies where measurements for individual subjects are collected relative to a benchmark or an occurring subject-specific event; and studies which follow a *rotating panel* study design [Fitzmaurice et al., 2004]. Some statistical techniques for the analysis of longitudinal data can easily account for unbalanced data due to study design. In contrast, unbalanced data due to missing data may pose the most dramatic difficulty [Gibbons et al., 2010]. In order to stress the

fact that an intended observations on a subject could not be observed, these data are often referred to as being *incomplete* [Fitzmaurice et al., 2004]. The presence of missing data can lead to a far more complex analysis, see Chapter 5. The advantage of longitudinal data to cross-sectional data, however, is that all available data from each subject can be used in the analysis, leading to increased statistical power, see Hedeker and Gibbons [2006] and Chapter 5.

A further analytical issue which frequently arises in observational longitudinal studies is caused by the collection of time-varying covariates [Diggle et al., 1996]. Time-varying covariates can play an important role in the investigation of causal relationships and come with a complication of the data analysis.

3.1.3 Simple Approaches for the Analysis of Longitudinal Data

All the sources of complexity in the analysis of longitudinal data demand special statistical tools. Over the last years several analysis techniques have been discussed in the statistical literature and many monographs offering detailed accounts of the models and their applications have been published [Diggle et al., 1996; Lindsey, 1999; Fitzmaurice et al., 2004; Molenberghs and Verbeke, 2005; Hedeker and Gibbons, 2006]. In the following, we provide a short overview of the historical approaches used in the analysis of longitudinal data.

One relatively simple but sometimes effective approach is the *derived variables analysis* [Diggle et al., 1996; Fitzmaurice et al., 2004; Hedeker and Gibbons, 2006]. For every subject the repeated measurements are reduced into summary measures; e.g. average across time, linear trend across time, last observation carried forward, change scores or computing the area under the curve [Hedeker and Gibbons, 2006]. Subsequently, traditional statistical techniques for the analysis of cross-sectional data, e.g. ANOVA, can be used to compare group means or to analyze covariate effects [Diggle et al., 1996]. In the case of two repeated observations per subject, change scores can be computed and ANCOVA can be used to investigate covariate effects [Hedeker and Gibbons, 2006]. In general, how-

ever, derived variable analysis is not able to investigate within-individual changes, and by construction this method can not handle time-varying covariates. Furthermore, unbalanced data lead to heteroscedasticity as the amount of information per subject varies. Fitzmaurice et al. [2004] states that 'thus, a simple univariate analysis cannot proceed without proper consideration of the covariance, the very feature of the data that these methods were developed to avoid having to specify.' The calculation of some derived variables, e.g change scores, is not even defined for incomplete data sets. Hence, this approach is generally too restrictive and does not address the primary goal of longitudinal data analysis, namely the characterisation of the change in the response variable over time and the factors that affect the change.

Two classical approaches and, according to Fitzmaurice et al. [2004], two of the earliest proposals for the analysis of Gaussian longitudinal data are the *repeated measures analysis of variance* (ANOVA), sometime referred to as the *univariate* or *mixed-model* ANOVA, and the *multivariate repeated measures analysis of variance* (MANOVA). The repeated measures ANOVA accounts for the correlation among repeated measurements of one subject by the inclusion of a subject-specific random intercept. This random intercept reflects the natural variation in individuals' response and the correlation arises because all repeated observation on one subject share the same random intercept. The inclusion of a random intercept only, instead of for example a random slope, leads to a *compound symmetry* covariance structure, which implies constant variances and covariances over time. These assumptions can be too restrictive in practice and do not account for (possibly existing) serial correlation. Moreover, this analysis can not handle continuous covariates. For example, time is used as a classifying variable; thus, repeated measures ANOVA can not readily handle unbalanced or incomplete data, although extensions to the unbalanced case exist [Hedeker and Gibbons, 2006].

The MANOVA approach is an extension of the classical ANOVA to handle multivariate response vectors [Fitzmaurice et al., 2004]. As the repeated measurements for each subject form a response vector, this approach can be used for the analysis of longitudinal

data. In comparison to the repeated measures ANOVA, this technique enables the specification of an unstructured and therefore more flexible covariance structure [Gibbons et al., 2010]. The main disadvantage of this approach, however, is that (by construction) it can only be used on complete data. Furthermore, MANOVA also requires balanced data and can not handle continuous covariates [Fitzmaurice et al., 2004].

3.1.4 Regression Models for Gaussian Data

The disadvantages of the simple approaches presented emphasize that in order for a model to be of real practical use for the analysis of longitudinal data, it needs to satisfy certain conditions. Firstly, it should be able to deal with unbalanced and incomplete data in a 'natural' way. Secondly, the inclusion of continuous and time-varying covariates should be straightforward. Thirdly, the model should account for the three different sources of variation and allow flexible specifications of correlation structures. Finally, the model should be able to characterize the within-individual change over time. In the following we will present regression models that meet these desirable features.

Most useful statistical models for the analysis of Gaussian longitudinal data can be classified either as *full multivariate models* or *linear random-effect models* [Laird and Ware, 1982]. They differ in the approach taken to account for the correlated nature of the data. The full multivariate models are also known as *marginal multivariate models*, whereas linear random-effect models are also referred to as *linear mixed models* [Molenberghs and Verbeke, 2005], *multi-stage random-effect models* [Laird and Ware, 1982; Diggle et al., 1996], *subject-specific models* [Molenberghs and Verbeke, 2005], *hierarchical linear models* [Davidian and Giltinan, 1995] or *mixed-effects regression models* [Hedeker and Gibbons, 2006].

In the full multivariate model, the repeated measurement vector for every subject is assumed to follow a multivariate normal distribution with a mean vector and a covariance matrix. In particular, the mean vector and the covariance matrix are modelled separately, possibly in terms of available covariates. Here, the covariance matrix accounts for the asso-

ciation structure among the repeated measurements of a subject. In contrast, the correlation in a linear random-effect model arises from the inclusion of unobserved subject-specific random effects and the within-individual heterogeneity. For this approach, we assume that the regression models for the response vectors of all subjects have the same form; however, the regression coefficients are subject-specific. For example, in many longitudinal studies with continuous data the responses can be assumed to follow a linear regression model (perhaps after adequate transformation). However, in general, the intercept or the slope vary from subject to subject. For instance, the aforementioned repeated measures ANOVA assumes that the intercept varies among subjects, reflecting the idea that the baseline measurement of subjects is expected to be subject-specific. This may be due to unmeasurable factors, e.g. genetic predisposition. General multi-stage random-effect models can include several random-effects and induce different covariance structures.

Both model families can handle unbalanced data and to a certain extent incomplete data. Likelihood-based inference for full multivariate and random-effect models are robust to ignorable missingness processes, see Chapter 5. They have the capacity to include continuous and time-varying covariates; and they account for the dependence structure of the data. However, they have limitations. The full multivariate models can be cumbersome in certain settings, as the computational complexity is closely linked to the parametric model assumed for the covariance matrix and the dimension thereof. Some modifications of full multivariate methods to semi-parametric methods overcome the excessive computational requirements, e.g. *pseudo-likelihood methods* and *generalized estimating equations*. These methods are robust to misspecification of the covariance structure and useful when interest focuses on mean parameters only [Diggle et al., 1996]. However, using these methods can incur efficiency loss and can lead to bias for partially observed data, see Chapter 5. A further downside of the full multivariate model is that it does not separate between-individual heterogeneity from within-individual heterogeneity. In particular, it does not specify individual characteristics, but sometimes these are of interest, e.g. when the prediction of subject-specific outcomes is the aim of the study. In contrast, random-effect models quan-

tify the between-individual heterogeneity and subject-specific characteristics. They usually involve fewer parameters than full multivariate models and are thus computationally more appealing. However, relative to the full multivariate model, they are limited by the special form assumed for the covariance structure [Laird and Ware, 1982]. This limitation will be discussed in more detail in Section 3.2 and Chapter 4.

We note that although these two model families are generally different, they can lead to the same inference for fixed effect parameters, i.e., the parameters associated with available covariates. This feature arises from the elegant properties of the multivariate normal distribution: each random-effect model implies a full multivariate model. However, these models are not equivalent, see Section 3.2.

3.1.5 Regression Models for Non-Gaussian Data

So far we focused on Gaussian data, but the aforementioned full multivariate and random-effect models can be extended to the case of discrete and categorical data. Assuming the single responses follow a distribution from the *exponential family*, a *generalized linear model* [McCullagh and Nelder, 1989] can be formulated for each component of the repeated measurement vector. Similar to the Gaussian case, there are different ways to account for the correlation among the observations of a subject. Firstly, a *marginal model* can be formulated. As in the case of the full multivariate model, the association structure is modelled separately from the mean through a marginal covariance structure. In this context, the covariance matrix can be modelled through a parametric approach (full likelihood approach) or by means of a working covariance matrix (semi-parametric approach), see *generalized estimating equations* (GEE) presented in Liang and Zeger [1986], and *pseudo-maximum likelihood methods* [Molenberghs and Verbeke, 2005]. As mentioned above, the latter methods imply that, under some regularity conditions, inference for the regression coefficients of interest is valid under misspecification of the covariance matrix.

Secondly, subject-specific random regression parameters can be incorporated in the generalized linear model, inducing correlation in the same fashion as for linear random-

effect models. In the literature these models are referred to as random-effect models, *generalized linear mixed models* [Molenberghs and Verbeke, 2005] or *hierarchical models* [Davidian and Giltinan, 1995].

Thirdly, a *conditional model* can be formulated. The conditional expectation of a component of the response vector is modelled in terms of available explanatory variables and a subset of past outcomes. Thus, the relation between the outcome and the predictors, but also the dependence structure are modelled through a single equation [Diggle et al., 1996]. Note that a conditional model can also be formulated for Gaussian outcomes.

All these model families can be seen as extensions of the *generalized linear models* introduced by McCullagh and Nelder [1989] for independent observations in the context of correlated data [Molenberghs and Verbeke, 2005].

We noted that in the case of Gaussian data, fixed effects parameters based on marginal and random-effect model have the same interpretation. In the case of non-Gaussian data, however, different assumptions about the source of correlation can lead to regression parameters with distinct interpretations [Diggle et al., 1996]. A marginal model measures population-averaged effects of explanatory variables on the mean response. In contrast, random-effect models analyze the effect of covariates on the mean response of single individuals conditional upon their subject-specific random effects. In conditional models, the covariate effects of explanatory variables on the mean response are measured conditional on the subjects' measurement history.

From the cited literature, it is obvious that marginal and random-effect models are more popular than conditional models. This might be due to the drawbacks in the interpretability of conditional models. To quote Molenberghs and Verbeke [2005]: 'Diggle et al. [1996] criticized the conditional approach because the interpretation of a fixed effect parameter (e.g., time evolution or treatment effect) of one response is conditional on other responses for the same subject, outcomes of other subjects, and the number of repeated measures. Not only may such parameters make answering the substantive question difficult, they are also ill founded when the number of measurements per subject is not constant.'

3.1.6 Regression Models for Non-Linear Relationships

The models presented so far assume that the mean response is a linear function of the unknown parameters or that a link-function exists such that a linear model for the transformed mean is adequate. However, some studies observe data that are intrinsically non-linear and fitting a linear or generalized linear model to the components of the response vector is inadequate. In many settings, transformations of the outcome of interest or covariates are proposed to establish linear relationships between the mean or transformed mean (generalized linear model) and the parameters of interest. Although the approach of data transformations is widely used in applied data analysis, it is not always possible to obtain a linear relation, see Section 3.3. In some cases, the research questions of interest can only be answered based on the original data. Davidian and Giltinan [1995] note that when the model for the mean response is non-linear it is often because of a meaningful empirical or theoretical relationship between the response and covariates, and it is desirable that this relationship be preserved in the data. Both marginal and random-effect models can be extended to account for a non-linear relationship of the mean response and covariates. Admittedly, there has been very little attention to marginal models in the literature [Molenberghs and Verbeke, 2005]. Non-linear random-effect models are frequently termed as *non-linear mixed models* Davidian and Giltinan [1993, 1995]; Vonesh and Chinchilli [1997]; Davidian and Giltinan [2003]; Molenberghs and Verbeke [2005]. This class of models has found many biological applications, such as pharmacokinetic analysis, rate of clearance of a drug, studies of growth to adult size and decay. However, we will show that they can be also very useful in the health-care context, see Section 3.5.

3.1.7 Exploratory Analysis

In statistical research, it is common practice to conduct exploratory data analysis before fitting a statistical model. Thus, it comes as a surprise that, although various monographs and papers on the analysis of longitudinal data have been published, only a few researchers discuss this aspect of the data analysis. Among the literature that addresses the challenge

of exploratory analysis, the book of Diggle et al. [1996] gives a few guideline for making an effective exploratory analysis for longitudinal data:

- instead of data summaries show as much as possible of the raw data;
- reveal patterns of scientific interest;
- highlight both cross-sectional and longitudinal features of the data;
- identify subjects with unusual patterns.

For graphical presentation *time plots* [Fitzmaurice et al., 2004], also informally referred to as *'spagetthi'-plots*, can be very useful. In these plots the outcome of interest is plotted against time. The repeated measurements of every subject are connected through lines to emphasize the change over time. As longitudinal studies usually involve a large number of subjects, these plots can be excessively busy. Diggle et al. [1996] suggest different ways to overcome this problem; one of which is to create a time plot for a random sample of patients. Such a plot was presented in Chapter 2, Figure 2.1. By connecting the repeated measurements of every subject, cross-sectional features can be distinguished from longitudinal features. These plots also enable the comparison of within-individual and between-individual heterogeneity, and can help in identifying unusual patterns, e.g. response profiles of subjects with missing observations might differ from the profiles of completers.

Most longitudinal studies also aim to investigate the effect of explanatory variables other than time on the outcome of interest. A scatter-plot of the response against explanatory variables can be useful. Furthermore, a modification of time plots, where time is replaced by another explanatory variable, can be created. In order to reveal typical patterns, non-parametric smoothing techniques, such as lowess, spline and kernel estimation can be used to fit a smooth curve to the data [Diggle et al., 1996]. Additionally, mean response profiles for different sub-groups of the population can be plotted, for example see Chapter 2, Figure 2.2.

We mentioned earlier that time plots can be useful for visualizing the dependence among the repeated measurements. Additionally, a scatter-plot of each pair of repeated

measurements [Fitzmaurice et al., 2004] or the empirical correlation matrix can be helpful in assessing the dependence structure. An alternative method, which has it origins in the analysis of spatially correlated data, is the *variogram*, which plots half-squared differences between pairs of residuals against the corresponding time lags [Diggle, 1988; Diggle et al., 1996; Hedeker and Gibbons, 2006].

3.1.8 Model Diagnostics

The methods presented in the last section are aimed at exploring longitudinal data before a statistical model is fitted. They should therefore be the starting point of any data analysis. In contrast, the final point of any data analysis should consist of model diagnostics to check the fitted model against the observed data. Similarly to the case of exploratory analysis, we note that model diagnostics have not been explored intensively in the literature.

Model diagnostic tools for longitudinal data ought to check for systematic departures in the data from the mean and the dependence structure modelled. Although interest usual lies in the mean model and associated parameters, Fitzmaurice et al. [2004] showed that a misspecification of the dependence structure can lead to incorrect inference for the mean parameters. Vonesh et al. [1996] note that very little work has been done in the area of model diagnostics with respect to the assumed dependence structure. This is partly due to the various levels of complexity in the analysis of longitudinal data, as set out in the thesis of Chiswell [2007]. For example, in order to fit a mixed model the investigator needs to decide how many, and which, parameter coefficients are assumed to be subject-specific random coefficients, as these induce correlation on the marginal level. Furthermore, sources for the within-individual variation have to be identified. This implies decisions about (i) the within-individual variances and whether these are homo- or heteroscedastic, and (ii) the existence of within-individual serial correlation. Given all these factors that influence the covariance model, Yang and Yue-Chen [2006] come to the conclusion that a single model diagnostic tool alone is not able to resolve all problems. Yang and Yue-Chen [2006] argue that a useful model diagnostic tool should check for random fluctuation in the residual

values; eight different testing procedures for randomness are presented in their paper.

In order to check the adequacy of the model for the mean response, Diggle et al. [1996] suggest superimposing the fitted mean response profiles on a time plot of the average observed response within each combination of treatment and time. Additionally, the fitted variogram can be superimposed on a plot of the empirical variogram. Fitzmaurice et al. [2004] also suggest the use of the empirical variogram to assess the validity of the variance assumptions. Graphical diagnostics using transformed residuals are recommended to explore the adequacy of the model for the mean response. Alternatively, Vonesh et al. [1996] present mean and covariance concordance correlation coefficients between fitted and observed measurements, which are similar to the R^2 criterion widely used for univariate linear regression models. For normally distributed data, the adequacy of the fitted covariance model is assessed via a pseudo-likelihood ratio test. A plot, similar to the quantile-quantile plot (Q-Q plot), to check the model fit for longitudinal data and that is able to account for an unbalanced design is presented in Park and Lee [2004]. The observation vector of each subject is summarized through a univariate *normalized residual* and based on the residuals for all subjects a Q-Q plot using the standard normal distribution can be constructed. This approach is shown to be effective in determining the adequacy of the mean model fit, and in the detection of outlying and influential observations. However, Park and Lee [2004] state that these model diagnostic tools are less effective as covariance model diagnostics. Alternatively, we propose performing a sensitivity analysis with respect to the covariance model chosen. We will illustrate this approach in Chapter 4.

For a more detailed account on model diagnostic tools for longitudinal data with particular focus on non-linear mixed models, we refer to Chiswell [2007].

3.1.9 Summary

In this section we have elaborated the objectives of longitudinal studies, and listed some of the advantages over cross-sectional studies. The price to pay for these advantages is that more complex data analysis techniques are required. Among other desirable features,

these techniques ought to account for the correlated nature of longitudinal data. We have discussed the limitations of simple approaches, such as repeated measures ANOVA and MANOVA. Full multivariate models and linear mixed models are suitable alternatives for the analysis of Gaussian longitudinal data; these will be discussed in more detail in Section 3.2. Different extension to non-Gaussian data and non-linear regression models have been discussed. Furthermore, we give a short account on exploratory and model diagnostic tools in the context of longitudinal data.

We have briefly mentioned that missing data poses a considerable challenge in the analysis of longitudinal data; and that likelihood-based inference for full multivariate and random-effect models are robust to ignorable missingness processes. Chapter 5 and Chapter 6 are devoted to these aspects.

We note that this chapter focuses mainly on modelling the mean response. In Chapter 4, we will discuss models for the covariance structure.

Finally, we admit that there are numerous additional topics on the analysis of longitudinal data that we are not going to discuss in this thesis, e.g. multivariate longitudinal data and non-parametric analysis techniques. We refer to Diggle et al. [1996] and Gibbons et al. [2010] for a discussion of these topics.

The remains of this chapter are organized as follows. We discuss full multivariate and linear-mixed models in Section 3.2. We then apply linear-mixed models to the CAST data set in Section 3.3 and discuss why these models are not suitable for this particular data set. Non-linear mixed models are presented as a valuable alternative in Section 3.4 and applied to the CAST data in Section 3.5.

3.2 Full Multivariate Models and Linear-Mixed Models

In this section, full multivariate and linear mixed models for Gaussian longitudinal data will be presented in detail.

3.2.1 Notation

Let $y_{i,j}$ denote the observation of subject $i \in \{1, ..., N\}$ at observation time $t_{i,j}$, $j \in \{1, ..., M_i\}$, and $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,M_i})^\top$ be the M_i-dimensional response vector of subject i. For a balanced design we denote $M_i = M$ for all $i \in \{1, ..., N\}$. The vector $\boldsymbol{y}_i$ is a realisation of the random vector $\boldsymbol{Y}_i$, where $\boldsymbol{Y}_i$ is assumed to follow a multivariate normal distribution. The joint outcome vector for all subjects is denoted by $\boldsymbol{y} = (\boldsymbol{y}_1^\top, ..., \boldsymbol{y}_N^\top)^\top$ and the associated random vector by $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, ..., \boldsymbol{Y}_N^\top)^\top$.

For subject i and observations time $t_{i,j}$, the q-dimensional ($q \in \mathbb{N}$) vector of explanatory variables (e.g. time, age, gender, treatment) is represented by $\boldsymbol{x}_{i,j}$. The covariate vectors for all observation times of subject i are collected through the matrix $X_i = (\boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,M_i})^\top$.

The mean and the variance of $Y_{i,j}$ are denoted by $\mathbb{E}(Y_{i,j}) = \mu_{i,j}$ and $\mathrm{Var}(Y_{i,j})$, respectively. The corresponding M_i-dimensional mean vector and the $M_i \times M_i$-dimensional covariance matrix of $\boldsymbol{Y}_i$ are represented by $\mathbb{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ and $\mathrm{Cov}(\boldsymbol{Y}_i) = \Sigma_i$, respectively.

In the CAST study presented in Chapter 2, the design implies $N = 559$, $M_i \in \{1, 2, 3, 4\}$ and $t_{i,j} = t_j \in \{0, 4, 12, 39\}$ for $i \in \{1, ..., 559\}$ and $j \in \{1, 2, 3, 4\}$. The explanatory variables of interest are given by the randomisation group, age and gender of the subjects. The randomisation groups are denoted by $s_i \in \{1, 2, 3, 4\}$ for Tubigrip, BKC, Aircast brace and Bledsoe boot, respectively. Note that in this thesis we mean *randomisation group* whenever we refer to *treatment groups*, because the analysis will be performed based on the intention-to-treat principle. The age and gender of subject $i \in \{1, ..., N\}$, are denoted by age_i and sex_i, respectively.

3.2.2 Full Multivariate Model

In the marginal multivariate model for Gaussian data, we assume all subjects are independent and that $\boldsymbol{Y}_i \sim \mathcal{N}_{M_i}(\boldsymbol{\mu}_i, \Sigma_i)$, where the mean $\boldsymbol{\mu}_i$ and the covariance matrix Σ_i are modelled separately.

Most longitudinal studies are interested in studying the change of the mean response

over time and in identifying covariates that influence the change. For this purpose a linear regression model

$$\mathbb{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i = X_i\beta,$$

for the mean response vector is formulated, where $\beta \in \mathbb{R}^q$ is a q-dimensional vector of unknown coefficients. In a separate step, a model for the covariance matrix Σ_i is formulated. It is convenient to set $\Sigma_i = \Sigma$, $i \in \{1, ..., N\}$, when the data are balanced and the covariance matrix does not depend on covariates. In general, however, the covariance matrix will depend on the subject i, at least through its dimension. Different assumptions regarding the covariance structure can be made, e.g. unstructured form, first-order autoregressive structure, compound symmetry structure or Toeplitz structure [Hedeker and Gibbons, 2006]. All these covariance structures, except the unstructured form, assume equal variances and a specific structure for the pairwise correlations. In contrast, the unstructured covariance structure allows all variances and covariances to be different. Hence, the number of parameters depends on the number of repeated measures per subject. For $M_i = M$, $\Sigma_i = \Sigma$, $i \in \{1, ..., N\}$, the number of covariance parameters under an unstructured covariance model is given by $\frac{M(M+1)}{2}$.

3.2.3 Linear Mixed Model

In Section 3.1.4, the linear mixed model was briefly introduced as a regression model that accounts for the correlation among the repeated measures of a subject through the inclusion of unobserved subject-specific random effects and a model for the within-individual heterogeneity. Following the derivations in Laird and Ware [1982]; Davidian and Giltinan [1995] and Diggle et al. [1996], we define a random-effect model for Gaussian data, i.e. a linear mixed model, in terms of a hierarchical two-stage model. In the first stage population parameters, individual effects and the within-subject heterogeneity are introduced, whereas the second stage accounts for the between-subject variation [Fitzmaurice et al., 2004].

Stage 1:

The response vector $\boldsymbol{Y}_i$ satisfies the mixed model

$$\boldsymbol{Y}_i = X_i\beta + A_i\boldsymbol{U}_i + \boldsymbol{\epsilon}_i,$$

where $\beta \in \mathbb{R}^q$ represents the parameter vector of fixed effects, $\boldsymbol{U}_i$ is a p-dimensional vector of random effects, A_i is a $M_i \times p$- dimensional design matrix linking $\boldsymbol{Y}_i$ to the random effects and $\boldsymbol{\epsilon}_i$ is a M_i-dimensional vector of *within-individual* errors. We assume:

- The error terms satisfy

  $$\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \tilde{\Sigma}_i),$$

  where $\tilde{\Sigma}_i$ is a $(M_i \times M_i)$-dimensional covariance matrix, accounting for the within-individual heterogeneity. In particular, this matrix ought to account for variation due to measurement error and serial correlation.

- The conditional mean vector $\tilde{\boldsymbol{\mu}}_i := \mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{U}_i)$ satisfies the linear mixed model

  $$\tilde{\boldsymbol{\mu}}_i = X_i\beta + A_i\boldsymbol{U}_i.$$

Stage 2:

In order to account for the between-individual variation we assume:

- The random vectors $\boldsymbol{U}_i$ are mutually independent and identically distributed for all $i \in \{1, ..., N\}$ with

  $$U_i \overset{iid.}{\sim} \mathcal{N}(\mathbf{0}, D), \tag{3.1}$$

  where D is a $p \times p$-dimensional covariance matrix and

- $\boldsymbol{U}_i$ are independent of $\boldsymbol{\epsilon}_i$ and the explanatory variables X_i for all $i \in \{1, ..., N\}$.

The motivation for the assumption shown in equation (3.1), is given in [Molenberghs and Verbeke, 2005, page 37]: 'because subjects are randomly sampled from a population of

subjects, it is natural to assume that the subject-specific regression coefficients [...] are randomly sampled from a population of regression coefficients. It is customary to assume the [...] $\boldsymbol{U}_i$ to be (multivariate) normal, but extensions can be formulated [...].'

Under these assumptions, the marginal mean vector and covariance matrix for $\boldsymbol{Y}_i$ are given by:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{Y}_i) &= \mathbb{E}\left[\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{U}_i)\right] = X_i\beta; \\
\mathrm{Cov}(\boldsymbol{Y}_i) &= \mathbb{E}\left[\mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{U}_i)\right] + \mathrm{Cov}\left[\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{U}_i)\right] \\
&= \tilde{\Sigma}_i + A_i D A_i^\top.
\end{aligned}$$

Therefore, every linear mixed model has a hierarchical but also a marginal model formulation. In Section 3.1.4, we discussed that parameters based on marginal and random-effect models generally have different interpretations. While parameters based on a marginal model have population-averaged interpretations, random-effect model parameters describe the evolution of the response variable based on a certain level of the subject-specific random variable U_i. Nevertheless, the parameter estimates for fixed effects obtained through both model families should be equal in the case of Gaussian data. This is due to the unique properties of the multivariate normal distribution. Let β^{RE} and β^{M} denote the fixed effects parameter vectors based on the hierarchical and marginal model formulation, respectively. The hierarchical model formulation yields

$$\begin{aligned}
\mathbb{E}(\boldsymbol{Y}_i) &= \mathbb{E}\left[\mathbb{E}(\boldsymbol{Y}_i|\boldsymbol{U}_i)\right] = X_i\beta^{RE}, \quad \text{and the marginal formulation} \\
\mathbb{E}(\boldsymbol{Y}_i) &= X_i\beta^{M}.
\end{aligned}$$

Assuming the rank of the matrix X_i is full for every i, i.e. $\mathrm{rank}(X_i) = q$ for all $i \in \{1, ..., N\}$, leads to $\beta^{RE} = \beta^{M}$. This connection of marginal and hierarchical models in the case of Gaussian models is so natural that it can be misleading. Even though every hierarchical model has a marginal model formulation, these two model families are not equivalent.

Molenberghs and Verbeke [2005] present an example where two different hierarchical models lead to the same marginal model. On the other hand, Molenberghs and Verbeke [2005] show that not every marginal model can be expressed in terms of a hierarchical model. For example, a marginal model that assumes a compound symmetry covariance structure with negative correlations cannot be implied by a hierarchical model.

The last observation we want to make concerns the within-individual heterogeneity modelled through $\tilde{\Sigma}_i$. Frequently, a special structure, the so-called *conditional independence model* is proposed [Laird and Ware, 1982], where

$$\tilde{\Sigma}_i = \sigma^2 I_{M_i},$$

I_{M_i} is the $M_i \times M_i$-dimensional identity matrix and $\sigma^2 \in \mathbb{R}^+$. While this assumption simplifies the likelihood construction to a great extent, it ignores serial correlation and emphasizes measurement error as only source for within-individual variation. Diggle et al. [1996] explain the popularity of this assumption by noting that in many applications the magnitude of serial correlation is dominated by the combination of between-individual correlation and measurement error. According to Molenberghs and Verbeke [2005], computational complexity when accounting for serial correlation is a further reason for adopting the simpler model.

3.2.4 Estimation and Inference

As seen in the last subsection, every hierarchical model has a marginal formulation. Thus, model fitting for full multivariate and linear mixed models can be based on the marginal model formulation. In this context, we will focus on the maximum likelihood approach for estimation and inference.

In the full multivariate and linear mixed model the response vector $\boldsymbol{Y}_i$ is assumed to follow a multivariate normal distribution with mean $X_i\beta$ and covariance matrix Σ_i and $\tilde{\Sigma}_i + A_i D A_i^\top$, respectively. Independent of the model choice we denote the covariance matrix

through $\Sigma_i(\tau)$, where τ is a vector of covariance parameters. For example, the compound symmetry structure only involves two different parameters, one for the variance and one for the correlation, whereas the unstructured form involves $\frac{M(M+1)}{2}$ parameters for a balanced data set.

We now consider estimating the mean parameter β and the covariance parameter τ jointly by maximizing the joint likelihood:

$$\begin{aligned} L_{\boldsymbol{Y}}(\beta,\tau) &= \prod_{i=1}^{N} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\,\pi)^{M_i}\,\det\left(\Sigma_i(\tau)\right)}}\,\exp\left\{-\frac{1}{2}(\boldsymbol{Y}_i - X_i\beta)^{\top}\,\Sigma_i^{-1}(\tau)\,(\boldsymbol{Y}_i - X_i\beta)\right\} \end{aligned} \tag{3.2}$$

Conditionally on τ, the maximum likelihood estimator for β is given by the weighted least-squares estimator [Laird and Ware, 1982; Diggle et al., 1996]

$$\hat{\beta}(\tau) = \left(\sum_{i=1}^{N} X_i^{\top}\,\Sigma_i^{-1}(\tau) X_i\right)^{-1} \sum_{i=1}^{N} X_i^{\top}\,\Sigma_i^{-1}(\tau)\,\boldsymbol{y}_i. \tag{3.3}$$

The maximum-likelihood estimator $\hat{\tau}$ for τ can be obtained by substituting $\hat{\beta}(\tau)$ into the likelihood (3.2) and subsequent maximization of $L_{\boldsymbol{Y}}(\hat{\beta},\tau)$ with respect to τ. The maximum-likelihood estimator for β is then given by $\hat{\beta} = \hat{\beta}(\hat{\tau})$. As noted by Laird and Ware [1982], the maximum likelihood estimate for τ fails to account for the degrees of freedom lost in estimating β and is thus biased downwards. Alternative estimating methods for the covariance parameters, such as the *restricted maximum likelihood* (REML) method, were proposed [Laird and Ware, 1982; Diggle et al., 1996]. The REML estimator is defined as a maximum likelihood estimator based on a linearly transformed set of data, such that the distribution of the transformed data does not depend on β [Diggle et al., 1996]. Maximizing the likelihood based on the transformed data set yields a consistent estimator $\tilde{\tau}$ for τ. The REML estimator for β is then obtained by substituting $\tilde{\tau}$ in equation 3.3.

Diggle et al. [1996] note that the distinction between maximum likelihood and REML is important only when the number of mean parameters q is relatively large com-

pared to the overall number of observation, i.e. $N_{tot} = \sum_{i=1}^{N} M_i$. In the applications relevant to this thesis, q is usually much smaller than N_{tot}. Hence, we will focus on the maximum likelihood approach and refer to Laird and Ware [1982] and Diggle et al. [1996] for further details regarding REML.

In order to conduct maximum-likelihood inference for the mean parameters, the moments of the asymptotically normally distributed estimator $\hat{\beta}(\hat{\tau})$ need to be calculated. Assuming the mean vector and covariance matrix were correctly specified we obtain [Molenberghs and Verbeke, 2005]:

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}(\hat{\tau})\right) &= \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau}) X_i\right)^{-1} \sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau})\, X_i \beta = \beta \quad \text{and} \\
\mathrm{Cov}\left(\hat{\beta}(\hat{\tau})\right) &= \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau}) X_i\right)^{-1} \sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau})\, \Sigma_i(\hat{\tau})\, \Sigma_i^{-1}(\hat{\tau})\, X_i \\
&\qquad \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau}) X_i\right)^{-1} \\
&= \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau}) X_i\right)^{-1}.
\end{aligned}$$

Note that the covariance matrix of $\hat{\beta}(\hat{\tau})$ does not account for the variability in estimating τ, and thus underestimates the variability. For this reason, the *t-test* is often used instead of the *z-test* / *Wald-test* for hypothesis testing [Molenberghs and Verbeke, 2005]. Several approaches for the estimation of the degrees of freedom exist [Fitzmaurice et al., 2004]; for example, it can be computed as the number of subjects minus the dimension of the random-effect vector [SAS/STAT, 1999]. Molenberghs and Verbeke [2005] note that for large sample sizes most estimation methods yield very similar results in terms of p-values. For small samples the estimation method of the degrees of freedom has to be chosen more carefully. In this work, we will focus on studies with large samples sizes and will use the estimation method proposed by SAS/STAT [1999], see above.

For the covariance parameters asymptotic Wald-type, likelihood ratio and score tests

can be used [Molenberghs and Verbeke, 2005]. We will focus on the t-test for the covariance parameters. For more details regarding this aspect we refer to Molenberghs and Verbeke [2005].

3.2.5 Inference for the Random Effects

In many longitudinal studies, inference focuses on the parameters of the marginal distribution, i.e. β and τ. However, in some applications, interest lies in predicting subject-specific outcomes or response profiles. Diggle et al. [1996] present an example, the *CD4+ study*, where investigators want to estimate the subject-specific evolution of *CD4+ cell counts* for the counselling of HIV-infected men.

In order to construct predictions but also to identify outliers [Molenberghs and Verbeke, 2005] and for model diagnostics, the subject-specific random-effect vectors U_i, $i \in \{1, ..., n\}$, can be estimated through their *empirical Bayes* (EB) estimates. The EB estimates $\hat{U}_i$ are calculated as means of the posterior distribution of U_i given $\hat{\beta}$ and $\hat{\tau}$, i.e. the distribution of $\boldsymbol{U}_i|\boldsymbol{Y}_i, \hat{\beta}, \hat{\tau}$. Let $\Sigma_i(\hat{\tau}) = \tilde{\Sigma}_i(\hat{\tau}) + A_i\, D(\hat{\tau})\, A_i^\top$ denote the maximum-likelihood estimator for the covariance matrix. Under the assumption of normality for $\boldsymbol{Y}_i$ and $\boldsymbol{U}_i$, Laird and Ware [1982] present a closed form solution for the EB estimate:

$$\hat{\boldsymbol{U}}_i = D(\hat{\tau})\, A_i\, \Sigma_i^{-1}(\hat{\tau}) \left(\boldsymbol{y}_i - X_i \hat{\beta}\right) \tag{3.4}$$

and its covariance matrix:

$$\begin{aligned} \mathrm{Cov}\left(\hat{\boldsymbol{U}}_i\right) \;=\; & D(\hat{\tau})\, A_i^\top \left\{ \Sigma_i^{-1}(\hat{\tau}) - \Sigma_i^{-1}(\hat{\tau})\, X_i \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1}(\hat{\tau})\, X_i \right) X_i^\top\, \Sigma_i^{-1}(\hat{\tau}) \right\} \\ & \times A_i\, D(\hat{\tau}). \end{aligned} \tag{3.5}$$

Using the covariance expressed through equation (3.5) to assess the precision of $\hat{\boldsymbol{U}}_i$ fails to account for the random variation of $\boldsymbol{U}_i$, and hence underestimates the variation in $\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i$.

Laird and Ware [1982] propose to use

$$\mathrm{Cov}\left(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i\right) = D(\hat{\tau}) - \mathrm{Cov}\left(\hat{\boldsymbol{U}}_i\right)$$

to assess the precision of the EB estimates.

A commonly observed feature of EB estimates is that they are shrunken towards the mean of the prior distribution of $\boldsymbol{U}_i$, i.e. $\mathbf{0}$. The degree of shrinkage depends on the interplay of within-subject heterogeneity and between-subject heterogeneity [Molenberghs and Verbeke, 2005]. This interplay can be illustrated through the following calculations [Fitzmaurice et al., 2004]:

$$\begin{aligned}
\hat{\boldsymbol{y}}_i &= X_i\hat{\beta} + A_i\ \hat{\boldsymbol{U}}_i \\
&\overset{(3.4)}{=} X_i\hat{\beta} + A_i\, D(\hat{\tau})\, A_i\, \Sigma_i^{-1}(\hat{\tau})\, \left[\boldsymbol{y}_i - X_i\hat{\beta}\right] \\
&= \left[I_{M_i} - A_i\, D(\hat{\tau})\, A_i\, \Sigma_i^{-1}(\hat{\tau})\right] X_i\hat{\beta} + A_i\, D(\hat{\tau})\, A_i\, \Sigma_i^{-1}(\hat{\tau})\, \boldsymbol{y}_i \\
&\overset{(*)}{=} \left[\tilde{\Sigma}_i(\hat{\tau})\, \Sigma_i^{-1}(\hat{\tau})\right] X_i\hat{\beta} + \left[I_{M_i} - \tilde{\Sigma}_i(\hat{\tau})\, \Sigma_i^{-1}(\hat{\tau})\right] \boldsymbol{y}_i,
\end{aligned}$$

where $\tilde{\Sigma}_i(\hat{\tau})$ denotes the estimated within-subject variation and the equality $(*)$ follows from:

$$\begin{aligned}
I_{M_i} = \Sigma_i(\hat{\tau})\, \Sigma_i^{-1}(\hat{\tau}) &= \left[A_i\, D(\hat{\tau})\, A_i + \tilde{\Sigma}_i(\hat{\tau})\right] \Sigma_i^{-1}(\hat{\tau}) \\
&= A_i\, D(\hat{\tau})\, A_i\, \Sigma_i^{-1}(\hat{\tau}) + \tilde{\Sigma}_i(\hat{\tau})\, \Sigma_i^{-1}(\hat{\tau}).
\end{aligned}$$

These derivations show that the predicted response profile for subject i, i.e. $\hat{\boldsymbol{y}}_i$, can be interpreted as a weighted average of the population-averaged profile and the observed data of subject i [Molenberghs and Verbeke, 2005]. The population-averaged profile is weighted more when the within-subject variation is large relative to the between-subject variation. On the other hand, the observed response is weighted more when the opposite holds. Thus, Molenberghs and Verbeke [2005] come to the conclusion that 'severe shrinkage is to be expected when the residual variability is large in comparison to the between-subject vari-

ability (modelled by the random effects), whereas little shrinkage will occur if the opposite is true'.

3.3 Linear Mixed Models and CAST

The CAST study, presented in Chapter 2, provides an example of a longitudinal study in which the change of the mean response over time and the effect of factors, such as treatment and age, on the change is of interest.

The time plot shown in Figure 2.1, page 13, revealed that the scores were usually a non-linear increasing function of time. The scores increase much faster at the beginning of the study than towards the end.

In general, the responses exhibited similarly shaped curves. However, the scores at the study points and the rates at which these scores were achieved varied across the subjects. This variation appears to be smaller at the end of the study period than at the beginning. In fact, we observe that later measurements are clustered towards the upper end of the range. In particular, different patients might have the same initial and the same final scores, but the rate at which they achieve the final score can differ substantially.

Our aim is to analyze the FAOSS-data through a random-effect model. In this context, we will assume an ignorable missingness mechanism, which implies that the missingness process can be ignored in a full-likelihood based analysis, see Chapter 5. More specifically, the likelihood contribution of a given subject i is obtained by integrating out the missing values from the density of $\boldsymbol{Y}_i$. As the scores are continuous, we assume that the measurements for each subject follow a normal distribution. Based on normality, a linear mixed model could be fitted. However, this model assumes a linear relation between the mean response and the parameter vector of interest, while the observed response profiles suggest a non-linear relation. The classical approach to solve this problem is to transform the data so that a linear regression model fits adequately.

Using the notation introduced in Section 3.2.1, we have tried the following trans-

formations for the responses $y_{i,j}$ and observation times $t_{i,j}$, respectively:

Transformation A:

$$y_{i,j} \mapsto \ln\left(100.5 - y_{i,j}\right) \tag{3.6}$$

Transformation B:

$$y_{i,j} \mapsto \ln\left(\frac{y_{i,j} + \frac{1}{2}}{100.5 - y_{i,j}}\right) \tag{3.7}$$

Transformation C:

$$y_{i,j} \mapsto \ln\left(y_{i,j} + \frac{1}{2}\right) \tag{3.8}$$

Transformation D:

$$t_{i,j} \mapsto \ln\left(t_{i,j} + \frac{1}{2}\right) \tag{3.9}$$

Transformation E:

$$t_{i,j} \mapsto \ln\left(t_{i,j} + 1\right). \tag{3.10}$$

In order to fit an adequate model based on the transformed data, we show the time plots for these five transformation in Figure 3.1 and Figure 3.2.

The time plots for these transformations suggest that a model linear in time will not be sufficient to describe the transformed data adequately. While the time-plots based on Transformation A, D and E suggest that a quadratic model in time might be appropriate, this is not the case for the data based on Transformations B and C. For the latter transformations, the bounded nature of the original score is carried forward to the transformed data.

Nevertheless, we attempted to fit linear mixed models, which are quadratic in time and included age-effects, to all transformed data sets. The mean model of interest for Transformations A, B and C is given by:

$$\begin{aligned}\mathbb{E}\left(Y_{i,j}^{trans}\right) &= \beta_0 + \beta_1\, a_i + U_{i,1} + \left([\beta_{21} + \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \beta_3\, a_i + U_{i,2}\right)\, t_{i,j} \\ &\quad + \left([\beta_{41} + \beta_{4,s_i}\, \mathbb{1}(s_i \neq 1)] + \beta_5\, a_i\right)\, t_{i,j}^2,\end{aligned} \tag{3.11}$$

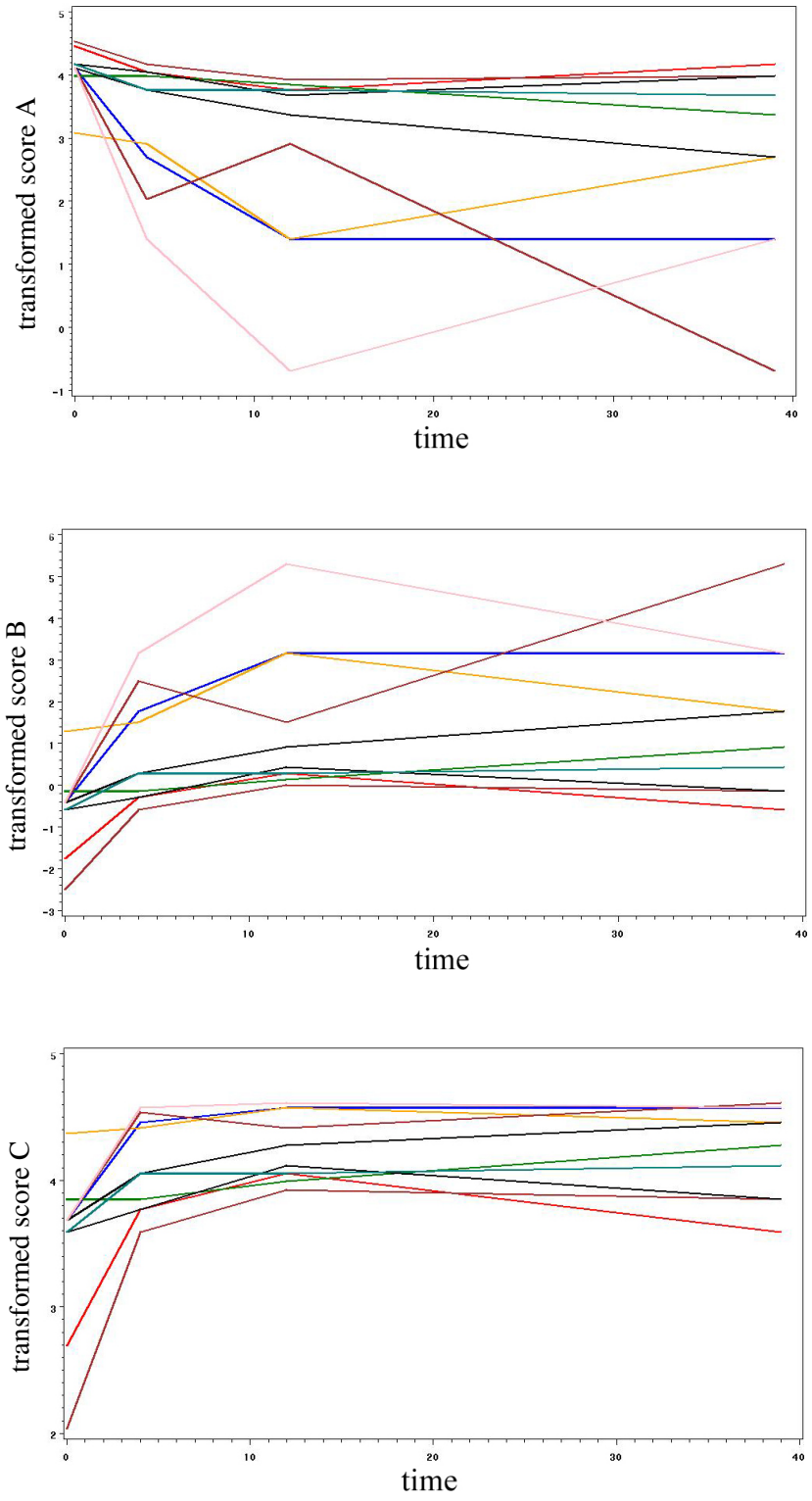


Figure 3.1: Time plots for Transformations A, B, C and a random subset of 10 patients.

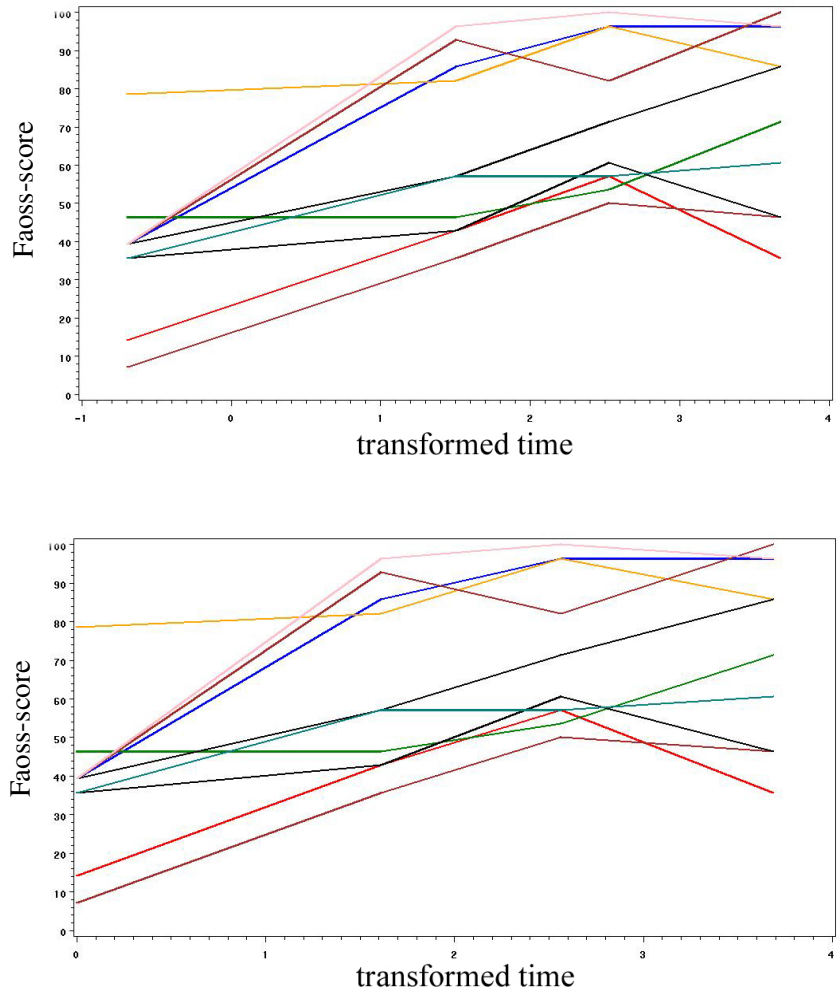


Figure 3.2: Time plots for Transformations D (top), E (bottom) and a random subset of 10 patients.

where $Y_{i,j}^{trans}$ is the transformed random variable of $Y_{i,j}$ based on the transformations $trans \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, $a_i = age_i - 27$ (age centered around median age) and $\mathbb{1}(s_i \neq 1)$ is one if $s_i \neq 1$ and zero otherwise, i.e. interest lies in the treatment contrasts compared to the standard treatment Tubigrip ($s_i = 1$). Model (3.11) allows for a subject-specific intercept (through $U_{i,1}$) and a subject-specific interaction with time (through $U_{i,2}$).

Regarding the covariance structure, we assume the serial correlations are negligible, see Section 3.2.3. This implies that the error of measurements are conditionally indepen-

dent, i.e.

$$Y_{i,j}^{trans}|\boldsymbol{U}_i \overset{ind.}{\sim} \mathcal{N}\left(\mathbb{E}(Y_{i,j}^{trans}), \sigma^2\right) \quad \text{for all} \quad j \in \{1, ..., M_i\},$$

where $\boldsymbol{U}_i = (U_{i,1}, U_{i,2})^\top$ and

$$\boldsymbol{U}_i \sim \mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} \sigma_{U_1}^2 & 0 \\ 0 & \sigma_{U_2}^2 \end{bmatrix}\right).$$

This assumption implies that the two subject-specific effects are independent. This simplification as well as the assumed conditional independence and homoscedasticity for the within-subject variances are likely to be violated. However, in this section we want to focus on modelling the mean. Chapter 4 is devoted to modelling the covariance structure.

Overall, the assumed covariance matrix for the transformed marginal response vector Y_i^{trans} is then given by:

$$\begin{aligned} V_i(\tau) &= \sigma^2 I_{M_i} + \begin{pmatrix} 1 & t_{i,1} \\ \vdots & \vdots \\ 1 & t_{i,M_i} \end{pmatrix} \begin{pmatrix} \sigma_{U_1}^2 & 0 \\ 0 & \sigma_{U_2}^2 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ t_{i,1} & \cdots & t_{i,1} \end{pmatrix}, \qquad (3.12) \\ &= \begin{pmatrix} \sigma^2 + \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,1}^2 & \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,1} t_{i,2} & \cdots & \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,1} t_{i,M_i} \\ \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,2} t_{i,1} & \sigma^2 + \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,2}^2 & \cdots & \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,2} t_{i,M_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,M_i} t_{i,1} & \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,M_i} t_{i,2} & \cdots & \sigma^2 + \sigma_{U_1}^2 + \sigma_{U_2}^2 t_{i,M_i}^2 \end{pmatrix} \end{aligned}$$

where $\tau = (\sigma^2, \sigma_{U_1}^2, \sigma_{U_2}^2)^\top$. Note that the covariance matrix $V_i(\tau)$ depends on i only through its dimension. In the case of missing observations, the covariance matrix is composed of the rows and columns that correspond to the observed measurements.

For the Transformations D and E, we will investigate the following mean models:

$$\mathbb{E}\left(Y_{i,j}\right) = \beta_0 + \beta_1 a_i + U_{i,1} + ([\beta_{21} + \beta_{2,s_i} \mathbb{1}(s_i \neq 1)] + \beta_3 a_i + U_{i,2}) \ln(t_{i,j} + c)$$

$$+ \left([\beta_{41} + \beta_{4,s_i}\, \mathbb{1}(s_i \neq 1)] + \beta_5\, a_i\right) \ln^2(t_{i,j} + c), \tag{3.13}$$

where $c \in \{\frac{1}{2}, 1\}$. We assume the same covariance structure as presented in equation (3.12), however, the observation times $t_{i,j}$ are replaced by the transformed observation times $\ln(t_{i,j}+c)$ for $c \in \{\frac{1}{2}, 1\}$.

Note that the models defined through equation (3.11) and equation (3.13) do not condition on the baseline score $Y_{i,0}$, they rather model the baseline score as part of the outcome vector Y_i. In retrospective, we admit that conditioning on the baseline score could lead to advantages in terms of statistical power and to more precise estimates of treatment effects.

The maximum-likelihood estimators for $\beta = (\beta_0, \beta_1, \beta_{2,1}, \beta_{2,2}, \beta_{2,3}, \beta_{2,4}, \beta_3, \beta_{4,1}, \beta_{4,2}, \beta_{4,3}, \beta_{4,4}, \beta_5)^\top$ and $\tau = (\sigma^2, \sigma^2_{U_1}, \sigma^2_{U_2})^\top$ based on Transformations A-E, models (3.11),(3.13) and an ignorable missingness process are shown in Table 3.1, on page 50. Note that the model based on Transformation C revealed that the inclusion of a subject-specific time-interaction is not necessary. Moreover, the analyses of all transformed data sets but Transformation D do not reject the null hypothesis $\beta_1 = 0$ at a significance level of 0.05.

The aim of the CAST study was to compare the effectiveness of Tubigrip to that of BKC, Aircast brace and Bledsoe boot. The effectiveness is measured in terms of the FAOSS-score change over time. In Model (3.11) and Model (3.13), the treatment effects on the average response profile are quantified through the parameters β_{2,s_i} and β_{4,s_i}, $s_i \in \{1, 2, 3, 4\}$. The results based on Transformation A and B suggest that the mean change of Tubigrip is significantly different to the average response evolution of the remaining treatments. In contrast, the results for Transformation C and E imply that solely the score profiles of Tubigrip and BKC differ, whereas the results for Transformation D yield that the average response evolutions are equal for all treatments. The results based on all transformations reveal significant age-effects on the time evolution.

Thus, the significance of treatment differences and age-effects can be derived based on the results presented in Table 3.1. However, in case of a significant treatment difference

		Transformation A			Transformation B			Transformation C			Transformation D			Transformation E		
Variable		Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.
Int.	β_0	3.99	0.03	<0.001	-0.30	0.05	<0.001	3.69	0.02	<0.001	48.79	0.65	<0.001	40.39	0.69	<0.001
Age on Int.	β_1	-	-	-	-	-	-	-	-	-	-0.15	0.06	0.010	-	-	-
Effects associated with linear time-effect																
Tubigrip	β_{21}	-0.12	0.01	<0.001	0.18	0.01	<0.001	0.06	0.004	<0.001	11.30	0.89	<0.001	15.30	1.35	<0.001
Con.: BKC	β_{22}	0.05	0.01	<0.001	-0.06	0.02	<0.001	-0.01	0.006	0.071	-1.39	1.23	0.261	-5.23	1.81	0.004
Con.: Aircast	β_{23}	0.04	0.01	0.003	-0.05	0.02	0.005	-0.01	0.006	0.160	-1.19	1.23	0.332	-2.91	1.80	0.107
Con.: Bledsoe	β_{24}	0.03	0.01	0.024	-0.04	0.02	0.030	-0.01	0.006	0.215	0.25	1.21	0.837	-2.08	1.77	0.241
Age	β_3	0.002	$4{\cdot}10^{-4}$	<0.001	-0.003	$6{\cdot}10^{-4}$	<0.001	$-9{\cdot}10^{-4}$	$2{\cdot}10^{-4}$	<0.001	-0.24	0.04	<0.001	-0.36	0.06	<0.001
Effects associated with quadratic time-effect																
Tubigrip	β_{41}	0.002	$3{\cdot}10^{-4}$	<0.001	-0.003	$3{\cdot}10^{-4}$	<0.001	-0.001	$1{\cdot}10^{-4}$	<0.001	-0.58	0.28	0.038	-1.10	0.37	0.004
Con.: BKC	β_{42}	-0.001	$3{\cdot}10^{-4}$	<0.001	0.002	$4{\cdot}10^{-4}$	<0.001	$3{\cdot}10^{-4}$	$1{\cdot}10^{-4}$	0.072	0.32	0.38	0.394	1.30	0.51	0.11
Con.: Aircast	β_{43}	-0.001	$3{\cdot}10^{-4}$	0.002	0.001	$4{\cdot}10^{-4}$	0.004	$2{\cdot}10^{-4}$	$1{\cdot}10^{-4}$	0.183	0.26	0.38	0.493	0.68	0.51	0.187
Con.: Bledsoe	β_{44}	-0.001	$3{\cdot}10^{-4}$	0.006	0.001	$4{\cdot}10^{-4}$	0.011	$2{\cdot}10^{-4}$	$1{\cdot}10^{-4}$	0.215	-0.06	0.37	0.871	0.56	0.50	0.264
Age	β_5	$-4{\cdot}10^{-5}$	$1{\cdot}10^{-5}$	<0.001	$6{\cdot}10^{-5}$	$1{\cdot}10^{-5}$	<0.001	$2{\cdot}10^{-5}$	$4{\cdot}10^{-6}$	<0.001	0.05	0.01	<0.001	0.07	0.02	<0.001
Variance Components																
Within	σ^2	0.57	0.02	<0.001	0.90	0.04	<0.001	0.10	0.004	<0.001	146.67	6.62	<0.001	151.40	6.64	<0.001
Between: Int.	$\sigma^2_{U_1}$	0.21	0.03	<0.001	0.53	0.06	<0.001	0.06	0.006	<0.001	123.60	11.34	<0.001	112.37	11.62	<0.001
Between: Time	$\sigma^2_{U_2}$	0.002	$2{\cdot}10^{-4}$	<0.001	0.002	$2{\cdot}10^{-4}$	<0.001	-	-	-	11.21	1.60	<0.001	13.04	1.91	<0.001

Table 3.1: Parameter estimates and standard errors for (β, τ) based on the five different transformations. The abbreviations *Int.* and *Con.* stand for intercept and contrast, respectively.

it is not obvious which treatment is more effective. Also, the significant age effects do not reveal whether older or younger people recover faster from ankle sprains. That is, the interpretation of the parameter estimates is not straightforward and demands further analyses or graphical displays. In order to investigate the nature of the treatment differences, we contrast the observed average score evolutions for Tubigrip and BKC with the fitted response profiles, see Figure 3.3, page 52. The plots show that BKC is more effective than Tubigrip for all transformations where a significant treatment difference was detected, i.e. Transformations A, B and E. However, especially the plots for Transformations A, B and C suggest that the models fit poorly from the twelfth week onwards. Admittedly, this is a rather lopsided comparison as the fitted curves of 27 year old patients are compared to the observed average evolutions of *all* patients.

In order to investigate the overall mean model fit of the proposed models, we use the diagnostic plots presented by Park and Lee [2004]. For incomplete data a *normalized residual* plot based on the *quantile-quantile* (Q-Q) plots is proposed. For every subject the normalized residual q_i^* is a univariate summary measure of the residual vector

$$\boldsymbol{r}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - X_i \hat{\beta},$$

with

$$\mathrm{Cov}(\boldsymbol{r}_i) = H \, \Sigma_i(\tau) \, H^\top,$$

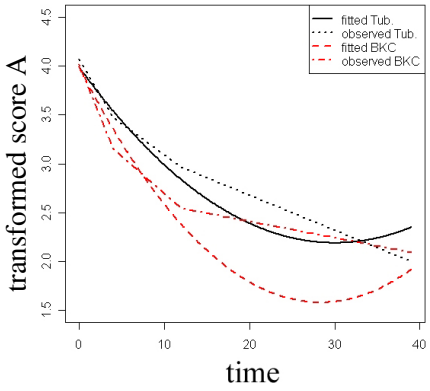
where $H = I_{M_i} - \left(\sum_{i=1}^N X_i^\top \, \Sigma_i^{-1}(\tau) X_i \right)^{-1} \sum_{i=1}^N X_i^\top \, \Sigma_i^{-1}(\tau)$ and $\Sigma_i(\tau)$ is the covariance matrix of the outcome vector $\boldsymbol{Y}_i$, see Section 3.2. The normalized residual is defined as

$$q_i^* = \frac{q_i^{1/4} - (M_i - \frac{1}{2})^{1/4}}{(8 \sqrt{M_i})^{-1/2}},$$

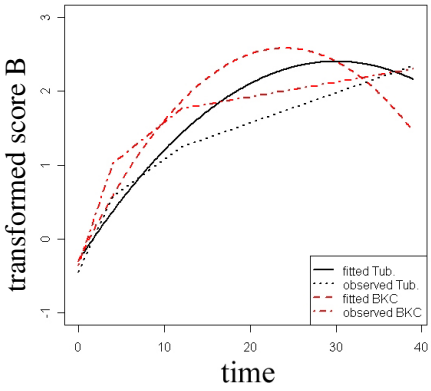
where

$$q_i = \boldsymbol{r}_i^\top \, \mathrm{Cov}(\boldsymbol{r}_i) \, \boldsymbol{r}_i.$$

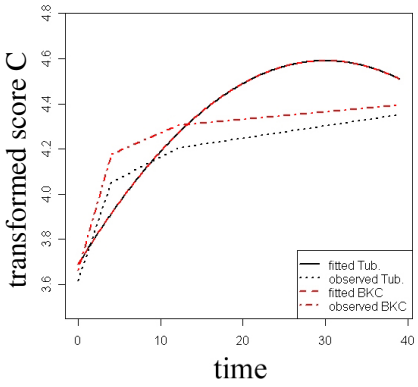
It can be shown that the q_i^*, $i \in \{1, ..., N\}$, follow approximately a standard normal distribu-



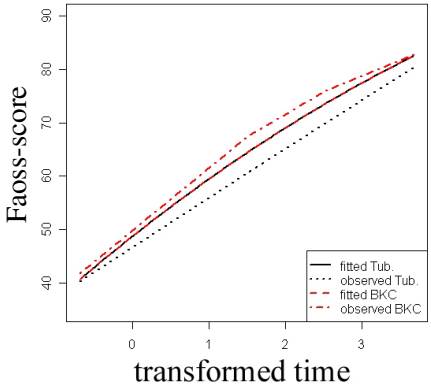
(a) Transformation A



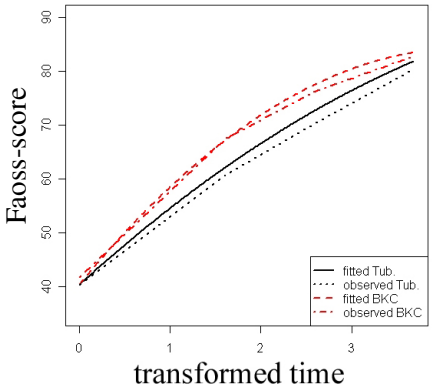
(b) Transformation B



(c) Transformation C



(d) Transformation D



(e) Transformation E

Figure 3.3: Observed average response profile versus fitted response profiles for Tubigrip, BKC and the different transformations. For the fitted response profiles we choose $age = 27$, which corresponds to the median age.

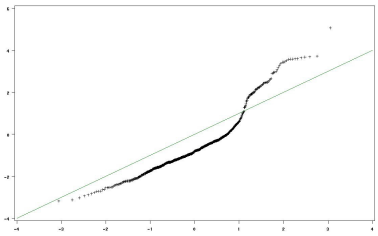
tion [Park and Lee, 2004]. Based on the normalized residuals, a Q-Q plot which plots the observed q_i^*, $i \in \{1, ..., N\}$, against the quantiles of the standard normal distribution can be constructed.

In order to simplify the calculations we replace the covariance matrix of the residuals, $\text{Cov}(\boldsymbol{r}_i)$, by the covariance matrix of the vector of error terms $\boldsymbol{Y}_i - X_i\beta$. Although these two covariance matrices are not identical, Fitzmaurice et al. [2004] state that for all practical purposes the covariance matrix of the residual vector can be approximated by

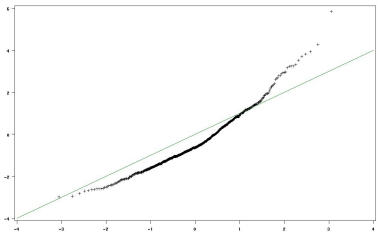
$$\text{Cov}(\boldsymbol{r}_i) \approx \text{Cov}(\boldsymbol{Y}_i - X_i\beta) = \Sigma_i(\tau).$$

Now, the observed normalized residuals can be calculated based on the estimated mean parameter vector $\hat{\beta}$ and the estimated covariance matrix $\Sigma_i(\hat{\tau})$. These residuals will depend on the covariance structure modelled and a poor choice is likely to affect the residuals. Because (in this section) we are mainly interested in checking the mean model, we would like to rule out sensitivity to the covariance structure chosen. In fact, Park and Lee [2004] note that the model diagnostic proposed is not very sensitive to covariance misspecification. Nevertheless, we used both the fitted and the sample covariance matrix (based on observed pairs), to approximate $\text{Cov}(\boldsymbol{r}_i)$ and to investigate the impact of the chosen covariance matrix. The resulting Q-Q plot pairs confirm the observation made by Park and Lee [2004], see Figure 3.4, page 54.

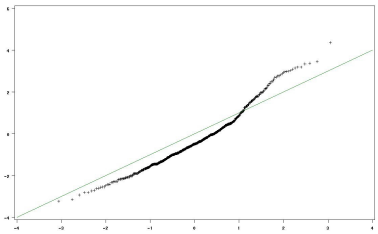
The Q-Q plots for the different transformations and mean models (3.11) and (3.13), reveal stretched *S-shape* patterns for Transformations A, B and C. The quantiles of the standard normal distribution near to zero, exceed the quantiles based on the observed normalized residuals. For the quantiles near one, however, the observed residuals exceed the expected quantiles of the standard normal distribution. This suggests that the distribution of the normalized residuals is negatively skewed or heavy left-tailed. In contrast, the Q-Q plots based on Transformations D and E show no drastic systematic departures; merely a lack of fit for the quantiles near zero is noticeable.



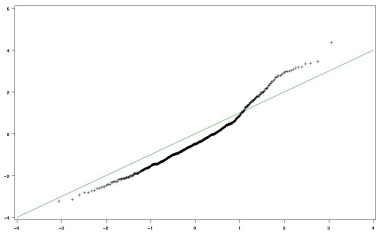
(a) Transf. A, fitted $\Sigma_i(\tau)$



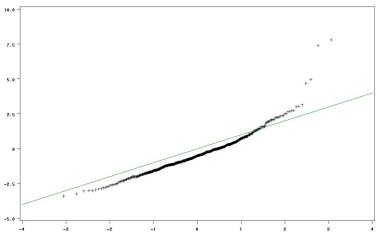
(b) Transf. A, observed $\Sigma_i(\tau)$



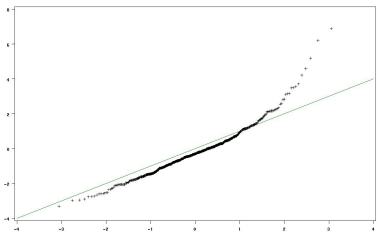
(c) Transf. B, fitted $\Sigma_i(\tau)$



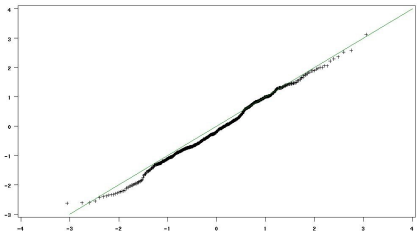
(d) Transf. B, observed $\Sigma_i(\tau)$



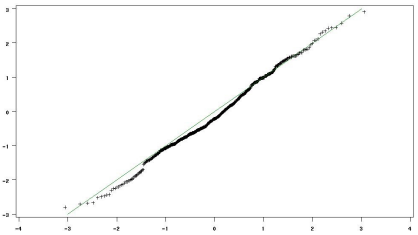
(e) Transf. C, fitted $\Sigma_i(\tau)$



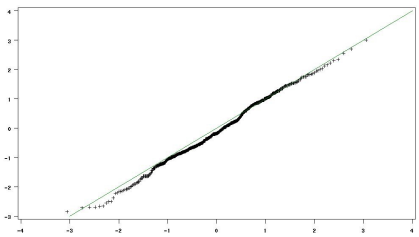
(f) Transf. C, observed $\Sigma_i(\tau)$



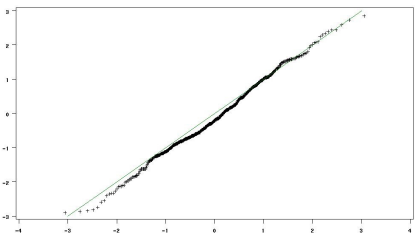
(g) Transf. D, fitted $\Sigma_i(\tau)$



(h) Transf. D, observed $\Sigma_i(\tau)$



(i) Transf. E, fitted $\Sigma_i(\tau)$



(j) Transf. E, observed $\Sigma_i(\tau)$

Figure 3.4: Q-Q plots for the five transformed data sets and fitted models using the normal distribution. The horizontal axes correspond to the expected quantiles of the standard normal distribution, whereas the vertical axes correspond to the observed normalized residuals from the fitted models. The dashed lines illustrate the distribution reference line.

Thus, the fitted models based on Transformations D and E seem to provide reasonable model fits. Transformations D and E differ only through an additional constant $c \in \{\frac{1}{2}, 1\}$, which was needed to resolve the issue with zeros when using a log-transformation. Now, although these transformations are very similar, the conclusions based on the parameter estimates, shown in Table 3.1, differ. While the model based on Transformation E reveals a significant treatment difference between Tubigrip and BKC, this difference is not detected in the analysis based on Transformation D. These contradicting results are caused by the transformation of the time line: the relative distance between the time points of interest is changed through the transformation. Let t_j^D and t_j^E denote the transformed observation times based on Transformation D and Transformation E, respectively. Then we obtain

Transformation D			**Transformation E**		
$t_1 = 0$	$\mapsto$	$t_1^D = -0.69$	$t_1 = 0$	$\mapsto$	$t_1^E = 0$
$t_2 = 4$	$\mapsto$	$t_2^D = 1.50$	$t_2 = 4$	$\mapsto$	$t_2^E = 1.61$
$t_3 = 12$	$\mapsto$	$t_3^D = 2.53$	$t_3 = 12$	$\mapsto$	$t_3^E = 2.56$
$t_4 = 39$	$\mapsto$	$t_4^D = 3.68$	$t_4 = 39$	$\mapsto$	$t_4^E = 3.69.$

As we are applying a monotone transformation the relative ordering of the observation times persists for both transformations. However, the distances between adjacent time points differ. For example $t_2^D - t_1^D = 2.19$ weeks whereas $t_2^E - t_1^E = 1.61$ weeks. In contrast, the distances between the last two times points are quite similar: $t_4^D - t_3^D = 1.15$ and $t_4^E - t_3^E = 1.13$. In the CAST study, the score changes much faster at the beginning of the study than towards the end, see Figure 2.1 on page 13. Clearly two transformations with differing effects on the time line at the beginning of the study but similar effects towards the end of the study lead to varying results, see also Figure 3.3(d) and Figure 3.3(e), page 52.

3.3.1 Summary

In this section we fitted linear mixed models under ignorable missingness (see Chapter 5) to the CAST data set. Exploratory analysis, shown in Chapter 2, revealed that the FAOSS-score evolves non-linearly over time. Also, the scores approach an upper limit as time increases. We tried five different transformations to reduce the effect of the boundedness and to improve the linearity of the data with respect to covariates. Three of these transformations transform the outcome data, whereas the remaining two transformations are applied to the observations times $t_{i,j}$, $i \in \{1, ..., N\}$ and $j \in \{1, ..., M_i\}$.

While some transformations suggest that a quadratic model in time might be appropriate, other transformations seem to carry forward the boundedness of the original data. Nevertheless, we fitted linear mixed models quadratic in time and age to the transformed data. A subject-specific intercept and a subject-specific time-interaction was included. The fitted models based on the different transformations of the outcome data, i.e. Transformations A, B and C, lead to a poor model fit. In particular, the fits at the last two time points, i.e. week 12 and week 39, were unsatisfactory.

The results based on the transformations of the time line, i.e. Transformations D and E, lead to adequate model fits. However, the conclusions vary substantially dependent on the transformation. The main aim of the CAST study was to investigate the effects of covariates on the change over time. Transforming the time line may change the relative distance of the time points; and an analysis based on such a transformation may not answer the original research question. Hence, extreme care should be taken when transforming the time line.

Furthermore, all transformations lead to parameter estimates that are relatively difficult to interpret. Our analyses required graphical displays to explore the nature of existing covariate effects.

A further limitation of the transformations presented is that none of them accounts for the bounded nature of the score. We noted previously that for every subject the original score approaches a final score as time increases. Generally, recovery from acute injuries

is not always complete, so the *achieved final score* might not be on the upper bound, i.e. 100 score points. This achieved final score is of interest to investigators, as it is likely to depend on covariates, such as age and gender. These effects cannot be explored based on the Transformations A, B and C, as the bounds of the score are fixed, and are part of the specification of the transformation. Similarly, we cannot investigate these effects based on Transformations D and E, because the bounded nature of the score is simply ignored.

Overall, the listed limitations discourage the use of any of these transformations. We are aware that the list of transformations explored here is not exhaustive, e.g. Carpenter et al. [2002] propose a transformation of the time line that might be suitable. Nevertheless, we refrain from exploring any other transformation. As an alternative, we will present a model for the outcome score on the original scale as a function of covariates in Section 3.5. The model is constructed for scores where the rate of recovery changes over time. In addition, the model is able to account for the bounded nature of the score and to investigate the effect of covariates on the achieved final score.

3.4 Non-Linear Mixed Models

In the last section, we have seen that linear mixed models are not always the adequate model choice for continuous longitudinal data. In some applications the mean response is intrinsically non-linear and there exists no suitable transformation of the data or covariates to transfer the problem to a linear regression problem. In such cases, *non-linear mixed models* can be a valuable alternative.

A non-linear mixed model for normally distributed longitudinal data is based on a framework that is very similar to that of the linear mixed model. The within- and between-subject variation is accommodated in a two-stage model. In the first stage, the systematic variation is characterised through a non-linear regression model with a model for the within-individual variation. Then, the second stage accounts for the between-individual variation through the inclusion of subject-specific parameters.

Davidian and Giltinan [1995] give the following definition of a non-linear mixed model for normally distributed longitudinal data:

Stage 1:

The systematic variation of the response for subject i at time point j is modelled through

$$Y_{i,j} = g(x_{i,j}, \theta_i) + \epsilon_{i,j}, \tag{3.14}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a non-linear function and $\epsilon_{i,j}$ is the random error term reflecting uncertainty in the response. With $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, ..., \epsilon_{i,M_i})^\top$ we assume

$$\mathbb{E}(\boldsymbol{\epsilon}_i|\theta_i) = 0; \qquad \text{and} \qquad \text{Cov}(\boldsymbol{\epsilon}_i|\theta_i) = \tilde{\Sigma}_i = \tilde{\Sigma}_i(\xi_1).$$

The covariance matrix $\tilde{\Sigma}_i$ reflects the within-individual variation and depends on variance and correlation parameters which are summarised through the parameter vector ξ_1. In line with linear mixed models, it is common to assume $\tilde{\Sigma}_i(\xi_1) = \sigma^2 I_{n_i}$, $\xi_1 = \sigma^2$, for the within-individual covariance structure, see Section 3.2.3. That is, the serial correlation is neglected and equal variances are assumed.

Stage 2:

The inter-individual variation is captured through the subject-specific regression parameter θ_i. The variation of θ_i might have different sources, such as individual characteristics (treatment group, age, gender etc.). But unexplained variation, due to natural variability among subjects, might also be a source. To account for these possibilities, the parameter of interest θ_i is modelled in dependence of known fixed quantities, say θ, and a random component $\boldsymbol{U}_i$ associated with the i-th subject. A general model is then given by

$$\theta_i = d(\theta, \boldsymbol{U}_i),$$

where $d(\cdot)$ is a vector-valued function. Furthermore, we assume

$$\boldsymbol{U}_i \overset{iid.}{\sim} \mathcal{N}(0, D(\xi_2));$$

and denote all covariance parameters accounting for within- and between-subject variation by $\xi = (\xi_1^\top, \xi_2^\top)^\top$. Note that, unless $d(\cdot)$ leads to an affine transformation of U_i, we are usually not able to establish the distribution of θ_i easily. Alternative distributional assumptions for U_i, e.g. U_i is gamma distributed, can be made [Davidian and Giltinan, 2003].

3.4.1 Estimation and Inference

In Section 3.2.3, we have seen that every hierarchical model has a marginal model formulation. Maximum-likelihood estimation was based on the marginal normal distribution and closed formed solutions for the mean parameters are available. However, a marginal model formulation does not always exist in the case of non-linear mixed models. The estimation of θ and ξ is usually much more complicated, especially when the random effects $\boldsymbol{U}_i$ enter the model equation (3.14) in a non-linear fashion. Different approaches to deal with the higher complexity in the parameter estimation are proposed in the literature [Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997]. In this thesis we will focus on the maximum-likelihood approach. Alternatively, Bayesian inference [Davidian and Giltinan, 1995, Chapter 8], inference based on individual estimates [Davidian and Giltinan, 1995, Chapter 5] or on linearisation [Davidian and Giltinan, 1995, Chapter 6] can be made.

For maximum-likelihood inference we want to base inference on the marginal distribution of $\boldsymbol{Y}$, i.e.

$$L_{\boldsymbol{y}}(\theta, \xi) = \prod_{i=1}^{N} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) = \prod_{i=1}^{N} \int f_{\boldsymbol{Y}_i|\boldsymbol{U}_i}(\boldsymbol{y}_i)\, f_{\boldsymbol{U}_i}(\boldsymbol{u}_i)\, d\boldsymbol{u}_i.$$

The calculation of the likelihood involves the integration over the random-effects U_i. The complexity of this integration depends on the dimension of U_i, and on the way the random effects U_i enter the model. The integration is more complex when the random effects enter

the model non-linearly. In particular, it is usually not possible to solve this integral explicitly. Approximative integration methods, such as *Gaussian-Quadrature* are frequently used to evaluate the likelihood.

Following the numerical evaluation of the likelihood, maximum-likelihood estimates can be calculated through numerical optimization routines, e.g. *quasi-Newton* or *Newton-Raphson methods*. Inference can be based on the same principles as presented in Section 3.2.4.

3.4.2 Inference for the Random Effects

As seen in Section 3.2.5, the random effects can be estimated through *empirical Bayes* (EB) estimates. The EB estimates are calculated based on the posterior distribution of $\boldsymbol{U}_i|\boldsymbol{Y}_i$, in which all unknown parameters have been replaced by their estimates. In the linear mixed model case, $\boldsymbol{U}_i|\boldsymbol{Y}_i$, $i \in \{1, ..., N\}$, are normally distributed and the EB estimates are calculated as the expectations of these distributions. A closed form for the EB estimates was presented in Section 3.2.5. However, in the case of non-linear mixed models the marginal distribution of $\boldsymbol{Y}_i$ may be non-normal and thus the distribution of $\boldsymbol{U}_i|\boldsymbol{Y}_i$ may be non-normal as well. According to Molenberghs and Verbeke [2005], it is then customary to define the EB estimates as the posterior modes of the posterior distribution

$$f_{\boldsymbol{U}_i|\boldsymbol{Y}_i,\hat{\theta},\hat{\xi}}(\boldsymbol{u}_i|\boldsymbol{y}_i,\hat{\theta},\hat{\xi}) = \frac{f_{\boldsymbol{Y}_i|\boldsymbol{U}_i,\hat{\theta},\hat{\xi}}\,(\boldsymbol{y}_i|\boldsymbol{u}_i,\hat{\theta},\hat{\xi})\, f_{\boldsymbol{U}_i|\hat{\xi}}\,(\boldsymbol{u}_i|\hat{\xi})}{\int f_{\boldsymbol{Y}_i|\boldsymbol{U}_i,\hat{\theta},\hat{\xi}}\,(\boldsymbol{y}_i|\boldsymbol{u}_i,\hat{\theta},\hat{\xi})\, f_{\boldsymbol{U}_i|\hat{\xi}}\,(\boldsymbol{u}_i|\hat{\xi})\,\mathrm{d}u_i}.$$

3.5 Non-Linear Mixed Models and CAST

In Section 3.3 we fitted linear mixed models to the CAST data set. We investigated several transformations and included higher-order time effects. All attempts led to unsatisfactory fits. Importantly, when using these transformations we were not able to investigate covariate effects on the final recovery level. In this context, the *final recovery level* is defined as the final achieved score and should not be confused with the upper bound (100 score points for

FAOSS). This distinction is important, because recovery from acute injuries is not always complete, so the score at final recovery might not be on the actual upper bound.

Based on exploratory analysis, given in Section 2.4, we propose a non-linear mixed model for the original score data that models the rate of recovery as a function of explanatory variables, and takes the bounded nature of the score into account. The investigations in this section will base on an ignorable missingness mechanisms, see Chapter 5. This mechanism implies that valid likelihood inference can be based on the observed data only, i.e. the likelihood contribution of a given subject i is obtained by integrating out the missing values from the density of $\boldsymbol{Y}_i$.

For an individual $i \in \{1, ..., N\}$ we propose the following mixed model for the outcome vector:

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{U}_i &\overset{indep.}{\sim} \mathcal{N}_{M_i}\left(\boldsymbol{\mu}_i, \sigma^2 I_{M_i}\right); \\ \boldsymbol{U}_i &\overset{iid}{\sim} \mathcal{N}_d(\mathbf{0}, D(\xi_2)), \quad i \in \{1, ..., N\}; \quad \text{and} \\ \mu_{i,j} &= g(\check{\boldsymbol{x}}_{i,j}, s_i, t_{i,j}, \theta_i) \quad \text{for } j \in \{1, ..., M_i\}, \end{aligned}$$

where $\boldsymbol{U}_i$, a random-effect vector, has a multivariate normal distribution with mean vector zero and covariance matrix $D(\xi_2)$, I_{M_i} is the M_i-dimensional identity matrix, $g(\cdot)$ is a non-linear function, $\boldsymbol{\mu}_i = (\mu_{i,1}, ..., \mu_{i,M_i})^\top$, $\check{\boldsymbol{x}}_{i,j}$ is a sub-vector of $\boldsymbol{x}_{i,j}$ excluding the observation times ($t_{i,j}$) and randomisation groups (s_i), and θ_i is the parameter vector of interest. For convenience we omit the i-subscript for θ_i in the derivation of the non-linear model. Note that the model presented makes use of the conditional independence assumption; with the notation of the last section we have $\tilde{\Sigma}_i(\xi_1) = \sigma^2 I_{M_i}$, $\xi_1 = \sigma^2$.

The FAOSS-score increases over time and is bounded, thus motivating our proposal that the recovery rate should change over time. We expect a very low rate of recovery when patients suffer from extreme symptoms. In particular, $y_{i,j} = 0$ implies a recovery rate of zero, e.g. for the FAOSS-score worst symptoms indicate a very swollen and stiff ankle, which delays the start of recovery. Additionally, we know that the recovery rate

is zero when the final recovery level is achieved. This means that the rate of recovery at a certain time depends on the distance of the current score to the lower bound and the final recovery level. In mathematical terms, we expect the rate of improvement at t, i.e. $g'(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)$, to be proportional to the current score, $g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)$, and the still achievable score $[\lim_{r\to\infty}\{g(\check{\boldsymbol{x}}_{i,j}, s_i, r, \theta)\} - g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)]$, where $\lim_{r\to\infty}\{g(\check{\boldsymbol{x}}_{i,j}, s_i, r, \theta)\}$ denotes the final recovery level. Hence, we are interested in solving the differential equation

$$g'(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta) = \kappa_{s_i}\, g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)\, [\lim_{r\to\infty}\{g(\check{\boldsymbol{x}}_{i,j}, s_i, r, \theta)\} - g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)], \tag{3.15}$$

where $\kappa_{s_i} \geq 0$ for $s_i \in \{1, ..., 4\}$ is the treatment-specific proportion-factor. The model proposed in equation (3.15) implies that $g(\cdot)$ is a monotone increasing function of time, because we assume $\kappa_{s_i} \geq 0$ and

$$0 \leq g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta) \leq \lim_{r\to\infty}\{g(\check{\boldsymbol{x}}_{i,j}, s_i, r, \theta)\}$$

for all $t \in \mathbb{R}_0^+$. In order to solve the differential equation, we reduce the problem to $\check{\boldsymbol{x}}_{i,j} = \{\,\}$, and define $g(t) := g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta)$ and $g_{max} := \lim_{r\to\infty}\{g(\check{\boldsymbol{x}}_{i,j}, s_i, r, \theta)\}$. As long as $0 < g(t) < g_{max}$, $t \in \mathbb{R}_0^+$, we obtain

$$\begin{aligned}
& g'(t) = \frac{\mathrm{d}g(t)}{\mathrm{d}t} &&= \kappa_{s_i}\, g(t)\, [g_{max} - g(t)] \\
\Leftrightarrow \quad & \frac{1}{g(t)\,[g_{max} - g(t)]}\, \mathrm{d}g(t) &&= \kappa_{s_i}\, \mathrm{d}t \\
\Rightarrow \quad & \int \frac{1}{g(t)\,[g_{max} - g(t)]}\, \mathrm{d}g(t) &&= \int \kappa_{s_i}\, \mathrm{d}t.
\end{aligned} \tag{3.16}$$

Then

$$\frac{1}{g(t)\,[g_{max} - g(t)]} = \frac{A}{g(t)} + \frac{B}{[g_{max} - g(t)]},$$

with $A, B \in \mathbb{R}$ implies

$$A\,[g_{max} - g(t)] + B\, g(t) = 1.$$

A solution of this equation is given by $A = B = \frac{1}{g_{max}}$ and equation (3.16) simplifies to

$$\begin{aligned} & \int \frac{1}{g_{max}} \left(\frac{1}{g(t)} + \frac{1}{g_{max} - g(t)} \right) \mathrm{d}g(t) && = && \int \kappa_{s_i} \, \mathrm{d}t \\ \Leftrightarrow \quad & \int \frac{1}{g(t)} \, \mathrm{d}g(t) \; + \; \int \frac{1}{g_{max} - g(t)} \, \mathrm{d}g(t) && = && g_{max} \, \kappa_{s_i} \, t + C_1, \end{aligned} \tag{3.17}$$

where C_1 is the integration constant. Integrating the left-hand side of equation (3.17) yields

$$\begin{aligned} & \ln\left(|g(t)|\right) - \ln\left(|g_{max} - g(t)|\right) && = && g_{max} \, \kappa_{s_i} \, t + C_2 \\ \Leftrightarrow \quad & \ln\left(\left| \frac{g(t)}{g_{max} - g(t)} \right| \right) && = && g_{max} \, \kappa_{s_i} \, t + C_2 \\ \Leftrightarrow \quad & \frac{g_{max} - g(t)}{g(t)} && = && \exp\left\{ -g_{max} \, \kappa_{s_i} \, t \right\} \cdot C_3, \end{aligned} \tag{3.18}$$

where C_i, $i \in \{2, 3\}$ are real-valued constants. In order to determine the integration constant C_3, we set $t = 0$ and denote the intercept, i.e. $g(0)$, by β_0. Then,

$$C_3 = \frac{g_{max} - \beta_0}{\beta_0}.$$

Substituting C_3 into equation (3.18) and solving for $g(t)$ yields

$$g(t) = \frac{g_{max}}{\exp\left\{ -g_{max} \, \kappa_{s_i} \, t \right\} \left(\frac{g_{max}}{\beta_0} - 1 \right) + 1}.$$

The above calculations were conducted for $0 < g(t) < g_{max}$. However, we note that $g(t) = 0$ and $g(t) = g_{max}$ solve the differential equation (3.15).

For the following considerations, we define $\beta_1 := g_{max}$ and $\beta_{2,s_i} := \kappa_{s_i} \, g_{max}$. Then, the differential equation (3.15) is solved through

$$g(t) = g(s_i, t, \theta) = g(\tilde{\boldsymbol{x}}_{i,j}, s_i, t, \theta) = \frac{\beta_1}{e^{-\beta_{2,s_i} \, t} \left(\frac{\beta_1}{\beta_0} - 1 \right) + 1},$$

where $\theta = (\beta_0, \beta_1, \beta_{2,1}, ..., \beta_{2,4})^\top$. In this model β_0 denotes the intercept, β_1 the final recov-

ery level, and β_{2,s_i} the treatment specific recovery rate. However, usually the scores and the rate of recovery depend on explanatory variables. Incorporating the covariates $\check{\boldsymbol{x}}_{i,j}$ is straightforward:

$$g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta) = \frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}) \cdot t\} \left(\frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}} - 1 \right) + 1}, \quad (3.19)$$

where $\mathbb{1}(s_i \neq 1)$ is one if $s_i \neq 1$ and zero otherwise, i.e. without loss of generality we assume that interest lies in the treatment contrasts compared to the standard treatment Tubigrip, i.e. $s_i = 1$. The parameter vector characterizing the mean of Y is then given by $\theta = (\beta_0, \beta_1, \beta_{2,1}, ..., \beta_{2,4}, \alpha_0^\top, \alpha_1^\top, \alpha_2^\top)^\top$. The interpretation of all parameters is straightforward:

- $\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}$ describes the intercept, where α_0 indicates the effect of the covariates on the intercept.

- $\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}$ describes the final recovery level and α_1 the covariate effects on this recovery level. This model accounts for the bounded nature of the score. As time increases, a limiting score which varies with $\check{\boldsymbol{x}}_{i,j}$ is achieved.

- $[\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}$ indicates the rate of improvement, i.e. how fast the final recovery level is achieved. This rate depends on the randomisation group s_i and the covariates $\check{\boldsymbol{x}}_{i,j}$. For $s_i \in \{2, ..., 4\}$ the parameters β_{2,s_i} denote the contrast to the treatment slope of $s_i = 1$, i.e. β_{21}.

Note that in order to increase statistical power and the precision of treatment effect estimates, we could condition on the baseline score $Y_{i,0}$ instead of including it as component of the outcome vector. In more detail, model (3.19) with $\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j} = \beta_0 + \alpha_0\, y_{i,0}$ could be fit to the outcome vector $\tilde{\boldsymbol{Y}}_i = (Y_{i,4}, Y_{i,12}, Y_{i,39})^\top$. Although this reformulation is generally appealing, we will focus on the four-dimensional outcome vector Y_i which includes the baseline score and model (3.19).

In order to capture the inter-individual variation, we extend this model to a *nonlinear mixed model* by including the subject-specific random effect U_i. Generally, the ran-

dom effect can be incorporated in a number of ways. For the current investigations we focus on a one-dimensional random effect $U_i \sim \mathcal{N}(0, \sigma_U^2)$ which is included in an additive manner:

$$g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta_i) = \frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}) \cdot t\} \left(\frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}} - 1 \right) + 1} + U_i,$$

where $\theta_i = (\theta^\top, U_i)^\top$. That is, we assume that the intercept and the upper bound vary across patients, but not the rate of recovery.

The covariance parameter for this model is given by $\xi = (\sigma, \sigma_U)^\top$. Note that this model can easily be reformulated in terms of a multivariate normal model with a compound symmetry covariance structure:

$$\begin{aligned} \mathbf{Y}_i &\sim \mathcal{N}_{M_i}(\tilde{\mu}_i, \Sigma_i), \text{ where} \\ \tilde{\mu}_{ij} &= \frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}) \cdot t_j\} \left(\frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}} - 1 \right) + 1}, \end{aligned} \tag{3.20}$$

$\tilde{\mu}_i = (\tilde{\mu}_{i,1}, ..., \tilde{\mu}_{i,M_i})^\top$; J_{M_i} is a $M_i \times M_i$ matrix with all elements unity and

$$\Sigma_i = \sigma^2 I_{M_i} + \sigma_U^2 J_{M_i}.$$

Note that the covariance matrix of subject i depends on i only through its dimension and is a sub-matrix of the $M \times M$-dimensional overall covariance matrix Σ. The sub-matrix Σ_i has the appropriate rows and columns for the observed measurement times. The assumed covariance structure implies equal correlation of any two different measurements on the same subject regardless of the length of the time interval between these measurements. However, the design of the CAST study had unequally spaced time points, and with repeated measurements, we expect higher correlation when the measurements are closer in time than when they are further apart. Additionally, with bounded data, correlations increase as measurements reach the bounds regardless of the time interval between measurements, further

Ignorable Missingness				
Variable	Parameter	Est.	SE	p-val.
Intercept	β_0	40.92	0.79	-
Final Recovery	β_1	82.76	1.00	-
Covariate Effects on Final Recovery Level				
Age-effect	$\alpha_{1,1}$	-0.24	0.07	<0.001
Gender-effect	$\alpha_{1,2}$	-5.28	1.44	<0.001
Treatment-Specific Recovery Rates				
Rate of Tubigrip	β_{21}	0.29	0.03	-
Contrast: BKC	β_{22}	-0.11	0.04	0.004
Contrast: Aircast	β_{23}	-0.07	0.04	0.072
Contrast: Bledsoe	β_{24}	0.0005	0.03	0.989
Covariate Effects on Recovery Rate				
Age-effect	$\alpha_{2,1}$	-0.01	0.001	< 0.001
Gender-effect	$\alpha_{2,2}$	-0.06	0.03	0.034
Variance Components				
Within - Variance	σ	13.63	0.26	-
Between - Variance	σ_U	12.00	0.51	-

Table 3.2: Overview of the parameter estimates and standard errors for the outcome model (3.20) based on the assumption of an ignorable missingness process. The p-values are reported only for the components of θ and ξ that might be zero.

complicating the situation. More flexible covariance structures for bounded continuous data, and the CAST data set in particular, are discussed in Chapter 4.

In the following, we fit model (3.20) to the CAST data set, assuming an ignorable missingness process. We adjust for the randomisation group s_i, age and gender of the patients. Let $a_i = age_i - 27$ (age centered around the median) and gender $sex_i \in \{f, m\}$ (f female, m male). Then

$$\check{\boldsymbol{x}}_{i,j} = \check{\boldsymbol{x}}_i = (a_i, \mathbb{1}(sex_i = f))^\top .$$

Furthermore, $\alpha_\ell = (\alpha_{\ell,1}, \alpha_{\ell,2})^\top$ for $\ell \in \{0, 1, 2\}$. The maximum likelihood estimates for θ and ξ can be calculated through the *Newton-Raphson* method, which is implemented in the `SAS`-procedure `NLMIXED` [SAS/STAT, 1999]. The parameter estimates for the fitted model are summarized in Table 3.2. Note that a primary analysis revealed insignificant age- and gender-effects on the intercept, i.e. $\alpha_0 = \mathbf{0}$. The interpretation of all parameters is straightforward. The intercept for an average patient, where the patient-specific quantity is zero, is given by $\hat{\beta}_0 = 40.92$. The final recovery level for an average person is given by

$\hat{\beta}_1 = 82.76$, but with a negative age-effect and the final recovery level for female patients is on average approximately 5 score points lower than for male patients. Using these point estimates and allowing for the age range of 16 to 72 implies that the final recovery level for male patients varied between 72 and 85 score points, whereas for an average female patient the final recovery level lay between 67 and 80. Furthermore, we observe a negative age effect on the rate of improvement, i.e. older participants recovered less quickly than younger patients. In addition, female patients recovered less quickly than male patients as $\hat{\gamma}_2 < 0$. Generally this means that the final recovery level for older and female patients was lower than for younger and male patients. In particular, this implies that older and female patients were less likely to recover completely from a severe sprain with the treatment options tested in this trial.

The standard deviations reflecting the within- and between-patient variations are of the same magnitude.

Regarding the treatment comparison, a significantly higher recovery rate of BKC compared to Tubigrip is detected. The rate of recovery for Aircast brace was only marginally higher than for Tubigrip. There was no significant difference in recovery rates between Tubigrip and Bledsoe boot. The fitted curves for the different randomisation groups for average male patients of age 37 or 66 are shown in Figure 3.5(a) and Figure 3.5(b) respectively. Independent of the randomisation group, patients ended at the same score. However, the rate at which they achieved the final recovery level differed substantially, in particular for older patients. Note that the fitted curves for Tubigrip and Bledsoe are indistinguishable due to the insignificant treatment difference. The dependence of the recovery rate and the final recovery level on age and gender is visualized for the randomisation group Tubigrip in Figure 3.5(c). In order to capture the interplay between the covariate effects of age, gender and randomisation groups with the bounded nature of the score, we present the average improvements between two adjacent time points dependent on these covariates, see Table 3.3. The estimates for the improvements underline the large effect of age and gender on the recovery. Comparing the score gain for 16 year old and 66 year old patients shows that

Age	Gender	Treatment	weeks 0-4		weeks 4-12		weeks 12-39	
			Est.	CI	Est.	CI	Est.	CI
16 years	female	Tubigrip	20.3	[15.9,24.8]	16.1	[13.1,19.1]	2.4	[0.1,4.6]
		BKC	26.2	[22.4,30.0]	12.0	[8.6,15.5]	0.6	[-0.03,1.2]
		Aircast	23.8	[20.1,27.5]	13.9	[10.8,17.0]	1.1	[0.1,2.0]
		Bledsoe	20.4	[16.3,24.5]	16.1	[13.2,18.9]	2.3	[0.3,4.4]
	male	Tubigrip	26.2	[21.9,30.4]	16.6	[13.1,20.1]	1.4	[0.1,2.7]
		BKC	32.0	[28.5,35.4]	11.8	[8.5,15.1]	0.3	[0.01,0.7]
		Aircast	29.7	[25.9,33.4]	13.8	[10.4,17.3]	0.6	[0.04,1.2]
		Bledsoe	26.2	[23.0,29.5]	16.5	[13.9,19.2]	1.4	[0.4,2.3]
21 years	female	Tubigrip	18.4	[14.4,22.5]	16.2	[13.9,18.5]	1.3	[0.5,5.6]
		BKC	24.4	[20.8,28.1]	12.4	[9.2,15.6]	0.4	[0.01,1.5]
		Aircast	22.0	[18.5,25.5]	14.2	[11.5,16.9]	0.6	[0.2,2.5]
		Bledsoe	18.5	[14.6,22.3]	16.1	[13.9,18.4]	1.3	[0.6,5.4]
	male	Tubigrip	24.2	[20.2,28.2]	17.0	[14.0,19.9]	0.8	[0.3,3.3]
		BKC	30.2	[26.8,33.5]	12.3	[9.2,15.4]	0.2	[0.03,0.8]
		Aircast	27.8	[24.1,31.5]	14.3	[11.1,17.6]	0.4	[0.1,1.5]
		Bledsoe	24.3	[21.1,27.4]	16.9	[14.6,19.2]	0.6	[0.7,2.9]
27 years	female	Tubigrip	16.2	[12.5,19.8]	16.0	[14.4,17.6]	4.0	[1.1,7.0]
		BKC	22.3	[18.8,25.9]	12.8	[9.9,15.7]	1.0	[0.1,2.0]
		Aircast	19.9	[16.5,23.2]	14.5	[12.2,16.8]	1.9	[0.4,3.3]
		Bledsoe	16.2	[12.5,19.9]	16.0	[14.3,17.6]	4.0	[1.0,7.0]
	male	Tubigrip	21.8	[18.1,25.6]	17.3	[14.9,19.6]	2.4	[0.6,4.2]
		BKC	28.0	[24.7,31.4]	12.9	[9.9,15.9]	0.6	[0.1,1.1]
		Aircast	25.6	[21.8,29.3]	14.9	[11.8,17.9]	1.1	[0.1,2.1]
		Bledsoe	21.9	[18.8,25.0]	17.3	[15.3,19.2]	2.4	[1.0,3.8]
37 years	female	Tubigrip	12.4	[9.5,15.4]	14.9	[13.8,16.1]	6.4	[2.8,9.9]
		BKC	18.9	[15.4,22.4]	13.3	[10.8,15.8]	1.7	[0.2,3.1]
		Aircast	16.3	[13.0,19.6]	14.5	[12.7,16.3]	3.0	[0.8,5.2]
		Bledsoe	12.5	[8.8,16.2]	15.0	[13.8,16.1]	6.3	[1.9,10.8]
	male	Tubigrip	17.9	[14.4,21.4]	17.3	[15.7,18.9]	3.9	[1.4,6.5]
		BKC	24.4	[20.8,28.0]	13.7	[10.8,16.7]	1.0	[0.1,1.9]
		Aircast	21.8	[17.8,25.9]	15.5	[12.7,18.4]	1.8	[0.2,3.4]
		Bledsoe	18.0	[14.6,21.3]	17.3	[15.7,18.9]	3.9	[1.4,6.4]
66 years	female	Tubigrip	2.3	[-1.3,5.9]	4.3	[-1.9,10.5]	10.9	[2.6,19.1]
		BKC	9.2	[4.4,13.9]	11.4	[9.0,13.9]	6.3	[-0.3,12.8]
		Aircast	6.3	[1.3,11.3]	9.7	[5.1,14.3]	10.1	[2.2,18.1]
		Bledsoe	2.4	[-3.3,8.0]	4.4	[-5.2,14.0]	11.0	[-1.3,23.3]
	male	Tubigrip	6.9	[1.9,11.8]	11.2	[5.8,16.5]	13.1	[5.2,21.0]
		BKC	14.1	[8.5,19.6]	14.2	[11.3,17.1]	4.1	[-0.9,9.0]
		Aircast	11.1	[4.7,17.5]	14.0	[11.3,16.8]	7.1	[-1.3,15.5]
		Bledsoe	6.9	[0.7,13.1]	11.2	[4.7,17.7]	13.0	[3.3,22.7]

Table 3.3: Overview of the average improvements (Est.) between two adjacent time points for different age groups, genders and randomisation groups. The age classes were classified according to the first percentile (16 years), the first (21 years), second (27 years) and third (37 years) quantiles and the 99th percentile (66 years). The confidence intervals are denoted by CI.

(a) Recovery for $age = 37$ years (b) Recovery for $age = 66$ years (c) Recovery for Tubigrip

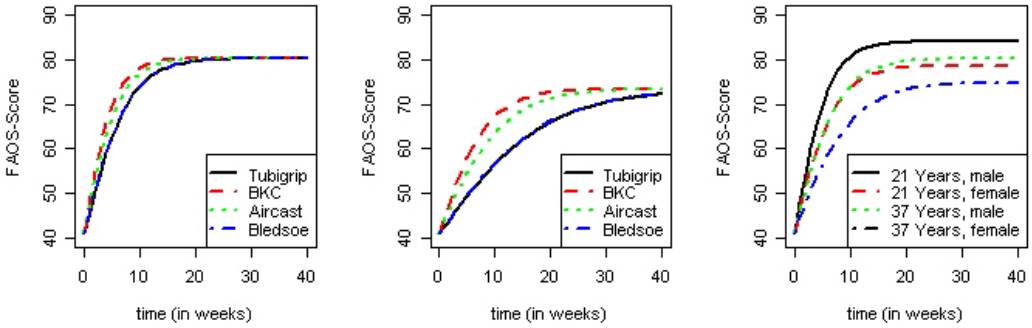


Figure 3.5: a) Fitted recovery curve versus time for the different randomisation groups and 37 year old male patients. b) Fitted recovery curve versus time for the different randomisation groups and 66 year old male patients. c) Fitted FAOS-score versus time for different genders, age classes and Tubigrip. The age groups were classified according to the first (21 years) and third (37 years) quantiles.

older patients recovered much more slowly than young patients. Furthermore, the effect of the bounded score stands out. In the last time interval, the improvement was generally much smaller than in the previous intervals. However, the effect of the bounded score depended also on age and gender.

For an average person, i.e. $\hat{U}_i = 0$, we are able to calculate the expected time to achieve a certain score S based on the fitted curves. For this purpose we need to rearrange equation (3.19) to solve for time $t(S)$:

$$t(S) = \frac{-1}{[\beta_{21} - \beta_{2,s_i}\,\mathbb{1}(s_i \neq 1)] + \alpha_2\, a_i + \gamma_2\,\mathbb{1}(sex_i = f)} \times \log\left(\frac{\beta_1 + \alpha_1\, a_i + \gamma_1\,\mathbb{1}(sex_i = f) - S}{\left(\frac{\beta_1 + \alpha_1\, a_i + \gamma_1\,\mathbb{1}(sex_i = f)}{\beta_0} - 1\right) \cdot S}\right)$$

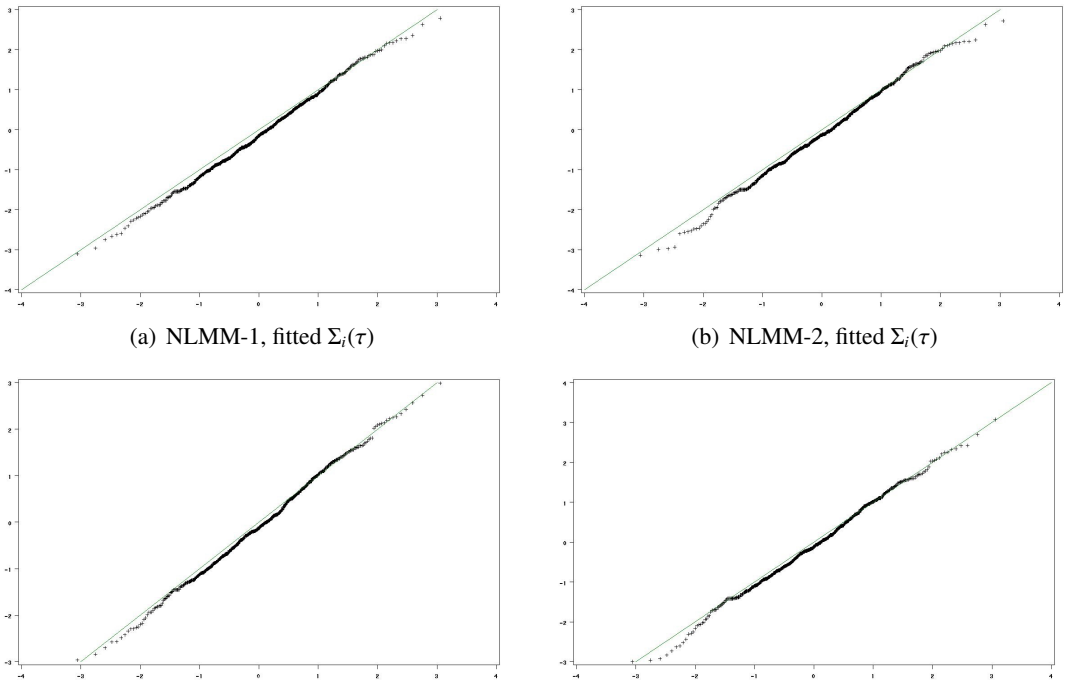
The estimated times to achieve a score of $S = 65$ are shown in Table 3.4. While a 16 year old male patient under Tubigrip needed approximately 4 weeks to achieve a score of 65, a man at the age of 37 years needed on average 2 weeks longer. The difference for 66 year old patients is even more drastic. However, care should be taken in reporting these results

		female		male	
Age	Treatment	Weeks	CI	Weeks	CI
16 years	Tubigrip	5.0	[3.6,6.3]	3.6	[2.8,4.3]
	BKC	3.5	[2.8,4.3]	2.7	[2.2,3.1]
	Aircast	4.0	[3.2,4.9]	3.0	[2.4,3.5]
	Bledsoe	4.9	[3.7,6.2]	3.5	[3.0,4.1]
21 years	Tubigrip	5.6	[4.1,7.1]	3.9	[3.1,4.7]
	BKC	3.9	[3.0,4.7]	2.9	[2.4,3.4]
	Aircast	4.5	[3.5,5.4]	3.3	[2.7,3.9]
	Bledsoe	5.6	[4.2,7.0]	3.9	[3.3,4.5]
27 years	Tubigrip	6.7	[4.9,8.4]	4.5	[3.6,5.4]
	BKC	4.4	[3.4,5.4]	3.2	[2.6,3.8]
	Aircast	5.2	[4.1,6.3]	3.7	[2.9,4.4]
	Bledsoe	6.6	[4.9,8.4]	4.5	[3.7,5.2]
37 years	Tubigrip	9.3	[6.8,11.8]	5.8	[4.5,7.1]
	BKC	5.6	[4.2,7.0]	3.9	[3.1,4.7]
	Aircast	6.8	[5.1,8.5]	4.5	[3.4,5.6]
	Bledsoe	9.3	[6.3,12.2]	5.7	[4.5,7.0]
66 years	Tubigrip	73.5	[-40.0,186.9]	18.4	[4.4,32.4]
	BKC	17.1	[4.7,29.5]	8.3	[4.2,12.5]
	Aircast	25.6	[2.2,49.1]	10.9	[3.9,18.0]
	Bledsoe	71.5	[-96.3,239.4]	18.2	[1.4,35]

Table 3.4: An overview of the expected number of weeks to reach a score of 65 for different genders, age groups and the four randomisation groups.

as for 66 year old female patients the maximum achievable score was nearly 68, which is close to $S = 65$ and thus leads to an imprecise estimation. This is reflected in the wide confidence intervals which even include negative values. In general, however, we believe that this information, together with the ability to quantify the expected final recovery level per age band and gender, could be of particular interest to patients.

Finally, the adequacy of the model for the mean response is assessed via model diagnostic plots (Figure 3.6, page 71) presented by Park and Lee [2004] and used in Section 3.3. We show the Q-Q plot for model (3.20) and in order to enable a comparison with the results in Section 3.3, we present also the results based on model (3.20) with $\alpha_{1,2} = \alpha_{2,2} = 0$, i.e. we do not adjust for the gender of patients. We denote these fitted non-linear mixed models by NLMM-1 (model (3.20)) and NLMM-2 (model (3.20) without gender), respectively. The plots using the fitted covariance matrix show small signs of skewness. However, using the



(a) NLMM-1, fitted $\Sigma_i(\tau)$

(b) NLMM-2, fitted $\Sigma_i(\tau)$

(c) NLMM-1, sample $\Sigma_i(\tau)$

(d) NLMM-2, sample $\Sigma_i(\tau)$

Figure 3.6: Q-Q plots for the fitted non-linear mixed models using the normal distribution. NLMM-1 corresponds to model (3.20) and NLMM-2 to model (3.20), where we do not adjust for gender. The results based on fitted and sample covariance matrtices are contrasted. The horizontal axes correspond to the expected quantiles of the standard normal distribution, whereas the vertical axes correspond to the observed normalized residuals from the fitted models. The dashed lines illustrate the distribution reference line.

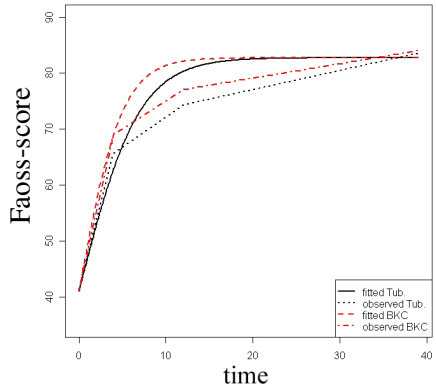
sample covariance matrix reveals an adequate fit of the models for the mean response. Note that independent of the chosen covariance matrix, the non-linear mixed models presented lead to better model fits than any transformation explored in Section 3.3. In addition, we contrast the observed average score evolution for Tubigrip and BKC with the fitted mean response profiles. For NLMM-1 the observed average evolutions of men are contrasted with the fitted response curves of 27 year old, male patients, see Figure 3.7(a), page 73. In the case of NLMM-2, the observed average evolutions of all patients are compared to the fitted mean response curves of 27 year old patients, Figure3.7(b), page 73. These plots reveal a relatively poor model fit at the twelfth week time point. This is caused by comparing the fitted curves of 27 year old patients to the observed evolutions of *all* patients. To present a fairer comparison, we predict the FAOSS-scores based on the presented models

and compare these to the observed average score evolutions. In this context, we can predict the outcomes of a given subject through the population-average mean response model (via $\tilde{\mu}_{i,j}$, see equation (3.20)) or through the subject-specific mean response model (via $\tilde{\mu}_{i,j} + \hat{U}_i$, where $\hat{U}_i$ is the EB estimate, see Section 3.4.2). Graphical displays of these comparisons are given in Figure 3.7, page 73. The NLMM-1 and NLMM-2 model are shown to predict the mean responses accurately. In this context, the predictions based on the subject-specific mean response models (Figure 3.7(e) and Figure 3.7(f)) are shown to be more accurate than those based on the population-average mean response models (Figure 3.7(c) and 3.7(d)). Overall, the non-linear mixed models lead to a satisfying fit.

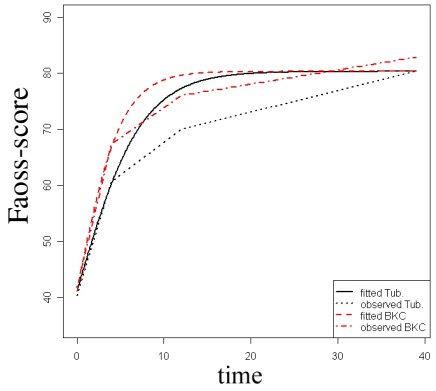
3.5.1 Summary

We have proposed a model for continuous, bounded longitudinal data which enables the investigation of covariate effects on the rate of change and the final recovery level. The model belongs to the class of non-linear mixed models which have found many biological applications, such as pharmacokinetic analysis, rate of clearance of a drug, studies of growth to adult size and decay [Davidian and Giltinan, 1995, 1993; Vonesh and Chinchilli, 1997; Davidian and Giltinan, 2003]. However, to the best of our knowledge they are not yet used in the health-care context.

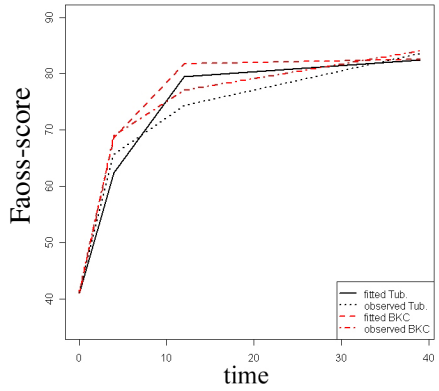
Although this specific model was motivated by the CAST study, we believe that it can be suitable for the analysis of other studies. In health-care and medical research, it is very common to measure physical or mental ability repeatedly over time, through questionnaires or scales. Based on the answers, scores can be derived for the study points. In many applications, these scores will have finite range, where one bound indicates 'no symptoms' and the other bound 'extreme symptoms'. Examples are the *Barthel Index* [Mahony and Barthel, 1965], the *Neck Disability Index* [Vernon and Mior, 1991], the *Foot and Ankle Outcome Score* (FAOS) [Roos et al., 2001] and visual analogue scales. In studies where we expect most patients to recover, we often observe that later measurements are clustered towards one end of the range. In this case, different patients might have the same initial and



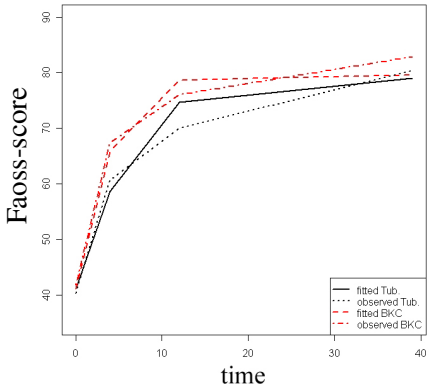
(a) NLMM-1, fitted curve



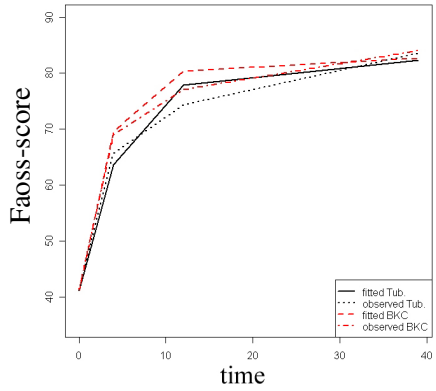
(b) NLMM-2, fitted curve



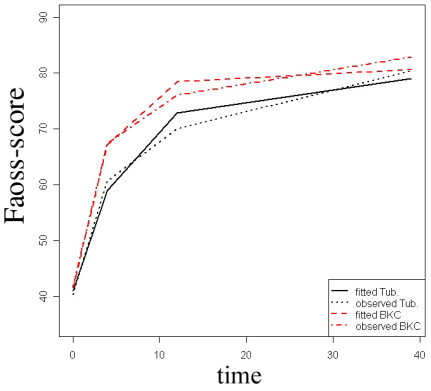
(c) NLMM-1, predicted, no $\hat{U}_i$



(d) NLMM-2, predicted, no $\hat{U}_i$



(e) NLMM-1, predicted, with $\hat{U}_i$



(f) NLMM-2, predicted, with $\hat{U}_i$

Figure 3.7: Observed average response profiles versus fitted response profiles for Tubigrip and BKC. For NLMM-1 the observed average evolutions of men are contrasted with the fitted response curves of 27 year old, male patients (Figure 3.7(a)). For NLMM-2, the observed average evolutions of all patients are compared to the fitted mean response curves of 27 year old patients (Figure 3.7(b)). In addition, the observed average response profiles are contrasted with the predicted mean response evolutions, where the predictions are calculated based on the population-average mean response models (without $\hat{U}_i$, Figure 3.7(c) & (d)) or based on the subject-specific mean response models (with $\hat{U}_i$, Figure 3.7(e) & (f)).

the same final scores. However, the rate at which they achieve the final score might differ substantially depending on explanatory variables, for example, treatment or age. The scores at final recovery can also be of scientific interest, as they can differ substantially between different ages and genders.

For a continuous and bounded score, the classical approach is to transform the data so that a linear regression model fits adequately. We claim that this approach, for example by using the log or logit transformation does not always resolve the problem of a non-linear relationship between the outcome score and covariates. We investigated several transformations for the CAST study, but a non-linear relation with time persisted due to the bounded nature of the outcome. Also the inclusion of higher order time effects did not lead to a satisfying fit. Importantly, using transformations we were no longer able to investigate covariate effects on the final recovery level, and the interpretation of covariate effects on the original score was severely complicated. For all these reasons, we believe that non-linear mixed models can be a valuable alternative.

In the specific case of the CAST study, the original analysis did not distinguish between the within- and between-individual variation [Lamb et al., 2009; Cooke et al., 2009]. Additionally, the repeated measurements were not modelled jointly: the treatment differences were investigated for every time point separately and it was not clear how to combine these estimates into a single inference. This situation was even more complicated as the data were clustered towards the end of the trial. Discrimination between treatments at these time points was practically impossible. Moreover, no satisfactory description of the score evolution over time for different age groups and genders was presented.

Our model was derived on medical research grounds and reflects the knowledge of experts in the specific research area. It enables a very flexible incorporation of exploratory variables and is easy to interpret, which is a valuable advantage over the alternative of data transformation. In addition, it accounts for the two different sources of variation and enables us to model the final recovery, which might be of particular interest to patients.

Fitting this model to the CAST data revealed that recovery was more rapid with

BKC than with Tubigrip. These results coincide with those of the original analysis, but add to it by allowing the estimation of the time to recovery in each of the groups. The results of this analysis re-enforce the results of the original CAST trial, insofar as it provides further that interventions that immobilise the ankle, such as below knee plaster, are more effective than those which permit early movement. These conclusions are contrary to previous studies [Bridgman et al., 2003]:'There is some evidence that interventions that immobilise the ankle, such as below knee plaster, are less effective than those which permit early movement.' Taking into account the considerable variation in the costs for each treatment, these results might be relevant for the UK National Health Service. 'The unit costs of devices and the labour to fit them is estimated to be, per each patient, £60 for boots (Bledsoe), £30 for braces (Aircast), £12 for below knee plaster casts, and £2 for Tubigrip (from retailers list prices, estimates of nurse time to fit devices in a pilot study, and standard NHS nursing costs [...])' [Bridgman et al., 2003].

Further, we show that older and female patients recovered substantially more slowly than younger and male participants. Also the scores at final recovery for older female patients were lower than for young male patients, suggesting that older female patients were less likely to recover completely from an acute soft tissue injury. We translated these findings into auxiliary information, such as the expected time to achieve a certain score for different patient groups.

Although we believe that our model is superior to standard analysis techniques applied in this field, we note that the model is limited by the covariance structure it assumed for the outcome vector. We worked with a compound symmetry covariance structure, which implied equal correlation of any two different measurements on the same subject regardless of the length of the time interval between these measurements. We will relax this assumption and discuss more suitable covariance structures for bounded, continuous longitudinal data in Chapter 4. Furthermore, we note that our analysis assumed an ignorable missingness process. A sensitivity analysis scrutinizing this assumptions will be presented in Chapter 5 and Chapter 6.

Chapter 4

Modelling the Covariance Matrix for Continuous Bounded Longitudinal Data

4.1 Introduction

The aim of the CAST study was to compare the effectiveness of four ankle sprain treatments, where effectiveness was measured in terms of the mean response vector. Thus, the scientific focus lies on the mean response, and in Chapter 3 we derived a model for the mean response vector while treating the covariance structure as a nuisance. Indeed, most longitudinal studies focus on regressing the mean response on available covariates, while the covariance structure is thought to be a 'nuisance parameter' [MacKenzie and Pan]. However, as noted by MacKenzie and Pan, in the case of Gaussian data the inferential challenge is symmetrical in mean and covariance. A misspecification of the covariance structure is able to bias inference for the mean vector of interest [Fitzmaurice et al., 2004], e.g. due to confidence intervals that are too narrow or wide and p-values that are too small or large. Therefore, the covariance structure should not be seen as a nuisance, but as an additional object which needs to be specified.

In the context of generalized estimating equations, standard errors for the mean vector of interest can be based on the so-called *sandwich* estimators [Liang and Zeger, 1986], which are robust to misspecifications of the covariance matrix [Fitzmaurice et al., 2004]. In principal, it is also possible to base inference for linear mixed models on the sandwich estimator. However, using generalized estimating equations can lead to great loss of efficiency for mean parameters. Also inference may be biased for partially observed data sets when the reason for data being missing is related to the outcome of interest, see Chapter 5. For these reasons we want to focus on likelihood-based inference.

In Section 3.2, we discussed several approaches to account for the correlation among the repeated measures in longitudinal studies. We presented the approach of full multivariate models, where the dependence structure is modelled through the covariance matrix; and the random-effect approach, where the inclusion of subject-specific random effects induces the correlation structure. As Laird and Ware [1982] note, full multivariate models are more flexible in modelling the covariance matrix than random-effect models. However, this comes with the cost of more parameters and decreased power in small samples.

According to the scientific literature, there are three broad approaches to modelling the covariance structure for full multivariate models. Firstly, an *unstructured covariance* matrix can be fitted to the data [Fitzmaurice et al., 2004]. The great advantage is that no assumptions about variances and covariances are made. However, fitting models with this covariance structure can be challenging, because the estimated matrix must be positive semi-definite. Different matrix decompositions which enable unconstrained parametrizations have been proposed by Pinheiro and Bates [1996]. A further challenge of the unstructured covariance models is that the number of covariance parameters, e.g. for a balanced study design: $\frac{M(M+1)}{2}$, increases rapidly with the number of observations times M. In particular, the covariance matrix may feature a structure that can be appropriately captured by fewer parameters, which themselves can be estimated more accurately than the unconstrained covariance matrix [Diggle et al., 1996]. This leads to the second approach, the so-called *covariance pattern models* which impose a structure on the covariance matrix.

Depending on the accuracy of the imposed structure, the precision of the estimation can be improved. Some of the most widely used covariance pattern models for longitudinal data are *compound symmetry* (see Section 3.5), *Toeplitz*, *autoregressive*, *banded* and *exponential* covariance models. For details regarding these covariance structures we refer to Fitzmaurice et al. [2004] and Hedeker and Gibbons [2006]. In practice, a covariance pattern model is selected from a set of candidate structures via a model selection criterion, e.g. likelihood ratio test, *Akaike* or *Bayesian information criterion* [Hedeker and Gibbons, 2006]. Pan and MacKenzie [2006] refer to this approach as *menu selection* and note that the choice out of a finite set of covariance structures may be not optimal. Alternatively, a data-driven regression modelling approach for the covariance matrix was proposed by Pourahmadi [1999]. This approach bases on a modified *Cholesky* decomposition and an unconstrained parameterization of the covariance matrix. We will discuss this methodology in Section 4.3.

Note that the presented covariance models for full multivariate models are generally not capable of distinguishing between-subject heterogeneity from within-subject heterogeneity. If both sources of variation or random-effect models are of scientific interest, a covariance structure can indirectly be imposed by the inclusion of random effects, see for example Section 3.5. However, the set of covariance structures that can be implied by random-effect models is limited, see Section 3.2. Increased flexibility may be reached by using one of the covariance model approaches for full-multivariate models to specify the within-individual variation matrix $\tilde{\Sigma}$. We note that it can be impossible to derive the marginal covariance matrix $\Sigma_i(\tau)$ when random effects enter the mean model equation in a non-linear way.

In this chapter, we will investigate suitable covariance structures for bounded, continuous longitudinal data. These investigations are motivated by the CAST study and we aim to explore the sensitivity of the inference for mean parameters based on different covariance structures. In this context, the mean parameters will be estimated based on the non-linear mixed model proposed in Section 3.5. An ignorable missingness mechanism will be assumed, that is valid likelihood-inference can be based on the observed data only,

see Chapter 5.

Before fitting appropriate models to the CAST data set, we explore the empirical covariance matrix, Cov_{emp}, and the empirical correlation matrix, $Corr_{emp}$, of the four dimensional response vectors Y_i, $i \in \{1, ..., N\}$. As we are confronted with missing data, we calculate these matrices based on the pairwise available cases [Little and Rubin, 2002]:

$$Cov_{emp} = \begin{pmatrix} 234.2 & 101.6 & 86.2 & 73.9 \\ & 389.5 & 257.8 & 196.2 \\ & & 404.6 & 260.0 \\ & & & 365.2 \end{pmatrix} \quad \text{and} \quad Corr_{emp} = \begin{pmatrix} 1 & 0.34 & 0.28 & 0.25 \\ & 1 & 0.65 & 0.52 \\ & & 1 & 0.68 \\ & & & 1 \end{pmatrix}. \tag{4.1}$$

We see that correlations decay with increasing time lags along the rows. Along the superdiagonals, however, the correlations increase with time due to the bounded nature of the score. Furthermore, the variances appear to increase with time until the last observation time is reached. This characteristic arises because most people recover towards the end of the study, and are thus clustered at the upper bound of the score, see Figure 2.2, page 14. Overall, the covariance structure implied by Cov_{emp} and $Corr_{emp}$ suggests an unusual covariance matrix, where none of the widely used covariance pattern models accounts for such a structure.

The most flexible approach to obtain a suitable structure would be to fit an unstructured covariance matrix. However, this structure involves eight additional parameters compared to the compound symmetry structure fitted in Section 3.5. Due to the models' complexity and convergence problems we were not able to fit this model. In a second attempt, we fit various covariance pattern models and covariance models induced by the inclusion of random effects. None of the fitted models is able to account for the features discussed previously. Therefore, we adopt the regression modelling approach presented by Pourahmadi [1999] and discussed in Pan and MacKenzie [2003, 2006]. This methodology is extended by allowing for ignorable missingness, see Chapter 5. Furthermore, instead of a polynomial model for the mean vector we fit the non-linear mixed model presented in Section 3.5. Results based on different model selection tools and details of our investigations

are given in the next sections.

4.2 First Attempts to Model the Covariance Structure

The non-linear mixed model introduced in Section 3.5 is of the form

$$\begin{aligned} \mathbf{Y}_i &\sim \mathcal{N}_{M_i}(\tilde{\mu}_i, \Sigma_i), \text{ where} \\ \tilde{\mu}_{ij} &= \frac{\beta_1 + \alpha_1^\top x_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\,\mathbb{1}(s_i \neq 1)] + \alpha_2^\top x_{i,j}) \cdot t_j\}\left(\frac{\beta_1 + \alpha_1^\top x_{i,j}}{\beta_0 + \alpha_0^\top x_{i,j}} - 1\right) + 1}, \end{aligned} \tag{4.2}$$

$i \in \{1, ..., N\}$, $j \in \{1, ..., M_i\}$; and the covariance matrix of subject i, i.e. Σ_i, depends on i only through its dimension and is a sub-matrix of the $M \times M$-dimensional overall covariance matrix Σ. Here we are interested in specifying this overall covariance matrix Σ. In this context we focus on model (4.2) with $x_{i,j} = \ln(age_i)$. We admit that the transformation of age is not consistent with the approach taken in Section 3.5; it reflects the development of the work throughout the course of the PhD.

Due to the bounded nature of the score, we argue that suitable covariance structures ought to incorporate a dependence on the mean score and the time lags. Amongst others we investigated the following covariance structures, assuming an ignorable missingness process, see Chapter 5.

Let J_M be a square matrix with all elements unity and $t_j \in \{0, 4, 12, 39\}$ for $j \in \{1, 2, 3, 4\}$. We further define the following two symmetric matrices:

$$A = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ & \sigma_2^2 & 0 & 0 \\ & & \sigma_3^2 & 0 \\ & & & \sigma_4^2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} & \rho^{t_4-t_1} \\ & 1 & \rho^{t_3-t_2} & \rho^{t_4-t_2} \\ & & 1 & \rho^{t_4-t_3} \\ & & & 1 \end{pmatrix},$$

with $\sigma_k^2 \in \mathbb{R}^+$, $k \in \{1,2,3,4\}$ and $\rho \in [-1,1]$. Based on these matrices we define

$$\Sigma_1 := A + \tau^2 B + D^2 J_M, \tag{4.3}$$

where $\tau^2, D^2 \in \mathbb{R}^+$. This covariance matrix accounts for correlations that vary with the lags of time through the matrix B. Furthermore, it allows for heteroscedasticity. Fitting model (4.2) with the covariance matrix $\Sigma = \Sigma_1$ yields that the hypothesis $D^2 = 0$ is not rejected at a significance level of 0.05. Moreover, $\hat{\rho} = 0.996$ with standard error 0.004. The corresponding fitted covariance and correlation matrices are given by

$$\hat{\Sigma}_{fit,1} = \begin{pmatrix} 340.9 & 151.4 & 146.6 & 131.5 \\ & 330.9 & 149.0 & 133.6 \\ & & 323.9 & 138.0 \\ & & & 311.2 \end{pmatrix} \quad \text{and} \quad \hat{Corr}_{fit,1} = \begin{pmatrix} 1 & 0.45 & 0.44 & 0.40 \\ & 1 & 0.46 & 0.42 \\ & & 1 & 0.43 \\ & & & 1 \end{pmatrix},$$

such that the correlations along the sub-diagonals and the variances decrease with time; these features contradict the structure seen and expected for bounded continuous longitudinal data.

In an attempt to recover the structure seen in equation (4.1), we tried to fit an alternative covariance model, where the variances allow for the bounded nature of the data. We assume the variances depend on the current average score, $\bar{y}_j$, and the still achievable average score, $\left(100 - \bar{y}_j\right)$; $\bar{y}_j$ is the average score of the observed scores at observation time t_j, $j \in \{1,2,3,4\}$. The matrix A in equation (4.3) is replaced by $\tilde{A} = diag(\tilde{a}_{11}, \tilde{a}_{22}, \tilde{a}_{33}, \tilde{a}_{44})$ where $\tilde{a}_{kk} = \bar{y}_k\ (100 - \bar{y}_k)$. However, this attempt failed due to convergence problems.

We fitted various other, simpler covariance pattern models, but none of the fitted models recovered the structure seen in the empirical covariance and correlation matrices, see equation (4.1). As noted in the introduction to this chapter, these methods model the covariance matrix in terms of fixed effects, i.e. the covariance of the marginal response vector is specified. Alternatively, we could focus on the mixed model formulation as the

random effects induce a covariance structure after marginalization. Admittedly, the covariance matrix defined through equation (4.3) could arise from a random-effect model where one-dimensional subject-specific random effects $U_i \sim \mathcal{N}_M(0, D^2)$ enter the mean model linearly and the within-individual variation is modelled through $\tilde{\Sigma}_i = A + \tau^2 B$.

In Section 3.5, we fitted a non-linear mixed model with compound symmetry covariance structure. We discussed why this structure is not suitable and will discuss an alternative random-effect model that allows the covariances to depend on the time lags:

$$\begin{aligned} Y_i|U_i &\overset{ind.}{\sim} \mathcal{N}_4\left(\mu_i, \sigma^2 I_4\right) \quad \text{and} \quad U_i \overset{ind.}{\sim} \mathcal{N}(0, D^2) \\ \text{with} \quad \mu_{ij} &= \tilde{\mu}_{ij} + w_{ij}^\top U_i \quad \text{and} \quad w_{ij} = (1 + t_j)^\phi. \end{aligned}$$

In particular, we make use of the conditional independence assumption. The resulting covariance matrix for the marginal outcome vector Y_i is then given by

$$\Sigma_2 := \sigma^2 I + D^2 \begin{pmatrix} 1 & 5^\phi & 13^\phi & 40^\phi \\ & 25^\phi & 65^\phi & 200^\phi \\ & & 169^\phi & 520^\phi \\ & & & 1600^\phi \end{pmatrix} \tag{4.4}$$

and we estimate $\hat{\sigma}^2 = 176.21\,(\pm 6.9)$, $\hat{D} = 100.06\,(\pm 13.15)$ and $\hat{\phi} = 0.131\,(\pm 0.02)$. Thus,

$$\hat{\Sigma}_{fit,2} = \begin{pmatrix} 276.3 & 123.6 & 140.1 & 162.4 \\ & 328.9 & 173.1 & 200.6 \\ & & 372.4 & 227.4 \\ & & & 439.8 \end{pmatrix} \quad \text{and} \quad \hat{Corr}_{fit,2} = \begin{pmatrix} 1 & 0.41 & 0.44 & 0.47 \\ & 1 & 0.49 & 0.53 \\ & & 1 & 0.56 \\ & & & 1 \end{pmatrix}.$$

The correlations along the rows and sub-diagonals increase with time, so do the variances.

Turning to the parameter vector θ, we note that the estimates $\hat{\theta}$ based on Σ_1, Σ_2 and compound symmetry (CS) vary slightly, see Table 4.1, page 84. The intercepts, treatment-specific recovery rates and the age-effects on the rates of recovery are practically indis-

tinguishable for all covariance structures. Although the standard errors vary slightly, the inference remains the same for all models investigated. However, it is noticeable that the final recovery level and the age-effect on this quantity are considerably different when using Σ_2.

From the two examples, Σ_1 and Σ_2, we see that modelling the covariance structure for bounded longitudinal data is very challenging. While the correlations along the rows decrease because the serial correlations decrease, the correlations along the sub-diagonals increase due to the bounded nature of the score. Furthermore, the variances are not constant over time. In order to account for these features, we exploit a more flexible and direct approach introduced in Pourahmadi [1999].

4.3 Data-Driven Regression Modelling Approach

In this section we present a covariance model that uses the modified Cholesky decomposition of the covariance matrix and which has a nice regression interpretation. It was first presented by Pourahmadi [1999] and refined by Pan and MacKenzie [2003].

Based on derivations in Pourahmadi [1999], we know that there exists a unique lower triangular matrix F_i with 1's as main diagonal entries and a unique diagonal matrix Q_i, such that the marginal covariance matrix Σ_i of Y_i can be decomposed to

$$\Sigma_i = F_i \, Q_i \, F_i^{\top}. \tag{4.5}$$

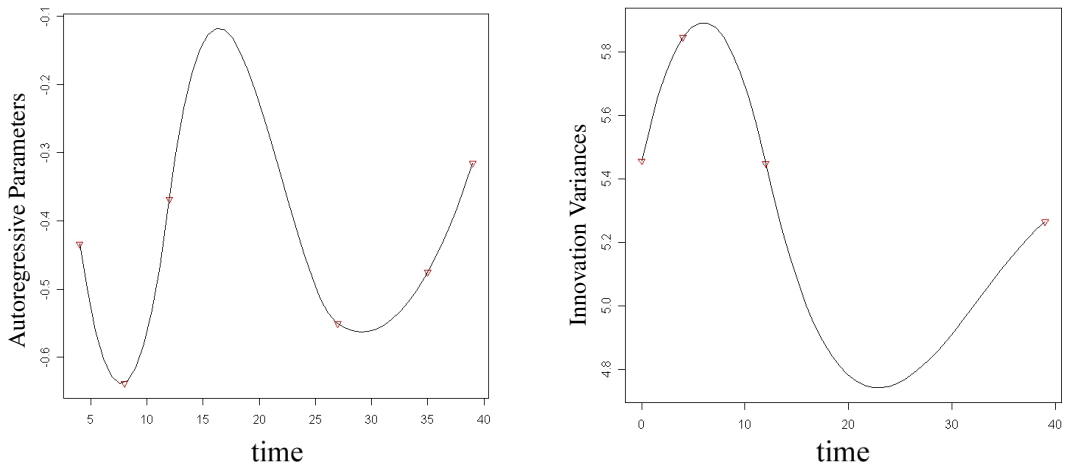
As discussed in Pourahmadi [1999]; Pan and MacKenzie [2003, 2006] all entries of the matrices F_i and Q_i have a statistical interpretation: the below-diagonal entries of the matrix F_i are the negatives of the autoregressive coefficients, $\varphi_{i,j,k}$ $(j > k)$, in

$$\hat{y}_{i,j} = \tilde{\mu}_{i,j} + \sum_{k=1}^{j-1} \varphi_{i,j,k} \, (y_{i,k} - \tilde{\mu}_{i,j}),$$

the linear square predictor of $y_{i,j}$ based on its predecessors $y_{i,j-1}, ..., y_{i,1}$. The diagonal

		CS			Σ_1			Σ_2		
Variable	Parameter	Est.	SE	P-val.	Est.	SE	P-val.	Est.	SE	P-val.
Intercept	β_0	41.04	0.78	-	41.17	0.80	-	40.82	0.71	-
Final Recovery	β_1	106.89	6.82	-	106.65	6.69	-	98.08	7.65	-
Treatment-Specific Recovery Rates										
Rate of Tubigrip	β_{21}	0.97	0.14	-	0.98	0.14	-	0.99	0.13	-
Contrast: BKC	β_{22}	-0.12	0.04	0.006	-0.12	0.04	0.006	-0.12	0.04	0.005
Contrast: Aircast	β_{23}	-0.06	0.04	0.145	-0.06	0.04	0.138	-0.05	0.04	0.161
Contrast: Bledsoe	β_{24}	-0.01	0.03	0.755	-0.01	0.03	0.795	-0.01	0.03	0.665
Covariate Effects on Final Recovery Level and Recovery Rate										
Age-effect on Max.	α_1	-7.28	2.05	< 0.001	-8.19	2.01	< 0.001	-5.78	2.30	0.012
Age-effect on Rate	α_2	-0.21	0.04	< 0.001	-0.22	0.04	< 0.001	-0.22	0.04	< 0.001
Variance Components										
	σ^2	186.10	7.33	< 0.001	-	-	-	176.21	6.86	< 0.001
	D^2	148.58	12.64	< 0.001	8.63	13.12	0.510	100.06	13.15	< 0.001
	σ_1^2	-	-	-	186.00	16.84	< 0.001	-	-	-
	σ_2^2	-	-	-	176.99	17.01	< 0.001	-	-	-
	σ_3^2	-	-	-	169.99	18.79	< 0.001	-	-	-
	σ_4^2	-	-	-	157.32	28.00	< 0.001	-	-	-
	τ^2	-	-	-	153.88	22.66	< 0.001	-	-	-
	ρ	-	-	-	0.996	0.004	< 0.001	-	-	-
	ϕ	-	-	-	-	-	-	0.13	0.02	< 0.001
Deviance	$-2\,\ell$	15339			15337			15291		

Table 4.1: Overview of the parameter estimates and standard errors (SE) of θ and the covariance parameters. All results are based on the assumption of an ignorable missingness process. The p-values are reported only for the components of θ that might be zero. The first row denotes which covariance structure was used. For the definition of Σ_1 and Σ_2 see equation (4.3) and (4.4), respectively.



(a) Sample Autoreg. Parameters (b) Sample Innovation Variances

Figure 4.1: Empirical regressogram based on the empirical covariance matrix. The triangles denote the observed $(-1)\,\varphi_{obs,j,k}$ and $\log(\sigma^2_{obs,j})$. The smooth curve was derived using local polynomial regression fitting and the R-program `scatter.smooth`.

entries of Q_i are the prediction (or innovation) variances $\sigma^2_{i,j} = \mathrm{Var}(y_{i,j} - \hat{y}_{i,j})$, where $j \in \{1, 2, 3, 4\}$. The advantage of this method compared to other decomposition methods is that the parameters $\varphi_{i,j,k}$ are unconstrained, while $\sigma^2_{i,j} \in \mathbb{R}^+$. In order to simplify modelling, we focus on $\log(\sigma^2_{i,j}) \in \mathbb{R}$. We note that this transformation could be problematic for variances near zero. For the CAST data set, however, this is not an issue, as variances are expected to be substantially larger than zero.

In line with the last section, we assume that Σ_i, depends on i only through its dimension and is a sub-matrix of the $M \times M$-dimensional overall covariance matrix Σ.

In the next step, we will suggest parametric models for the covariance parameters $\varphi_{j,k}$ and σ^2_j. We use *empirical regressograms*, introduced in Pourahmadi [1999], to inform our model choice. Based on the empirical covariance matrix (equation (4.1)) the unique matrices F and Q are specified. Then, the negative sample autoregressive parameters $\varphi_{obs,j,k}$ are plotted against the time lags, $|t_j - t_k| \in \{0, 4, 12, 27, 35, 39\}$, and the log-innovation variances $\log(\sigma^2_{obs,j})$ are plotted versus time $t_j \in \{0, 4, 12, 39\}$. The resulting plots are shown in Figure 4.1. The fitted smooth curves were derived using local polynomial regression fitting. The empirical regressograms and the fitted smooth curves suggest that $\varphi_{j,k}$ and

$\log(\sigma_j^2)$ are polynomial functions of the lags and time, respectively. For the logarithm of the innovation variances a cubic function and for the autoregressive parameters a polynomial of degree 5 seem reasonable. Note that polynomials of these degrees would lead to exact interpolations of the covariance parameters.

However, according to Pan and MacKenzie [2003], we should be cautious with the regressogram-based model selection, as a model that is not optimal may be proposed. An alternative approach based on the *Bayesian Information Criterion* (BIC) is suggested. For this method, the profile likelihoods by saturating the parameter sets in pairs are investigated. Let $d_\varphi \in \{0, 1, ..., 5\}$ denote the degree of the negative autoregressive parameters and $d_\sigma \in \{0, 1, 2, 3\}$ the degree of the log-innovation variances. In our context, the BIC for covariance model selection depends on the chosen degrees and is defined by

$$\text{BIC}(d_\varphi, d_\sigma) = -2\, l_{max}(d_\varphi, d_\sigma) + n_p\, \log(N),$$

where $l_{max}(d_\varphi, d_\sigma)$ is the maximised log-likelihood for the specific pair (d_φ, d_σ) and $n_p = d_\varphi + d_\sigma + 2$ is the number of covariance parameters. According to Pan and MacKenzie [2003], the optimum pair $(d_\varphi^*, d_\sigma^*)$ may be found using the two following searches:

$$\begin{aligned} d_\varphi^* &= argmin_{d_\varphi} \left\{ \text{BIC}(d_\varphi, \max(d_\sigma) = 3) \right\} \quad \text{and} \\ d_\sigma^* &= argmin_{d_\sigma} \left\{ \text{BIC}(\max(d_\varphi) = 5, d_\sigma) \right\}. \end{aligned}$$

However, as emphasized by Fitzmaurice et al. [2004] and Hedeker and Gibbons [2006], BIC extracts a very large penalty for the addition of parameters. To quote Fitzmaurice et al. [2004]: ‘ In general, we do not recommend the use of BIC for covariance model selection as it entails a high risk of selecting a model that is too simple or parsimonious for the data at hand.’ Therefore, we will contrast the results based on the BIC searches with those based

on the *Akaike Information Criterion* (AIC), which is defined as

$$\mathrm{AIC}(d_\varphi, d_\sigma) = -2\, l_{max}(d_\varphi, d_\sigma) + 2\, n_p$$

and known to overfit, see Hurvich et al. [1989]. The general polynomials we are considering are given by

$$\begin{aligned} \varphi_{j,k} &= \gamma_0 + \gamma_1 (t_j - t_k) + \ldots + \gamma_{(d_\varphi+1)} (t_j - t_k)^{d_\varphi} \quad \text{and} \\ \log(\sigma^2_\ell) &= \delta_0 + \delta_1\, t_\ell + \ldots + \delta_{(d_\sigma+1)}\, t_\ell^{d_\sigma}, \end{aligned} \tag{4.6}$$

where $j \in \{2,3,4\}$, $k \in \{1,2,3,4\}$ with $j > k$ and $\ell \in \{1,2,3,4\}$. We define $\gamma = (\gamma_0, ..., \gamma_{(d_\varphi+1)})^\top$ and $\delta = (\delta_0, ..., \delta_{(d_\sigma+1)})^\top$.

Based on the marginal model

$$Y_i \sim \mathcal{N}_4 (\tilde{\mu}_i, \Sigma)$$

with $\tilde{\mu}_i$ specified in equations (3.19), (3.20) and the covariance matrix Σ determined through equations (4.5) and (4.6), we can calculate the log-likelihood. Numerical optimization routines are used in an attempt to maximize the joint log-likelihood. However, using the parametrization showed in (4.6) our optimization method fails to converge. As suggested in Pan and MacKenzie [2006], it is beneficial to represent the polynomials in orthogonal form. This re-parametrization has three main advantages, see Hedeker and Gibbons [2006]. Firstly, it avoids collinearity problems that can result from using multiples of t_j as regressors. Secondly, using orthogonal polynomials has the advantage of putting the polynomials on the same scale. That is, the estimated coefficients of the polynomial are standardized and their relative contribution can be compared based on their magnitude. Finally, orthogonal polynomials circumvent the fact that for higher-degree polynomials it gets increasingly difficult to estimate the regression coefficients as the coefficients and their standard errors become vanishingly small.

Based on the work described in Bock [1975] and Hedeker and Gibbons [2006], we re-parametrize the polynomials defined in (4.6) through orthogonal polynomials. In a first step we calculate the lag (S_{auto}) and time (S_{innov}) matrices based on the original metric:

$$S_{auto} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 8 & 12 & 27 & 35 & 39 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 4^{d_\varphi} & 8^{d_\varphi} & 12^{d_\varphi} & 27^{d_\varphi} & 35^{d_\varphi} & 39^{d_\varphi} \end{pmatrix} \text{ and } S_{innov} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 4 & 12 & 39 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 4^{d_\sigma} & 12^{d_\sigma} & 39^{d_\sigma} \end{pmatrix}.$$

The first rows correspond to the intercept. Note that these matrices simply capture the information of the regressors in the polynomials defined above, see equation (4.6). Based on these matrices and linear algebra the orthogonal polynomial values O_{auto} and O_{innov} can be calculated. Let $S \in \{S_{auto}, S_{innov}\}$ and O be the corresponding orthogonal transformation. Then we perform the following steps to obtain O:

Step 1: Compute the symmetric matrix $S^\top S$.

Step 2: Obtain the upper-triangular Cholesky factor H of $S^\top S$.

Step 3: Calculate the inverse of H, i.e. H^{-1}.

Step 4: Compute O as $O = S\, H^{-1}$.

To distinguish between the two representations, i.e. the original time metric S and the orthogonal polynomial time metric O, we denote the associated parameter vectors based on the orthogonal parametrization as γ_{orth} and δ_{orth}.

Using the orthogonal transformation we are now able to maximize the likelihood and to perform the BIC- and AIC-based search for the optimal model. The values of the maximized log-likelihoods and the corresponding BIC and AIC values are summarized in Table 4.2. The significance of the differences in the AIC and BIC values for two nested models can be determined based on the χ^2-distribution. For the difference of the BIC- and AIC-values based on $l_{max}(d_\varphi = 5, d_\sigma = u)$, $l_{max}(d_\varphi = 5, d_\sigma = v)$; $u, v \in \mathbb{N}$ and $v > u$ we

Search for d_φ^*				
d_φ	$l_{max}(d_\varphi, d_\sigma = 3)$	n_p	BIC$(d_\varphi, d_\sigma = 3)$	AIC$(d_\varphi, d_\sigma = 3)$
0	-7654.64	5	15340.86	15319.28
1	-7637.38	6	15312.65	15286.76
2	-7634.47	7	15313.15	15282.94
3	-7627.25	8	15305.02	15270.50
4	-7626.88	9	15310.60	15271.76
5	-7605.72	10	15274.59	15231.44
Search for d_σ^*				
d_σ	$l_{max}(d_\varphi = 5, d_\sigma)$	n_p	BIC$(d_\varphi = 5, d_\sigma)$	AIC$(d_\varphi, d_\sigma = 3)$
0	-7617.83	7	15279.87	15249.66
1	-7612.89	8	15276.30	15241.78
2	-7612.68	9	15282.20	15243.36
3	-7605.72	10	15274.59	15231.44

Table 4.2: BIC- and AIC-based search for the optimal pair $(d_\varphi^*, d_\sigma^*)$. The log-likelihood values and the corresponding AIC, BIC values are shown.

obtain

$$
\begin{aligned}
&-2\left[l_{max}(d_\varphi = 5, d_\sigma = u) - l_{max}(d_\varphi = 5, d_\sigma = v)\right] \\
&\qquad = \begin{cases} \mathrm{BIC}(d_\varphi = 5, d_\sigma = u) - \mathrm{BIC}(d_\varphi = 5, d_\sigma = v) + \log(N)\,(v-u) \\ \\ \mathrm{AIC}(d_\varphi = 5, d_\sigma = u) - \mathrm{AIC}(d_\varphi = 5, d_\sigma = v) + 2\,(v-u) \end{cases} \\
&\qquad\qquad \sim \chi^2_{(v-u)}.
\end{aligned}
\tag{4.7}
$$

A similar expression can be obtained for testing the difference based on $l_{max}(d_\varphi = u, d_\sigma = 3)$ and $l_{max}(d_\varphi = v, d_\sigma = 3)$; $u, v \in \mathbb{N}$ and $v > u$. Based on this test, the AIC-based and BIC-based searches yield that the optimal model is specified by the degree pair $(d_\varphi^*, d_\sigma^*) = (5, 3)$. This is the same pair we would have chosen based on the regressogram.

The optimal model involves the estimation of ten covariance parameters. It specifies the exact polynomial interpolation to the autoregressive parameters and log-innovation variances. In particular, this model is equivalent to modelling an unstructured covariance structure. The parameter estimates for θ, γ_{orth}, δ_{orth} and the back-transformed γ and δ based on this model are shown in Table 4.3. Here, the back-transformation of γ_{orth} and δ_{orth} can

		Orthogonal Metric				Original Metric		
Variable		Est.	SE	P-val.		Est.	SE	P-val.
Intercept	β_0	41.11	0.68	-				
Final Recovery	β_1	108.47	7.89	-				
Treatment-Specific Recovery Rates								
Rate of Tubigrip	β_{21}	1.06	0.15	-				
Contrast: BKC	β_{22}	-0.12	0.05	0.014				
Contrast: Aircast	β_{23}	-0.05	0.04	0.266				
Contrast: Bledsoe	β_{24}	-0.005	0.04	0.906				
Covariate Effects on Final Recovery Level and Recovery Rate								
Age-effect on Max.	α_1	-8.75	2.35	< 0.001				
Age-effect on Rate	α_2	-0.24	0.04	< 0.001				
Variance Components								
	$\gamma_{orth,0}$	0.76	0.04	< 0.001	γ_0	-2.34	0.49	< 0.001
	$\gamma_{orth,1}$	-0.30	0.04	< 0.001	γ_1	1.25	0.20	< 0.001
	$\gamma_{orth,2}$	-0.10	0.05	0.071	γ_2	-0.17	0.03	< 0.001
	$\gamma_{orth,3}$	-0.22	0.05	< 0.001	γ_3	0.01	0.001	< 0.001
	$\gamma_{orth,4}$	-0.006	0.06	0.923	γ_4	-0.0002	$3.6 \cdot 10^{-5}$	< 0.001
	$\gamma_{orth,5}$	0.37	0.06	< 0.001	γ_5	$2.17 \cdot 10^{-6}$	$3.3 \cdot 10^{-7}$	< 0.001
	$\delta_{orth,0}$	11.05	0.07	< 0.001	δ_0	5.50	0.06	< 0.001
	$\delta_{orth,1}$	-0.23	0.07	0.001	δ_1	0.12	0.04	0.001
	$\delta_{orth,2}$	-0.02	0.07	0.778	δ_2	-0.01	0.004	< 0.001
	$\delta_{orth,3}$	0.24	0.07	0.002	δ_3	0.003	$7.3 \cdot 10^{-5}$	< 0.001

Table 4.3: Overview of the parameter estimates and standard errors (SE) of θ and the covariance parameters γ_{orth}, δ_{orth}, γ and δ. The parameter estimates of θ are not affected by the orthogonal transformation and are therefore only listed once. All results are based on the assumption of an ignorable missingness process. The p-values are reported only for the components of θ that might be zero.

be calculated through

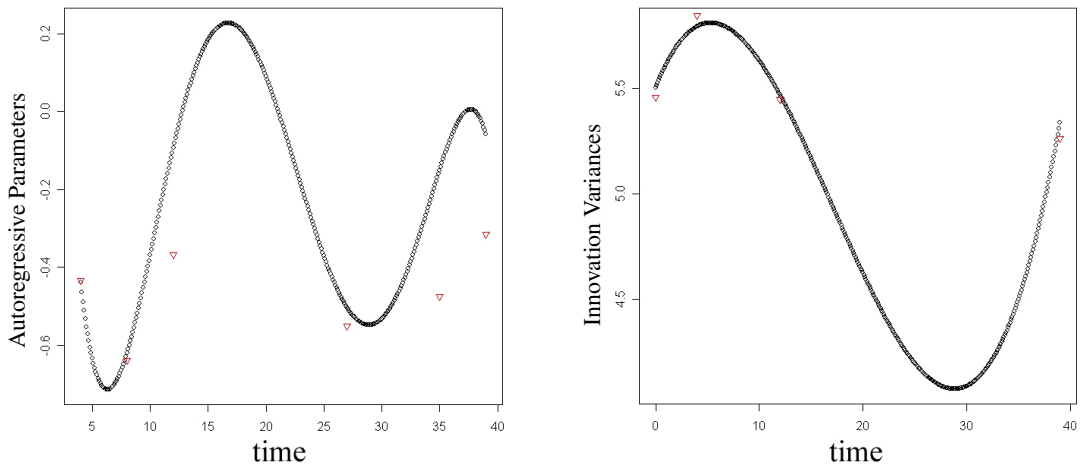
$$\gamma = H^{-1}\,\gamma_{orth} \quad \text{and} \quad \delta = H^{-1}\,\delta_{orth}.$$

The corresponding standard errors can be derived based on

$$\text{Cov}(\gamma) = H^{-1}\text{Cov}(\gamma_{orth})(H^{-1})^\top \quad \text{and} \quad \text{Cov}(\delta) = H^{-1}\text{Cov}(\delta_{orth})(H^{-1})^\top,$$

see Hedeker and Gibbons [2006].

The fitted covariance and correlation matrices based on this model and the decom-



(a) Fitted Autoreg. Parameters (b) Fitted Innovation Variances

Figure 4.2: Fitted polynomials based on the model with $(d^*_\varphi, d^*_\sigma) = (5, 3)$. The triangles denote the observed $(-1)\,\varphi_{obs,j,k}$ and $\log(\sigma^2_{obs,j})$.

position in equation (4.5) are then given by

$$\hat{\Sigma}_{fit} = \begin{pmatrix} 244.96 & 107.25 & 89.58 & 75.85 \\ & 374.73 & 242.73 & 185.19 \\ & & 396.38 & 241.53 \\ & & & 361.69 \end{pmatrix} \text{ and } \hat{Corr}_{fit} = \begin{pmatrix} 1 & 0.35 & 0.29 & 0.25 \\ & 1 & 0.63 & 0.50 \\ & & 1 & 0.64 \\ & & & 1 \end{pmatrix}.$$

It is remarkable how similar these results are, compared with the empirical covariance and correlation matrices shown in equation (4.1). Based on these results, we were able to fit an unstructured covariance matrix, using the fitted covariances as starting values in our optimization routine. Not surprisingly, the parameter estimates for θ and the fitted covariance matrix are identical (for up to two decimal places). In fact, the parameter estimate $\hat{\theta}$ based on this method are very close to the ones based on a compound symmetry structure. The age-effect on the final recovery score and the final recovery score itself are slightly different. However, overall the resulting conclusions are the same with just the effect sizes and associated standard errors varying slightly.

The fitted regressograms visualize the model fit, see Figure 4.2. We see that the

model fit is very good for the log-innovation variances, but less good for the autoregressive parameters. The good overall fit can be explained by results shown in Pan and MacKenzie [2003]. According to these, misspecification is dominated mainly by the difference in the innovation variances. Misspecification of the autoregressive coefficients appears to be less important.

Although the fitted model leads to a satisfying fit, we were interested in investigating whether the covariances also depend on covariates other than time. The covariate of interest is age, because the final recovery score varies by age. Thus, we expect the effect of skewness and the correlations to depend on this variable. However, incorporating this variable in the model for the innovation variances and autoregressive parameters did not improve significantly the AIC or BIC values.

4.4 Summary

In this section we have seen that modelling the covariance structure for bounded longitudinal data is challenging. This is due to three likely features of the covariance matrix. Firstly, the correlations along the rows decrease because serial correlations decrease over time. Secondly, the correlations along the sub-diagonals increase due to the bounded nature of the score. Finally, the variances are rarely constant over time. In fact, due to randomisation and inclusion criteria we expect the variances to be small at baseline. The variability is expected to increase throughout the study. However, as time passes most patients achieve their final recovery level and are clustered towards the upper bound of the score. Thus, the variation decreases as the bounds are reached.

We show that finding a suitable covariance pattern model that accounts for these characteristics is not straightforward. The same holds for covariance structures that are implied by random-effect models. The approach introduced in Pourahmadi [1999] and described in Pan and MacKenzie [2003, 2006] appears to be a flexible and suitable approach to model the covariance structure of bounded longitudinal data. This approach is extended

by using the non-linear mixed model proposed in Section 3.5 for the outcome process and by allowing for missing values. Different model selection tools are used. Applying this method to the CAST data recovers an unstructured covariance structure.

The aim of this chapter was to explore the sensitivity of the inference presented in Section 3.5 to the choice of the covariance model. The parameter estimates of interest for the CAST study, i.e. the treatment differences and age-effects, remain very stable under all covariance matrices investigated. The inference remains the same, solely the effect sizes vary a little.

So far we focused on the challenge of modelling the mean and covariance of longitudinal bounded data, while ignoring the missing data issue. The next two chapters are devoted to this aspect of the data analysis.

Chapter 5

Missing Data and CAST

5.1 Introduction

In Chapter 3 we mention that one challenge in the analysis of longitudinal data arises through unbalanced data, i.e. the observation times are not common to all subjects. Unbalanced data can arise due to the study design or due to *missing data*. By missing data we mean that intended measurements are not taken, leading to missing entries in the outcome vector Y_i. Generally, missing values can occur on independent and dependent variables. However, in this thesis we will focus on issues that arise when the dependent variable is missing. Regardless of the cause, unbalanced data raise technical difficulties in the analysis. However, unbalanced data due to missing data additionally raise 'deeper conceptual issues, since we have to ask *why* the values are missing and more specifically whether their being missing has any bearing on the practical questions posed by the data' [Diggle et al., 1996]. Take, for example, a clinical study to evaluate the efficacy of a new drug where all patients for whom the drug is not effective withdraw after the baseline assessment. Analyzing the data for the complete cases only would overstate the treatment effect and thus lead to an incorrect inference. Lack of efficacy is only one potential cause of missing data; others are: patients skip visits for practical or administrative reasons, patients move away, equipment failure, other medical conditions not related to the primary outcome, unaccept-

able side effects, the primary outcome might reach a pre-specified benchmark or patients may consider themselves to have fully recovered [Carpenter et al., 2002; Molenberghs and Kenward, 2007; Nakash et al., 2008].

Longitudinal studies usually aim to make inference for population quantities, e.g. the mean change over time, and with missing data the scientific interest remains the same. That is, we are usually interested in: What would have happened had there been no missing values? Drawing this inference can be quite complicated. In general, three potential problems arise with missing data: loss of power due to reduced information, complication in data handling and analysis, and bias due to differences between the observed and unobserved data [Horton and Lipsitz, 2001]. The extent of these problems depends on the proportion of missingness, the *missing data pattern* and most importantly the strength of the relationship between the unobserved outcomes and the probability of dropout [Wood et al., 2006].

Generally, a missing data pattern describes which measurements are observed and which are missing. We distinguish between two different patterns:

- *Monotone missingness patterns* occur when subjects are observed without interruption from the beginning of the study until a given point in time, when they quit the study and do not return. In the case of longitudinal studies, monotone patterns are often referred to as *dropout* or *attrition*.

- In the case of *non-monotone missingness patterns* the measurements can be observed and missing on any occasion.

Non-monotone missingness patterns usually raise more difficulties than monotone patterns, see Chapter 6.

Let us now turn to the relationship between the outcome of interest and the probability of missing values. Rubin [1976] proposed a framework for this probabilistic relationship. Let $r_{i,j}$ be a realisation of the random variable $R_{i,j}$, which indicates whether the outcome of subject $i \in \{1, ..., N\}$ at time point $j \in \{1, ..., M\}$, i.e. $y_{i,j}$, was observed, $r_{i,j} = 1$,

or missing, $r_{i,j} = 0$. We refer to $\boldsymbol{R}_i = (R_{i,1}, ..., R_{i,M})^\top$, $i \in \{1, ..., N\}$, as the *missing data indicator* of subject i. Accordingly, $\boldsymbol{R} = (\boldsymbol{R}_1^\top, ..., \boldsymbol{R}_N^\top)^\top$ denotes the missingness indicator for all subjects. Rubin [1976] developed a taxonomy to classify different *missingness mechanisms*, i.e. to describe the relationship between the missingness indicator $\boldsymbol{R}_i$, the outcome vector $\boldsymbol{Y}_i$ and covariates. He distinguishes between the *missing completely at random* (MCAR), the *missing at random* (MAR) and the *missing not at random* (MNAR) mechanisms. Before introducing these mechanisms, we briefly want to discuss likelihood-based inference for incomplete data.

Generally, the appropriate starting point for likelihood-based analysis is the density of the full data $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$, $i \in \{1, ..., N\}$ [Molenberghs and Kenward, 2007]:

$$f_{(\boldsymbol{Y}_i, \boldsymbol{R}_i)}(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \theta, \phi) = f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i | X_i, \theta)\, f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \phi), \tag{5.1}$$

where X_i and W_i denote the matrices of explanatory variables affecting the outcome and missingness mechanism, respectively. The associated parameter vectors are given by θ and ϕ, respectively. Assume that $\boldsymbol{Y}_i$ is not fully observed and thus can be partitioned into the observed, $\boldsymbol{Y}_{i,obs}$, and missing part, $\boldsymbol{Y}_{i,mis}$. Then, the density in equation (5.1) depends on missing values and inference needs to be based on the *observed data density*:

$$\begin{aligned} f_{(\boldsymbol{Y}_{i,obs}, \boldsymbol{R}_i)}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i | X_i, W_i, \theta, \phi) \quad = \quad & \int f_{\boldsymbol{Y}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis} | X_i, \theta) \\ & \times f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i | \boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, W_i, \phi)\, \mathrm{d}\boldsymbol{y}_{i,mis}. \end{aligned} \tag{5.2}$$

This density generally requires a model for the outcome process and the missingness mechanism. In most cases interest lies in the outcome process, while the missingness mechanism itself is a nuisance. It has been shown that under certain circumstances the missingness mechanism can be ignored and no model is required.

The classification of different missingness mechanisms proposed by Rubin [1976] is based on the conditional model for $\boldsymbol{R}_i|\boldsymbol{Y}_i$, and hence is based on the idea that subjects are selected to have missing values by their response $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,obs})$ and explanatory

variables W_i.

A missingness process is said to be *MCAR*, when missingness may be related to available covariates but is conditionally independent of the outcome $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,obs})$. Mathematically, this is equivalent to:

$$f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \phi) = f_{\boldsymbol{R}_i}(\boldsymbol{r}_i|W_i, \phi),$$

where W_i corresponds to the set of relevant explanatory variables. Note that the original definition of MCAR [Rubin, 1976] does not allow for a dependence on covariates. However, nowadays statistical literature often does not distinguish between the original MCAR definition and the covariate-dependent MCAR definition given in equation (5.1), see Carpenter et al. [2002]; Wood et al. [2006]; Molenberghs and Kenward [2007].

Under MCAR, the incomplete data set can be seen as a random subsample of the complete data set, which would have been observed without missingness [Little and Rubin, 2002]. In a clinical trial context, the assumption of MCAR corresponds to a response being missing for any reason, possibly noted at baseline, which (conditional on baseline covariates in the model) is not associated with their post-randomization response [Carpenter et al., 2002]. This reason could be skipping a visit due to another appointment, moving away or presence of non-study related medical conditions.

Regarding the observed data density, equation (5.2), MCAR implies:

$$\begin{aligned} f_{\boldsymbol{Y}_{i,obs},\boldsymbol{R}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi) &\overset{MCAR}{=} \int f_{\boldsymbol{Y}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}|X_i, \theta)\, f_{\boldsymbol{R}_i}(\boldsymbol{r}_i|W_i, \phi)\, \mathrm{d}\boldsymbol{y}_{i,mis} \\ &= f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)\, f_{\boldsymbol{R}_i}(\boldsymbol{r}_i|W_i, \phi), \end{aligned}$$

i.e. if the parameter sets θ and ϕ are distinct, valid inference for θ can be drawn based on $f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)$. By valid we mean that likelihood-inference based on the observed data $\boldsymbol{Y}_{obs}$ only has the same properties as inference based on the complete data Y, i.e. consistent estimators, confidence intervals with the correct coverage and tests with the correct size. If θ and ϕ overlap, basing inference on $f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)$ leads to consistent but not fully

efficient estimates [Carpenter et al., 2002]. In particular, inference for θ does not require a model for $\boldsymbol{R}_i$, i.e. the missingness mechanism is *ignorable*.

The conditional independence assumption of a MCAR missingness mechanism is very strong and rarely realistic. A missingness process which uses less restrictive assumptions is the *MAR* mechanism. In this case, missingness may be related to covariates and observed measurements $\boldsymbol{Y}_{i,obs}$ but is conditionally independent of the missing outcomes $\boldsymbol{Y}_{i,mis}$. More specifically, a MAR missingness process holds if

$$f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \phi) = f_{\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs}}(\boldsymbol{r}_i|y_{i,obs}, W_i, \phi).$$

That is, missingness occurs for reasons which are associated with the observed outcome of interest, e.g. adverse events, total recovery from illness or reaching a pre-specified benchmark.

The observed data density in equation (5.2) simplifies to

$$\begin{aligned} f_{\boldsymbol{Y}_{i,obs},\boldsymbol{R}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi) &\overset{MAR}{=} \int f_{\boldsymbol{Y}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}|X_i, \theta)\, f_{\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs}}(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs}, W_i, \phi)\, \mathrm{d}\boldsymbol{y}_{i,mis} \\ &= f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)\, f_{\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs}}(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs}, W_i, \phi). \end{aligned}$$

As in the case of a MCAR process, valid inference for the parameter vector θ can be drawn based on $f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)$, subject to loss of efficiency if the *separability* or *distinctness condition* for the parameters θ and ϕ does not hold [Molenberghs and Kenward, 2007]. Again, inference for θ does not require a model for the missingness mechanism.

Note that the concept of *ignorability* under MAR, i.e. ignoring the missingness mechanism, is only valid for likelihood-based (and thus Bayes-based) inference. Frequentist techniques, such as the *generalized estimating equations* (GEE), usually require the stronger MCAR assumption.

Finally, if the missingness probability depends on unknown quantities, i.e.

$$f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \phi) = f_{\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs},\boldsymbol{Y}_{i,mis}}(\boldsymbol{r}_i|y_{i,obs}, y_{i,mis}, W_i, \phi)$$

the missingness process is termed *MNAR*. This missingness mechanism might take place in clinical studies, where adverse events or total recovery occur after the last documented outcome, but prior to the next planned one. Following an example in the handbook of the SAS-procedure MI [SAS/STAT, 1999], consider a trivariate data set with variables Y_1 and Y_2 fully observed, and a variable Y_3 that has missing values. MAR assumes that the probability that Y_3 is missing for an individual can be related to the individuals values of variables Y_1 and Y_2, but not to its value of Y_3. On the other hand, if a complete case and an incomplete case for Y_3 with exactly the same values for variables Y_1 and Y_2 have systematically different values, then there exists a response bias for Y_3, and MAR is violated. Note that without additional information of Y_3, it is impossible to test the MAR assumption against MNAR [Molenberghs et al., 1998].

The observed data density under MNAR does not simplify and we remain with

$$f_{\boldsymbol{Y}_{i,obs},\boldsymbol{R}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi) \overset{MNAR}{=} \int f_{\boldsymbol{Y}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}|X_i, \theta) \times f_{\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs},\boldsymbol{Y}_{i,mis}}(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,obs}, W_i, \phi)\, \mathrm{d}\boldsymbol{y}_{i,mis},$$

where we cannot simplify the integral. In particular, we cannot ignore the missingness process but need to model the measurement and missingness process jointly. MNAR missingness processes are therefore often referred to as *informative* or *non-ignorable* missingness.

Methods modelling the measurement and missingness process jointly, such as *selection models*, *pattern-mixture models* and *shared parameter models* have been proposed and will be discussed in Section 5.4. All these models base on untestable assumptions for the conditional distribution of $Y_{i,mis}|Y_{i,obs}$, see Section 5.4.

Although the assumption of MAR can be realistic for certain settings, in most applications it is impossible to exclude the possibility of MNAR. Hence, it is unwise to rely on the precise conclusions of an analysis based on a particular MAR or MNAR model. Many researchers recommend exploring the stability of the conclusions across a range of different MAR and MNAR models through a *sensitivity analysis*. Carpenter et al. [2002] suggest:

'The MNAR models need to be selected to build an envelope of conclusions, bounded by the results of the MAR model and the *worst case* MNAR model. Clearly, selection of the latter and the form of the MNAR model both depend on scientific judgement.'

In the following we will discuss different ways for handling longitudinal studies with missing values. Many monographs and papers focusing on reviewing issues that arise with missing data in longitudinal studies were published in the recent years, for example Little [1995]; Schafer and Graham [2002]; Little and Rubin [2002]; Carpenter et al. [2002]; Molenberghs et al. [2004]; Molenberghs and Kenward [2007]; Diggle et al. [2007]; Daniels and Hogan [2008]; Graham [2009].

According to Carpenter et al. [2002] we can distinguish four different analysis approaches:

1. Perform the analysis only on those subjects who complete the trial;
2. Analyse only the available data;
3. Use a single or multiple imputation technique to replace the missing observations with plausible values, then analyse the completed data set(s); and
4. Model the repeated data and missingness process jointly.

The first option yields a *complete case analysis*. The second option can be realised through the *direct likelihood approach*, which is the likelihood-based way of using available information only [Molenberghs and Kenward, 2007]. Other, mostly nonparametric, methods of using only the observed data are available [Little and Rubin, 2002]. *Single* and *multiple imputation* techniques are well known [Rubin, 1987, 1996; Schafer, 1997, 1999; Horton and Lipsitz, 2001; Little and Rubin, 2002]. According to Carpenter et al. [2002], the fourth option 'is usually the most complex computationally, but it is also the most useful, as it elucidates the often unexpectedly subtle assumptions behind the other methods, and allows the sensitivity of the conclusions to assumptions about the missing data mechanism to be assessed'.

A short review of popular methods to handle missing data in longitudinal data is presented in the following sections. We will focus on frequentist approaches and refer to Daniels and Hogan [2008] for the discussion of missing data methods under the Bayesian paradigm. In Section 5.2 we review *ad hoc* methods to handle missing data, while in Section 5.3 and Section 5.4 we review more principled methods under the assumptions of MAR or MNAR, respectively. We conclude this chapter by performing a sensitivity analysis for the CAST data set in Section 5.5 and by summarizing our findings in Section 5.6.

5.2 A Review of Simple Missing Data Methods

In this section we will discuss *simple* methods to handle missing data in longitudinal studies. By 'simple' we mean ad-hoc methods that edit the incomplete data sets to produce *completed* data sets that can be analyzed by standard methods and software for balanced data. We will argue why these methods should be avoided and review more suitable methods in Section 5.3 and Section 5.4.

5.2.1 Complete Case Analysis

In a *complete case analysis* all participants with missing values are discarded from the sample and the missingness process is ignored. This method is also known as *case deletion* or *listwise deletion* and is the default method for handling missing values in some statistical techniques, e.g. MANOVA.

Advantages of a complete case analysis are that it is easy to communicate between statisticians and non-statisticians and that the implementation is straightforward. Furthermore, it can be applied to monotone and non-monotone missingness.

A complete case analysis yields consistent estimators under a MCAR missingness process [Molenberghs and Kenward, 2007]. However, depending on the number of discarded subjects the estimators can be very inefficient. Generally, this approach is not cost-effective and it can lead to substantial bias in case of a MAR or a MNAR process.

5.2.2 Last Observation Carried Forward

Another common, albeit problematic, strategy is the *last observation carried forward* (LOCF) method. In this approach, missing values for each participant are substituted by the most recent available value. This leads to the assumption that a patient sustains a specific level after drop-out and may cause an unrealistic response profile. In particular, longitudinal studies aim to investigate the change over time in the response variable and factors that influence that change. An analysis using LOCF may not answer the original research question or bias the inference. Further, LOCF does not distinguish between imputed and observed data and thus usually overestimates the precision of the estimates.

This method is widely used in the pharmaceutical industry and advertised by some public authorities, e.g. the *European Medicines Agency* (EMEA) [EMEA, 2009]. The reason for this is that LOCF is thought to lead to a *conservative* analysis, that is an analysis that underestimates treatment differences. However, many researchers have shown that even in the setting of MCAR a bias (positive or negative) can occur which leads to a risk of overstating the magnitude of a treatment effect. This is shown via an example in Molenberghs and Kenward [2007] and the authors conclude: '[...] even under the unrealistically strong assumption of MCAR, we see that the bias in the LOCF estimator typically does not vanish and, even more importantly, the bias can be positive or negative, and can even induce an apparent treatment effect when there is none.' Thus, it is surprising that EMEA [2009] states: 'LOCF only produces unbiased estimates under the MCAR assumption [...].'

5.2.3 Single Imputation Methods

In the last section we have discussed the LOCF approach, where every missing value is filled in by the last available response. Alternative methods to *fill in* the incomplete data are available and generally referred to as *single imputation methods*. Examples are: *imputing unconditional means; imputing from unconditional distribution; imputing conditional means; imputing from a conditional distribution; hot deck imputation; baseline observation carried forward; worst case value imputation* and many more, see Schafer and Gra-

ham [2002]; Wood et al. [2006]; Molenberghs and Kenward [2007]; EMEA [2009]. All these methods produce apparently complete data sets. However, basing inference on these *completed data sets* can be misleading, as the data distribution and relationships can be distorted [Schafer and Graham, 2002]. For example the point estimates can be biased if the imputation model is wrong. On the other hand, even if a correct imputation model is provided, the standard errors may fail to account for the added uncertainty due to missingness. Molenberghs and Kenward [2007] list three potential pitfalls of single imputation techniques:

- 'The performance of imputation techniques is unreliable. Situations where they do work are difficult to distinguish from situations where they prove misleading.
- Imputation often requires *ad hoc* adjustments to obtain satisfactory point estimates.
- This methods fail to provide simple, correct estimators of precison.'

The latter limitation is caused by not distinguishing between *observed* and *imputed* values. This can lead to underestimated standard errors, p-values that are too small and confidence intervals which are too narrow [Little and Rubin, 2002]. The accommodation of the added uncertainty due to non-response can, among others, be achieved through resampling methods, such as *bootstrap* and *jackknife*, or *multiple imputation*. We refer to [Little and Rubin, 2002, Chapter 5] for resampling methods and will discuss multiple imputation in Section 5.3.3.

A few words of caution regarding imputation methods were already given by Dempster and Rubin [1983]: 'The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where the standard estimators applied to the real and imputed data have substantial biases.' Nevertheless, single imputation techniques, especially LOCF, remain very popular.

5.2.4 Summary

In this section we have reviewed *simple* methods to handle missing data in longitudinal studies and discussed why these should be avoided. While the complete case analysis and some single imputation techniques are valid analysis techniques under the unrealistic assumption of MCAR, LOCF can lead to incorrect inference even in this case. In particular, '[...] when there are missing values, simple methods of analysis do not necessarily imply simple, or even accessible, assumptions, and without understanding properly the assumptions being made in an analysis we are not in the position to judge its validity or value' [Molenberghs and Kenward, 2007].

Thus, it comes as a surprise that these techniques are still widely used and wrongly advocated by authorities such as the EMEA. Wood et al. [2006] reviewed 71 published papers in major medical journals and the methods adopted to handle missing data. Their findings for 34 studies with repeated measures are:

- 46% of the trials used complete case analysis in the primary analysis;
- 15% used LOCF;
- 18% used single imputation techniques: worst case, nearest value and regression imputation;
- 12% used repeated measures ANOVA;
- one trial used multiple imputation (see Section 5.3.3); and
- one trial was analysed through GEE.

Furthermore, 29% of the repeated measures studies performed a sensitivity analysis. However, as noted by Wood et al. [2006]: 'The most common form of sensitivity analysis was LOCF when the primary analysis adopted a complete case analysis.' Given all the limitations of complete case analysis and LOCF this is a rather questionable sensitivity analysis. In the next sections we will review more suitable approaches.

5.3 A Review of Missing Data Methods under MAR and Ignorability

In this section we will present methods that are suitable when a MAR mechanism holds. All presented approaches but one do not explicitly model the missingness process.

5.3.1 Direct Likelihood Approach

The direct likelihood approach is the likelihood-based way of using the available information only, while ignoring the missing data mechanism [Molenberghs and Kenward, 2007]. Inference for the parameter of interest θ can be based on the likelihood

$$\begin{aligned} L_{\boldsymbol{Y}_{obs}}(\theta|\boldsymbol{Y}_{obs}, X) &= \prod_{i=1}^{N} f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta) \\ &= \prod_{i=1}^{N} \int f_{\boldsymbol{Y}_{i,obs},\boldsymbol{Y}_{i,mis}}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,obs}|X_i, \theta)\, \mathrm{d}\boldsymbol{y}_{i,mis}. \end{aligned}$$

Evaluating this likelihood requires integration over the missing data, which can be very complex in the case of non-monotone missingness patterns. The subsequent maximisation with respect to the parameter of interest θ usually requires numerical techniques, such as the *Newton-Raphson* or *Fisher-Scoring* method. Both optimization methods involve calculating the matrix of second derivatives of the likelihood. According to Little and Rubin [2002], the entries in this matrix tend to be complicated functions of θ. Alternative computing strategies for incomplete data problems, which simplify the integration and the maximization, such as the *Expectation Maximization* (EM) algorithm or *Multiple Imputation* were proposed.

5.3.2 Expectation-Maximization Algorithm

The *expectation-maximization algorithm* (EM algorithm) was proposed by Dempster et al. [1977] and is a convenient and widely applicable computational technique that can be used in the case of ignorable missingness where the observed data likelihood is difficult to compute [Molenberghs and Kenward, 2007].

The fundamental idea behind this algorithm is to associate a complete data problem for which maximum likelihood estimation is computationally tractable with the given incomplete data problem. Each iteration of the EM algorithm consists of two steps, the *expectation step* (E-step) and the *maximization step* (M-step). The EM algorithm is easy to implement and converges reliably. Furthermore, the Hessian matrix does not have to be calculated. However, the algorithm is slow to converge if the fraction of missingness is large and in some problems the likelihood in the M-step turns out to have no closed form which complicates the maximization substantially [Molenberghs and Kenward, 2007]. A major drawback of the EM algorithm is that additional steps are required to compute standard errors [Little and Rubin, 2002, Chapter 9].

Detailed accounts on the EM algorithm are given in Dempster et al. [1977], [Little and Rubin, 2002, Chapter 8] and [Molenberghs and Kenward, 2007, Chapter 8].

5.3.3 Multiple Imputation

In Section 5.2 we have reviewed the approach of *filling in* missing values by plausible values based on the observed data. We have discussed that 'a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests' [Schafer, 1999]. Rubin [1987] presented a more principled method to create apparently complete data sets and inference that properly reflects uncertainty due to non-response in the case of an ignorable missingness mechanism. This approach is referred to as *multiple imputation* (MI). It is a simulation-based technique where missing values are replaced by $k > 1$ *Bayesian* draws from the conditional distribution of $Y_{i,mis}$ given $Y_{i,obs}$ and relevant covariates X_i, creating k completed data sets [Molenberghs and Kenward, 2007]. The *data augmentation algorithm*, introduced by Tanner and Wong [1987], is a convenient way to create multiple imputations from this conditional distribution [Schafer, 1997]. This algorithm consists of two steps, the *imputation step* (I-step) and the *posterior step* (P-step). In the I-step the missing data are drawn based on the observed data, covariates and the current parameter estimate. Then, in the P-step, the updated parameter estimate is drawn

based on the current (imputed) data. The resulting imputed data sets define a Markov Chain, which under some conditions, converges to the stationary distribution of $Y_{i,mis}|Y_{i,obs}, X_i$ for all $i \in \{1, ..., N\}$ [Tanner and Wong, 1987]. Other approaches to create multiple imputations are discussed in Little and Rubin [2002] and Molenberghs and Kenward [2007].

The k completed data sets can then be analysed using standard methods such as GEE or (generalized) linear mixed models, resulting in k sets of parameter estimates and associated covariance matrices. Rubin [1987] presents rules to combine these k estimates using simple arithmetic to produce overall estimates and confidence intervals that adequately incorporate missing data uncertainty.

One advantage of the MI approach is that the model used to create the imputed data sets and the model used subsequently to analyze the 'completed' data sets can be considered separately [Schafer, 1999; Molenberghs and Kenward, 2007]. According to Schafer [1999], the MI leads to valid inferences with perhaps some loss of power, when the imputer's model makes fewer assumptions than the analyst's model. Furthermore he states that 'the only serious danger of inconsistency arises when the imputer makes more assumptions than the analyst and these additional assumptions are unwarranted. For example, consider a situation where a variable is imputed under no-interactions regression model and the analyst subsequently looks for evidence of interactions; if interactions are present, then the MI estimates will be biased toward null values.' In addition, the imputation model should allow for all covariates that are known to be predictive for missingness [Molenberghs and Kenward, 2007].

Note that MI is most commonly used under an ignorable missingness mechanism. However, extensions to account for informative missingness settings exist, see Molenberghs and Kenward [2007] and citations therein.

A detailed review of MI is beyond the scope of this thesis. Many papers and monographs discussing the properties and different extensions to the classical multiple imputation approach were published in recent years [Rubin, 1987, 1996; Little and Rubin, 2002; Schafer, 1997, 1999; Horton and Lipsitz, 2001; Molenberghs and Kenward, 2007].

5.3.4 Inverse Probability Weighting

In Section 5.2 we have discussed that analyzing complete cases only usually leads to inefficient estimators and incorrect inference under MAR or MNAR mechanisms. One way of removing the bias under MAR is by reweighting the information of the completers, see for example Horvitz and Thompson [1952] and Alho [1990].

This approach has found application in combination with the semi-parametric generalized estimating equations. GEE usually require the strong assumption of MCAR to yield valid inference. However, by incorporating specific weights these can be extended to yield valid inference under a MAR missingness process. The GEE is formulated based on the complete cases only, but the contribution of every *complete case* is weighted. For a given subject and point in time, the weight is derived from the inverse probability of dropping out at that measurement occasion [Molenberghs and Kenward, 2007]. The resulting estimating equation is referred to as *weighted generalized estimating equation* (WGEE). Given a correct outcome and missingness model, the WGEE leads to consistent estimators under MAR. However, the estimators remain inefficient as only the information of completers is used. In order to improve the efficiency, the WGEE were extended by adding a term of expectation zero. This additional term is a function of the partially observed data, i.e. the data of non-completers. The additional term does not affect the unbiasedness and leads to *fully efficient* estimators, see Molenberghs and Kenward [2007] and citations therein. By fully efficient estimator, we mean that given the chosen semi-parametric model the estimator attains the minimum variance possible among all consistent asymptotically normal estimators [Molenberghs and Kenward, 2007].

Overall, efficient WGEE estimators require three models: (1) one model for the outcome of interest Y, (2) one model for the missingness mechanism and (3) one model for the partially observed data, i.e. a model for $Y_{mis}|Y_{obs}$ which is compatible with the model for the outcome of interest in (1) [Molenberghs and Kenward, 2007]. A misspecification of model (1) usually leads to inconsistent estimators for all parameters. However, given a correct specification of model (1) and '*either* model (2) or model (3) is wrong, but not

both, the estimators in model (1) are still consistent' [Molenberghs and Kenward, 2007]. Therefore, the resulting estimates are termed *doubly robust*.

In this thesis we have not focused on GEE for the analysis of longitudinal data. Thus, we also neglected WGEE and double robust estimators in the forthcoming analysis of the CAST data set. We believe that these approaches can be more robust to misspecifications than the fully parametric approaches presented, e.g. the direct likelihood approach. However, WGEE require a model for the missingness mechanism and we agree with Schafer and Graham [2002] who note: 'We acknowledge that these weighting techniques may be useful in some circumstances. However, as a general principle, we also believe that a researcher's time and effort are probably better spent building an intelligent model for the data rather than building a good model for the missingness, especially if departures from MAR are not a serious concern.'

5.4 A Review of Missing Data Methods under Non-ignorable or Informative Missingness

All missing data methods presented in Section 5.3 but the inverse probability weighting approach focus on ignorable missingness. Inference for the parameter of interest θ is based on the likelihood

$$L_{\boldsymbol{Y}_{obs}}(\theta|\boldsymbol{Y}_{obs}, X) = \prod_{i=1}^{N} f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta),$$

where the function $f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|X_i, \theta)$ is obtained by integrating $y_{i,mis}$ out of the complete data density $f_{(\boldsymbol{Y}_{i,obs}, \boldsymbol{Y}_{i,mis})}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}|X_i, \theta)$.

In the case of informative missingness, the missingness process is not ignorable and inference for θ needs to be based on the joint likelihood of $\boldsymbol{Y}_{obs}$ and $\boldsymbol{R}$:

$$L_{(\boldsymbol{Y}_{obs}, \boldsymbol{R})}(\theta, \phi|\boldsymbol{Y}_{obs}, R, X, W) = \prod_{i=1}^{N} \int f_{(\boldsymbol{Y}_i, \boldsymbol{R}_i)}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi)\, \mathrm{d}\boldsymbol{y}_{i,mis}.$$

That is, we need to explicitly model the missingness in addition to the model for the out-

come $\boldsymbol{Y}$.

Three different model families, basing on different factorisations of the joint density $f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi)$, are distinguished: *selection models*, *pattern-mixture models* and *shared-parameter models* [Molenberghs and Kenward, 2007]. A selection model factorises the joint density $f_{\boldsymbol{Y}_i,\boldsymbol{R}_i}(\cdot)$ into the marginal outcome density $f_{\boldsymbol{Y}_i}(\cdot)$ and the missingness model, conditional on the measurements $f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\cdot)$. 'The term selection model comes from the econometric literature [Heckman, 1976] and it can be seen that a subject's missing values are 'selected' through the probability model, given their measurements whether observed or not' [Kenward, 1998]. In a pattern-mixture model, we use the alternative factorisation, where the joint density of the full data is factorised into the marginal missingness density $f_{\boldsymbol{R}_i}(\cdot)$ and the measurement process, conditional on the missingness pattern $f_{\boldsymbol{Y}_i|\boldsymbol{R}_i}(\cdot)$. Thus, 'pattern-mixture models stratify the population by the pattern of dropout, implying a model for the whole population that is a mixture over patterns' [Little, 1995]. In a shared parameter model, the density of the full data is modelled through the incorporation of random effects, which drive both the outcome and the missingness process. These model families will be reviewed in the next sections.

5.4.1 Selection Models

Selection models were first used in the econometric literature [Heckman, 1976, 1979] and base on the following factorisation of the joint density:

$$f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, \boldsymbol{r}_i|X_i, W_i, \theta, \phi) = f_{\boldsymbol{Y}_i}(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}|X_i, \theta)\; f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}, W_i, \phi)\,.$$

In order to fit a selection model, we need to model the marginal outcome process and the conditional missingness process. The conditional missingness process specifies the relation between the probability of an observation being missing and the outcome of interest and covariates. For example, the probability of dropping out may depend on observed quantities only, leading to MAR, or on unobserved quantities, yielding a MNAR process. In particular,

the taxonomy of the different missing data mechanisms introduced by Rubin [1976] is based on the selection model framework, see Section 5.1. Typically a logistic regression model [Diggle and Kenward, 1994; Kenward, 1998; Carpenter et al., 2002] or a probit model [Heckman, 1979; Little, 1995] is formulated for the hazard of dropping out.

Based on the selection model factorisation and different models for the conditional missingness process, some researchers have attempted to test a MAR null hypothesis against a MNAR alternative [Diggle and Kenward, 1994]. However, it has been shown that these tests are very sensitive to the model assumptions, see discussion below. More generally, Molenberghs et al. [2008] have shown that an assessment of MAR versus MNAR based on the observed data only is not possible.

Fitting selection models is usually computationally demanding, especially in the case of non-monotone missingness patterns. One main difficulty arises through the required integration over the missing values. Furthermore, likelihood surfaces tend to be flat or awkwardly shaped which makes selection models difficult to use [Molenberghs and Kenward, 2007].

However, as mentioned above, the technical difficulties do not pose the only drawback of selection models. Like all models that attempt to model MAR or MNAR missingness, selection models are based on untestable assumptions: 'It is assumed in the modelling approach taken here that the relationship among the measurements from a subject are the same whether or not some of these measurements are unobserved due to dropout. It is this assumption, combined with the adoption of an explicit model linking outcome and dropout probability, that allows us to infer something about the MNAR nature of the dropout process. Given the dependence of the inferences on untestable assumptions, care is needed in the interpretation of the analysis' [Molenberghs and Kenward, 2007].

Thus, although in parametric selection models all parameters are usually identifiable, this identification is driven by the parametric assumptions for the conditional model $f_{\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}}(\cdot)$ (implied by the model for the complete data $\boldsymbol{Y}_i$) and the explicit missingness process model $f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}(\cdot)$. This can be seen from the following relationship, where we sup-

press dependence on covariates [Kenward, 1998]:

$$\begin{aligned} f_{(\boldsymbol{Y}_{i,obs},\boldsymbol{R}_i)}\left(\boldsymbol{y}_{i,obs},\boldsymbol{r}_i|\theta,\phi\right) &= \int f_{\boldsymbol{Y}_i}\left(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|\theta\right) f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}\left(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\phi\right) \mathrm{d}\boldsymbol{y}_{i,mis} \\ &= f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|\theta) \\ &\quad \times \int f_{\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,mis}|\boldsymbol{y}_{i,obs},\theta) f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}\left(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\phi\right) \mathrm{d}\boldsymbol{y}_{i,mis} \\ &= f_{\boldsymbol{Y}_{i,obs}}(\boldsymbol{y}_{i,obs}|\theta)\ \mathbb{E}_{\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}}\left[f_{\boldsymbol{R}_i|\boldsymbol{Y}_i}\left(\boldsymbol{r}_i|\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\phi\right)\right]. \end{aligned}$$

For a simple example, Kenward [1998] has shown that inference for the missingness process can critically depend on the assumed distribution for $\boldsymbol{Y}_{i,mis}|\boldsymbol{Y}_{i,obs}$. Changing just the tails of this distribution can lead to substantially different conclusions, see also Daniels and Hogan [2008]. The fact that selection models are highly sensitive to parametric assumptions and that the identifiability problem is 'masked' [Molenberghs and Kenward, 2007] has led some researchers to avoid such modelling. However, we note that all models for informative missingness are based on untestable assumptions and agree with Kenward [1998] that these models could be very useful in the context of a sensitivity analysis, where we want to assess the sensitivity of our conclusions to various plausible assumptions about the reasons for missingness and the complete data model. The main appeal of selection models is that they fit nicely with the classification of missing data mechanisms. Furthermore, in most settings they have an intuitive appeal: 'It is natural to think about how a participant's data values influence his or her probability of dropping out [...]' [Graham, 2009]. Additionally, selection models enable direct inference for the parameter of interest θ because the marginal outcome process is modelled. This is not the case for pattern-mixture models as will be discussed in the next section.

We will fit selection models to the CAST data set and provide an extension to the 'traditional' selection model in Chapter 6.

5.4.2 Pattern-Mixture Models

The pattern-mixture model (PMM) factorisation of the joint density $f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(\cdot)$ is given by

$$f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis},\boldsymbol{r}_i|X_i,W_i,\theta,\phi) = f_{\boldsymbol{Y}_i|\boldsymbol{R}_i}(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|\boldsymbol{r}_i,X_i,\theta)\; f_{\boldsymbol{R}_i}(\boldsymbol{r}_i|W_i,\phi)\,. \tag{5.3}$$

The sample is essentially stratified by the observed patterns and different models are fitted to each pattern. Fitting different models for different dropout patterns is generally appealing; e.g. the average evolution over time for completers and non-completers may be quite different. However, the price to pay for this generalisation is that pattern-mixture models are by construction under-identified. Several approaches to overcome the under-identification were proposed, some of which will be touched on below.

A further drawback of PMMs is that the missing data terminology introduced by Rubin [1976] is not naturally applicable to pattern-mixture models. For monotone missingness a link between MAR missingness processes and a certain type of identifying restriction has been developed [Molenberghs et al., 1998]. Generally, research on pattern-mixture models focuses on monotone missingness, see Molenberghs et al. [1998]; Molenberghs and Kenward [2007] and citations therein. We will also focus on monotone missingness in this section.

For a longitudinal study with M different observation times and monotone missingness, we are able to distinguish between M different dropout patterns, where the pattern denotes how many observations were made. Let $d \in \{1,...,M\}$ denote the missingness pattern and P be the associated random variable. Then, suppressing the subscript i, covariates and parameters, the joint density of $Y = (Y_1, Y_2, ..., Y_M)^\top$ and $P = d$ is given by

$$\begin{aligned} f_{(\boldsymbol{Y},P)}(y_1,...,y_m,P=d) &= f_{\boldsymbol{Y}|P}(y_1,...,y_M|P=d)\, f_P(d) \\ &\overset{(*)}{=} f_d(y_1,...,y_M)\, f_P(d) \\ &= f_d(y_1,...,y_d)\, f_d(y_{d+1},...,y_M|y_1,...,y_d)\, f_P(d) \end{aligned}$$

$$= \underbrace{f_d(y_1, ..., y_d)}_{known} \prod_{s=0}^{M-d-1} \underbrace{f_d(y_{M-s}|y_1, ..., y_{M-s-1})}_{unknown} \underbrace{f_P(d)}_{known}, \quad (5.4)$$

where $(*)$ follows from defining $f_{\boldsymbol{Y}|P}(y_1, ..., y_M|P = d) := f_d(y_1, ..., y_M)$. Following Little [1994], the inestimable parameters for the unknown conditional densities in equation (5.4) are set equal to functions of the parameters describing the distribution of the completers. Using these identifying restrictions is not the only way to overcome under-identification. Alternatives, such as using patterns as explanatory variables, are discussed in Molenberghs and Kenward [2007].

We will illustrate the approach of using identifying restrictions for the CAST study design. Assume a given subject withdraws from the study after the assessment at week four, i.e. $d = 2$, $y_{obs} = (y_0, y_4)^\top$ and $y_{mis} = (y_{12}, y_{39})^\top$. With the notation $f_{\boldsymbol{Y}|P}(y_1, ..., y_M|P = d) := f_d(y_1, ..., y_M)$ the distribution of $\boldsymbol{Y} = (Y_0, Y_4, Y_{12}, Y_{39})^\top$ for the second missingness pattern is given by

$$\begin{aligned} f_2(y_0, y_4, y_{12}, y_{39}) &= f_2(y_0, y_4)\, f_2(y_{12}, y_{39}|y_0, y_4) \\ &= \underbrace{f_2(y_0, y_4)}_{known} \underbrace{f_2(y_{12}|y_0, y_4)\, f_2(y_{39}|y_0, y_4, y_{12})}_{unknown}. \end{aligned}$$

In order to identify the unknown conditional densities of unobserved components given a set of observed components, we use identifying restrictions:

$$f_2(y_{12}|y_0, y_4) = \omega f_3(y_{12}|y_0, y_4) + (1 - \omega) f_4(y_{12}|y_0, y_4),$$

and

$$f_2(y_{39}|y_0, y_4, y_{12}) = f_4(y_{39}|y_0, y_4, y_{12})$$

where unavailable information is borrowed from patterns for which the required information is available.

Different choices for ω were proposed [Little, 1993, 1994; Molenberghs et al., 1998; Kenward et al., 2003; Molenberghs and Kenward, 2007] and yield different missingness mechanisms, e.g.

- **complete case missing value** (CCMV):

  $\omega = 0$, i.e. information which is unavailable is always borrowed only from the model for the completers;

- **neighbouring case missing value** (NCMV):

  $\omega = 1$, i.e. information which is unavailable is always borrowed from the first pattern which observes the missing observation;

- **available case missing value** (ACMV):

  Here we use all the available cases to calculate the weight, i.e.

  $$\omega = \frac{\alpha_3 f_3(y_0, y_4)}{\alpha_3 f_3(y_0, y_4) + \alpha_4 f_4(y_0, y_4)},$$

  where α_3 and α_4 are the fractions of observations in pattern $d = 3$ and $d = 4$, respectively. Molenberghs et al. [1998] showed that for longitudinal data with monotone missingness, MAR is equivalent to the available missing value case. This equivalence does not hold in case of non-monotone missingness.

- **non-future dependence missing value**:

  This missingness mechanism describes a MNAR process in which missingness is allowed to depend on past measurements and on the present, possibly unobserved outcome, but not on future ones. We refer to Kenward et al. [2003] for details regarding this identifying restriction.

Using one of these identifying restrictions Thijs et al. [2002] and Molenberghs and Kenward [2007] present an overall strategy for fitting pattern-mixture models. We will touch upon this for the specific case of the CAST data set:

1. Specify and fit a model to the pattern-specific identifiable densities, $f_1(y_0)$, $f_2(y_0, y_4)$, $f_3(y_0, y_4, y_{12})$ and $f_4(y_0, y_4, y_{12}, y_{39})$. For patterns $P \in \{2, 3, 4\}$ the parameter vector of interest, θ_P, is estimated. Then all of them are grouped in one vector Ξ with distribution G.

2. To properly account for the uncertainty with which these parameters are estimated, we draw a realisation from this distribution.

3. In this case, there are only four patterns. Dependent on the pattern, identification takes the following form:

   - $f_4(y_0, y_4, y_{12}, y_{39}) = f_4(y_0, y_4, y_{12}, y_{39})$;
   - $f_3(y_0, y_4, y_{12}, y_{39}) = f_3(y_0, y_4, y_{12}) f_4(y_{39}|y_0, y_4, y_{12})$;
   - with $\omega_{12,3} + \omega_{12,4} = 1$ we set

$$\begin{aligned} f_2(y_0, y_4, y_{12}, y_{39}) &= f_2(y_0, y_4)[\omega_{12,3} f_3(y_{12}|y_0, y_4) + \omega_{12,4} f_4(y_{12}|y_0, y_4)] \\ &\quad \times f_4(y_{39}|y_0, y_4, y_{12}); \end{aligned}$$

   - with $\omega_{4,2} + \omega_{4,3} + \omega_{4,4} = 1$ and $\omega_{12,3} + \omega_{12,4} = 1$ we set

$$\begin{aligned} f_1(y_0, y_4, y_{12}, y_{39}) &= f_1(y_0)[\omega_{4,2} f_2(y_4|y_0) + \omega_{4,3} f_3(y_4|y_0) + \omega_{4,4} f_4(y_4|y_0)] \\ &\quad \times [\omega_{12,3} f_3(y_{12}|y_0, y_4) + \omega_{12,4} f_4(y_{12}|y_0, y_4)] \\ &\quad \times f_4(y_{39}|y_0, y_4, y_{12}) \end{aligned}$$

   Select an identification method of choice and calculate the weights ω_{ij}.

4. Using this identification method, determine all the unknown conditional distributions of the unobserved outcomes, given the observed ones.

5. Using standard multiple imputation methodology, we draw multiple imputations for the unobserved components, given the observed ones and the pattern-specific densi-

ties ($\Rightarrow L$ completed data-sets per pattern).

6. Repeat all the steps except the first S times ($\Rightarrow L \times S$ completed data-sets per pattern).

7. For every pattern analyse the $L \times S$ multiply imputed sets of data using the models of choice.

8. Inference for each pattern can be conducted in a multiple imputation way.

We see that fitting pattern-mixture models yields pattern-specific parameter estimates. However, these are rarely of scientific interest. In most cases, we are aiming to provide inference for population parameters across all missingness patterns. Drawing this inference based on pattern-mixture models is not precluded: For each parameter the overall estimate can be calculated as the weighted average of the pattern specific estimates. The weights are estimated as the pattern probabilities, i.e. the proportion of subjects in a given pattern. The associated standard errors can be calculated through the *delta-method* [Molenberghs and Kenward, 2007].

The fitting strategy also reveals that generally a large number of parameters is involved in estimating pattern-mixture models. Assume we would like to use the ACMV restrictions. Then we need to estimate the pattern-specific parameter of interest $\hat{\theta}_P$, $P \in \{2, 3, 4\}$ based on $f_2(y_0, y_4)$, $f_3(y_0, y_4, y_{12})$ and $f_4(y_0, y_4, y_{12}, y_{39})$, respectively. In the case of the CAST study, the outcome model is quite complex and the number of subjects dropping out after week 4, i.e. with drop-out pattern $P = 2$, is not large enough to estimate the pattern-specific estimate $\theta_{P=2}$. Thus, we are not able to borrow information from pattern 2 in order to identify parameters in pattern 1. This observation is confirmed by Hogan and Laird [1997] and Molenberghs and Kenward [2007] who noted that, to estimate the large number of parameters in a pattern-mixture model, one has to fulfil the delicate condition that each dropout pattern occurs sufficiently often. In Section 5.5, we will therefore focus on CCMV restrictions, where all unavailable information is imputed based on $\hat{\theta}_{P=4}$.

In Section 5.4.1, we discussed that conclusions based on selection models are very sensitive to the parametric assumptions on the full data $\boldsymbol{Y}$ and the conditional missingness

model $\boldsymbol{R}|\boldsymbol{Y}$. Pattern-mixture models do not suffer from this substantial sensitivity. However, the assumptions underlying the identifying restrictions are not less strong. To cite Graham [2009]: 'Estimation of population effects is possible through identifying restrictions, and the observed data provide no evidence whatsoever to support or contradict these assumptions.' Nevertheless, we believe that Molenberghs et al. [1998] have a point by stating that the pattern-mixture approach 'is more honest, because parameters for which the data provide information are clearly distinguished from parameters for which there is no information at all.' In line with the case of selection models, researchers recommend the use of pattern-mixture models within a sensitivity analysis.

We conclude this section by noting that in the case of a MCAR mechanism pattern-mixture and selection models coincide, because the outcome process and missingness process are independent.

5.4.3 Shared Parameter Models

Until now we have seen two ways of factorising the joint density of the outcome and missingness process, namely the selection model and the patten-mixture model. A third way is to write the joint distribution in terms of a latent variable $\boldsymbol{B}_i$, which drives both the outcome and the missingness process:

$$f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i)}(\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, C_i, \theta, \psi) = \int f_{(\boldsymbol{Y}_i,\boldsymbol{R}_i,\boldsymbol{B}_i)}(\boldsymbol{y}_i, \boldsymbol{r}_i, \boldsymbol{b}_i|X_i, W_i, C_i, \theta, \phi, \xi)\, \mathrm{d}b_i,$$

where $\boldsymbol{B}_i$ follows a parametric model, $\boldsymbol{B}_i|C_i \sim \mathcal{P}(\xi)$, where C_i is a set of covariates. These models are called *shared parameter models*. In this work we will not focus on these models, we refer to Molenberghs and Kenward [2007] and citations therein.

5.5 Sensitivity Analysis for CAST

In this section, we want to explore the sensitivity of conclusions based on the non-linear mixed model presented in Section 3.5 and different approaches to handle missing data.

Before using any approach to deal with missing data, it is worth reflecting which missingness mechanism may be operating and which explanatory variables influence the missingness. The latter point is of special interest under the MNAR assumption, where the missingness process has to be modelled. We know from the thesis [Nakash, 2007] which explanatory variables are more likely to influence the response of participants in the CAST trial. The analysis of the effect of age on response in CAST demonstrates a positive effect with the best responders older than 44. Furthermore, females were generally better at responding than males, a pattern which has been noted in the health related survey literature. There were no significant differences on the type of employment. Furthermore, no effect in terms of the treatment received as part of CAST between the responders and non-responders was detected. These results are based on a logistic regression analysis performed for each follow up time point separately.

We now turn to the relation between missingness and the outcome of interest. As mentioned in Section 2.5, patients with a low baseline score tend to not return their questionnaire at the 4 week follow-up point. Also, participants with a high 4 week score or a high 12 week score show the tendency to not return their questionnaire for the following point in time. Thus, missingness seems to depend on the observed outcome of interest, leading to a MAR missingness process. However, we already discussed that a formal distinction between MNAR and MAR based on the observed data only is not possible. In fact, we have reasons to believe that missingness also depends on unobserved scores, i.e. that the operating missingness process is MNAR. A study to investigate response issues surrounding the CAST trial revealed that patients who considered themselves to have made full recovery did not return their questionnaires [Nakash et al., 2008].

In order to investigate the sensitivity of conclusions for CAST, we will use different methods to handle missing data. We will perform: a complete case analysis (CC); a last observation carried forward analysis (LOCF); an analysis based on multiple imputation (MI, 5 imputations); an analysis using the direct likelihood approach (DL) and finally we will fit a pattern-mixture model (PMM) using the CCMV identifying restrictions. The imputations

for the MI approach were created separately for each randomisation group and are based on an imputation model that accounts for the age of patients. The imputations were created using the software package `MI` (`SAS`), which assumes that $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ is normally distributed. We are aware that this imputation model leaves room for discussion. In Chapter 6 we will also fit selection models.

For all approaches except the pattern-mixture model the following mixed model will be assumed for the outcome vector Y_i, $i \in \{1, ..., N\}$:

$$\begin{aligned} \mathbf{Y}_i | U_i &\overset{ind.}{\sim} \mathcal{N}_{M_i}\left(\mu_i, \sigma^2 I_{M_i}\right); \qquad U_i \overset{iid}{\sim} \mathcal{N}(0, D^2); \quad \text{and} \\ \mu_{i,j} &= g(\check{\boldsymbol{x}}_{i,j}, s_i, t_{i,j}, \theta_i) \quad \text{for } j \in \{1, ..., M_i\}, \end{aligned}$$

where U_i, a subject-specific effect, has a normal distribution with mean zero and variance D^2, I_{M_i} is the M_i-dimensional identity matrix, $\mu_i = (\mu_{i,1}, ..., \mu_{i,M_i})^\top$, θ_i the parameter vector of interest and $g(\cdot)$ the model function of interest:

$$g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta_i) = \frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}) \cdot t\} \left(\dfrac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}} - 1\right) + 1} + U_i, \tag{5.5}$$

where $\check{\boldsymbol{x}}_{i,j} = \check{\boldsymbol{x}}_i = age_i$. We note that in previous chapters we have transformed the covariate age. We refrain from doing so in this chapter; this inconsistency reflects the development of the work throughout the course of the PhD. We also note that for convenience we assume a compound symmetry covariance structure in this section, see Section 3.5 and Chapter 4. The resulting parameter estimates for DL, CC, LOCF, MI and outcome model (5.5) are summarized in Table 5.1 on page 121.

We observe that based on the DL, CC and MI approaches, intercepts and final recovery scores are nearly equal. The standard deviations for the parameters obtained by the CC analysis are somewhat larger due to the reduced information. The estimated intercept and the final recovery score obtained for the LOCF analysis differ substantially from the remaining results. We observe a lower final recovery level because potentially lower scores

Variable		DL			CC			LOCF			MI		
		Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.
Intercept	β_0	40.07	2.26	-	40.14	2.86	-	43.92	2.51	-	40.09	2.29	-
Final Recovery	β_1	89.73	2.33	-	89.68	2.92	-	83.24	2.48	-	89.75	2.20	-
Treatment-Specific Recovery Rates													
Rate of Tubigrip	β_{21}	0.42	0.06	-	0.45	0.07	-	0.27	0.05	-	0.40	0.06	-
Contrast: BKC	β_{22}	-0.13	0.04	0.003	-0.10	0.04	0.025	-0.15	0.04	< 0.001	-0.13	0.04	0.002
Contrast: Aircast	β_{23}	-0.06	0.04	0.128	0.01	0.04	0.792	-0.07	0.03	0.035	-0.06	0.04	0.089
Contrast: Bledsoe	β_{24}	-0.01	0.04	0.701	0.01	0.04	0.833	-0.05	0.03	0.111	-0.01	0.04	0.711
Age Effects on Intercept, Final Recovery Level and Recovery Rate													
Age-effect on Int.	α_0	0.04	0.07	0.613	0.05	0.09	0.555	-0.08	0.08	0.337	0.04	0.07	0.577
Age-effect on Max.	α_1	-0.33	0.07	< 0.001	-0.32	0.09	< 0.001	-0.26	0.08	< 0.001	-0.33	0.07	< 0.001
Age-effect on Rate	α_2	-0.01	0.001	< 0.001	-0.01	0.002	< 0.001	-0.003	0.001	0.011	-0.01	0.001	< 0.001
Variance Components													
Within-Var.	σ	13.64	0.26	-	13.55	0.29	-	14.24	0.25	-	13.58	0.25	-
Between-Var.	D	12.20	0.51	-	12.33	0.59	-	14.99	0.55	-	12.74	0.54	-

Table 5.1: Overview of the parameter estimates, standard errors and p-values for different approaches (DL: direct likelihood approach; CC: complete case analysis; LOCF: last observation carried forward; MI: multiple imputation (5 imputations)). The p-values are reported only for the components of θ that might be zero.

are carried forward to the endpoint. The lower intercept may be due to potentially lower values at the second point in time, which leads to an adjustment of the intercept.

For the recovery rate of Tubigrip, we observe quite similar estimates based on the DL, CC and MI analysis. The results based on the LOCF analysis differ, suggesting a considerably lower recovery rate.

All approaches detect a significant difference in the rate of improvements between Tubigrip and BKC. Only the LOCF analysis confirms a significant treatment difference between Tubigrip and Aircast. The MI approach suggests a marginally significant treatment difference between Tubigrip and Aircast. None of the methods finds a significant treatment difference for the Bledsoe boot. We note that the standard errors of the contrasts resulting from the LOCF analysis are the smallest across all approaches.

For all methods, the null hypothesis $\alpha_0 = 0$ is not rejected at a significance level of 5%. The remaining *age*-effect parameters are very similar for the DL, CC and MI approach, but are somewhat different for the LOCF approach. In line with results seen in previous chapters, older patients achieve a lower final recovery level than younger patients. Furthermore, all approaches confirm that older participants recover less quickly than younger patients.

The inter- and intra-individual variance parameters and their standard deviations are nearly identical for the DL, the CC and the MI approach. The estimates are larger for the LOCF analysis.

Let us now turn to pattern-mixture models. In this context we focus on the data set with monotone missingness. We obtained the data set with monotone missingness by deleting $y_{i,k}$ for all $k > j$ when $r_{i,j} = 0$, $i \in \{1, ..., N\}$. Attempts to fit a pattern-mixture model based on outcome model (5.5) to this modified data set failed, as the pattern-specific parameter $\theta_{P=4}$ used to identify unknown conditional densities could not be estimated due

to convergence problems. We therefore decided to fit the following, slightly simpler model:

$$g(\check{\boldsymbol{x}}_{i,j}, s_i, t, \theta_i) = \frac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)]) \cdot t + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}\} \left(\dfrac{\beta_1 + \alpha_1^\top \check{\boldsymbol{x}}_{i,j}}{\beta_0 + \alpha_0^\top \check{\boldsymbol{x}}_{i,j}} - 1 \right) + 1} + U_i, \tag{5.6}$$

where $\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)] + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}) \cdot t\}$ in the denominator of the model equation (5.5) is replaced by $\exp\{-([\beta_{21} - \beta_{2,s_i}\, \mathbb{1}(s_i \neq 1)]) \cdot t + \alpha_2^\top \check{\boldsymbol{x}}_{i,j}\}$. The estimates for each pattern and parameter are shown in Table 5.2, page 124. Note that the intercepts and final recovery levels accounting for age and based on model (5.6) are no longer given by $\beta_0 + \alpha_0 \cdot age_i$ and $\beta_1 + \alpha_1 \cdot age_i$, respectively. Table 5.3 gives the estimated intercept and final recovery level for an average individual, i.e. with $age = \text{median}(age) = 27$ and $\hat{U}_i = 0$.

Comparing the parameter estimates and associated standard errors across all patterns reveals that especially the variance components and standard errors vary. Given the varying sample sizes for the different patterns this is not surprising.

The intercepts of pattern 1 and pattern 2 are considerably smaller than those of patterns 3 and 4, see Table 5.3. Due to the small number of patients in pattern 2 and pattern 3, the associated standard errors are substantially larger than for the remaining patterns. The final recovery levels of the first three patterns are very similar. The other estimates are noticeably different, however, there seems to be no consistent structure. This is not very surprising, as the different outcome evolutions are assumed for each pattern.

Regarding the treatment contrasts we note that, in the second pattern, the differences between Tubigrip and BKC or Aircast are much larger than in the other patterns. This confirms our assumption that participants who recover faster drop out earlier than others. Furthermore, only pattern 2 and pattern 4 confirm a significant treatment difference between the recovery rates of Tubigrip and BKC. In pattern 2, also a significant treatment contrast for Tubigrip versus Aircast is suggested. For all other patterns this difference remains insignificant. None of the patterns reveal a difference in the recovery rate of Tubigrip and Bledsoe boot. Note that in order to get an overall treatment effect we would need to

Variable		Pattern 1			Pattern 2			Pattern 3			Pattern 4		
		Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.
Intercept$^{(*)}$	β_0	42.05	4.73	-	45.27	17.71	-	43.37	11.73	-	40.31	4.39	-
Final Recovery$^{(*)}$	β_1	86.89	4.91	-	90.34	8.15	-	89.18	7.18	-	89.96	2.40	-
Treatment-Specific Recovery Rates													
Rate of Tubigrip	β_{21}	0.262	0.08	0.04	0.176	0.05	< 0.01	0.251	0.11	0.04	0.254	0.04	< 0.01
Contrast: BKC	β_{22}	-0.147	0.08	0.10	-0.421	0.17	0.01	-0.126	0.15	0.40	-0.112	0.05	0.04
Contrast: Aircast	β_{23}	-0.065	0.06	0.29	-0.483	0.16	< 0.01	-0.131	0.13	0.32	-0.003	0.05	0.95
Contrast: Bledsoe	β_{24}	-0.031	0.09	0.76	-0.073	0.09	0.43	-0.039	0.14	0.78	-0.025	0.05	0.61
Age Effects on Intercept, Final Recovery Level and Recovery Rate													
Age-effect on Int.	α_0	0.454	0.10	< 0.01	0.300	0.68	0.66	-0.327	0.48	0.53	-0.183	0.58	0.75
Age-effect on Max.	α_1	-0.289	0.16	0.10	-0.317	0.28	0.27	-0.345	0.27	0.20	-0.364	0.07	< 0.01
Age-effect on Rate	α_2	-0.035	0.009	< 0.01	-0.029	0.05	0.56	0.014	0.04	0.75	0.011	0.04	0.76
Variance Components													
Within-Var.	σ	6.38	0.93	-	10.06	0.80	-	11.86	0.73	-	13.67	0.29	-
Between-Var.	D	8.24	0.98	-	11.74	1.95	-	14.19	1.94	-	12.32	0.59	-
# Subjects		94			39			50			376		

Table 5.2: Overview of the parameter estimates, standard errors and p-values for the different patterns assuming monotone missingness and using CCMV identifying restrictions (4 imputations). Note that the intercept and final recovery level for this modified model (5.6) have a different interpretation to those obtained based on model (5.5), see discussion in Section 5.5 and Table 5.3. The p-values are reported only for the components of θ that might be zero.

Parameter	Pattern 1	Pattern 2	Pattern 3	Pattern 4
Intercept	36.38	37.71	42.05	41.65
Final Recovery Level	79.09	81.78	79.87	85.05

Table 5.3: Estimated intercepts and final recovery scores for the different patterns, $age_i = 27$ and $\hat{U}_i = 0$ using a pattern-mixture model with CCMV identifying restrictions and five imputations.

combine the results for all patterns, taking into account the sample size of each pattern. Related work can be found in Molenberghs and Kenward [2007].

5.6 Summary

In this chapter, we have reviewed the issue of missing data in the analysis of longitudinal studies. We present the statistical framework introduced by Rubin [1976] and popular methods to handle missing data. Their respective merits and drawbacks are depicted. In this context, we argue why 'ad hoc' methods should be avoided and present alternatives, such as the direct likelihood approach or the multiple imputation approach.

In the introduction and Section 5.4, we stress that it is not possible to distinguish MAR and MNAR missingness processes based on the observed data only. We discuss that a sensitivity analysis, where the stability of the conclusions is investigated under different assumptions, is a valuable approach to analyze incomplete data.

In Section 5.5 we perform a sensitivity analysis for the CAST data set, where the results based on a complete case analysis, the last observation carried forward approach, the direct likelihood approach, multiple imputations and a pattern-mixture model are compared. This sensitivity analysis will be continued in the next chapter, where we aim to account for informative missingness through selection models.

Chapter 6

Adjusting for Missingness through the Reminder Process

6.1 Introduction

In Chapter 5 we have discussed that under *non-ignorability* and MNAR, the measurement and missingness processes need to be modelled jointly. Methods such as *pattern-mixture models*, *shared parameter models* and *selection models* have been proposed for this case. We have reviewed these model families and applied pattern-mixture models to the CAST data set in Section 5.4 and Section 5.5, respectively.

In this chapter, we want to resume the sensitivity analysis started in Section 5.5. In this context, we aim to account for informative missingness through selection models.

In order to fit a selection model, we need to formulate models for the marginal measurement process and the conditional missingness process, see Section 5.4.1. Assuming a monotone missingness pattern, a logistic model for the dropout process in combination with a multivariate normal linear model for the measurement process was proposed [Diggle and Kenward, 1994]. The assumption of monotone missingness has been relaxed by Baker [1995] and Troxel et al. [1998]. However, Baker [1995] discusses models for repeated binary data and one of the main challenges of selection models - the integration over the

missing data - reduces to calculating simple sums. In contrast, Troxel et al. [1998] analyze continuous longitudinal data. A logistic and probit model for the missingness process and a multivariate normal linear model for the outcome of interest are proposed. The missing data model allows the probability of non-response to depend on current and previous outcomes, see also Diggle and Kenward [1994] and Baker [1995]. However, in order to facilitate the integration and the construction of the likelihood, a first-order Markov dependence structure for the measurement vector is chosen.

In the CAST study 10% of the patients exhibit a non-monotone missingness pattern. Although the methodology presented in this chapter is able to account for non-monotone missingness patterns, we will focus on monotone missingness. In particular, we deleted all those observations that were made after a patient failed to return a previous questionnaire. Discussion for the non-monotone case can be found in Section 6.4 and Section 9.2, respectively.

In the following sections, the traditional selection model is extended in three ways. Firstly, none of the aforementioned approaches includes additional information about the missingness process, which can be very helpful in obtaining a better understanding of the missing data mechanism [Wood et al., 2006]. This information usually consists of proxy outcomes [Jackson et al., 2010], follow-up studies on a sample of non-responders [Cooke et al., 2009], collection of the reasons for dropout or extended retrieval efforts. The additional information we will use is of the last type. More precisely, we use the number and nature of attempts made to contact initial non-responders, see Section 2.5. Following the ideas in Alho [1990] and Wood et al. [2006], we will use a multinomial model for the reminder process. Alho [1990] focuses on studies with a single time point and a logistic regression model is used to analyze the response probabilities at each contact attempt. Based on these probabilities, a Horvitz-Thompson type estimator for the sample moments is proposed. The same assumptions are made by Wood et al. [2006], but different estimation methods are discussed: conditional likelihood method; EM algorithm and a Bayesian approach using MCMC methods. These approaches will be extended for the longitudinal

case.

Secondly, the models discussed by Alho [1990] and Wood et al. [2006] assume that the effect of covariates and outcomes on the reminder probabilities is the same at all contact attempts. Our model extension allows for covariate or outcome effects to vary between different reminder process categories.

Thirdly, instead of a multivariate linear model, we fit the non-linear mixed model presented in Section 3.5.

This chapter is arranged as follows. In Section 6.2, we present the selection model framework, in which we use the missingness indicator or the number and nature of attempts to account for informative missingness. Using this model, the impact of missingness on the rate of improvement is evaluated for different missingness processes in Section 6.3. Concluding remarks are given in Section 6.4.

6.2 Selection Models and CAST

In the following subsections, we propose a selection model for continuous longitudinal data to adjust for informative missingness when initial non-responders are re-approached several times.

6.2.1 Notation

Let $z_{i,j} \in \{0, 1, ..., K\}$, denote the reminder category which is a realisation of the random variable $Z_{i,j}$. Thus, we distinguish $K + 1$ reminder categories. The vector of reminders for subject i is denoted by $\mathbf{Z_i} = (Z_{i,1}, ..., Z_{i,M})^\top$ and for all subjects by $\mathbf{Z} = (\mathbf{Z}_1^\top, ..., \mathbf{Z}_N^\top)^\top$. Because for CAST the attempt information is only available for $N^* = 553$ patients out of the $N = 559$ patients, we focus on this sub-sample. Further, we have $M = 4$, $K = 5$ and $j \in \{1, 2, 3, 4\}$. The documented levels for $z_{i,j} \in \{0, 1, 2, 3, 4, 5\}$ were illustrated in Section 2.5.

6.2.2 Selection Models

Suppose the complete data $\mathbf{Y}$ follows the parametric model $\mathcal{P}(\theta)$ and $\mathbf{R}$ follows the parametric model $\mathcal{P}(\phi)$. We partition the vector $\mathbf{Y}$ into the observed, $\mathbf{Y}_{obs}$, and unobserved part, $\mathbf{Y}_{mis}$. If the missingness process is non-ignorable or informative we need to base inference for θ on the joint likelihood of $\mathbf{Y}_{obs}$ and the missingness process $\mathbf{R}$. A selection model factorises the joint density of the measurement process and the response mechanism into the marginal measurement process and the response process, conditional on the measurements. Thus, the joint likelihood for (θ, ϕ) based on $\mathbf{Y}_{obs}$ and $\mathbf{R}$ is given by

$$L_{\mathbf{Y}_{obs},\mathbf{R}}(\theta,\phi) = \prod_{i=1}^{N} \int f_{\boldsymbol{Y_i}}(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}|X_i, \theta)\, f_{\boldsymbol{R_i}|\boldsymbol{Y_i}}(\mathbf{r}_i|W_i, \mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}, \phi)\, \mathrm{d}\mathbf{y}_{i,mis}. \quad (6.1)$$

As

$$z_{i,j} \in \{0, 1, 2, 3, K-1\} \Leftrightarrow r_{i,j} = 1 \quad \text{and} \quad z_{i,j} = K \Leftrightarrow r_{i,j} = 0$$

we can extend the selection model by adjusting for non-response through $z_{i,j}$ rather than $r_{i,j}$. This approach is motivated by the hypothesis that subjects who reply after several reminders might be more similar to non-responders, than those who reply at the first attempt. In particular, we can see $r_{i,j}$ as a special case of $z_{i,j}$. The extension of the likelihood in equation (6.1) to adjust for the reminder process is straightforward. Let $\mathbf{Z}$ follow the parametric model $\mathcal{P}(\psi)$. We then simply need to replace $\mathbf{r}_i$ by $\mathbf{z}_i$ and ϕ by ψ in equation (6.1). Fitting these selection models requires a model for the outcome vector Y_i and models for the conditional response process $R_i|Y_i$ and the conditional reminder process $Z_i|Y_i$. We use the non-linear mixed model proposed in Section 3.5 for the marginal outcome process and various plausible regression models for the conditional response and the conditional reminder process. All models that attempt to model MAR or MNAR missingness are based on untestable assumptions. Thus, our modelling framework has to be seen in the context of a sensitivity analysis, where we want to assess the sensitivity of our conclusions to various plausible assumptions about the reasons for missingness.

6.2.3 Outcome Model

The outcome model is motivated by the CAST study and was presented in Section 3.5. For an individual $i \in \{1, ..., N^*\}$, the following mixed model for the outcome vector $\boldsymbol{Y}_i$ will be assumed:

$$\begin{aligned} \mathbf{Y}_i|U_i &\overset{ind.}{\sim} \mathcal{N}_M\left(\mu_i, \sigma^2 I_M\right); \qquad U_i \overset{iid}{\sim} \mathcal{N}(0, D^2); \quad \text{and} \\ \mu_{i,j} &= g(\breve{\boldsymbol{x}}_{i,j}, s_i, t_{i,j}, \theta_i) \quad \text{for } j \in \{1, ..., M\}, \end{aligned}$$

where U_i, a subject-specific effect, has a normal distribution with mean zero and variance D^2, I_M is the M-dimensional identity matrix, $\mu_i = (\mu_{i,1}, ..., \mu_{i,M})^\top$, θ_i the parameter vector of interest and $g(\cdot)$ the model function of interest given in equation (5.5), page 120 with $\breve{\boldsymbol{x}}_{i,j} = \breve{\boldsymbol{x}}_i = age_i - 27$ and $\alpha_0 = \mathbf{0}$. We note that for convenience we assume a compound symmetry covariance structure in this section.

6.2.4 Reminder Process Model

The reminder process instituted in the CAST study was illustrated in Section 2.5. At first glance the geometric and Poisson models seem realistic in capturing the characteristics of the attempt process. However, the non-monotonic frequencies in the reminder categories discourages the use of these models, see Table 2.2. Following ideas of Alho [1990] and Wood et al. [2006] we will therefore focus on a multinomial model for the attempt process.

We develop a model for a single subject, and use the assumed independence between subjects to give the complete model. For the time points $j \in \{2, ..., M\}$ let $p_{j,0}$ be the probability of responding at the first attempt. For $k \in \{1, ..., K-1\}$ let $p_{j,k}$ denote the probability of responding to the k-th attempt, given that the subject has not responded earlier. From the study design, we know that all subjects reply without any chasing at the baseline assessment, i.e. $p_{1,0} = 1$ for all subjects.

The unconditional probabilities $\mu_{j,k}$ of replying to attempt $k \in \{1, ..., K-1\}$ at time

point $j \in \{2, ..., M\}$ are then given by:

$$\mu_{j,0} = p_{j,0}; \quad \mu_{j,1} = p_{j,1}\,(1 - p_{j,0}); \quad \ldots \quad \mu_{j,K-1} = p_{j,K-1} \prod_{k=0}^{K-2} (1 - p_{j,k}).$$

Furthermore, the probability of not replying at time point $j \in \{2, ..., M\}$, i.e. $z_j = K$, is given by $\mu_{j,K} = 1 - \sum_{k=0}^{K-1} \mu_{j,k}$. Corresponding to these probabilities, we redefine the random variable Z_j in terms of an indicator random vector. Let $\boldsymbol{V}_j$ be a $(K+1)$-dimensional random vector, where the ℓ-th component is defined as

$$V_{j,\ell} = \begin{cases} 1, & \text{if attempt } Z_j = \ell - 1; \\ 0, & \text{otherwise} \end{cases}$$

for $\ell \in \{1, ..., K+1\}$. All information about $\mathbf{Z}$ is now captured through the indicator matrix $V = (\mathbf{V}_2, ..., \mathbf{V}_M)^\top$ and the likelihood based on $\boldsymbol{Y}_{obs}$ and V can be derived by replacing $R_{i,j}$ with $V_{i,j}$ and ϕ with ψ in equation (6.1). We can write

$$\mathbf{V}_j | \mathbf{Y} \sim \textit{Multinomial}\left(1, \mu_{j,0}, ..., \mu_{j,K}\right). \tag{6.2}$$

A generalized linear model for either $\mu_{j,k}$ or $p_{j,k}$ can be formulated. The marginal probability $\mu_{j,k}$ determines the chance of replying to the k-th attempt. In contrast, formulating a model for the conditional probability $p_{j,k}$ investigates the effect of covariates on replying to the k-th attempt, given the previous attempts were unsuccessful. As the attempt process evolves over time, it is sensible to explore the latter case.

The generalized linear models we propose for $p_{j,k}$, $j \in \{2, ..., M\}$, $k \in \{0, 1, ..., K-1\}$ are given by

$$\text{logit}(p_{j,k}) = \psi_{0k} + \boldsymbol{\psi}_1^\top \breve{\boldsymbol{w}}_j + \psi_2\, t_j + \psi_3\, y_{j-1} + \psi_4\, y_j, \tag{6.3}$$

where attempt-varying intercepts (ψ_{0k}), covariates ($\breve{\boldsymbol{w}}_j$), and observation times (t_j) are in-

cluded. Further, y_{j-1} is the previous outcome and y_j the current score. Under monotone missingness, this general model allows for different assumptions about the missingness mechanisms; a MNAR model is implied by $\psi_4 \neq 0$, a MAR model by $\psi_4 \equiv 0$ and, conditioned on covariates, a MCAR model is implied by $\psi_3 \equiv 0 \equiv \psi_4$.

In line with earlier work, this model assumes that the covariate and outcome effects on $p_{j,k}$ are common to all attempts, see Alho [1990] and Wood et al. [2006]. Only the intercept varies across the different reminder categories. This assumption can be too restrictive in practice and can lead to a misspecified model, see Wood et al. [2006]. Our model can be expanded to allow the covariate and outcome effect to vary between the different reminder categories. However, the inclusion of attempt-varying covariate or outcome effects leads to a large number of parameters and high computational complexity. Parameter estimation and inference can be hindered if the sample size of some reminder categories is too small. We underline the feasibility of fitting a model with attempt-varying covariate or outcome effects by considering a MNAR model where we include an attempt-varying gender effect, see Section 6.3.3.

The model in (6.3) assumes that the model for $\mathbf{V}_j|\mathbf{Y}$ does not depend on later observations y_ℓ, where $\ell > j$. We regard this as a sensible assumption for most settings.

With bounded scores and highly correlated scores at adjacent occasions, we will usually observe a high correlation between the estimates for ψ_3 and ψ_4. This problem will always persist when including previous and current scores linearly. For example, with two perfectly positively correlated scores at adjacent time points, a distinction between the proposed MAR and MNAR models is not possible. Therefore, we will also consider a different parametrization proposed by Diggle and Kenward [1994]:

$$\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1^\top \breve{\mathbf{w}}_j + \psi_2\, t_j + \psi_3^* \left[y_{j-1} + y_j\right] + \psi_4^* \left[y_j - y_{j-1}\right] \tag{6.4}$$

where the estimates for ψ_3^* and ψ_4^* are usually less correlated. Again, this model allows for different assumptions regarding the missingness processes: a MCAR model is implied by

$\psi_3^* \equiv 0 \equiv \psi_4^*$, a MAR model by $\psi_3^* \equiv -\psi_4^*$ and a MNAR model by $\psi_3^* \neq -\psi_4^*$.

In order to account for the within-patient correlation across the attempts at the different observation times, we need to extend this model, see Section 6.2.6.

6.2.5 Missingness Process Model

We now consider modelling the missingness process, conditional on the outcome of interest. In the spirit of regression modelling, we propose the following logistic linear model for all $i \in \{1, ..., N^*\}$, $j \in \{2, ..., M\}$:

$$R_{i,j} = 1|\mathbf{Y}_i \sim Bernoulli(\rho_{i,j}), \text{ with} \tag{6.5}$$

$$\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1^\top \tilde{\boldsymbol{w}}_{i,j} + \phi_2\, t_j + \phi_3\, y_{i,j-1} + \phi_4\, y_{i,j} \tag{6.6}$$

where $\tilde{\boldsymbol{w}}_{i,j}$ denotes covariates we wish to include in the missingness process model. As above, t_j are the observation times, $y_{i,j-1}$ the previous outcome and $y_{i,j}$ the current score. We do not specify a model for $R_{i,1}$ as all scores are observed at baseline. This model corresponds to a MNAR model if $\phi_4 \neq 0$ and to a MCAR model (conditioned on covariates) if $\phi_3 \equiv 0 \equiv \phi_4$. As we focus on monotone missingness, a MAR missingness process is obtained by setting $\phi_4 \equiv 0$, but this holds for monotone missingness only. In the case of non-monotone missingness, it is a less than trivial manner to construct sensible MAR models [Molenberghs et al., 2008]. We will also consider a model similar to that given in equation (6.4), see Section 6.3.2.

6.2.6 Full Model under Monotone Missingness

For monotone missingness, we can now construct the joint likelihood of $\mathbf{Y}_{obs}, \mathbf{R}$ and $\mathbf{Y}_{obs}, V$ respectively. The derivations will be shown for a selection model that uses the reminder process (via $\mathbf{V}_{i,j}$) to account for missingness. The likelihood using the missingness indicator process $R_{i,j}$ can be derived by replacing $\mathbf{V}_{i,j}$ by $R_{i,j}$ and ψ by ϕ in all following equations.

The observed data likelihood contribution of a certain subject is given by:

$$f_{(\mathbf{Y}_{i,obs},V_i)}\left(\mathbf{y}_{i,obs}, v_i|X_i, W_i, \theta, \psi\right) = \int f_{\mathbf{Y}_i}\left(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}|X_i, \theta\right) f_{V_i|\mathbf{Y}_i}\left(v_i|W_i, \mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}, \psi\right) \mathrm{d}\mathbf{y}_{i,mis}.$$

Here, we illustrate the derivations for $M = 4$ and by assuming that dropout for the subject of interest occurs after the second measurement time, i.e. $\mathbf{y}_{i,obs} = (y_{i,1}, y_{i,2})^\top$. The derivations for other cases follow straightforwardly. We obtain

$$\begin{aligned} f\left(\mathbf{y}_{i,obs}, v_i|X_i, W_i, \theta, \psi\right) = & \int\int f\left(y_{i,4}|y_{i,3}, y_{i,2}, y_{i,1}, X_i, \theta\right) f\left(y_{i,3}|y_{i,2}, y_{i,1}, X_i, \theta\right) f\left(y_{i,2}|y_{i,1}, X_i, \theta\right) \\ & \times f\left(y_{i,1}|X_i, \theta\right) f\left(\mathbf{v}_{i,4}|\mathbf{v}_{i,3}, \mathbf{v}_{i,2}, \mathbf{v}_{i,1}, W_i, \mathbf{y}_i, \psi\right) f\left(\mathbf{v}_{i,3}|\mathbf{v}_{i,2}, \mathbf{v}_{i,1}, W_i, \mathbf{y}_i, \psi\right) \\ & \times f\left(\mathbf{v}_{i,2}|\mathbf{v}_{i,1}, W_i, \mathbf{y}_i, \psi\right) \mathrm{d}y_{i,4}\, \mathrm{d}y_{i,3}. \end{aligned}$$

For the sake of clarity we suppress subscripts, referring to the relevant distributions, in the notation of the densities. In the case of monotone missingness we observe

$$\mathbf{v}_{i,j} = (0, 0, 0, 0, 0, 1) \Longrightarrow \mathbf{v}_{i,j+1} = (0, 0, 0, 0, 0, 1) \tag{6.7}$$

for $j \in \{2, 3\}$ and $j + 1 \in \{3, 4\}$. Therefore,

$$f\left\{\mathbf{v}_{i,4} = (0, 0, 0, 0, 0, 1)|\mathbf{v}_{i,3} = (0, 0, 0, 0, 0, 1), \mathbf{v}_{i,2}, \mathbf{v}_{i,1}, W_i, \mathbf{y}_i, \psi\right\} = 1.$$

Rearranging the observed likelihood yields

$$\begin{aligned} f\left(\mathbf{y}_{i,obs}, v_i|X_i, W_i, \theta, \psi\right) = & f\left(y_{i,2}|y_{i,1}, X_i, \theta\right) f\left(y_{i,1}|X_i, \theta\right) f\left(\mathbf{v}_{i,2}|\mathbf{v}_{i,1}, W_i, y_{i,2}, y_{i,1}, \psi\right) \\ & \times \int f\left(y_{i,3}|y_{i,2}, y_{i,1}, X_i, \theta\right) f\left(\mathbf{v}_{i,3}|\mathbf{v}_{i,2}, \mathbf{v}_{i,1}, W_i, y_{i,3}, y_{i,2}, \psi\right) \\ & \times \underbrace{\int f\left(y_{i,4}|y_{i,3}, y_{i,2}, y_{i,1}, X_i, \theta\right) \mathrm{d}y_{i,4}}_{=1}\, \mathrm{d}y_{i,3}, \end{aligned} \tag{6.8}$$

i.e. the integrals reduce to one-dimensional integrals for $i \in \{1, ..., N^*\}$. We note that the

likelihood terminates after the time of the first missing observation due to the relation shown in equation (6.7). In particular, implication (6.7) does not hold for non-monotone missingness; and we are confronted with multidimensional integrals.

The likelihood contribution shown in equation (6.8) stresses that we ought to consider the dependence structure across the reminders at the different time points for a given subject. When modelling the reminder process such a dependence structure can be included in various ways, e.g. by formulating a random-effect model for $p_{j,k}$. However, this would require the computation of further integrals which complicates the evaluation of the likelihood. Alternatively, we can account for the dependence by formulating a model for $\mathbf{V}_{i,j}$ conditional on $\mathbf{V}_{i,j-1}, ..., \mathbf{V}_{i,1}$. For simplification, we decide to model $\mathbf{V}_{i,j}$ conditional on $\mathbf{V}_{i,j-1}$; that is, we extend the models given in equation (6.3) and equation (6.4) by adding the term $\psi_{5,k}\, \mathbb{1}\left(z_{i,j-1} = k\right)$, which indicates which attempt category $z_{i,j-1} \in \{0, ..., K-1\}$ was observed at the previous point in time. We note that modelling the missingness process R_i under the assumption of monotone missingness does not require the incorporation of a dependence structure.

The integral in the likelihood contribution of subject i, shown in equation (6.8), can be solved through an adaptive Romberg-type integration technique. This approach produces a quick, rough estimate of the integration result and then refines the estimate until achieving the prescribed accuracy [SAS/STAT, 1999]. The maximum likelihood estimates for θ and ψ (or ϕ) can then be calculated through the Newton-Raphson ridge optimization method which is implemented in the subroutine `call nlpnrr` in `proc IML` [SAS/STAT, 1999].

6.3 Results for CAST under Monotone Missingness

In this section we explore the impact of missingness on the estimated rate of recovery through a sensitivity analysis. Further, we aim to investigate covariate and outcome effects on the reminder process and which reminders are most effective. In this context, we focus on monotone missingness and adjust for missingness by modelling the reminder process,

Monotone and Ignorable Missingness				
Variable	Parameter	Est.	SE	p-val.
Intercept	β_0	41.01	0.78	-
Maximum	β_1	80.05	0.85	-
Rate of Tubigrip	β_{21}	0.27	0.03	-
Contrast: BKC	β_{22}	-0.12	0.04	0.006
Contrast: Aircast	β_{23}	-0.05	0.04	0.174
Contrast: Bledsoe	β_{24}	-0.004	0.04	0.915
Age-effect on Max.	α_1	-0.30	0.07	< 0.001
Age-effect on Rate	α_2	-0.006	0.001	< 0.001
Within - Variance	σ^2	186.1	7.33	-
Between - Variance	D^2	148.2	12.7	-

Table 6.1: Overview of the parameter estimates and standard errors for the outcome model (3.20) with $\check{\boldsymbol{x}}_{i,j} = a_i$ based on the assumption of an ignorable missingness process. The p-values are reported only for the components of θ that might be zero.

see Section 6.2.4. We contrast this model with the traditional selection model, where we adjust for missingness by modelling the missingness process, see Section 6.2.5.

For CAST we focus on $\check{\boldsymbol{x}}_{i,j} = \check{\boldsymbol{w}}_{i,j} = \tilde{\boldsymbol{w}}_{i,j} = age_i - 27 = a_i$, i.e. age centered around the median age. Different assumptions for the missingness mechanisms will be made and the results will be compared with those obtained based on the assumption of ignorability, see Table 6.1.

6.3.1 Results using the Reminder Process Model

Using the notation in Section 6.2.4, we investigated the following logistic regression models for the conditional reminder process probabilties $p_{j,k}$:

MCAR$_p$: $\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1 a_i + \psi_2 t_j + \psi_{5,k} \mathbb{1}\left(v_{i,j-1} = k\right)$;

MAR$_p$: $\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1 a_i + \psi_2 t_j + \psi_3 y_{j-1} + \psi_{5,k} \mathbb{1}\left(v_{i,j-1} = k\right)$;

MNAR$_p$-1: $\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1 a_i + \psi_2 t_j + \psi_3 y_{j-1} + \psi_4 y_j + \psi_{5,k} \mathbb{1}\left(v_{i,j-1} = k\right)$;

MNAR$_p$-2: $\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1 a_i + \psi_2 t_j + \psi_4 y_j + \psi_{5,k} \mathbb{1}\left(v_{i,j-1} = k\right)$; and

MNAR$_p$-3: $\text{logit}(p_{j,k}) = \psi_{0k} + \psi_1 a_i + \psi_2 t_j + \psi_3^* \left[y_{j-1} + y_j\right] + \psi_4^* \left[y_j - y_{j-1}\right] + \psi_{5,k} \mathbb{1}\left(v_{i,j-1} = k\right)$.

where $k \in \{0, 1, 2, 3, 4\}$ and $j \in \{2, 3, 4\}$, i.e. $t_j \in \{4, 12, 39\}$. Note that here MCAR denotes a mechanism where missingness is allowed to depend on covariates but not on the outcome of interest. Initial analysis shows that the inclusion of $\psi_{5,k}\, \mathbb{1}\left(v_{i,j-1} = k\right)$ is not necessary, as these parameters are equal and so can all be absorbed into the intercept. Furthermore, the age-effect on the intercept of the non-linear mixed model, i.e. α_0, was shown to be not significant. The results for the simpler model with α_0, $\psi_{5,k}$ omitted and the case of monotone missingness are shown in Table 6.2.

The estimates for the outcome model parameters, i.e. θ, are practically identical under all reminder processes investigated, including the estimates under the assumption of an ignorable missingness process. The treatment effect of Tubigrip, the treatment differences and the associated p-values are essentially equal for all models. In line with the results shown in Section 3.5, all approaches detect that BKC is significantly better than Tubigrip and that Bledsoe is not measurably different from Tubigrip. For the treatment difference of Aircast and Tubigrip we observe p-value= 0.2, while in Section 3.5 a marginal difference was detected (p-value= 0.07). This is likely due to the exclusion of the covariate gender in the model fitted here.

Older participants achieve a lower final recovery level than younger participants, as $\hat{\alpha}_1 < 0$. Furthermore, all models confirm that older participants recover less fast than younger patients.

For the reminder vector $\mathbf{Z}$, given the outcome vector $\mathbf{Y}$ and covariates, the results vary substantially under the different assumptions for the missingness processes. Especially, the MNAR$_p$-3 process leads to different conclusions from the other models.

Under MCAR$_p$ we observe that sending a second questionnaire is least effective for the return of questionnaires. The other reminder categories appear to be equally effective. We observe a positive age-effect, i.e. the probability of replying at a certain attempt increases with age. Furthermore, the probability of replying at a certain attempt decreases as time passes. The age- and time-effects persist under MAR$_p$, MNAR$_p$-1 and MNAR$_p$-2; the effect sizes are practically identical, and likelihoods similar.

		$MCAR_p$			MAR_p			$MNAR_p$-1			$MNAR_p$-2			$MNAR_p$-3		
Variable		Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.
Outcome Process																
Intercept	β_0	41.01	0.78	-	41.01	0.78	-	41.01	0.78	-	41.01	0.78	-	41.01	0.78	-
Maximum	β_1	80.05	0.85	-	80.05	0.85	-	79.98	0.85	-	79.95	0.85	-	80.36	0.85	-
Rate for Tubigrip	β_{21}	0.28	0.03	-	0.28	0.03	-	0.27	0.03	-	0.27	0.03	-	0.28	0.03	-
Contrast: BKC	β_{22}	-0.12	0.04	0.006	-0.12	0.04	0.006	-0.12	0.04	0.005	-0.12	0.04	0.005	-0.12	0.04	0.007
Contrast: Aircast	β_{23}	-0.05	0.04	0.174	-0.05	0.04	0.174	-0.05	0.04	0.171	-0.05	0.04	0.171	-0.05	0.04	0.180
Contrast: Bledsoe	β_{24}	-0.004	0.04	0.915	-0.004	0.04	0.915	-0.005	0.04	0.899	-0.005	0.04	0.893	-0.002	0.04	0.960
Age-effect on Max.	α_1	-0.30	0.07	<0.001	-0.30	0.07	<0.001	-0.30	0.07	<0.001	-0.30	0.07	<0.001	-0.31	0.07	<0.001
Age-effect on Rate	α_2	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001
Within - Variance	σ^2	186.41	7.33	-	186.41	7.33	-	186.40	7.33	-	186.39	7.33	-	186.39	7.33	-
Between - Variance	D^2	148.17	12.67	-	148.17	12.68	-	148.22	12.69	-	148.26	12.68	-	148.68	12.75	-
Conditional Reminder Process																
Intercept: 1st Quest.	ψ_{00}	-0.14	0.07	0.045	0.02	0.11	0.879	0.14	0.15	0.348	0.12	0.14	0.391	0.39	0.16	0.014
Intercept: 1st Tel.	ψ_{01}	0.09	0.08	0.312	0.24	0.12	0.046	0.37	0.16	0.020	0.35	0.15	0.023	1.19	0.18	<0.001
Intercept: 2nd Quest.	ψ_{02}	-0.61	0.12	<0.001	-0.45	0.15	0.002	-0.33	0.18	0.062	-0.35	0.17	0.047	0.46	0.19	0.018
Intercept: 2nd Tel.	ψ_{03}	0.07	0.13	0.607	0.23	0.16	0.147	0.35	0.19	0.061	0.34	0.18	0.068	1.23	0.21	<0.001
Intercept: Core Resp.	ψ_{04}	13.02	20.10	0.517	14.87	29.13	0.610	13.32	26.59	0.617	22.43	74.45	0.763	15.49	54.04	0.774
Age-effect	ψ_1	0.008	0.004	0.022	0.007	0.004	0.042	0.006	0.004	0.071	0.007	0.004	0.068	0.002	0.004	0.666
Time-effect	ψ_2	-0.009	0.003	<0.001	-0.006	0.003	0.047	-0.006	0.003	0.051	-0.007	0.003	0.012	0.011	0.003	<0.001
Previous Score	ψ_3	-	-	-	-0.003	0.002	0.073	-0.002	0.002	0.417	-	-	-	-	-	-
Current Score	ψ_4	-	-	-	-	-	-	-0.003	0.002	0.213	-0.004	0.002	0.041	-	-	-
Average Score	ψ_3^*													-0.013	0.001	<0.001
Improvement	ψ_4^*													0.023	0.002	<0.001
Deviance	-2ℓ	19357.72			19354.48			19352.92			19353.58			18933.3		

Table 6.2: Parameter estimates, standard errors and deviances for the outcome model (3.20) with $\tilde{\boldsymbol{x}}_{i,j} = a_i$, the reminder process and the different missing data mechanisms, see Section 6.3.1. The p-values are reported only for the components of (θ, ψ) that might be zero.

The MAR_p results suggest that the reminder process, and therefore the missingness process, depends on the outcome of interest (p-value= 0.07). The probability of returning a questionnaire decreases with the score at the prior occasion: patients with high scores at the previous observation times tend to return the questionnaires only after several attempts or not at all. This result is in line with quantitative findings presented in Nakash et al. [2008], which suggest that patients who considered themselves to have made fully recovery, did not return their subsequent questionnaire. Furthermore, we observe that phone calls are most effective for the retrieval of questionnaires. In contrast, sending a second questionnaires is least effective. The same conclusions are carried forward to the models under the MNAR_p-1 and the MNAR_p-2 assumption. However, the effect sizes and the associated p-values vary across the models.

For the MNAR_p-1 model, no significant effect of current or previous score on the response probabilities is found. As mentioned in Section 6.2.6, this is likely due to the high correlation of scores at adjacent occasions. The empirical correlations based on the observed data are given by: $Corr_{emp}(\boldsymbol{y}_{\bullet,0}; \boldsymbol{y}_{\bullet,4}) = 0.34$, $Corr_{emp}(\boldsymbol{y}_{\bullet,4}; \boldsymbol{y}_{\bullet,12}) = 0.65$ and $Corr_{emp}(\boldsymbol{y}_{\bullet,12}; \boldsymbol{y}_{\bullet,39}) = 0.68$, where $\boldsymbol{y}_{\bullet,j}$ is the vector of all observations made at time $t_j \in \{0, 4, 12, 39\}$.

As scores reach the final recovery level, $Y_{i,j}$ and $Y_{i,j-1}$ are highly correlated, so that after allowing for the effect of one score, there is little additional effect of the other score. At the final recovery level, the scores are effectively interchangeable.

Removing the previous outcome, that is fitting the model under a MNAR_p-2 process, reveals a negative effect of the current score on the probability of replying. The effect size is comparable with the effect of the previous score, $\hat{\psi}_3$, in the MAR_p model.

The alternative parametrization, i.e. MNAR_p-3, suggests that the reminder process depends on the mean score and the improvement of the score between two adjacent time points. The probability of replying decreases with the mean but increases with the improvement. In contrast to the previous models, the age effect is shown to be not significant and the time effect is positive. Furthermore, phone calls are again most effective for the retrieval

of questionnaires. The deviance suggests that the MNAR$_p$-3 leads to the best fit.

6.3.2 Results using the Missingness Process Model

When adjusting for monotone missingness through the missingness indicator $R_{i,j}$, we explore the impact of dropout on the rate of improvement under the following assumptions:

MCAR$_r$: $\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1 a_i + \phi_2 t_j$;

MAR$_r$: $\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1 a_i + \phi_2 t_j + \phi_3 y_{i,j-1}$;

MNAR$_r$-1: $\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1 a_i + \phi_2 t_j + \phi_4 y_{i,j}$;

MNAR$_r$-2: $\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1 a_i + \phi_2 t_j + \phi_3 y_{i,j-1} + \phi_4 y_{i,j}.$; and

MNAR$_r$-3: $\text{logit}(\rho_{i,j}) = \phi_0 + \phi_1 a_i + \phi_2 t_j + \phi_3^* (y_{i,j-1} + y_{i,j}) + \phi_4^* (y_{i,j} - y_{i,j-1})$.

The estimated outcome parameters (Table 6.3) are consistent with those obtained by modelling the reminder process. The intercepts for the missingness processes vary substantially across the assumed models. This is not surprising, as we include more covariates to explain the missingness process. The probability of replying increases with age for all investigated models, but the time effect is not significant.

The MAR$_r$ model suggests that the probability of replying at a certain time does not depend on the score at that time (p-value= 0.12). Note that this result is contrary to the conclusions based on modelling the reminder process. Including both previous and current scores using MNAR$_r$-1 is not informative. The MNAR$_r$-2 model finds a marginal effect of the current score on the missingness probabilities: as the score increases, the probability of replying increases. Under the assumption of MNAR$_r$-3, we obtain that missingness depends positively on the average score (p-value= 0.06) but not on the improvement of the scores at adjacent observations times. Thus the covariate effects differ from the findings in Section 6.3.1. However, once an effect is included in only one of these two models (e.g. the time effect), the remaining parameter estimates become difficult to compare. Therefore,

		$MCAR_r$			MAR_r			$MNAR_r$-1			$MNAR_r$-2			$MNAR_r$-3		
Variable		Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.	Est.	SE	p-val.
Outcome Process																
Intercept	β_0	41.01	0.78	-	41.01	0.78	-	41.00	0.78	-	41.00	0.78	-	41.00	0.78	-
Maximum	β_1	80.05	0.85	-	80.05	0.85	-	79.65	0.93	-	79.63	0.88	-	79.65	0.93	-
Rate for Tubigrip	β_{21}	0.28	0.03	-	0.28	0.03	-	0.27	0.03	-	0.27	0.03	-	0.27	0.03	-
Contrast: BKC	β_{22}	-0.12	0.04	0.006	-0.12	0.04	0.006	-0.12	0.04	0.005	-0.12	0.04	0.004	-0.12	0.04	0.005
Contrast: Aircast	β_{23}	-0.05	0.04	0.174	-0.05	0.04	0.174	-0.06	0.04	0.161	-0.06	0.04	0.160	-0.06	0.04	0.161
Contrast: Bledsoe	β_{24}	-0.004	0.04	0.915	-0.004	0.04	0.915	-0.008	0.04	0.829	-0.008	0.04	0.824	-0.008	0.04	0.829
Age-effect on Max.	α_1	-0.30	0.07	<0.001	-0.30	0.07	<0.001	-0.29	0.07	<0.001	-0.29	0.07	<0.001	-0.29	0.07	<0.001
Age-effect on Rate	α_2	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001	-0.006	0.001	<0.001
Within - Variance	σ^2	186.41	7.34	-	186.41	7.34	-	186.68	7.38	-	186.71	7.38	-	186.68	7.38	-
Between - Variance	D^2	148.17	12.70	-	148.17	12.73	-	148.66	12.72	-	148.70	12.78	-	148.66	12.76	-
Conditional Missingness Process																
Intercept	ϕ_0	1.79	0.12	<0.001	1.52	0.21	<0.001	1.08	0.43	0.013	1.06	0.38	0.006	1.08	0.44	0.014
Age-effect	ϕ_1	0.03	0.01	<0.001	0.03	0.01	<0.001	0.03	0.01	<0.001	0.03	0.01	<0.001	0.03	0.01	<0.001
Time-effect	ϕ_2	0.007	0.006	0.228	0.001	0.007	0.870	0.001	0.007	0.871	0.001	0.006	0.844	0.001	0.01	0.871
Previous Score	ϕ_3	-	-	-	0.007	0.004	0.119	0.001	0.007	0.942	-	-	-	-	-	-
Current Score	ϕ_4	-	-	-	-	-	-	0.011	0.01	0.268	0.012	0.01	0.051	-	-	
Average Score	ϕ_3^*	-	-	-	-	-	-	-	-	-	-	-	-	0.006	0.003	0.061
Improvement	ϕ_4^*	-	-	-	-	-	-	-	-	-	-	-	-	0.005	0.01	0.513
Deviance	-2ℓ	16422.36			16419.90			16418.69			16418.70			16418.69		

Table 6.3: Parameter estimates, standard errors and deviances for the outcome model (3.20) with $\tilde{\boldsymbol{x}}_{i,j} = a_i$, the response process and the different missing data mechanisms defined in Section 6.3.2. The p-values are reported only for the components of (θ, ϕ) that might be zero.

MNAR$_p$-3			
current score	previous score	μ_{j5}	
		$t_j = 12$	$t_j = 39$
60	50	0.07	0.04
	60	0.13	0.08
90	60	0.08	0.04
	90	0.35	0.26

MNAR$_r$-3			
current score	previous score	$1 - \rho_j$	
		young	old
60	50	0.17	0.11
	60	0.17	0.11
90	60	0.14	0.09
	90	0.12	0.08

Table 6.4: Overview of the probabilities of not replying for different age groups and previous / current scores based on the point estimates obtained from fitting the MNAR$_p$-3 reminder and the MNAR$_r$-3 missingness models. The age groups were classified according to the first (21 years) and third (37 years) quantile.

we should refrain from comparing the separate parameter estimates for the reminder and the missingness process directly. For further discussion regarding the disagreement of the results for the two types of models, we refer to Section 6.3.4.

For illustration, we show the probabilities of not replying for different age groups and low/high scores under the MNAR$_p$-3 and the MNAR$_r$-3 model, see Table 6.4.

6.3.3 Model Extension for the Reminder Process Model

The reminder process models investigated so far assume that covariate and outcome effects are the same across all reminder categories. In this section, we will relax this assumption and allow our reminder process model to include attempt-varying gender effects, i.e.

$$\text{MNAR}_p - A : \text{logit}(p_{i,j,k}) \quad = \quad \tau_{0k} + \tau_1 a_i + \tau_2 t_j + \tau_3 y_{i,j} + \tau_{4k} \mathbb{1}_{female}(sex_i). \quad (6.9)$$

The results for the outcome and reminder process (Table 7.1) are consistent with those seen in Section 6.2.4. Female patients seem to be more likely to reply at the first two attempts than male patients. However, this effect is not significant. It is not clear whether a gender effect does not exist or whether the corresponding sample sizes were simply too small to detect a significant effect. We modify model (6.9) such that $\tau_{40} = \tau_{41}$ and $\tau_{42} = \tau_{43} = \tau_{44}$. We refer to this model as MNAR$_p$-B and the results are given in Table 7.1. The latter analysis confirms that female patients are more likely to respond at the first two attempts than male patients. This effect remains not detectable or not existent for the other reminder

categories.

We conclude that, although in principle it is possible to fit such a model, inference will depend strongly on the sample size per reminder category.

6.3.4 Comparison of the Investigated Selection Models

In addition to the previous models, we have also investigated the following model for the missingness process:

$$\text{MNAR}_r : \text{logit}(\rho_{i,j}) = \lambda_0 + \lambda_1 a_i + \lambda_2 t_j + \lambda_3 y_{i,j} + \lambda_4 \mathbb{1}_{female}(sex_i),$$

where we have included gender as a covariate. This analysis results in a non-significant gender-effect, but suggests a significant effect of age on the likelihood of being a responder. In contrast, the reminder process probabilities revealed no significant effect of age.

We already noted that covariate effects on the missingness and reminder probabilities are difficult to compare when covariates (here: time, age, gender) are included only in one of the two models. Nevertheless, we make an attempt to determine what the differences between the two types of models are that led to these disagreements. Theoretically speaking, we believe that it is possible that covariate effects are included in the reminder process but not in the missingness process (here: time and gender) and the other way around (here: age).

For simplicity, we distinguish only three reminder categories: prompt reply ($z = 0$), one reminder ($z = 1$) and non-responder ($z = 2$). We focus on a model where the effect of a single, one-dimensional and continuous covariate $w_{i,j}$ is of interest. The concept can be extended to more reminders and categorical covariates.

Assume that $w_{i,j}$ has a common and significant effect on the reminder probabilities $p_{i,j,k}$. In particular, we assume that all subjects who reply promptly have a low value for $w_{i,j}$ and all patients who reply after one reminder have a large value for $w_{i,j}$. All subjects with average values for $w_{i,j}$ are assumed to be non-responders. Pooling all responders into

		MNAR$_p$-A			MNAR$_p$-B				MNAR$_r$		
Variable	Parameter	Est.	SE	p-val.	Est.	SE	p-val.	Parameter	Est.	SE	p-val.
Outcome Process											
Intercept	β_0	41.01	0.78	-	41.01	0.78	-	β_0	41.00	0.78	-
Maximum	β_1	79.96	0.85	-	79.96	0.85	-	β_1	79.59	0.88	-
Rate for Tubigrip	β_{21}	0.27	0.03	-	0.27	0.03	-	β_{21}	0.27	0.03	-
Contrast: BKC	β_{22}	-0.12	0.04	0.005	-0.12	0.04	0.005	β_{22}	-0.12	0.04	0.004
Contrast: Aircast	β_{23}	-0.05	0.04	0.171	-0.05	0.04	0.171	β_{23}	-0.06	0.04	0.160
Contrast: Bledsoe	β_{24}	-0.005	0.04	0.895	-0.005	0.04	0.895	β_{24}	-0.009	0.04	0.816
Age-effect on Max.	α_1	-0.30	0.07	< 0.001	-0.30	0.07	< 0.001	α_1	-0.29	0.07	< 0.001
Age-effect on Rate	α_2	-0.01	0.001	< 0.001	-0.01	0.001	< 0.001	α_2	-0.01	0.001	< 0.001
Within - Variance	σ^2	186.39	7.33	-	186.39	7.33	-	σ^2	186.79	7.37	-
Between - Variance	D^2	148.24	12.71	-	148.24	12.67	-	D^2	148.76	12.73	-
Conditional Reminder and Missingness Processes											
Intercept: 1st Quest.	τ_{00}	0.02	0.16	0.896	0.041	0.16	0.941	Intercept: λ_0	-0.91	0.40	0.023
Intercept: 1st Tel.	τ_{01}	0.23	0.17	0.172	0.24	0.17	0.152	Age: λ_1	-0.03	0.01	<0.001
Intercept: 2nd Quest.	τ_{02}	-0.43	0.20	0.031	-0.39	0.20	0.045	Time: λ_2	-0.001	0.007	0.924
Intercept: 2nd Quest.	τ_{03}	0.33	0.21	0.120	0.29	0.21	0.161	y_j: λ_3	-0.01	0.01	0.035
Intercept: Core Resp.	τ_{04}	10.45	14.49	0.471	12.61	18.87	0.504	Gender: λ_4	-0.19	0.17	0.255
Age-effect	τ_1	0.005	0.004	0.155	0.005	0.004	0.154				
Time-effect	τ_2	-0.007	0.003	0.008	-0.007	0.003	0.009				
Current Score	τ_3	-0.004	0.002	0.081	-0.003	0.002	0.094				
1st Quest. $*$ Female	τ_{40}	0.17	0.11	0.113	0.19	0.09	0.32				
1st Tel. $*$ Female	τ_{41}	0.22	0.14	0.126	-	-	-				
2st Quest. $*$ Female	τ_{42}	0.13	0.22	0.547	0.04	0.16	0.793				
2st Tel. $*$ Female	τ_{43}	-0.07	0.25	0.777	-	-	-				
Core Resp. $*$ Female	τ_{44}	2.14	8.28	0.796	-	-	-				
Deviance	$-2\,\ell$	19348.44			19348.88			$-2\,\ell$	16417.38		

Table 6.5: Overview of the parameter estimates, standard errors and p-values for the models introduced in Section 6.3.3 and Section 6.3.4.

one category, as is done when modelling the missingness process, could then hinder the estimation of a significant covariate effect.

In contrast, it is possible that subjects within the reminder categories $z \in \{0, 1\}$ are very homogeneous with regard to $w_{i,j}$, so that no significant effect of $w_{i,j}$ is detectable. Or alternatively, the sample sizes of the reminder categories are too small to provide evidence for an effect. In both cases, pooling all responders into one category can lead to a significant effect; in particular, when the attempt-specific intercepts for these reminder categories are different. In the case of very homogeneous reminder categories $z \in \{0, 1\}$ with respect to $w_{i,j}$, the covariate $w_{i,j}$ might act as surrogate for the different reminder categories.

In addition, including attempt-varying coefficients in the reminder process can lead to further variations of the conclusions. Take the example where subjects who reply promptly have a low value for $w_{i,j}$ and all patients who reply after one reminder have a large value for $w_{i,j}$. Now assume that the covariate effect for the first reminder category ($z = 0$) is significantly positive and the effect for the second category is significantly negative. Dependent on the attempt-specific intercept the conclusions can vary when modelling only the missingness probabilities.

Overall, we note that using the richer information of the reminder probabilities enables a more accurate choice of covariates which induce missingness. However, this statement is conditional upon having large enough sample sizes to detect significant effects for all reminder categories.

6.4 Summary

We have proposed a selection model for continuous longitudinal data to adjust for non-ignorable or informative missingness when initial non-responders are re-approached several times. In addition, we have contrasted this model with the traditional selection model framework, where we adjust for missing data by modelling the missingness process.

The models presented combine the non-linear mixed model presented in Section

3.5 for the underlying outcome model with logistic regression models for the missingness and the reminder processes.

For the reminder process, we model the probability of replying at a certain attempt, given not having replied earlier, through a multinomial model. We use logistic regression to explore the dependence of the response probabilities on covariates as well as the outcome of interest. We investigate models which assume the same covariate and outcome effect across all reminder categories. However, we also discuss the limitation of this assumption and expand our model to allow for attempt-varying covariate or outcome effects.

We perform a sensitivity analysis, where we investigate the impact of missingness on the rate of improvement for different model families and under different assumptions about the missingness process. We focus on the case of monotone missingness patterns. The simplicity of the model fitting described relies heavily on this assumption. As soon as we relax this assumption, we are confronted with multi-dimensional integrals. Attempts to run the extended `SAS` code which accounts for non-monotone missingness failed due to slowness. The calculation of the likelihood in every iteration step requires the computation of several hundred integrals and every iteration step ran for several hours. Therefore, we moved to the Bayesian paradigm to fit models based on non-monotone missingness using `WinBUGS`. However, the complicated outcome model and correlated parameters make the model fitting very difficult, see Section 9.2.

For CAST, the conclusions that recovery is slower, and less satisfactory with age, and more rapid with BKC than Tubigrip do not alter materially across all models investigated.

Depending on whether the reminder process or the missingness process is explored, we find that the probabilities of replying decrease or increase with the observed outcome at the current or previous occasions. Due to the high correlation between the scores at adjacent time points, problems arise when including current and previous scores jointly. The MNAR_p-3 reminder model suggests that the improvement and the average score, rather than the actual scores, affect the missingness process and this model leads to the best fit.

When modelling the missingness process, only a marginal effect of the average score was found. Overall, these results suggest that missingness depends on the outcome of interest.

In general, we observe different covariate effects for the reminder and missingness processes. However, also within the two model families conclusions depend on the assumed missingness mechanism. For example, for all reminder process models, except MNAR$_p$-3, we observe a (sometimes marginal) positive effect of age on the response probabilities. The MNAR$_p$-3 suggests that age does not have a significant effect. We argue that a direct comparison of the results for the two model families is difficult, as different covariates are included in the corresponding models. We claim it is possible that covariate effects are included in the reminder process but not in the missingness process and the other way around, see Section 6.3.4.

Our investigations suggest that using the richer information of the reminder process enables a more accurate choice of covariates, which induce missingness, than modelling the missingness process. This holds under the condition that the sample sizes of all reminder categories are large enough to detect significant effects. A further advantage of modelling the reminder process versus the traditional selection model is the ability to incorporate the dependence across the reminders at the different observation times for a given patient.

Regarding the reminder process, we observe that phone calls are most effective, while sending a second questionnaire without further telephone chasing appears to be least effective in retrieving questionnaires. Such insight is important to understand the effectiveness of reminder systems and to improve the design of future studies.

Overall, the outcome parameters of interest are estimated very robustly across all models investigated. However, we believe that care has to be taken with such conclusions. The identification of all models presented is driven by untestable parametric assumptions on the marginal outcome model, the conditional missingness and the conditional reminder process, respectively. It is not clear to which extent these conclusions would differ under other assumptions; e.g. other covariance structure for the marginal outcome process, use of the probit link-function for the reminder and missingness probabilities, incorporation of

explanatory variables such as occupation in the missingness and reminder process models. Furthermore, we include the previous and current score linearly in the reminder and missingness process. Other functional relationships might lead to differing conclusions. In particular, we *estimate* the model parameters ψ_i for $i \in \{3, 4\}$ relating the selection model to missing responses. This estimation can be sensitive to the parametric assumptions on the marginal outcome model, see Kenward [1998]. Alternatively, we could *fix* these parameters, estimate the model and investigate the sensitivity of conclusions to a range of plausible values for ψ_i, $i \in \{3, 4\}$. For the traditional selection model where we adjust for missing data by modelling the missingness process, this approach was adopted by Carpenter et al. [2002].

Despite the listed shortcomings of our analysis , we believe that the model families explored are valuable for understanding treatment and covariate effects on the outcome as well as the inclination to reply. More efficient algorithms would facilitate extensions to non-monotone missingness patterns and wider use of these models.

Part II

Dose-Finding Studies and Missing Data

Chapter 7

Missingness and Dose-Finding Studies with Recurrent Event Data

7.1 Introduction

A clinical trial is a research study and usually categorized into Phase I, Phase II and Phase III clinical trials. In the first phase the experimental drug is tested the first time on human beings. The main goal is to prove the safety of the new drug. In a Phase II clinical trial we aim to determine the efficacy of the new drug compared to placebo or an active comparator. Moreover, we want to identify the dose-response relationship and the target dose for which the drug can be shown to be simultaneously safe and effective. Using this dose we then move to the confirmatory Phase III trial.

In this part of the thesis we will focus on dose-finding studies in the context of Phase II trials. The primary outcome of interest consists of the number of events per subject within a specified study period. As with all clinical trials which observe measurements repeatedly over time, we are confronted with missing data, see Chapter 5. In fact, the studies of interest observe pain-related outcomes. Thus, we expect patients to drop out for reasons that are related to the outcome of interest.

Given the importance of selecting the adequate dose, we carefully consider the in-

terplay between recurrent event data modelling, dose selection and different missingness mechanisms. This work is published in Akacha and Benda [2010].

This chapter is organized as follows. The remainder of this section is devoted to reviews on dose-finding studies, recurrent event data analysis and missing data issues. In Section 7.2 we introduce the study which motivated our work. Section 7.3 discusses different methods for handling missing data in recurrent event data studies. In Section 7.4, we investigate regression models for recurrent event data analyses and endpoint-analyses, that is only the data at the end of the study are analyzed. Concluding remarks are given in Section 7.5.

7.1.1 Dose-Finding Studies: A Brief Review

Typically, the aim of a clinical trial is to determine the efficacy of a new drug compared to placebo or an active comparator. In the special case of *dose-finding studies* the interest lies in identifying the *dose-response relationship* and the *target dose* for which the drug can be shown to be simultaneously as effective as a comparator and safe. A good understanding of the dose-response relationship is crucial in clinical drug development. A dose which is too low will hinder the proof of efficacy and a dose which is too high could result in safety issues. In fact, one of the main reasons for the high discontinuation rate of Phase III clinical trials was found to lie in a poor understanding of the dose-response relationship and consequently in an inadequate dose selection [Bretz et al., 2008].

Several approaches have been proposed for the efficacious planning and analysis of a dose-finding study. A methodology that combines formal hypothesis testing for dose response with flexible modeling of the dose-response relationship and estimating a target dose, i.e. a *minimum effective dose* (MED) that produces a clinical relevant effect was proposed in Bretz et al. [2006] and Pinheiro et al. [2006]. This concept of selecting the best model while controlling the familywise error rate and the subsequent target dose estimation is an extension of the ideas proposed in Tukey et al. [1985]. The estimation of the MED based on a given model can be regarded as a calibration problem which can be seen as

a reverse process to regression, i.e. the estimation of a value for an independent variable that yields an expected outcome for the dependent variable equal to a predefined value. An overview on the classical calibration problem is described in Osborne [1991]. There is also an extensive literature on calibration problems related to dose estimation, e.g. Filloon [1995]; Hsu and Berger [1999]; Tamhane and Logan [2002]; Morales et al. [2006]; Budtz-Jœrgensen [2007] and Bretz et al. [2008]. See also Forkman [2008] and Dette et al. [2008] regarding designing aspects of non-linear calibration problems.

Usually, the efficacy is measured relative to placebo. In many applications, however, a comparison with available medications is desired in order to identify the dose beyond which the new drug provides a better efficacy outcome than the competitor drug. This chapter focuses on the estimation of a target dose defined as the dose for which the expected response is equal to that of a competitor group. More generally, the target dose could be defined as the dose that yields an expected response which equals a given function of the average control effect. If for example this function is given by an additive constant it may correspond to the smallest dose for which the expected response is not inferior to that of a control group. The latter modification would be straightforward and may be used in planning non-inferiority trials with an adequate dose. In contrast to the literature cited above, the reference value for the outcome of the target dose must be estimated from the current study, i.e. embedded in the dose-response model, where - depending on the chosen parametrization- either the response of the comparator drug or the target dose itself represents an additional parameter to estimate.

7.1.2 Recurrent Event Data Analysis: A Brief Review

The dose-finding studies that motivated our work seek to analyze processes which generate events repeatedly over time. Such processes are referred to as *recurrent event processes*. Examples include seizures in epileptic studies, hot-flushes postmenopausal women suffer from or flares in gout studies.

Statisticians involved in these studies are usually interested in understanding the

underlying event occurrence process. This includes the investigation of the rate at which events occur, the inter-individual variation and most importantly, the relationship between the event occurrence and explanatory variables such as treatment or dose.

The modelling of recurrent events can be approached in a number of ways. The two most common ways are through *event counts* and *gaps* or *waiting times* between two events. According to Cook and Lawless [2007], models and methods based on counts are often useful when individuals frequently experience the events of interest, and the events are 'incidental' in the sense that their occurrence does not materially alter the process itself, either directly or through resulting interventions. In contrast, analyses based on waiting times are often relevant when events are relatively infrequent, some type of individual renewal occurs after an event, or when prediction of the time to the next event is of interest. The applications in mind for this work are of the former nature, and the canonical approach for the analysis of event counts is the *Poisson process*.

Poisson processes can be defined in various ways, one of which is via the *intensity function*, see Andersen et al. [1993] and Appendix A.3. It gives the instantaneous probability of an event occurring at a certain point in time. Modelling the dependence of recurrent events on explanatory variables can be achieved by specifying the associated intensity as a function of those variables. Corresponding work and different models have been presented in Andersen and Gill [1982]; Lawless [1987]; Andersen et al. [1993]; Lindsey [1995] and Cook and Lawless [2007].

7.1.3 Recurrent Event Data and Missingness: A Brief Review

The studies that motivated our work focus mainly on the number of events that occur by the end of the study period. Therefore, modelling the counts through a *generalized linear model* as introduced in McCullagh and Nelder [1989] might seem sufficient. However, in recurrent event data studies and more generally in studies with repeated measurements, incomplete data due to missed visits or dropouts are quite common. Hence, the endpoint of interest may be missing. In many situations, however, information about the counting process prior

to dropout is available, e.g. through patient diaries. This knowledge can be incorporated in a recurrent event data analysis.

For the specific case of recurrent event data, the literature mainly distinguishes between two different missingness mechanisms: the *conditionally independent censoring mechanism* and the *dependent* or *informative censoring mechanism*. The conditionally independent censoring corresponds to the general concept of data being 'missing at random' in the terminology of Little and Rubin [2002], see Cook and Lawless [2007]. It has been shown that valid likelihood inference can then be based on the observed data process by using the concept of *risk sets*, see Andersen et al. [1993]; Robins and Rotnitzky [1995] and Cook and Lawless [2007]. Under informative censoring, the censoring mechanism depends on quantities which are unknown prior to dropout. In this case, joint modelling of the recurrent events and censoring mechanism may be necessary, see Cook and Lawless [2007]. Joint parametric or semi-parametric models, similar in spirit to shared parameter models, have been discussed in Lancaster and Intrator [1998] and Wang et al. [2001]. A widely used method to adjust for informative censoring is based on *inverse probability of censoring weights*, see Robins and Rotnitzky [1995]; Miloslavsky et al. [2004] and Cook and Lawless [2007]. For a more thorough bibliographic note on censoring mechanisms in the case of recurrent event data we refer to [Cook and Lawless, 2007, Chapter 2 and Chapter 7].

Although these methods to adjust for missingness in the case of recurrent event data exist, we aim to explore and apply methods that are generally used for longitudinal data. This approach is advantageous, because these methods are usually better known to the clinical community. We hope this research will make it easier for clinicians to abandon the widely used complete case analysis or last observation carried forward approach in favor of more suitable models. To the best of our knowledge there is no such review of several missing data methods for the analysis of recurrent event data or any work on missingness in the case of dose-finding studies. Different missing data methods with the main focus on repeated measurement studies were discussed in Chapter 5.

Regulatory aspects of missing data in clinical trials are described in the *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use* (ICH) E9 guideline [ICH-E9, 1998], and the EMEA Points to Consider on Missing Data [EMEA, 2001]. Recently, a revised EMEA guideline has been published as a draft, EMEA [2009], giving a more detailed insight in potential issues related to the different methods of missing data handling. Whilst, from a regulatory point of view, the main focus appears to lie in confirmatory Phase III trials, the undesired effects of missing data are also relevant in dose-finding studies. As described in EMEA [2009], bias is the most important concern. In dose-finding, this may lead either to over- or underestimation of the treatment effect or a dose-response slope. Overestimation would, in general, lead to an underestimation of a target dose, and vice versa. Consequently, the assumed therapeutic window would be either too narrow or too wide. In the first case, the development program might erroneously be stopped or continued with an unnecessarily high dose that has a potentially poor safety profile, whereas in the latter case an ineffective dose might be put forward to Phase III.

7.1.4 Notation

Suppose m independent subjects are randomized into a trial and that each subject experiences a type of recurrent event.

Let n_i be a realisation of the random variable $N_i(T)$ and denote the number of events over the complete study period $[0, T]$ for the i-th subject, $i \in \{1, ..., m\}$. The event times for subject i are denoted by $0 < t_{i,1} < ... < t_{i,n_i} \leq T$ and the corresponding 'time of occurrence' random variables by $T_{i,1}, ..., T_{i,n_i}$.

Let x_{i,j_i} be the vector of explanatory variables for subject i and event time $j_i \in \{t_{i,1}, ..., t_{i,n_i}\}$. It is assumed that x_{i,j_i} are time-independent explanatory variables. Therefore, $x_{i,j_i} = x_i$ for all $j_i \in \{t_{i,1}, ..., t_{i,n_i}\}$. Let s_i denote the treatment for subject i. Assuming the study design involves ℓ doses, $q_1, ..., q_\ell$, of the new drug and one comparator (C), then $s_i \in \{q_1, ..., q_\ell, C\}$. The target dose of interest is denoted by $\eta \in \mathbb{R}$. Dependent on the

context, we denote the entire parameter vector of interest by $\theta \in \mathbb{R}^{p_1}$, $p_1 \in \mathbb{N}$ or $\vartheta \in \mathbb{R}^{p_2}$, $p_2 \in \mathbb{N}$.

Furthermore, let $N_i = \{N_i(T), T_{i,1}, ..., T_{i,n_i}\}$ denote the complete recurrent event data information for subject i and let $t_{i,d} \in (0, T]$ indicate the dropout time for subject $i \in \{1, ..., m\}$. For non-completers the dropout time is defined as the last observed event time. Then $N_{i,obs} = \{N_i(t_{i,d}), T_{i,1}, ..., T_{i,d}\}$ denotes the observed part and $N_{i,mis} = \{N_i(T), T_{i,d+1}, ..., T_{i,n_i}\}$ the missing part of the recurrent event data sequence. For clarity, in this work we assume that once patients drop out, they do not return to the study.

7.2 An Example: The Gout Study

The study that motivated our work is an upcoming dose-finding study, which investigates a new compound for the prophylaxis of signs and symptoms of acute flares in chronic gout patients. An active controlled Phase II study will be carried out: 350 patients suffering from chronic gout are randomized into one of six treatment groups. Five of these groups correspond to different doses of the new drug, the sixth is an active control. One hundred patients are randomized into the active control group, whereas 50 patients are assigned to one of the five doses of the new drug. For similar studies see Borstad et al. [2004] and Becker et al. [2005].

The primary objective is to determine the target single dose of the new compound that leads to the same efficacy as the active-control, with respect to the mean number of gout flares occurring within 16 weeks of randomization.

Given the primary outcome, an endpoint-analysis is to be performed. However, given the painful nature of gout flares many patients are expected to drop out. Safety concerns for high doses and adverse events are further expected reasons for dropout.

Performing an endpoint-analysis without adjustment for missingness would imply no inclusion of information about non-completers. The underlying assumption is an identical evolution for completers and non-completers. Given the reasons for dropout, this is

rather unrealistic.

Fortunately, all patients are asked to record the date they suffered from a gout flare. Using this information, the number of flares by the time of dropout will be known and can be included in the modelling framework, as will be shown in Section 7.4. Explanatory variables of interest are age, ranging between 16 years and 72 years, and treatment.

7.3 Approaches to the Analysis of Recurrent Event Data with Dropout

As described in the introduction, the studies that motivated this work focus mainly on the number of events that occur by the end of a specific study period. In practice, however, the primary endpoint is not always observed, but information about the counting process before dropping out is usually available. Hence, the remaining derivations of this chapter are based on this assumption.

In the following subsections the focus will lie on different methods to deal with missing data in this specific case. Five approaches will be discussed: complete case analysis; two single imputation techniques; the direct likelihood approach and pattern-mixture models.

7.3.1 Complete Case and Imputation-Based Procedures

In a complete case analysis (CC) all participants with missing data are simply discarded and the missingness process is not explicitly incorporated, i.e. we exclude all patients for whom n_i is missing.

Another common strategy is to fill in missing values based on the observed measurements. We will explore two single imputation techniques: last observed rate carried forward (LORCF) and last count carried forward (LCCF). For the shortcomings of complete case analysis and single imputation techniques we refer to Section 5.2.

LORCF makes the implicit assumption that events for a specific patient occur at the

same average rate before and after dropout. The missing true endpoint n_i is replaced by the rounded value of $n_i(t_{i,d}) + (T - t_{i,d}) \frac{n_i(t_{i,d})}{t_{i,d}}$, where $n_i(t_{i,d})$ is the number of events by drop-out.

For LCCF, the missing outcome n_i is simply replaced by $n_i(t_{i,d})$. Replacing every missing endpoint with the imputed value leads to a 'completed' data set.

Based on the 'completed' data sets a generalized linear model for count data can be formulated, see Section 7.4.1.

7.3.2 Direct-Likelihood Approach

The direct likelihood (DL) approach is valid under ignorability, see Section 5.3.1 and Little and Rubin [2002]. Information from non-completers is incorporated when fitting a specific model and estimating the target dose of interest. Let $N_{i,obs} = \{N_i(t_{i,d}), T_{i,1}, ..., T_{i,d}\}$ denote the observed part and $N_{i,mis} = \{N_i(T), T_{i,d+1}, ..., T_{i,n_i}\}$ the missing part of the recurrent event data sequence. Likelihood inference for the parameter of interest θ can then be based on the observed likelihood

$$L_{obs}(\theta|N_{obs}) = \prod_{i=1}^{m} f_{N_{i,obs}}(n_{i,obs}|\theta),$$

where $f(\cdot|\theta)$ denotes a parametric model for the recurrent event data sequence. Specific models which enable the estimation of the target dose of interest will be discussed in Section 7.4.2.

7.3.3 Analysis using Pattern-Mixture Models

In the case of pattern-mixture models (PMM), we assume that the distribution differs according to the underlying missingness pattern, see Section 5.4.2.

Let d_i denote the missingness pattern for subject i and P_i be the associated random variable. In case of continuous time recurrent event data the missingness pattern d_i is equivalent to the dropout time $t_{i,d}$. The joint distribution of N_i and P_i is then given by

$$f_{(N_i,P_i)}(n_i, t_{i,d}) \quad = \quad f_{N_i|P_i}(n_i|t_{i,d})\, f_{P_i}(t_{i,d})$$

$$= f_{N_{i,obs}|P_i}(n_{i,obs}|t_{i,d})\, f_{N_{i,mis}|N_{i,obs},P_i}(n_{i,mis}|n_{i,obs}, t_{i,d})\, f_{P_i}(t_{i,d}), \qquad (7.1)$$

where the first and third factors can be modelled through the data. We can use identifying restrictions to determine the unknown conditional densities of unobserved components, given a set of observed quantities, see Section 5.4.2 and Little [1994]. For other ways to overcome the under-identification, we refer to Molenberghs and Kenward [2007].

Different identifying restrictions yield different missingness mechanisms. In our setting each pattern (except the pattern of the completers) consists of one patient only. Therefore, estimating the pattern-specific identifiable densities, as required in the fitting procedure for the discrete time PMMs, see Section 5.4.2 and Thijs et al. [2002], is not always feasible. We will use the *complete case missing value restrictions* (CCMV), where information which is unavailable is borrowed from the model for the complete cases, see Section 5.4.2 and Molenberghs et al. [1998].

In order to be able to borrow information using other identifying restrictions, patients with similar dropout times have to be clustered into one group or pattern. Once the clustering is done, identifying restrictions for the unknown conditional densities need to be specified. Although very interesting, tackling this problem is beyond the scope of this thesis and we will focus on the CCMV restrictions.

From equation (7.1), PMMs rely on modelling the whole recurrent event data sequence. A corresponding model will be presented in Section 7.4.2 and this section will be revisited in Section 7.4.3.

7.4 Model Specification

In this section, we formulate regression models that enable the target dose selection and the investigation of covariate effects on the endpoints n_1,...,n_m, taking into account that some endpoints are missing. Some of the missing data handling methods discussed in Section 7.3 edit the incomplete data set in order to obtain 'completed' data sets (CC, LORCF, LCCF). A statistical analysis based on these approaches requires a model for the endpoints only.

In contrast, the direct likelihood approach requires a model for the whole recurrent event data sequence. In order to create multiple imputations of the endpoints based on a pattern-mixture model, we need to model the whole recurrent event data sequence. Based on the 'completed' data sets an endpoint analysis can be performed. That is, fitting a pattern-mixture model in this context requires two models.

In Section 7.4.1 we discuss a model suitable for an endpoint analysis, while Section 7.4.2 is devoted to models for the whole recurrent event data sequence. Section 7.4.3 discusses how these models can be combined to fit a pattern-mixture model. We develop the models for a single subject. In view of the assumed independence between subjects, it is then easy to build the complete models.

7.4.1 Models for Count Data

The probability of n events occurring in the interval $[0, T]$ can be modelled by a Poisson distribution with mean $\Lambda_x(\vartheta)$, i.e.

$$\mathbb{P}\left(N(T) = n\right) = \frac{\exp\left(-\Lambda_x(\vartheta)\right)\Lambda_x^n(\vartheta)}{n!},$$

where ϑ denotes the parameter of interest and x the explanatory variables.

Following the ideas in McCullagh and Nelder [1989], a generalized linear model will be formulated to model the mean $\Lambda_x(\vartheta)$ as a function of the dose and other explanatory variables. The canonical link function is used and the following relationship for the expectation of $N(T)$, $\mathbb{E}\left[N(T)\right]$, is assumed:

$$\ln\left\{\mathbb{E}\left[N(T)\right]\right\} \;=\; \ln\left[\Lambda_x(\vartheta)\right] \;=\; \alpha_0 + \alpha_1\,\ln(age) + \alpha_2\,g(s,\eta),$$

where η is the target dose of interest and $\vartheta = (\alpha_0, \alpha_1, \alpha_2, \eta)^\top$. Further explanatory variables such as gender and occupation can be included. The function $g(\cdot, \eta)$ quantifies the dose-response relationship between the endpoint $N(T)$ and the dose, treated as a continuous variable. This work will focus on linear and log-linear dose-response relations. An

extension of the models for other dose-response relations [Bretz et al., 2005] is straightforward. Now suppose subject i and j share the same covariates. The target dose is then defined as

$$\eta = \left\{ q \in \mathbb{R} \,:\, \mathbb{E}[N_i(T) \,|\, s_i = q] = \mathbb{E}[N_j(T) \,|\, s_j = C] \right\}.$$

The following parametrizations for $g(\cdot, \eta)$ are proposed for the estimation of η, $s \in \{q_1, ..., q_\ell\}$:

$$\begin{aligned} &\text{linear model:} && g(s,\eta) &&= s - \eta; \quad \text{and} \\ &\text{log-linear model:} && g(s,\eta) &&= \ln(s) - \ln(\eta). \end{aligned}$$

For $s = C$ the value $g(C, \eta)$ is set to zero, i.e.

$$g(C,\eta) = 0 \quad \text{and} \quad g(\eta,\eta) = 0,$$

enabling the selection of the target dose defined as the dose that leads to the same expected number of events as the comparator.

As defined, the Poisson model features the constraint of equal mean and variance, but the number of events usually varies beyond what can be explained through available covariates. To take this inter-individual variation into account, a model in which the regression parameters vary across different patients will be considered, see Chapter 3. The following *generalized linear mixed model* is proposed:

$$\begin{aligned} N(T)|U \;&\sim\; Poisson\left[U\,\Lambda_x(\vartheta)\right]; \quad U \sim Gamma\left(\zeta^{-1}, \zeta\right) \qquad \text{and} \\ \ln\left\{\mathbb{E}\left[N(T)|U\right]\right\} \;&=\; \ln\left[U\,\Lambda_x(\vartheta)\right] \;=\; \alpha_0 + \alpha_1\,\ln(age) + \alpha_2\,g(s,\eta) + \ln\left(U\right), \end{aligned}$$

where $g(\cdot, \eta)$ is defined above, $Poisson(\cdot)$ denotes the Poisson distribution and $Gamma(\cdot, \cdot)$ the gamma distribution, see Appendix A.2 for the parametrization used. Assuming the U's are gamma-distributed random variables with mean 1 and variance ζ implies that subjects with U greater than one are more likely to experience an event than those with U less

than one. In particular, the marginal distribution of $N(T)$ is given by a negative-binomial distribution ($\mathcal{NB}(\cdot,\cdot)$), see for example Diggle et al. [1996]:

$$N(T) \sim \mathcal{NB}\left(\frac{1}{\zeta}, \frac{1}{1+\zeta\,\Lambda_x(\vartheta)}\right) \quad \text{with} \quad \ln\left[\Lambda_x(\vartheta)\right] = \alpha_0 + \alpha_1\,\ln(age) + \alpha_2\,g(s,\eta);$$

such that the overdispersion coefficient is given by $1+\zeta\,\Lambda_x(\vartheta)$. We note that the overdispersion depends on the explanatory variables.

The joint likelihood $L_N(\vartheta,\zeta)$ of $n_1,...,n_m$ is then given by

$$\begin{aligned} L_N(\vartheta,\zeta) &:= \prod_{i=1}^{m} L_{N_i}(\vartheta,\zeta) \\ &= \prod_{i=1}^{m} \frac{\Gamma(n_i+\frac{1}{\zeta})}{\Gamma(\frac{1}{\zeta})} \left(\frac{\zeta\Lambda_{x_i}(\vartheta)}{\zeta\Lambda_{x_i}(\vartheta)+1}\right)^{n_i} \left(\frac{1}{\zeta\Lambda_{x_i}(\vartheta)+1}\right)^{\frac{1}{\zeta}}. \end{aligned}$$

Note that we consider this likelihood for all 'completed' data sets, i.e. under the complete case, LORCF and LCCF approach. Furthermore, we can use this likelihood in the context of the PMM analysis, once the missing endpoints were imputed, see Section 7.4.3 for more details.

The maximum likelihood estimates for ϑ and ζ can be calculated through the dual quasi-Newton algorithm, implemented for example in `NLMIXED`, SAS/STAT [1999]. The estimates obtained will depend substantially on the procedure used to deal with the missing data and the chosen dose-response model. This inference can be severely biased, as will be shown in Chapter 8.

7.4.2 Models for Recurrent Event Data

For inference based on the direct likelihood and the PMM approach, a model for the whole event data sequence $N = (N(T), T_1, ..., T_n)$ is required.

We assume that events occur in continuous time. For simplification, it is often assumed that events occur according to a Poisson process $\{N(t), t \geq 0\}$ with (time-dependent) intensity function $\lambda_x(t,\theta)$, see Appendix A.3. The intensity function specifies the instan-

taneous probability of an event occurring at a certain point in time. According to Lawless [1987], the corresponding *cumulative intensity function* is given by

$$\Lambda_x(t,\theta) = \int_0^t \lambda_x(w,\theta)\ \mathrm{d}w,$$

where θ quantifies the relationship between the intensity and covariates x. The joint distribution of the count n and the event times $0 \leq t_1, ..., t_n \leq T$ can be shown to be

$$\mathbb{P}\left[N(T) = n, T_1 = t_1, ..., T_n = t_n\right] = \exp\left[-\Lambda_x(T,\theta)\right] \prod_{j=1}^{n} \lambda_x\left(t_j,\theta\right)$$

via *product integration*, see Kalbfleisch and Prentice [1980]; Andersen et al. [1993] and Cook and Lawless [2007]. However, this model ignores the overdispersion. Therefore, we consider instead a model with varying event rates across patients, where

$$\lambda_x(t,\theta|U=u) = u\,\lambda_x(t,\theta) \tag{7.2}$$

and u is a realization of the random variable U. Here, the extended model of interest is given by

$$N(t)|U \sim Poisson\left[\Lambda_x(t,\theta|U)\right] \qquad \text{and} \qquad U \sim Gamma\left(\zeta^{-1},\zeta\right),$$

where

$$\Lambda_x(t,\theta|U=u) = \int_0^t u\,\lambda_x(w,\theta)\ \mathrm{d}w = u\,\Lambda_x(t,\theta).$$

Note that the intensity, implicitly given in equation (7.2), belongs to the conditional counting process $\{N|U=u\}$ and not to the marginal process N.

The contribution of a specific subject to the joint likelihood for the outcome 'n events occur at times $t_1, ..., t_n$', with U specified as above, is then given by

$$L_N(\zeta,\theta) = \int f_{N(T),T_1,...,T_n,U}(n,t_1,...,t_n,u)\,\mathrm{d}u$$

$$
\begin{aligned}
&= \int f_{N(T),T_1,\ldots,T_n|U}(n, t_1, \ldots, t_n)\ f_U(u)\,\mathrm{d}u \qquad (7.3)\\
&= \prod_{j=1}^{n}\left(\frac{u\,\lambda_x\left(t_j,\theta\right)}{u\,\Lambda_x\left(T,\theta\right)}\right) n!\ \int \frac{\exp\left\{-u\,\Lambda_x\left(T,\theta\right)\right\}\left(u\,\Lambda_x\left(T,\theta\right)\right)^n}{n!}\ f_U(u)\,\mathrm{d}u\\
&= n!\left(\prod_{j=1}^{n}\frac{\lambda_x\left(t_j,\theta\right)}{\Lambda_x\left(T,\theta\right)}\right)\underbrace{\frac{\Gamma(n+\frac{1}{\zeta})}{n!\,\Gamma(\frac{1}{\zeta})}\left(\frac{\zeta\Lambda_x\left(T,\theta\right)}{\zeta\Lambda_x\left(T,\theta\right)+1}\right)^n\left(\frac{1}{\zeta\Lambda_x\left(T,\theta\right)+1}\right)^{\frac{1}{\zeta}}}_{f(\theta,\zeta)},
\end{aligned}
$$

where $\Gamma(\cdot)$ denotes the gamma function (see Appendix A.1) and $f\left(\theta,\zeta\right)$ is the density of a negative-binomial distributed random variable with mean $\Lambda_x\left(T,\theta\right)$ and variance $\Lambda_x\left(T,\theta\right)+\zeta\,\Lambda_x^2\left(T,\theta\right)$. In particular, the overdispersion coefficient is given by $1+\zeta\,\Lambda_x(T,\theta)$, and thus depends on explanatory variables.

We now want to discuss suitable models for the intensity function $\lambda_x\left(t,\theta\right)$. In the literature it is very common to assume covariates affect the intensity through a *multiplicative model* of the form

$$\lambda_x\left(t,\theta\right) = \lambda_0(t,\delta)\,h(x,\beta), \qquad (7.4)$$

where $\theta=(\beta,\delta)^\top$ is the parameter vector of interest, $\lambda_0(\cdot)$ is the *baseline intensity function* and $h(\cdot)$ is a positive-valued function of x and β, see Andersen et al. [1993] and Lawless [1995]. Then, the corresponding cumulative intensity function is given by

$$\Lambda_x\left(t,\theta\right) = \Lambda_0\left(t,\delta\right)\ h(x,\beta), \quad \text{where} \quad \Lambda_0\left(t,\delta\right) = \int_0^t \lambda_0\left(w,\delta\right)\ \mathrm{d}w.$$

The case where $\lambda_0(\cdot)$ is left arbitrary (a semi-parametric model) is distinguished from the case where the baseline function is specified up to a parameter vector δ (a full parametric model). As implied by the notation in equation (7.4), the focus of this work lies in fully parametric models. Further discussion on semi-parametric models can be found in Lawless [1987] and Lin et al. [2001].

Model (7.4) implies that the intensity functions associated with any two sets of

covariate values, say x_i and x_j, are proportional over time, i.e.

$$\begin{aligned} & \lambda_{x_i}(t,\theta) = \lambda_0(t,\delta)\, h(x_i,\beta) \quad \text{and} \quad \lambda_{x_j}(t,\theta) = \lambda_0(t,\delta)\, h(x_j,\beta) \\ \Leftrightarrow \quad & \frac{\lambda_{x_i}(t,\theta)}{\lambda_{x_j}(t,\theta)} = \underbrace{\frac{h(x_i,\beta)}{h(x_j,\beta)}}_{=:c} \quad \Leftrightarrow \quad \lambda_{x_i}(t,\theta) = c \cdot \lambda_{x_j}(t,\theta)\,. \end{aligned}$$

According to Lin et al. [2001], this restriction may be too strong in practice. Alternative non-multiplicative models such as *additive* and *time transform* models can be formulated, see Cook and Lawless [2007] and citations therein. Alternatively, the model (7.4) could be extended such that the baseline rates depend on explanatory variables: $\lambda_x(t,\theta) = \lambda_0(t,x,\delta)\, h(x,\beta)$. The effect of covariates on the intensity remains easy to interpret; however, interpreting this effect on the cumulative intensity function, i.e. the mean of the counting process, can become quite difficult. In this work we chose to adopt a multiplicative model, despite knowing that the proportionality assumption might be too stringent. We are willing to make this assumption, in particular, because our focus does not lie on the modelling aspect. Our key goal is to investigate the impact of dropout, which will be based on a simulation study. In this simulation study we will employ the same model for the simulations as for the analysis of those simulated data sets. Thus, our main challenge is not to find an appropriate model for a real data set. However, we acknowledge that exploring the effect of using different simulation and analysis models should be considered in future investigations.

For full parametric multiplicative models, different models for the baseline function and the function $h(\cdot)$ may be proposed. It is often convenient to choose the function $h(\cdot)$ according to the *Andersen-Gill Model* [Andersen and Gill, 1982; Cook and Lawless, 2002]:

$$h(x,\beta) = \exp\left(x^\top \beta\right)$$

since no restrictions on the values of the regression parameter β are needed.

Here, we will focus on a slight extension of the Andersen-Gill-Model. We assume

$$h(x,\beta) = \exp\left[k(x,\beta)\right], \tag{7.5}$$

where $k(\cdot)$ is a (not necessarily linear) function of x and β. One possible model with $x = (\ln(age), z)^\top$ is

$$k(x,\beta) \quad = \quad \beta_0 \, \ln(age) + \beta_1 \, g(s,\eta). \tag{7.6}$$

Different choices for $g(\cdot,\eta)$ were discussed in Section 7.4.1.

Using the extended fully parametric Andersen-Gill model with $h(x,\beta)$ specified in equation (7.5) leads to a time dependence solely through the baseline intensity function. Different assumptions about the function $\lambda_0(\cdot)$ can be made. If the rate of events is expected to be constant over the whole study period, it is sensible to choose $\lambda_0(\cdot)$ as a constant function. This yields a *homogeneous Poisson process*. Alternatively, a monotone decreasing or increasing rate function may be suitable and can be realized through a *Weibull rate function*, see Lawless [1987]. In the case of a seasonal illness, a periodic rate function might be considered.

The baseline functions that we are going to investigate are given by:

$$\begin{aligned} &\text{Constant:} \quad \lambda_0(t,\delta) = \delta, \quad \delta \in \mathbb{R}^+ \\ &\text{Weibull :} \quad \lambda_0(t,\delta) = \delta_0 \, \delta_1 \, t^{\delta_1 - 1} \quad \text{with} \quad \delta = (\delta_0, \delta_1)^\top \in \mathbb{R}^+ \times \mathbb{R}^+. \end{aligned}$$

Note that the Weibull intensity function is monotone decreasing for $0 < \delta_1 < 1$, constant for $\delta_1 = 1$ and monotone increasing for $\delta_1 > 1$.

Depending on the chosen intensity function the likelihood in equation (7.3) will become more or less cumbersome; with a constant function the likelihood simplifies substantially. Leaving out the density of the negative-binomial distribution in equation (7.3)

leads to

$$\begin{aligned} \text{Constant:} \qquad & \prod_{j=1}^{n} \frac{\lambda_0\left(t_j, \delta\right)}{\Lambda_0\left(T, \delta\right)} &&= \frac{1}{T^n} \\ \text{Weibull:} \qquad & \prod_{j=1}^{n} \frac{\lambda_0\left(t_j, \delta\right)}{\Lambda_0\left(T, \delta\right)} &&= \frac{\delta_1^n (\prod_{j=1}^{n} t_j)^{\delta_1 - 1}}{T^{n\,\delta_1}}. \end{aligned}$$

In summary, the following model for the recurrent event data sequence N is considered:

$$\begin{aligned} N(t)|U &\sim Poisson\left[U\,\Lambda_x(t,\theta)\right]; \quad U \sim Gamma\left(\zeta^{-1}, \zeta\right) \qquad \text{and} \\ \ln\left\{\mathbb{E}\left[N(t)|U\right]\right\} &= \beta_0\,\ln(age) + \beta_1\,g(s,\eta) + \ln\left[\Lambda_0\left(t,\delta\right)\right] + \ln\left(U\right), \end{aligned} \tag{7.7}$$

where $\theta = (\beta, \eta, \delta)^\top$. The constant and Weibull intensity functions will be considered in the simulation study presented in Chapter 8.

For the DL approach we base inference on $L_{N_{obs}}(\theta, \zeta) = \prod_{i=1}^{m} L_{i,obs}(\theta, \zeta)$, where

$$L_{i,obs}(\zeta,\theta) = \tilde{n}_{i,obs}! \left[\prod_{j=1}^{\tilde{n}_{i,obs}} \frac{\lambda_{x_i}\left(t_{i,j},\theta\right)}{\Lambda_{x_i}\left(T,\theta\right)}\right] \frac{\Gamma(\tilde{n}_{i,obs} + \frac{1}{\zeta})}{\tilde{n}_{i,obs}!\Gamma(\frac{1}{\zeta})} \left[\frac{\zeta\Lambda_{x_i}\left(T,\theta\right)}{\zeta\Lambda_{x_i}\left(T,\theta\right)+1}\right]^{\tilde{n}_{i,obs}} \left[\frac{1}{\zeta\Lambda_{x_i}\left(T,\theta\right)+1}\right]^{\frac{1}{\zeta}}$$

with $\tilde{n}_{i,obs} := n_i(t_{i,d})$ if patient i drops out and $\tilde{n}_{i,obs} := n_i(T)$ for completers.

In order to fit a pattern-mixture model using the CCMV restrictions we first need to estimate (θ, ζ) based on the likelihood (7.3) for the complete cases only. Based on the resulting estimates the missing values are replaced by multiple imputations, see Section 7.3.3 and Section 7.4.3.

For both approaches the dual quasi-Newton method is used to maximize the corresponding likelihoods.

7.4.3 Pattern-Mixture Models: Revisited

We now revisit the pattern-mixture models described in Section 7.3.3, and aim to identify the conditional model $f_{N_{mis}|N_{obs},P}(n_{mis}|n_{obs}, t_d, X, \theta, \zeta)$.

As we are observing an overdispersed Poisson process, we incorporated an unob-

servable subject-specific random effect $U \sim Gamma(\zeta^{-1}, \zeta)$ in the model framework proposed in Section 7.4.2. Conditional on this random variable the event counts were assumed to follow a Poisson process. As here the interest lies in the unconditional counting process N, we note that this is no longer a Poisson process with rate function $u\,\lambda_x(t,\theta)$. Instead, using the notation in Section 7.4.2, the rate function of the unconditional counting process N is given by

$$\tilde{\lambda}_x(t,\zeta,\theta) = \frac{1+\zeta\,N(t-)}{1+\zeta\,\Lambda_x(t,\theta)}\,\lambda_x(t,\theta), \tag{7.8}$$

where $N(t-)$ is the number of events that occur in the time interval $[0,t)$, see Andersen et al. [1993].

In order to identify $f_{N_{mis}|N_{obs},P}(n_{mis}|n_{obs},t_d,X,\theta,\zeta)$ we use the CCMV identifying restrictions, where we only have to distinguish between the set of completers (=: I) and the set of non-completers (=: I^c). The parameters of interest in the likelihood contribution (7.3) are θ and ζ. Assume that these were estimated through $\hat{\theta}_{comp}$ and $\hat{\zeta}_{comp}$ for the complete cases, using the model in (7.7). Then, the rate function for the completers $\ell \in I$ is determined through

$$\tilde{\lambda}_{x_\ell}(t,\hat{\zeta}_{comp},\hat{\theta}_{comp}) = \frac{1+\hat{\zeta}_{comp}\,N_\ell(t-)}{1+\hat{\zeta}_{comp}\,\Lambda_{x_\ell}(t,\hat{\theta}_{comp})}\,\lambda_{x_\ell}(t,\hat{\theta}_{comp}).$$

For $i \in I^c$, we then identify

$$N_i(T) = N_i(t_{i,d}) + N_{comp}\left\{(t_{i,d},T]\right\}|N_i(t_{i,d}), \tag{7.9}$$

where $N_{comp}\left\{(t_{i,d},T]\right\}|N_i(t_{i,d})$ is a draw for the number of events occurring in the interval $(t_{i,d},T]$, based on the counting process with intensity function $\tilde{\lambda}_{x_i}(t,\hat{\zeta}_{comp},\hat{\theta}_{comp})$, $t \in (t_{i,d},T]$.

As it is cumbersome to draw from $N_{comp}\left\{(t_{i,d},T]\right\}|N_i(t_{i,d})$, we choose to draw the event times and to count the event occurrences by the end of the study. Given an event

occurred at time t_{j-1}, the waiting time w_j for event t_j can be simulated based on

$$B_j = \int_{t_{j-1}}^{t_{j-1}+w_j} \tilde{\lambda}_{x_i}(t, \hat{\zeta}_{comp}, \hat{\theta}_{comp})\mathrm{d}t,$$

where B_j is standard exponential distributed, see Cook and Lawless [2007].

Solving this integral for the constant and Weibull rate functions is straightforward. For the constant rate function we obtain:

$$\begin{aligned}
B_j &= \int_{t_{j-1}}^{t_{j-1}+w_j} \frac{1+\zeta N(t-)}{1+\zeta \Lambda_x(t,\theta)} \lambda_x(t,\theta)\mathrm{d}t \\
&= \int_{t_{j-1}}^{t_{j-1}+w_j} \frac{1+\zeta N(t-)}{1+\zeta \delta\, t\, h(x,\beta)} \delta\, h(x,\beta)\mathrm{d}t \\
&= (1+\zeta N(t_{j-1}))\, \delta\, h(x,\beta) \int_{t_{j-1}}^{t_{j-1}+w_j} \frac{1}{1+\zeta \delta\, t\, h(x,\beta)}\mathrm{d}t \\
&= (1+\zeta N(t_{j-1}))\, \delta\, h(x,\beta)\, \frac{1}{\zeta} \cdot \ln\left(|1+\zeta \delta\, t\, h(x,\beta)|\right) \Big|_{t_{j-1}}^{t_{j-1}+w_j}.
\end{aligned}$$

Reformulation of this equation yields

$$w_j = \frac{\exp\left(\frac{B_j}{\frac{1}{\zeta}+N(t_{j-1})} + \ln\left(|1+\zeta \delta\, t_{j-1}\, h(x,\beta)|\right)\right) - 1}{\zeta \delta\, h(x,\beta)} - t_{j-1},$$

where $B_j \sim Exp(1)$. We note that the denominator turns zero when $\zeta = 0$, i.e. when dealing with the classical homogeneous Poisson process with intensity function $\lambda_x(t,\theta) = \delta\, h(x,\beta)$. However, in that case it is well known that the waiting times are exponentially distributed with parameter $\delta\, h(x,\beta)$. In fact, by using *l'Hôspital's rule* (l'HR) for the calculation of limits, we obtain this result. Let

$$\xi := \exp\left(\frac{B_j}{\frac{1}{\zeta}+N(t_{j-1})} + \ln\left(|1+\zeta \delta\, t_{j-1}\, h(x,\beta)|\right)\right),$$

then

$$\begin{aligned}
\lim_{\zeta\to 0} w_j &= \lim_{\zeta\to 0} \frac{\exp\left(\frac{B_j}{\frac{1}{\zeta}+N(t_{j-1})} + \ln\left(|1+\zeta\,\delta\,t_{j-1}\,h(x,\beta)|\right)\right) - 1}{\zeta\,\delta\,h(x,\beta)} - t_{j-1} \\
&\overset{\text{l'HR}}{=} \lim_{\zeta\to 0} \frac{\left(\frac{B_j}{\zeta^2\,[\zeta^{-1}+N(t_{j-1})]^2} + \frac{\delta\,t_{j-1}\,h(x,\beta)}{1+\zeta\,\delta\,t_{j-1}\,h(x,\beta)}\right)\cdot\xi}{\delta\,h(x,\beta)} - t_{j-1} \\
&= \frac{B_j + \delta\,t_{j-1}\,h(x,\beta)}{\delta\,h(x,\beta)} - t_{j-1} \\
&= \frac{B_j}{\delta\,h(x,\beta)},
\end{aligned}$$

i.e. the inter-arrival times are exponentially distributed with parameter $\lambda_x(t,\theta) = \delta\,h(x,\beta)$.

Similarly, we obtain the following expression for the waiting times under a Weibull rate function, i.e. $\lambda_x(t,\theta) = \delta_0\,\delta_1\,t^{\delta_1-1}\,h(x,\beta)$:

$$w_j = \left(\frac{\exp\left\{\frac{B_j}{\frac{1}{\zeta}+N(t_{j-1})} + \ln\left(|1+\zeta\,\delta_0\,t_{j-1}^{\delta_1}\,h(x,\beta)|\right)\right\} - 1}{\zeta\,\delta_0\,h(x,\beta)}\right)^{\frac{1}{\delta_1}} - t_{j-1}.$$

Now, for each patient i from the set I^c, multiple draws of $N_i(T)$ are created according to the identifying restriction defined in equation (7.9). In our case we will use 5 imputations. Then, the missing endpoints for subjects $i \in I^c$ are replaced by these multiple draws, creating several 'completed' data sets for all subjects $i \in I^c$. Subsequently, the generalized linear mixed model, introduced in Section 7.4.1, is fitted to these 'completed' data sets. The same model is fitted separately to the data of all completers $\ell \in I$. For each parameter the final estimate is a weighted average of the pattern specific estimates. The weights are estimated as the pattern probabilities, see Section 5.4.2 and Thijs et al. [2002] for details about fitting procedures for pattern-mixture models.

7.5 Summary

This chapter was motivated by a dose-finding study which explores a recurrent event data process over a time period of several weeks. In practice, investigators are not interested in modelling the whole recurrent event sequence. Instead, they focus on the number of events occurring in a specific time interval. Due to missingness this endpoint is not observed for all patients and the classical approach of investigators is to perform a complete case analysis where all non-completers are discarded from the analysis.

In our study of interest, however, dropout is expected to be outcome related and the underlying assumption justifying a complete case analysis, i.e. MCAR, is violated. Therefore, other techniques to handle missing data and their performances have to be investigated. Some classical techniques for dealing with missing data in the case of repeated measurements were discussed in Section 7.3. Some of these methods require models for the whole recurrent event process, whereas others enable an endpoint analysis for the number of events occurring by the end of the study. Regression models which enable the target dose selection and the investigation of the effect of other explanatory variables were laid out in Section 7.4. The recurrent events are modelled as overdispersed Poisson process data, with dose and age as regressors. We examine constant and time-varying rate functions.

In order to compare the performances of the proposed missing data methods a scenario evaluation study will be presented in Chapter 8.

Chapter 8

Simulation Study

The key goal of this chapter is to investigate the performance of several missing data handling methods for the target dose selection in a recurrent event data study. To allow a direct quantitative comparison of the methods presented in Section 7.3 under the same conditions and using the same performance metrics, a simulation study, motivated by the setting of the example in Section 7.2, was undertaken.

As the data for the motivating study have not yet been collected, the investigations will be performed from a study design point of view. This investigation is necessary for deciding on the primary analysis techniques and the justification of the chosen sample size. In clinical research, these have to be specified in study protocols prior to the arrival of the data. Furthermore, choosing the adequate analysis option enables a prompt analysis of the data upon availability.

In Section 8.1, we present the design of the simulation study based on the case study introduced in Section 7.2. Section 8.2 presents the results of the simulation study for a large sample size and Section 8.3 investigates whether these results also hold for small sample sizes. Finally, concluding remarks are given in Section 8.4.

8.1 Simulation Assumptions

In this simulation study, we will explore various scenarios, using different:

- rate functions: constant and (decreasing) Weibull;
- dose-response profiles: linear and log-linear;
- missingness mechanisms (see page 175);
- dropout-rates: 20% – 50%;
- overdispersion coefficients: dependent on covariates $1 + \phi \Lambda_{x_i}(T, \vartheta) \in [1.1, 1.4]$, which corresponds to the range of $10\% - 40\%$ overdispersion.

We assume that five different doses $\{0.25, 0.5, 1, 2, 3\}$ of the new drug and one comparator are used in the trial. We investigate the impact of dropout on the target dose estimation and the most robust analysis method when missingness occurs.

Analyses based on a sample size of 350 subjects revealed a bias of 5% when analysing the complete simulated data sets, i.e. the data sets before introducing missingness. In order to reduce the bias due to finite sampling, a large sample size of 3500 subjects was chosen first to compare the different analysis methods, see Section 8.2. Subsequently in Section 8.3, a sample size of 350 was used to confirm the results for a more realistic setting.

Each of the doses of the new drug is assigned to one-seventh of the subjects (500 or 50, respectively) and the comparator to two-sevenths (1000 or 100, respectively). The study period is 112 days. We simulate the data using the target dose $\eta = 2$ and patients are assumed to suffer from three events on average. All the aforementioned assumptions for the simulation study were chosen in consultation with clinicians which were involved in planning the clinical study presented in Section 7.2.

In order to simulate recurrent event data based on the model given in equation (7.7), Section 7.4.2, we perform the following steps for each scenario, i.e. choice of explanatory variables, rate function etc.:

Step 1: For each subject $i \in \{1, ..., m\}$ we draw the random effect U_i and calculate the intensity function $\lambda_{x_i}(t, \theta | U_i = u)$, see equation (7.2).

Step 2: In Section 7.4.2, we assume that the event counts, given U_i, occur according to a Poisson process.

For a constant intensity function we are then dealing with a homogeneous Poisson process. Hence, the inter-arrival times are exponentially distributed with parameter $\lambda_{x_i}(t, \theta | U_i = u) = u\ \delta\, h(x, \beta)$. Drawing from this distribution is then straightforward using statistical software such as SAS.

In the case of the Weibull intensity function, the cumulative distribution function for the waiting time $E = w$ to the next event, given an event occurred at time t, has to be calculated directly. Let $N(t)$ denote the number of events occurring in the time interval $[0, t]$. Then,

$$\begin{aligned} F_E(w) &= \mathbb{P}(E \leq w) \\ &= \mathbb{P}(N(t+w) - N(t) \geq 1) \\ &= 1 - \mathbb{P}(N(t+w) - N(t) < 1) \\ &= 1 - \mathbb{P}(N(t+w) - N(t) = 0). \end{aligned}$$

Now, we know that $N(t + w) - N(t)$ is Poisson distributed with mean

$$\begin{aligned} \Lambda_x(t + w, \theta | U) - \Lambda_x(t, \theta | U) &= U\, h(x, \beta) \int_0^w \lambda_0(t + v, \theta)\ \mathrm{d}v \\ &= U\, h(x, \beta)\, \delta_0 \left[(t + w)^{\delta_1} - t^{\delta_1} \right]. \end{aligned}$$

Thus,

$$F_E(w) = 1 - \exp\left(-U\, h(x, \beta)\, \delta_0 \left[(t + w)^{\delta_1} - t^{\delta_1} \right] \right).$$

We can now simulate from this distribution by using the inverse CDF method. That is,

we sample from the uniform distribution on the interval [0, 1], i.e. $v \sim Uniform(0, 1)$, and find w, such that $F_E(w) = v$. Solving this equation yields:

$$w = \left(\frac{\ln(1-v)}{-U\, h(x,\beta)\, \delta_0} + t^{\delta_1} \right)^{\frac{1}{\delta_1}} - t.$$

Depending on the used intensity function, we draw the waiting times according to corresponding distribution and count the number of events by the end of the study, i.e. 112 days.

For each scenario 1000 data sets are created.

Step 3: Subsequently, we introduce missingness by applying the following missing data mechanisms:

$$\underline{\textbf{MAR}}: \qquad \text{Subject drops out after } t_j \quad \sim \quad Bernoulli\left[p(t_j)\right], \tag{8.1}$$

$$\text{where} \quad logit\left[p(t_j)\right] = \gamma_{0,s} + \gamma_{1,s}\, \ln(age) + \gamma_{2,s}\, s + \gamma_{3,s}\, \frac{N(t_j)}{t_j}, \quad \text{and}$$

$\gamma_{k,s} < 0$ for $k \in \{0, 1\}$ and $\gamma_{k,s} > 0$ for all $k \in \{2, 3\}$. That is, the probability of dropping out decreases with age but increases with the assigned dose (due to adverse events) and number of events.

$$\underline{\textbf{MNAR}}: \qquad \text{Subject drops out after } t_j \quad \sim \quad Bernoulli\left[p(t_j)\right], \tag{8.2}$$

$$\text{where} \quad logit\left[p(t_j)\right] = \gamma_{0,s} + \gamma_{1,s}\, \ln(age) + \gamma_{2,s}\, s + \gamma_{3,s}\, \frac{N(t_j)}{t_j} + \gamma_{4,s}\, \frac{N(T) - N(t_j)}{T - t_j}, \quad \text{and}$$

$\gamma_{k,s} < 0$ for $k \in \{0, 1\}$ and $\gamma_{k,s} > 0$ for all $k \in \{2, 3, 4\}$. Here, additionally, the probability of dropping out increases with the potentially unobserved number of events by the end of the trial.

Moreover, we distinguish two cases. Firstly, for $k \in \{0, 1, 3, 4\}$

$$\gamma_{k,s} = \gamma_k \quad \text{for } s \in \{0.25, ..., 3, C\} \quad \text{and} \quad \gamma_{2,s}\, s = \begin{cases} \gamma_2\, s & \text{for } s \in \{0.25, ..., 3\}; \\ \gamma_2\, \eta & \text{for } s = C; \end{cases}$$

i.e. the missingness processes for the comparator and the target dose group are identical. Secondly, for $k \in \{0, 1, 3, 4\}$

$$\gamma_{k,s} = \begin{cases} \gamma_k & \text{for } s \in \{0.25, ..., 3\}; \\ \gamma_k^* \neq \gamma_{k,\eta} & \text{for } s = C; \end{cases} \quad \text{and}$$

$$\gamma_{2,s}\, s = \begin{cases} \gamma_2\, s & \text{for } s \in \{0.25, ..., 3\}; \\ \gamma_2^* \neq \gamma_2\, \eta & \text{for } s = C; \end{cases}$$

i.e. the missingness processes for the comparator and the target dose group differ.

Based on the simulated data sets we present the results of the simulation study in the next section. As the results under a constant or decreasing Weibull rate are identical from a qualitative point of view, we will focus on the results for the constant rate. For this case a total of 12 scenarios were investigated. The different compositions are given in Table 8.1.

8.2 Results using a Large Sample Size

The results of the simulation study, using the mean bias of the estimated target dose as our performance metric, are given in Table 8.2 and Table 8.5. The mean (mean(η)) and median (median(η)) of the estimated target doses, the standard errors ($\hat{\sigma}_\eta$) and the 90% range for the estimates ($\hat{\gamma}$) are presented. The pattern-mixture model was fitted using the CCMV identifying restrictions and 5 imputations, see Section 7.4.3.

For the first eight scenarios, see Table 8.1, a homogeneous Poisson process with a linear dose-response relationship was assumed. Different dropout rates are considered. The

Simulation scenarios assuming a constant rate function

	Rate		Missingness		Contrast		Dropout-Rate
Scenario	linear	log-linear	MAR	MNAR	same miss.	diff. miss.	
1	X		X		X		22.10%
2	X			X	X		21.69%
3	X		X		X		46.06%
4	X			X	X		45.51%
5	X		X			X	25.88/36.71/21.55%
6	X			X		X	26.27/35.13/22.72%
7	X		X			X	54.18/71.04/47.43%
8	X			X		X	56.53/70.74/50.84%
9		X	X		X		22.86%
10		X		X	X		22.49%
11		X	X			X	26.50/36.69/22.41%
12		X		X		X	27.04/35.09/23.82%

Table 8.1: Overview of the different scenarios for the simulation study using a constant rate function. In case of different missingness rates for the new drug and the comparator, the column 'Dropout-Rate' consists of the dropout rates for the complete data set/ only for the comparator/ only for the experimental drug.

first four scenarios, which assume the same dropout rate and process for the target dose group and the comparator, lead to estimates which are very close to the true value $\eta = 2$. The estimates obtained by analyzing the complete data sets (CD) are the most accurate ones. However, in the case of MAR, it is remarkable how close these are to the direct likelihood (DL) estimates (see Scenario 1 and Scenario 3). Due to the reduced information, standard errors are slightly higher than those obtained from the CD analysis. Even in the case of MNAR, where ignoring the missingness process is usually not valid, we obtain acceptable estimates, including for Scenario 4, where almost half of the endpoints are missing. Also the estimates obtained by the complete case analysis (CC), the last count carried forward technique (LCCF) and the pattern-mixture models (PMM) are within a 2.5% range of the true value. The target dose is usually slightly underestimated. The standard errors are marginally higher than those of the complete data. The mean of the estimates under the last observed rate carried forward (LORCF) approach is highly biased due to outliers. We observe two reasons for outlying estimates. Firstly, imputing the missing endpoint through the rounded value of $n_i(t_{i,d}) + (T - t_{i,d}) \frac{n_i(t_{i,d})}{t_{i,d}}$ can yield very large numbers of events and will affect the estimation of the target dose. Secondly, outliers occur due to numerical issues:

Simulation results for constant rate and linear dose-response relation.
True value of the target dose used in the simulation study is $\eta = 2$.

Scenario	Parameter	CD	CC	LORCF	LCCF	DL	PMM
1	mean(η)	2.002	1.985	-84.60	1.997	2.003	1.997
1	median(η)	1.998	1.983	1.968	1.995	2.000	1.994
1	$\hat{\sigma}_\eta$	0.07	0.07	2382.1	0.07	0.08	0.07
1	$\hat{\gamma}$	[1.89,2.12]	[1.86,2.12]	[-11.78,6.41]	[1.88,2.13]	[1.88,2.13]	[1.88,2.12]
2	mean(η)	2.002	1.987	-84.45	1.997	2.003	1.997
2	median(η)	1.998	1.986	1.966	1.999	2.002	1.996
2	$\hat{\sigma}_\eta$	0.07	0.08	2417.8	0.08	0.08	0.08
2	$\hat{\gamma}$	[1.89,2.12]	[1.86,2.12]	[-11.75,7.18]	[1.87,2.12]	[1.88,2.13]	[1.87,2.13]
3	mean(η)	1.999	1.981	-881.9	1.973	1.999	2.006
3	median(η)	1.999	1.978	1.967	1.977	1.999	2.005
3	$\hat{\sigma}_\eta$	0.07	0.09	27576	0.09	0.08	0.09
3	$\hat{\gamma}$	[1.89,2.10]	[1.84,2.13]	[-3.44, 6.56]	[1.83,2.11]	[1.88,2.19]	[1.87,2.15]
4	mean(η)	1.999	1.972	-989.9	1.921	1.992	1.992
4	median(η)	1.999	1.970	1.959	1.919	1.991	1.994
4	$\hat{\sigma}_\eta$	0.07	0.09	30962	0.10	0.08	0.11
4	$\hat{\gamma}$	[1.89,2.10]	[1.82,2.12]	[-3.79, 6.38]	[1.76,2.09]	[1.87,2.12]	[1.83,2.15]
5	mean(η)	2.000	2.403	-0.982	2.494	2.000	2.404
5	median(η)	1.999	2.403	1.720	2.492	1.999	2.403
5	$\hat{\sigma}_\eta$	0.07	0.10	72.69	0.10	0.08	0.09
5	$\hat{\gamma}$	[1.88,2.12]	[2.25,2.57]	[-2.77,5.90]	[2.35,2.65]	[1.87,2.14]	[2.26,2.56]
6	mean(η)	2.000	2.473	-4.915	2.626	2.083	2.481
6	median(η)	1.999	2.474	1.496	2.623	2.081	2.478
6	$\hat{\sigma}_\eta$	0.07	0.10	144.51	0.10	0.08	0.10
6	$\hat{\gamma}$	[1.88,2.12]	[2.32,2.63]	[-4.85,5.38]	[2.47,2.78]	[1.96,2.22]	[2.32,2.64]
7	mean(η)	2.002	3.279	-561.21	3.082	2.000	3.237
7	median(η)	2.001	3.274	1.821	3.079	2.001	3.235
7	$\hat{\sigma}_\eta$	0.07	0.15	15117	0.11	0.09	0.13
7	$\hat{\gamma}$	[1.89,2.11]	[3.03,3.54]	[-10.56,6.09]	[2.91,3.26]	[1.85,2.14]	[3.03,3.46]
8	mean(η)	2.002	3.591	-614.04	3.602	2.085	3.586
8	median(η)	2.001	3.586	1.758	3.604	2.08	3.584
8	$\hat{\sigma}_\eta$	0.07	0.18	15187	0.20	0.09	0.20
8	$\hat{\gamma}$	[1.89,2.11]	[3.30,3.91]	[-9.92,4.83]	[3.34,3.88]	[1.93,2.24]	[3.30,3.89]

Table 8.2: Simulation results for a constant rate with a linear dose-response relationship. The amount of missingness and the missingness process vary according to Table 8.1. In the column 'CD' (complete data) the data sets were analyzed before introducing missingness. Moreover, 'CC' refers to the complete case analysis, ' LORCF' to last observed rate carried forward, 'LCCF' to the last count carried forward imputation and 'DL' to the direct likelihood approach. The estimate $\hat{\sigma}_\eta$ denotes the standard error of $\hat{\eta}$ and $\hat{\gamma}$ the 90% range of the η-estimates.

the dual quasi-Newton algorithm used to fit the generalized linear model defined in Section 7.4.1 did not converge for every simulated data set. Therefore, we use the more outlier-robust median, in order to compare the performance of this imputation technique to the other approaches. This shows that the LORCF estimates lie within a 2.5% range of the true value.

The results of Scenario 3 and Scenario 4 using the LORCF, LCCF and CC approach suggest that estimating the target dose in the given setting is very robust. But which component of the setting leads to these results? The remaining parameter estimates show some severe biases for LORCF, CC, LCCF and PMM.

Assuming a linear dose-response relation and the model introduced in Section 7.4.1 yields

$$\ln\{\mathbb{E}[N(T)|U]\} = \alpha_0 + \alpha_1 \ln(age) + \alpha_2 (s - \eta) + \ln(U).$$

For fitting purposes we reparametrise $\alpha_0 = \ln(\rho T)$. In comparison, the model for the whole recurrent event data sequence, Section 7.4.2, gives

$$\ln\{\mathbb{E}[N(T)|U]\} = \beta_0 \ln(age) + \beta_1 (s - \eta) + \ln(\delta T) + \ln(U).$$

In particular, both parametrisations imply $\rho = \delta$, $\alpha_1 = \beta_0$, $\alpha_2 = \beta_1$ and the same subject-specific parameter ζ and target dose η in both models. The estimates for these estimates and Scenario 3 and Scenario 4 are given in Table 8.3. The estimates for these parameters are more sensitive to dropout; the estimates for $\alpha_1 = \beta_0$, $\delta = \rho$ and ζ are biased for some methods. In particular, we observe a higher bias under the assumption of MNAR. The parameter $\alpha_1 = \beta_0$ is severely biased in the case of LORCF, but also for the CC, LCCF and PMM approaches. The estimates for ρ and the target dose using the LORCF method are also biased. This is very likely due to outlying measurements. Moreover, the estimation of the overdispersion parameter ζ seems to be biased in all approaches. A biased estimate for ζ implies that the variance estimate of $N(T)$ is biased, which is in line with the results presented in Gustafson [2001]. However, the estimation of ζ could be affected by the finite

Scenario	Parameter	True	CD		CC		LORCF	
			EST	SE	EST	SE	EST	SE
3	$\beta_0 = \alpha_1$	0.01	0.011	0.02	0.022	0.04	-0.051	0.61
3	$\beta_1 = \alpha_2$	-0.4	-0.401	0.01	-0.445	0.02	-0.451	0.34
3	$\delta = \rho$	0.02	0.020	0.002	0.014	0.002	101.74*	2909*
3	ζ	0.015	0.015	0.01	0.0003	0.002	3.130	0.50
3	η	2	1.999	0.07	1.981	0.09	-881.907*	27577
4	$\beta_0 = \alpha_1$	0.01	0.011	0.02	0.019	0.04	-0.055	0.63
4	$\beta_1 = \alpha_2$	-0.4	-0.401	0.01	-0.396	0.02	-0.462	0.35
4	$\delta = \rho$	0.02	0.020	0.002	0.014	0.003	120.40	3473
4	ζ	0.015	0.015	0.01	$2.94 \cdot 10^{-6}$	$1.9 \cdot 10^{-5}$	3.239	0.51
4	η	2	1.999	0.07	1.972	0.09	-989.9	30962

Scenario	Parameter	True	LCCF		DL		PMM	
			EST	SE	EST	SE	EST	SE
3	$\beta_0 = \alpha_1$	0.01	0.018	0.03	0.012	0.03	0.016	0.04
3	$\beta_1 = \alpha_2$	-0.4	-0.312	0.01	-0.400	0.02	-0.380	0.02
3	$\delta = \rho$	0.02	0.014	0.002	0.020	0.002	0.019	0.01
3	ζ	0.015	0.0005	0.002	0.016	0.01	0.0002	0.001
3	η	2	1.974	0.09	1.999	0.08	2.006	0.09
4	$\beta_0 = \alpha_1$	0.01	0.015	0.04	0.011	0.03	0.013	0.05
4	$\beta_1 = \alpha_2$	-0.4	-0.263	0.01	-0.379	0.01	-0.335	0.02
4	$\delta = \rho$	0.02	0.014	0.003	0.019	0.003	0.019	0.02
4	ζ	0.015	$2.74 \cdot 10^{-6}$	$2.1 \cdot 10^{-5}$	$7.98 \cdot 10^{-6}$	$6.1 \cdot 10^{-5}$	$5.50 \cdot 10^{-6}$	$2.1 \cdot 10^{-4}$
4	η	2	1.921	0.10	1.992	0.08	1.992	0.11

Table 8.3: Mean estimates and standard errors for the parameters involved in the models of Scenario 3 and Scenario 4, given in Table 8.1.

sample size and the small value we chose for the simulation study ($\zeta = 1/65 \approx 0.0154$). To rule out the influence of ζ on the results obtained, simulations were conducted with the value $\zeta = 1$. The results, given in Table 8.4, suggest similar findings to those presented in Table 8.2, but that the bias of the target dose estimation increases with the overdispersion parameter.

Note that all approaches except the direct likelihood and the LORCF method underestimate the parameter ζ. This is due to the dropout probabilities, which depend on explanatory variables and the number of events at a certain point in time, thus reducing the heterogeneity for the complete cases or imputed data sets. In contrast, the overdispersion parameter ζ is highly overestimated in the case of LORCF which is related to the outlying imputations mentioned earlier. Only the results yielded by the direct likelihood approach are very accurate.

Parameter	CD	CC	LORCF	LCCF	DL	PMM
same missingness: 21.14%						
mean(η)	1.992	1.971	-4.399	1.984	1.999	1.980
median(η)	1.994	1.964	1.821	1.982	2.001	1.980
$\hat{\sigma}_\eta$	0.12	0.14	72.61	0.13	0.14	0.13
$\hat{\gamma}$	[1.79,2.19]	[1.75,2.20]	[-10.17,5.30]	[1.78,2.21]	[1.78,2.22]	[1.77,2.20]
different missingness: 24.24% (total) - 32.44% ($s = C$) - 20.95% ($s \in \{0.25, 0.5, 1, 2, 3\}$)						
mean(η)	1.992	3.092	-28.90	2.911	2.004	3.098
median(η)	1.987	3.085	1.682	2.906	2.004	3.090
$\hat{\sigma}_\eta$	0.12	0.20	542.90	0.17	0.15	0.19
$\hat{\gamma}$	[1.79,2.19]	[2.78,3.45]	[-8.93,4.94]	[2.64,3.21]	[1.75,2.26]	[2.80,3.43]

Table 8.4: Mean and median estimates for the target dose, the corresponding standard error and the 90% range in case of MAR and $\zeta = 1$.

Overall the direct likelihood estimates for Scenario 3, that is MAR, are satisfactory and near the true values. In the case of MNAR a higher bias is observable.

These results lead to the hypothesis that the estimation of the target dose is very robust, because the same missingness process was applied for patients receiving the comparator and the target dose. The responses for both treatment groups are 'equally biased' and will therefore still yield the correct estimate for the target dose.

In order to investigate this hypothesis, data sets according to Scenarios 5 - 8 in Table 8.1 were simulated and analyzed. The dropout mechanisms for the comparator and the different doses of the new drug have the same functional form but different parameters, see equations (8.1) and (8.2), which lead to a higher dropout rate in the comparator group. Comparing the resulting target dose estimates for these scenarios, a remarkable bias for the CC, LCCF and PMM approaches is observed. In the case of Scenario 5 and Scenario 6, where we have an overall dropout rate of about 26%, these approaches overestimate η by approximately 25%. With 50% missingness, this bias increases to approximately 75%.

The overestimation of the target dose in the complete case analysis is due to the nature of the simulated missingness processes. Patients who suffer from many events are more likely to drop out than those with only a few events. As the dropout rate in the comparator group is higher than for the other dose groups, it appears that the CC analysis

discards relatively more patients assigned to the comparator group. The remaining patients in the comparator group present a subsample with relatively fewer events, thus leading to an overestimation of the target dose. Now, as we used the CCMV restrictions for the pattern-mixture models, this bias is also incorporated in the PMM estimates.

In the case of LCCF the last observed count is carried forward. Therefore, the number of events is underestimated for all non-completers and the target dose overestimated.

In contrast, the target dose is underestimated in the case of LORCF by at most 25%. The substantial bias is slightly surprising, because the data were simulated via a homogeneous Poisson process and the LORCF approach imputes the missing values based on this assumption.

Note that the true target dose η is not included in the 90% range of estimates for either CC, LCCF or PMM and that the bias increases with ζ in the case of different missingness processes too, see Table 8.4.

Let us now turn to Scenarios 9-12 and their results, which are shown in Table 8.1 and Table 8.5, respectively.

Again, the data were simulated according to a homogeneous Poisson process but with a log-linear dose-response relation. The results are very similar in spirit to those discussed previously. A negligible bias can be observed for Scenarios 9 and 10, where the same missingness process is assumed for the comparator and the target dose group. For Scenario 11 and 12 the bias is up to 55% for 20% missingness, compared to a maximum of 30% in Table 8.2. This bias increases to up to 600% in case of 50% missingness (results omitted). The DL approach performs very well in the MAR case but leads to a bias of around 10% in the case of MNAR. This bias is nevertheless considerably smaller than that of the remaining techniques.

These results show that the impact of missingness on the estimated target dose does not only depend on the assumed missingness process, the dropout rate and the inter-individual variation, but also on the assumed dose-response relationship.

Simulation results for constant rate and log-linear dose-response relation

Scenario	Parameter	CD	CC	LORCF	LCCF	DL	PMM
9	mean(η)	2.007	1.971	$7.3 \cdot 10^{10}$	1.989	2.011	1.974
9	median(η)	1.993	1.963	1.868	1.977	2.003	1.966
9	$\hat{\sigma}_\eta$	0.14	0.15	$8.1 \cdot 10^{11}$	0.15	0.16	0.15
9	$\hat{\gamma}$	[1.80,2.26]	[1.75,2.22]	[$1.5 \cdot 10^{-4}$,987.5]	[1.77,2.25]	[1.78,2.28]	[1.76,2.23]
10	mean(η)	2.007	1.970	$6.2 \cdot 10^{10}$	1.984	2.008	1.970
10	median(η)	1.993	1.961	1.839	1.973	1.995	1.963
10	$\hat{\sigma}_\eta$	0.14	0.15	$7.1 \cdot 10^{11}$	0.15	0.15	0.15
10	$\hat{\gamma}$	[1.80,2.26]	[1.75,2.23]	[$1.1 \cdot 10^{-5}$,594.7]	[1.75,2.26]	[1.78,2.28]	[1.74,2.24]
11	mean(η)	2.008	3.012	$4.3 \cdot 10^{10}$	3.372	2.010	2.968
11	median(η)	2.005	3.004	1.605	3.347	2.007	2.953
11	$\hat{\sigma}_\eta$	0.13	0.28	$5.4 \cdot 10^{11}$	0.33	0.16	0.28
11	$\hat{\gamma}$	[1.79,2.24]	[2.58,3.51]	[$9.3 \cdot 10^{-3}$,44.05]	[2.87,3.95]	[1.76,2.29]	[2.54,3.46]
12	mean(η)	2.008	3.278	$4.5 \cdot 10^{10}$	4.135	2.201	3.267
12	median(η)	2.005	3.261	1.299	4.095	1.999	3.239
12	$\hat{\sigma}_\eta$	0.13	0.33	$5.8 \cdot 10^{11}$	0.46	0.18	0.34
12	$\hat{\gamma}$	[1.79,2.24]	[2.78,3.90]	[$6.4 \cdot 10^{-3}$,24.19]	[3.46,4.98]	[1.93,2.51]	[2.76,3.86]

Table 8.5: Simulation results for a constant rate and a log-linear dose-response relationship. The amount of missingness and the missingness process vary according to Table 8.1. The estimate $\hat{\sigma}_\eta$ denotes the standard error of $\hat{\eta}$ and $\hat{\gamma}$ the 90% range of the η-estimates.

8.3 Results using a Realistic Sample Size

In Section 8.2 we chose a relatively large sample size and found that the bias depends on the missingness processes in the comparator and the target dose group. The DL approach was shown to perform best among the investigated methods. Here, we investigate how these results differ for a more realistic sample size. The sample size used is 350, see Section 8.1, and results are shown in Table 8.6.

We observe a bias increase of approximately 25% compared to the results with the larger sample size and larger standard errors. However, from a qualitative point of view the findings remain the same, see Section 8.2.

Parameter	CD	CC	LORCF	LCCF	DL	PMM
Linear Rate						
same missingness: 22.75%						
mean(η)	2.012	1.999	-51.82	2.006	2.013	2.001
median(η)	1.994	1.975	1.590	1.990	1.998	1.977
$\hat{\sigma}_\eta$	0.22	0.25	446.1	0.24	0.24	0.24
$\hat{\gamma}$	[1.69,2.40]	[1.63,2.44]	[-51.52,4.70]	[1.66,2.40]	[1.66,2.40]	[1.64,2.43]
different missingness: 25.88% (total) - 36.65% ($s = C$) - 21.56% ($s \in \{0.25, 0.5, 1, 2, 3\}$)						
mean(η)	2.006	2.418	-273.6	2.502	2.004	2.420
median(η)	1.994	2.396	1.528	2.477	1.987	2.399
$\hat{\sigma}_\eta$	0.22	0.29	6412	0.29	0.24	0.29
$\hat{\gamma}$	[1.67,2.39]	[1.99,2.92]	[-9.24,4.09]	[2.08,3.00]	[1.61,2.42]	[1.99,2.91]
Log-Linear Rate						
same missingness: 22.86%						
mean(η)	2.096	2.051	$1.7{\cdot}10^{10}$	2.071	2.111	2.057
median(η)	2.027	1.980	1.410	1.998	2.030	1.985
$\hat{\sigma}_\eta$	0.51	0.53	$2.7{\cdot}10^{11}$	0.53	0.56	0.53
$\hat{\gamma}$	[1.40,3.01]	[1.34,3.03]	[$4{\cdot}10^{-18}$,90.27]	[1.34,3.12]	[1.39,3.24]	[1.35,3.06]
different missingness: 26.47% (total) - 36.76% ($s = C$) - 22.34% ($s \in \{0.25, 0.5, 1, 2, 3\}$)						
mean(η)	2.058	3.184	$5.2{\cdot}10^{9}$	3.571	2.056	3.193
median(η)	1.977	2.977	1.0752	3.308	1.946	2.981
$\hat{\sigma}_\eta$	0.47	1.07	$9.4{\cdot}10^{9}$	1.26	0.56	1.09
$\hat{\gamma}$	[1.43,2.87]	[1.91,5.19]	[$4{\cdot}10^{-14}$,43.1]	[2.12,5.81]	[1.35,3.03]	[1.92,5.21]

Table 8.6: Mean and median estimates for the target dose, the corresponding standard error and the 90% range in case of MAR and a sample size of 350 subjects.

8.4 Summary

The performed simulation study suggests that the estimation of the target dose is very robust if the same dropout mechanisms apply for the comparator and the target dose group. However, care should be taken because other model parameters can be substantially biased. This holds in particular for the parameters which estimate subject-specific characteristics.

Further studies showed that this robustness is lost as soon as the dropout processes differ for the comparator and the target dose group. A dropout rate as little as 25% was able to severely bias the target dose when performing the 'traditional' complete case analysis. In the case of a log-linear dose-response and a homogenous Poisson process the target dose was overestimated by more than 50%. The pattern-mixture model performs similarly

badly. This is not surprising as all unavailable information is borrowed from completers. Moreover, both single imputation techniques perform poorly in this case. The bias introduced by the last count carried forward approach is comparable to that of a complete case analysis. The target dose is always overestimated and dependent on the rate function, dose-response relationship, dropout rate and mechanism the bias varies between 25% and 600%. In contrast, the target dose is usually underestimated in the case of the last observed rate carried forward approach. The bias ranges between 25% and 75%. Outliers and numerical problems hinder an adequate dose selection. As expected, the direct likelihood approach performed well, with a bias of maximal 5% in the case of a MAR process and of up to 10% in the case of a MNAR mechanism. These results suggest that the direct likelihood approach provides a certain protection against bias for the MNAR process used.

In light of these results, we remark that the direct likelihood performed best among the missing data handling approaches proposed in Section 7.3. This approach yields the smallest bias, but also small standard errors which justify the sample size of 350 patients in our specific case, see Table 8.6. Hence, we recommend the inclusion of this method in the study protocols. Furthermore, we discourage the use of complete case analysis and single imputation techniques. The poor performance of the pattern-mixture models is simply due to the identifying restrictions used. Ideally, we would like to consider other possibilities to borrow available information.

We observe that the magnitude of bias increases with the overdispersion, the dropout-rate and for decreasing sample sizes. Furthermore, the magnitude depends highly on the missingness process and the modelling framework, i.e. the underlying rate function and the assumed dose-response model.

We acknowledge that there are several limitations to our investigations of the impact of missingness on the dose-selection for recurrent event data studies. Firstly, we define the dropout time for non-completers as the last observed event time. We employ this definition, because using patient diaries it is difficult to observe the precise time of dropout. In fact, the real dropout time lies in the interval $(t_{i,d}, T]$ and more suitable approaches, e.g. by

modelling the time to dropout after t_{id} using survival models, might be preferable. We note that an alternative definition of the dropout times would alter the results of all methods except the complete case analysis and the last count carried forward method.

Secondly, our simulation results are based on using the same simulation and analysis models. In practice, however, this will rarely be the case and we expect that the impact of dropout also depends on the misspecification between 'true' and 'fitted' model.

Thirdly, the scope of our investigations is limited to cases where the same functional form of the missingness process is assumed for the new drug and the comparator drug. Solely the dropout rates differ. The impact of missingness is expected to be more severe if the functional form and the included covariates differ for the two treatment groups.

Finally, the models investigated in the last two chapters assume a specific dose-response relationship (linear or log-linear). In many applications, however, this relation is not known and needs to be specified, see Bretz et al. [2005]. The selection of a suitable dose-response relation itself is likely to be sensitive to dropout.

Chapter 9

Conclusions and Future Directions

9.1 Summary and Conclusions

This thesis is concerned with the analysis of repeated measurements studies and issues arising when data are incomplete.

The first part is motivated by the CAST study which observes bounded and continuous longitudinal data. The aim of this study is to investigate the change over time in the outcome score and factors that influence this change.

In Chapter 2, time plots reveal that the outcome of interest evolves non-linearly over time. Also, the scores approach an upper limit as time increases. To fit a linear-mixed model in Chapter 3, different transformations are used in an attempt to reduce the effect of the boundedness and to improve the linearity of the data with respect to covariates. Three of these transformations transform the outcome score, whereas the remaining two transformations are applied to the observation times. The fitted models based on the different transformations of the outcome score lead to a poor model fit. In contrast, the results based on the transformations of the observation times lead to adequate model fits. The inference, however, is shown to be very sensitive to the chosen transformation. We come to the conclusion that an analysis based on a transformation of the observation times may not answer the original research question. In addition, models based on transformations cannot investi-

gate the dependence of bounds on covariates, as the bounds need to be specified prior to the transformation. Using transformations can also complicate the interpretation of covariate effects on the original score.

In Chapter 3, we therefore present a valuable alternative to transforming data. A non-linear mixed model for the mean score on the original scale as a function of covariates is proposed. The model is constructed for continuous data where the rate of change is not constant over time. It enables a very flexible incorporation of exploratory variables and is easy to interpret, which is a valuable advantage over the alternative of data transformations. For CAST, it allows us to model the final recovery level, which might be of particular interest to patients. Although the non-linear mixed model is motivated by the CAST data set, it could be applied to other studies, e.g. a longitudinal study using the *Neck Disability Index* [Vernon and Mior, 1991].

For CAST, the non-linear mixed model reveals that recovery was more rapid with BKC than with Tubigrip. The rate of recovery for Aircast brace was only marginally higher than for Tubigrip. There was no significant difference in recovery rates between Bledsoe boot and Tubigrip. Further, we show that older and female patients recovered substantially more slowly than younger and male participants. Also, the score at final recovery for older female patients was lower than for young male patients, suggesting that older female patients were less likely to recover completely from an acute soft tissue injury. We translate these findings into auxiliary information, such as the expected time to achieve a certain score for different patient groups.

The fitted non-linear mixed model in Chapter 3 is limited by the covariance structure it assumes. We discuss alternative structures for the covariance matrix of bounded continuous longitudinal data in Chapter 4. Investigations of the study design and empirical covariance and correlation matrices suggest that a suitable covariance structure should meet three characteristics. Firstly, correlations should be larger when the measurements are closer in time than when they are further apart. Additionally, correlations are expected to increase as measurements reach the bounds regardless of the time interval between mea-

surements. Finally, the variances are rarely constant over time. Finding a covariance model that meets these features proves to be very difficult. Several covariance pattern models and random-effect models are fitted to the data. However, none of the these structures lead to a satisfying fit. Therefore, we adopt the data-driven regression approach introduced by Pourahmadi [1999]. Instead of a polynomial model for the outcome process, we chose the non-linear mixed model presented in Chapter 3. Furthermore, we allow for an ignorable missingness process and we argue that the AIC should be used for model selection purposes. Applying this method to the CAST data recovers an unstructured covariance structure. For CAST, the inference based on a compound symmetry structure and unstructured covariance matrix is the same; the effect sizes and standard errors vary only a little.

After having specified a statistical model for the outcome vector of interest, we turn to the challenge of missing values in repeated measurement studies. In Chapter 5 the statistical framework for the analysis of incomplete data is reviewed and a few popular methods to handle missing data and their underlying assumptions are reviewed. We stress that a distinction between MAR and MNAR missingness processes based on the observed data only is not possible. Moreover, we discuss how a sensitivity analysis, where the stability of the conclusions is investigated under different assumptions, is a valuable approach to analyze incomplete data. A sensitivity analysis for the CAST data set is performed, where the results based on a complete case analysis (CC), the last observation carried forward approach (LOCF), the direct likelihood approach (DL), multiple imputations (MI) and a pattern-mixture model are compared. While the conclusions based on the CC, MI and DL analysis are quite similar, they are substantially different for the LOCF approach.

This sensitivity analysis is continued in Chapter 6, where we account for informative missingness through selection models. The traditional selection model is extended by adjusting for missingness through the number and nature of reminders made to contact initial non-responders. Using this model the impact of missingness on the rate of change is evaluated in a sensitivity analysis. We contrast this model with the traditional selection model, where we adjust for missingness by modelling the missingness process.

Our investigations suggest that using the richer information of the reminder process enables a more accurate choice of covariates, which induce missingness, than modelling the missingness process. This holds under the condition that the sample sizes of all reminder categories are large enough to detect significant effects. A further advantage of modelling the reminder process versus the traditional selection model is the ability to incorporate the dependence across the reminders at the different observation times for a given patient.

Regarding the reminder process, we observe that phone calls are most effective. For the score data, the conclusions that recovery is slower, and less satisfactory with age, and more rapid with BKC than Tubigrip do not alter materially across all models investigated.

Overall we believe that the robustness of the rate of improvement parameters is closely related to the functional form of the non-linear mixed model. The score changes much faster at the beginning of the study than towards the end. At the same time most patients drop out in the last weeks. That is, the non-linear mixed model appears to be able to accurately pick up the change because most patients have available data at the crucial observation times. We expect that conclusions based on transformations and linear-mixed models are more sensitive to missing values. However, this conjecture requires further investigation.

The second part of the thesis is motivated by a dose-finding study which explores a recurrent event data process over a time period of several weeks. Investigators involved in this study are interested in the number of events occurring in a specific time interval. Due to missingness this endpoint is not observed for all patients and the classical approach of investigators is to perform a complete case analysis where all non-completers are discarded from the analysis.

In our study of interest, however, dropout is expected to be outcome related and the underlying assumption justifying a complete case analysis, i.e. MCAR, is violated. In Chapter 7, we discuss alternative techniques to handle missing data: last observed rate carried forward, last count carried forward, direct likelihood approach and pattern-mixture

models. Some of these methods require models for the whole recurrent event process, whereas others enable an endpoint analysis for the number of events occurring by the end of the study. Regression models which enable the target dose selection and the investigation of the effect of other explanatory variables are presented. In this context, the target dose is defined as the dose for which the expected response is equal to that of a competitor group. The reference value for the outcome of the target dose is estimated from the current study, i.e. embedded in the dose-response model.

The recurrent events are modelled as over-dispersed Poisson process data, with dose and age as regressor. Constant and Weibull rate functions are examined.

In order to investigate the impact of dropout on the target dose selection, and to compare the performances of the proposed missing data methods, a scenario evaluation study is performed in Chapter 8. The results of this study suggest that the estimation of the target dose is very robust if the same dropout mechanisms apply for the comparator and the target dose group. However, we observe that other model parameters can be substantially biased. This holds in particular for the overdispersion parameter.

Further studies showed that this robustness is lost as soon as the dropout processes differ for the comparator and the target dose group. A dropout rate as small as 25% leads to a severe bias for the target dose selection when performing the traditional complete case analysis. Of the methods explored, the direct-likelihood approach performs best, even when a MNAR mechanism holds. Hence, we recommend the inclusion of this method in the study protocols. Furthermore, we discourage the use of complete case analysis and single imputation techniques. The fitted pattern-mixture models also perform poorly, which is likely due to the identifying restrictions used.

We observe that the magnitude of bias increases with the overdispersion, the dropout rate and for decreasing sample sizes. Furthermore, the magnitude strongly depends on the missingness process and the modelling framework, i.e. the underlying rate function and the assumed dose-response model.

The work presented answered some of the questions raised by the two studies that motivated our work. Clearly, at least as many questions remain unanswered and thus require future research. Some of these questions concern assumptions which have to be scrutinized. Others concern extensions and generalisation of results, especially for the case of non-monotone missingness patterns. And some others pose questions which currently do not seem to draw attention in the broad research community. We will list some of these areas for future research in the next section.

9.2 Future Work

The following list could serve as an agenda for future work. Most of these aspects have been already mentioned in the summaries of each chapter.

(i) The inference in the first part of the thesis is heavily based on the assumption of normally distributed data. However, the distribution of the data at later time points appears to be skewed, see Figure 2.2. Suitable transformations or the use of another distribution for the components $Y_{i,j}$ need to be further explored.

(ii) The random effects enter the non-linear mixed model (Chapter 3) in a linear fashion. This assumption enables the formulation of an associated full multivariate model and simplifies the computational tasks to a great extent. In the case of random effects that enter the model equation in a non-linear way, we would be confronted with additional integrations in the evaluation of the likelihood functions.

(iii) All random effects in Part I of the thesis are normally distributed, while all random effects in Part II are assumed to be gamma distributed. Alternative distributions could be used.

(iv) In the first part we focus on the symptoms sub-scale of the CAST data. Future work could focus on all five sub-scales and make use of analysis methods for multivariate

longitudinal data. In this context, the correlation between the different scores needs to be accounted for.

(v) All fitted pattern-mixture and selection models are based on the assumption of monotone missingness patterns. In practice, however, we usually observe non-monotone missingness patterns. An extension allowing for non-monotone patterns would be of great value.

We have attempted to fit the selection model presented in Chapter 6 for the non-monotone case. The calculation of the likelihood requires the computation of several hundred integrals for each iteration step. A program written in SAS needed several hours for each iteration step. Therefore, we moved to the Bayesian paradigm using `WinBUGS`. We use the outcome model, reminder process model and missingness process model presented in Chapter 6. The following priors are used

$$
\begin{aligned}
\beta_0 &\sim Uniform\,[0, 100]\,; \\
\beta_1 &\sim Uniform\,[0, 100]\,; \\
\beta_{2,trt_i} &\sim \mathcal{N}\,(0, 1000) \quad \text{for } C \in \{1, 2, 3, 4\}; \\
\psi_{0,k} &\sim \mathcal{N}\,(0, 1000) \quad \text{for } k \in \{1, 3, 4\}; \\
\psi_j &\sim \mathcal{N}\,(0, 1000) \quad \text{for } j \in \{1, 3\}; \\
\psi_j^* &\sim \mathcal{N}\,(0, 1000) \quad \text{for } j \in \{4, 5\}; \\
\phi_j &\sim \mathcal{N}\,(0, 1000) \quad \text{for } j \in \{0, 1, 2, 3, 4\}; \\
\frac{1}{\sigma^2} &\sim \Gamma\,(0.001, 0.001) \quad \text{and} \\
\frac{1}{D^2} &\sim \Gamma\,(0.001, 0.001)\,.
\end{aligned}
$$

Unfortunately, the complicated outcome model and correlated parameters make the model fitting very difficult. This work is currently in progress.

(vi) The sensitivity analysis presented in Chapter 6 is based on numerous assumptions

which cannot be verified from the data. It is not clear to what extent our conclusions would differ under other assumptions; e.g., other covariance structures for the marginal outcome process, use of the probit link-function for the reminder and missingness probabilities or incorporation of explanatory variables such as occupation in the missingness and reminder process models. Furthermore, we include the previous and current score linearly in the reminder and missingness process. Other functional relationships might lead to differing conclusions and thus need further investigation.

(vii) In the last section, we mention that the robustness of the rate of improvement parameters may be closely related to the functional form of the non-linear mixed model. A comparison of results based on a suitable transformation and linear mixed models with our non-linear mixed model could be very interesting.

(viii) In the second part of this thesis, we define the dropout time as the last observed event time. Clearly, the real dropout time lies in the interval $(t_{id}, T]$ and more suitable approaches allowing for a more general definition of the dropout time should be explored.

(ix) The model for the counting process in Chapter 7 is based on a proportional model for the intensity function and the inclusion of a random effect in a multiplicative manner. These assumptions might be too strict in practice and alternatives should be investigated.

(x) The simulation study presented in Chapter 8 focuses on cases where the same functional form of the missingness process is assumed for the new drug and the competitor drug. Only the dropout rates differ. In practice, however, the functional forms for the two treatments might differ and we expect a more severe bias in the target dose selection if this is the case.

(xi) All investigated models in Chapter 7 and 8 assume a specific dose-response relationship. However, this relationship is rarely known and needs to be specified based on

the data. This specification itself is likely to be sensitive to dropout. This aspect needs to be further explored.

(xii) The pattern-mixture model for recurrent-event data presented in the second part of this thesis was fitted by using the CCMV identifying restrictions. In order to be able to borrow information using other identifying restrictions, patients with similar dropout times have to be clustered into one group or pattern. The clustering could be based on the four quartiles of the dropout-times.

(xiii) One aspect that we have not touched upon before concerns the assumed baseline rate functions in our investigations. The constant and Weibull rate functions lead to a monotone rate. With drugs for gout, however, it is commonly observed that the rate at which flares occur increases in the first weeks of the study, but tends to decrease afterwards. One way to model this is given by a Gaussian rate function

$$\lambda_0(t, \delta) = \frac{1}{\delta_1} \exp\left\{-\frac{(t-\delta_0)^2}{2\delta_1^2}\right\} + \delta_2,$$

where $\delta = (\delta_0, \delta_1, \delta_2)^\top$ and $\delta_1, \delta_2 \in \mathbb{R}^+$. We refrained from using this rate function in the presented work because simulating a counting process based on this rate is complicated.

(xiv) Finally, we note that the literature review revealed several relevant research areas which seem to attract very little attention. Some of these are: identifiability issues for non-linear mixed models; model diagnostic tools for non-linear mixed models; sample size calculations for non-linear mixed models; multiple imputation for recurrent event data and sensitivity of inference to data transformations.

Appendix A

Supplementary Material

A.1 Gamma Function

For a complex number z with positive real part, the gamma function $\Gamma(z)$ is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t)\, \mathrm{d}t.$$

A.2 Gamma Distribution

A random variable U is said to be gamma distributed with shape parameter λ and scale parameter α, if its density function is given by

$$f_U(u) = \frac{u^{\lambda-1}}{\Gamma(\lambda)\,\alpha^\lambda} \exp\left(-\frac{u}{\alpha}\right), \; u \geq 0.$$

We write $U \sim \Gamma(\lambda, \alpha)$. It can be shown that $\mathbb{E}(U) = \lambda\,\alpha$ and $Var(U) = \lambda\,\alpha^2$.

A.3 Poisson Process

A *counting process* $\{N(t), t \geq 0\}$ is a stochastic process that represents the number of events that occur up to time $t \in \mathbb{R}_0^+$. We speak of a *Poisson process* with rate function $\lambda(t) > 0$ if

(i) $N(0) = 0$;

(ii) the number of events that occur in disjoint time intervals are independent;

(iii) the counting process has *unit jumps*, i.e. for all t:

$$\begin{aligned} \mathbb{P}\left(N(t+h) - N(t) = 1\right) &= \lambda(t)h + o(h) \quad \text{and} \\ \mathbb{P}\left(N(t+h) - N(t) \geq 2\right) &= o(h), \end{aligned}$$

where $\lim_{h \to 0} \frac{o(h)}{h} = 0$.

Bibliography

M. Akacha and N. Benda. The impact of dropouts on the analysis of dose-finding studies with recurrent event data. *Statistics in Medicine*, 29:1635 – 1646, 2010.

M. Akacha and J. L. Hutton. Analysing the rate of change in a longitudinal study with missing data, taking into account the number of contact attempts. URL http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2010/paper10-08. 2010.

M. Akacha, T. C. O. Fonseca, and S. Liverani. First cladag data mining prize: Data mining for longitudinal data with different marketing campaigns. URL http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2009/paper09-46. 2009.

J. M. Alho. Adjusting for nonresponse bias using logistic regression. *Biometrika*, 77:617–624, 1990.

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics, 1993.

S. Avci and U. Sayli. Comparison of the results of short-term rigid and semi-rigid cast immobilization for the treatment of grade 3 inversion injuries of the ankle. *Injury*, 29:581–584, 1998.

S.G. Baker. Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics*, 51:1042–1052, 1995.

M.A. Becker, R. Schumacher, R.L. Wortmann, P.A. MacDonald, D. Eustace, W.A. Palo, J. Streit, and N. Joeph-Ridge. Febuxostat compared with allopurinol in patients with hyperuricemia and gout. *The New England Journal of Medicine*, 353:2450–2461, 2005.

R.D. Bock. *Multivariate statistical methods in behavioral research*. McGraw-Hill, New York, 1975.

G.C. Borstad, L.R. Bryant, M.P. Abel, D.A. Scroggie, M.D. Harris, and J.A. Alloway. Colchicine for prophylaxis of acute flares when initiating allopurinol for chronic gout arthritis. *The Journal of Rheumatology*, 31:2429–2432, 2004.

F. Bretz, J. C. Pinheiro, and M. Branson. Combining multiple comparisons and modeling techniques on dose-response studies. *Biometrics*, 61:738–748, 2005.

F. Bretz, J. C. Pinheiro, and M. Branson. *Dose finding in drug development*, chapter : Analysis of dose-response studies, pages 146–171. Springer: New York, 2006.

F. Bretz, J. Hsu, J. C. Pinheiro, and Y. Liu. Dose finding - A challenge in statistics. *Biometrical Journal*, 50:480–504, 2008.

S.A. Bridgman, D. Clement, A. Downing, G. Walley, I. Phair, and N. Maffulli. Population based epidemiology of ankle sprains attending accident and emergency units in the West Midlands of England, and a survey of UK practice for severe ankle sprains. *Emergency Medicine Journal*, 20: 508–510, 2003.

E. Budtz-Jœrgensen. Estimation of the benchmark dose by structural equation models. *Biostatistics*, 8:675–688, 2007.

J. Carpenter, S. Pocock, and C.J. Lamm. Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine*, 21:1043–1066, 2002.

K. E. Chiswell. *Model diagnostics for the nonlinear mixed effects model with balanced longitudinal data*. PhD thesis, North Carolina State University. URL `http://repository.lib.ncsu.edu/ir/handle/1840.16/4290`, 2007.

R.J. Cook and J.F. Lawless. Analysis of repeated events. *Statistical Methods in Medical Research*, 11:141–166, 2002.

R.J. Cook and J.F. Lawless. *The statistical analysis of recurrent events*. Springer, 2007.

M. W. Cooke, J. L. Marsh, M Clark, R. A. Nakash, R. M. Jarvis, J. L. Hutton, A Szczepura, S Wilson, and S. E. Lamb. Treatment of severe ankle sprain: A pragmatic randomised controlled trial comparing the clinical effectiveness and cost-effectiveness of three types of mechanical ankle support with tubular bandage. The CAST trial. *Health Technology Assessment*, 13: No.13, 2009.

M.J. Daniels and J.W. Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman & Hall/CRC, 2008.

M. Davidian and D.M. Giltinan. Some general estimation methods for non-linear mixed models. *Journal of Biopharmaceutical Statistics*, 3:23–55, 1993.

M. Davidian and D.M. Giltinan. *Nonlinear models for repeated measurement data*. Chapman & Hall, 1995.

M. Davidian and D.M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:387–419, 2003.

A.P. Dempster and D.B. Rubin. *Incomplete data in sample surveys (Volume 2): Theory and Bibliography*, chapter : Introduction, pages 3–10. New York: Academic Press, 1983.

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

H. Dette, F. Bretz, A. Pepelyshev, and J. C. Pinheiro. Optimal design for dose-finding studies. *Journal of the American Statistical Association*, 103:1225–1237, 2008.

P. J. Diggle. An approach to the analysis of repeated measurements. *Biometrics*, 44:959–971, 1988.

P. J. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49–93, 1994.

P.J. Diggle, K.-Y. Liang, and S.L. Zeger. *Analysis of longitudinal data*. Oxford University Press, 1996.

P.J. Diggle, D. Farewell, and R. Henderson. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Journal of the Royal Statistical Society. Series C*, 56:1–31, 2007.

EMEA. *Points to consider on missing data. Doc. Ref. EMEA/CPMP/EWP/1776/99, Committee for Medicinal Products in Human Use (CHMP).* London, 15 November 2001.

EMEA. *Guidline on missing data in confirmatory clinical trials. Doc. Ref. EMEA/CPMP/EWP/1776/99 Rev.1 Corr, Committee for Medicinal Products for Human Use (CHMP).* London, 23 April 2009.

T.G. Filloon. Estimating the minimum therapeutically effective dose of a compound via regression modelling and percentile estimation. *Statistics in Medicine*, 14:925–932, 1995.

G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied longitudinal analysis*. Wiley Interscience, 2004.

J.A. Forkman. A method for designing nonlinear univariate calibration. *Technometrics*, 50:479–486, 2008.

R.D. Gibbons, D. Hedeker, and S. DuToit. Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, 6:79–107, 2010.

J. W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

P. Gustafson. On measuring sensitivity to parametric model misspecification. *Journal of the Royal Statistical Society. Series B*, 63:81–94, 2001.

J.J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492, 1976.

J.J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

D. Hedeker and R.D. Gibbons. *Longitudinal data analysis*. Wiley Interscience, 2006.

J.W. Hogan and N.M. Laird. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, 16:239–258, 1997.

N.J. Horton and S.R. Lipsitz. Multiple imputation in practice: Comparison of software packages for regression models with missing data. *The American Statistician*, 55 (3):244–254, 2001.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

J.C. Hsu and R.L. Berger. Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association*, 94:468–482, 1999.

C.M. Hurvich, , and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

ICH-E9. *ICH Topic E9. Statistical principles of clinical trials,*, 1998. URL `http://www.emea.eu.int`.

D. Jackson, I. R. White, and L. Morven. How much can we learn about missing data?: An exploration of a clinical trial in psychiatry. *Journal of the Royal Statistical Society: Series A*, 173: 593–612, 2010.

J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data*. Wiley Series in Probability and Mathematical Statistics, 1980.

J. Karlsson and L. Peterson. Evaluation of ankle joint function: The use of a scoring scale. *The Foot*, 1:15–19, 1991.

M. G. Kenward. Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, 17:2723–2732, 1998.

M. G. Kenward, G. Molenberghs, and H. Thijs. Pattern-mixture models with proper time dependence. *Biometrika*, 90:53–71, 2003.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometris*, 38:963–974, 1982.

S. E. Lamb, J. L. Marsh, J. L. Hutton, R. A. Nakash, and M. W. Cooke. Mechanical supports for acute, severe ankle sprains: A pragmatic, multi-centre, randomised controlled trial. *Lancet*, 373: 575–581, 2009.

S.E. Lamb, R.A. Nakash, E.J. Withers, M.Clark, J.L. Marsh, S.Wilson, J.L. Hutton, A. Szczepura, J.R. Dale, and M.W. Cooke; Collaborative Ankle Support Trial research team. Clinical and cost

effectiveness of mechanical support for severe ankle sprains: Design of a randomised controlled trial in the emergency department. *BMC Musculoskeletal Disorders*, 6:1471–2474, 2005.

T. Lancaster and O. Intrator. Panel data with survival: Hospitalization of HIV-positive patients. *Journal of the American Statistical Association*, 93:46–53, 1998.

J.F. Lawless. Regression methods for Poisson process data. *Journal of the American Statistical Association- Theory and Methods*, 82:808–815, 1987.

J.F. Lawless. The analysis of recurrent events for multiple subjects. *Journal of the Royal Statistical Society, Applied Statistics*, 44:487–498, 1995.

K.-Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73-1:13–22, 1986.

D.Y. Lin, J. Wei, and Z. Ying. Semiparametric transformation models for point processes. *Journal of the American Statistical Association- Theory and Methods*, 96:620–628, 2001.

F. Linde, I. Hvass, U. Juergensen, and F. Madsen. Early mobilizing treatment in lateral ankle sprains. Course and risk factors for chronic painful or function-limiting ankle. *Scandinavian Journal of Rehabilitation Medicine*, 18:17–21, 1986.

J.K. Lindsey. *Modelling frequency and count data*. Oxford University Press, 1995.

J.K. Lindsey. *Models for repeated measurements*. Oxford University Press, 1999.

R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.

R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

R.J.A. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471–483, 1994.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley Interscience, 2002.

G. MacKenzie and J. Pan. Optimal joint mean-covariance modelling. URL `www.staff.ul.ie/mackenzieg/{R}esearch_/{S}tatistical_/{C}opy_of_gil_pan_torino.pdf`.

F. Mahony and D. Barthel. Functional evaluation: The Barthel index. *Maryland State Medical Journal*, 14:56–61, 1965.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

M. Miloslavsky, S. Keles, M.J. Van Der Laan, and S. Butler. Recurrent event analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society. Series B*, 66:239–257, 2004.

G. Molenberghs and M. G. Kenward. *Missing data in clinical studies*. Wiley, 2007.

G. Molenberghs and G. Verbeke. *Models for discrete longitudinal data*. Springer, 2005.

G. Molenberghs, Michiels, M. G. Kenward, and P. J. Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52:153–161, 1998.

G. Molenberghs, H. Thijs, I. Jansen, and C. Beuckens. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5:445–464, 2004.

G. Molenberghs, C. Beunckens, C. Sotto, and M.G. Kenward. Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society. Series B*, 70:371–388, 2008.

K.H. Morales, J.G. Ibrahim, C.J. Chen, and L.M. Ryan. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association*, 101:9–17, 2006.

R. A. Nakash. *A study of response and non-response to postal questionnaire follow-up in clinical trials*. PhD thesis, University of Warwick, Warwick Medical School, 2007. URL `http://wrap.warwick.ac.uk/2407/1/{WRAP}_{THESIS}_{N}akash_2007.pdf`.

R.A. Nakash, J.L. Hutton, S.E. Lamb, S. Gates, and J. Fisher. Response and non-response to postal questionnaire follow-up in a clinical trial - a qualitative study of the patient's perspective. *Journal of Evaluation in Clinical Practice*, 14:226–235, 2008.

C. Osborne. Statistical calibration: A review. *International Statistical Review*, 59:309–336, 1991.

J. Pan and G. MacKenzie. On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90:239–244, 2003.

J. Pan and G. MacKenzie. Regression models for covariance structures in longitudinal data. *Statistical Modelling*, 6:43–57, 2006.

T. Park and S.-Y. Lee. Model diagnostic plots for repeated measures data. *Biometrical Journal*, 4: 441–452, 2004.

J. C. Pinheiro and D.M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.

J. C. Pinheiro, B. Bornkamp, and F. Bretz. Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics*, 16: 639–656, 2006.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parametrisation. *Biometrika*, 86:677–690, 1999.

J.M. Robins and A. Rotnitzky. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82:805–820, 1995.

E. Roos, S. Brandsson, and J. Karlsson. Validation of the foot and ankle outcome score for ankle ligament reconstruction. *Foot & Ankle International*, 22(10):788–794, 2001.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

D.B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

SAS/STAT. *Users Guide, Version 8. Cary, NC: SAS Institute Inc.*, 1999.

J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, 1997.

J.L. Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8:3–15, 1999.

J.L. Schafer and J. W. Graham. Missing data: Our view of the state of art. *Psychological Methods*, 7:147–177, 2002.

G.R. Schapp, G. de Keizer, and K. Marti. Inversion trauma of the ankle. *Archives of Orthopaedic and Trauma Surgery*, 108:273–275, 1989.

A.C. Tamhane and B.R. Logan. Multiple test procedures for identifying the minimum effective and safe dose of drugs. *Journal of the American Statistical Association*, 97:293–301, 2002.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550, 1987.

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

A.B. Troxel, D.P. Harrington, and S.R. Lipsitz. Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47:425–438, 1998.

J.W. Tukey, J.L Ciminera, and J.F. Heyse. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics*, 41:295–301, 1985.

H. Vernon and S. Mior. The Neck Disability Index: A study of reliability and validity. *Journal of Manipulative and Physiological Therapeutics*, 14:409–415, 1991.

E.F. Vonesh and V.M. Chinchilli. *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker, Inc., 1997.

E.F. Vonesh, V.M. Chinchilli, and P. Kewei. Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, 52:572–587, 1996.

M.-C. Wang, J. Qin, and C. Chin-Tsang. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96:1057–1065, 2001.

A. M. Wood, I. R. White, and M. Hotopf. Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A*, 169:525–542, 2006.

K.H. Yang and C. Yue-Chen. A note on model diagnostics in longitudinal data analysis. *Computational Statistics*, 21:571–587, 2006.