Extending mixed effects models for longitudinal data before and after treatment

Oliver Thomas Stirrup

A thesis submitted to University College London for the degree of
DOCTOR OF PHILOSOPHY

MRC Clinical Trials Unit
Institute of Clinical Trials and Methodology

UNIVERSITY COLLEGE LONDON
December 2016

I, Oliver Thomas Stirrup, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

For the analysis of longitudinal biomedical data in which the timing of observations in each patient is irregular and in which there is substantial loss to follow-up, it is important that statistical models adequately describe both the patterns of variation within the data and any relationships between the variable of interest and time, clinical characteristics and response to treatment. We develop novel statistical models motivated by the analysis of pre- and post-treatment CD4 cell counts from HIV-infected patients, using the UK Register of Seroconverters and CASCADE datasets.

The addition of stochastic process components, specifically Brownian motion, to standard linear mixed effects models has previously been shown to improve model fit for pre-treatment CD4 cell counts. We review and further develop computational techniques for such models, and also propose the use of a more general 'fractional Brownian motion' process in this setting. Residual diagnostic plots for such models, based on a marginal multivariate normal distribution, show very heavy tails, and we address this issue by further extending the model to allow between-patient differences in variability over time.

It is known from the literature that response to treatment in HIV-patients is dependent on their baseline CD4 level at initiation. In order to further investigate the factors that determine the characteristics of recovery in CD4 counts, we develop a framework for the combined modelling of pre- and post-treatment CD4 cell counts in which key features of the response to treatment for each patient are dependent on a latent variable representing the unobserved 'true' baseline value, conditioned on all pre-treatment data for each patient. We further develop the model structure to account for uncertainty in the exact time of seroconversion for each patient, by integration of the log-likelihood function over all possible dates.

Acknowledgements

I am grateful to my supervisors, Andrew Copas and Ab Babiker, for their support and guidance, and also to James Carpenter for his advice in the early stages of the project. I would like to thank the Medical Research Council for making this project possible by providing me with a PhD Studentship.

I gratefully acknowledge the work of the steering committee members of the CASCADE and UK Register of Seroconverters studies and colleagues at clinical centres, and thank them for allowing use of these datasets. I also thank all the participants of the CASCADE and UK Register of Seroconverters studies for allowing their routine clinical data to be included. I thank and acknowledge the following collaborators from the CASCADE study for their feedback on the applied analysis presented in this thesis: Andrew Phillips, M. John Gill, Ronald Geskus, Giota Touloumi, James Young and Heiner Bucher.

Thank you to Truly Johnston, my partner when I started this project and now my wife, for unwavering support, love and faith in my ability.

Contents

List of Figures

List of Tables

Abbreviations

AIC, Akaike information criterion;
AIDS, acquired immune deficiency syndrome;
ANCOVA, analysis of covariance;
ART, antiretroviral therapy;
BIC, Bayesian information criterion;
CASCADE, Concerted Action on SeroConversion to AIDS and Death in Europe;
CI, confidence interval;
Cor, correlation;
Cov, covariance;
E, expectation;
ECME, expectation/conditional maximisation either;
EM, expectation–maximisation;
HAART, highly active antiretroviral therapy;
HCV, hepatitis C virus;
HIV, human immunodeficiency virus type-1;
IDU, injecting drug user;
IGLS, iterative generalised least squares;
INSTI, integrase strand transfer inhibitor;
IOU, integrated Ornstein–Uhlenbeck;
IQR, interquartile range;
MAR, missing at random;
MCAR, missing completely at random;
MNAR, missing not at random;
MVN, multivariate normal;
N, normally distributed;
NNRTI, non-nucleoside reverse transcriptase inhibitor;
NRTI, nucleoside/nucleotide analog reverse transcriptase inhibitor;
PI, protease inhibitor;
PKPD, pharmacokinetic/pharmacodynamic;
RCT, randomised controlled trial;
REML, restricted maximum likelihood;
RIGLS, restricted iterative generalised least squares;
RNA, ribonucleic acid;
START, Strategic Timing of AntiRetroviral Treatment;
Var, variance;
VL, viral load.

1 Introduction

Longitudinal data, in which repeated observations have been recorded over time for each individual, require specialised statistical techniques that account for the resulting lack of independence between measurements. This increases the required complexity of any statistical analysis, but has the advantage that patterns of between- and within-patient variability can be investigated and quantified. In this thesis, we propose extensions to existing methodologies for the analysis of longitudinal data and apply these to CD4 cell count data from human immunodeficiency virus type-1 (HIV) positive patients. More specifically we analyse CD4 cell count data from 'HIV seroconverters', patients in whom the timing of HIV seroconversion (the appearance of HIV-specific antibodies in the blood) can be well estimated, providing a natural zero time-point for statistical models.

Linear mixed effects models are particularly common for the analysis for longitudinal data. These models represent an extension of the linear regression framework, in which model coefficients for predictive variables are permitted to vary randomly between individuals or groups. The use of linear mixed effects models for the analysis of longitudinal data was proposed and formalised by Laird and Ware[1], and thorough reviews of this topic are given by Verbeke and Molenberghs[2] and Diggle *et al.*[3] among others. Non-linear mixed effects models, in which the function for the expectation of the outcome variable may be non-linear in both the fixed effect parameters and the random effect terms, are also widely used in biomedical research[4–7].

Mixed effects models, both linear and non-linear, have a number of appealing characteristics: they account for the dependency in datasets that results from multiple observations being obtained from each individual over time, they can be fitted to unbalanced datasets in which there are missing data or in which each individual has been observed at irregular time-points, and the theoretical basis of the model can be easily reported, understood and interpreted. One additional practical benefit is that linear mixed effects models can now be readily implemented in any of the major statistical software packages, and non-linear mixed effects models can also be implemented in most. However, the software implementations of these modelling frameworks nonetheless impose restrictions on the structures of the statistical models that can be fitted, necessitating statistical assumptions that may be questionable in some situations relating to the analysis of biomedical data. The use of more complex models to account for patterns of variability in the data may allow more information to be gained and more accurate statistical inference regarding model parameters.

In the context of analysing CD4 T-cell counts in HIV patients Taylor *et al.*[8] found that the addition of non-stationary stochastic process components to linear mixed effects models for pre-treatment data led to a substantial improvement in model fit,

but this extension is not available as an option in any of the major statistical software packages and these types of models have not gained widespread use in practice. The mixed effects modelling framework has also been extended to allow for differences in the overall level of variability between patients, for example Wang and Fan[9] used a linear mixed model generalised to follow a multivariate-t distribution to analyse CD4 counts in HIV patients. In this thesis we explore these augmented mixed effects models, discussing and developing computational approaches to the maximum likelihood estimation of model parameters and proposing and applying further novel extensions.

We also consider the analysis of a longitudinally monitored biomarker following treatment initiation, for which the characteristics of response to treatment are dependent on the value of the biomarker at initiation. In this setting, there has been debate in the literature as to whether the baseline measurement should be included as an outcome variable within a parametric model[10;11] or whether it should be included as an independent predictive variable[12]. In this thesis we develop a novel modelling framework in which the 'true' baseline value is treated as a latent variable, with a distribution conditioned on all available pre-treatment data, with post-treatment observations modelled as following a distribution that is dependent on this baseline value. We incorporate the potential for non-stationary stochastic processes and heavy-tailed distributions within this framework.

The combined modelling framework is applied to pre- and post-treatment CD4 cell count data from 'HIV seroconverters', allowing the time interval from seroconversion to treatment initiation to be considered as a factor in the analysis of post-treatment characteristics. However, the definition of 'well estimated date of seroconversion' for the datasets concerned includes patients with an interval between last negative and first positive test for HIV of up to 3 years, with the estimated date of seroconversion in most cases set to be the mid-point between these tests. Although we retain this simplifying assumption in some of the analyses presented, we also develop a model for pre- and post-treatment data in which uncertainty in the exact date of seroconversion is taken into account.

The motivation for the work presented in this thesis is to develop statistical models that better reflect the structure and the patterns of within- and between-patient variability that are observed in the data under investigation. The implementation of models that more fully describe the data has the potential firstly to allow more robust inferences regarding questions of clinical interest and secondly to allow investigation of characteristics of the data that are otherwise ignored.

The array of techniques available for the analysis of longitudinal data has expanded greatly in recent years, and so there is a need to place some restrictions on the scope of the thesis. We focus on statistical models for continuous outcome vari-

ables, and so do not consider the 'generalised linear mixed model' framework[13] that allows for Bernoulli or other non-normal conditional distributions for the outcome variable. We extend mixed effects models based on parametric functions rather than differential equations, and so the literature relating to 'dynamic' or pharmacokinetic/ pharmacodynamic (PKPD) modelling[14] is not reviewed except where this overlaps with that for function-based non-linear mixed effect models. We consider only parametric models, rather than semi-/non-parametric modelling techniques[15] that have been developed for longitudinal data. We employ maximum likelihood estimation for the parameters of the models developed, and Bayesian approaches to model fitting are not considered other than as a point for discussion.

In Chapter 2, we review the characteristics of linear mixed models that incorporate stochastic processes in addition to the random effect terms and propose efficient methods for obtaining maximum likelihood estimates of model parameters. We also review both the characteristics of models based on the multivariate-t distribution and the computational approaches available for maximum likelihood estimation of more complex non-linear statistical models.

In Chapter 3, we review the available methods for using residual diagnostics of linear mixed models to evaluate the plausibility of modelling assumptions. We review and critique the ways in which these techniques have been generalised for mixed effects models based on the multivariate-t distribution, and propose novel residual diagnostic plots that could be used in this setting; the methodology developed is applied in later chapters.

In Chapter 4, a novel extension of the linear mixed model combining a fractional Brownian motion process and a multivariate-t distribution is applied to a large dataset of pre-treatment CD4 counts in HIV-positive patients from the Concerted Action on SeroConversion to AIDS and Death in Europe (CASCADE)[16] collaboration of seroconverter cohorts. In addition, a patient cohort simulation is presented to assess the implications of the model, and a separate simulation study is reported that demonstrates the potential for substantial biases in parameter estimates when overly simplistic models are used in the presence of missing data.

In Chapter 5 a new framework is developed for the combined modelling of pre- and post-treatment longitudinal data, which is demonstrated through application to CD4 counts before and after initiation of highly active antiretroviral therapy (HAART) using data from the UK Register of Seroconverters Cohort. As in Chapter 4, fractional Brownian motion processes and multivariate-t distributions are included in the model to capture the patterns of within- and between-patient variation that can be observed in the raw data. Simulations are presented to explore the potential for biases in the observed baseline value of a biomarker at treatment initiation relative to the true underlying value and to demonstrate that the modelling framework pro-

posed is capable of identifying non-linear relationships between the true baseline value and the characteristics of response to treatment.

The model structure developed in Chapter 5 is then applied to a larger dataset of pre- and post-treatment CD4 counts from the CASCADE collaboration in Chapter 6, with the additional inclusion of patient and drug regimen characteristics that could potentially predict response to HAART. This work is further developed in Chapter 7, in which the models are extended to allow for uncertainty in the exact timing of seroconversion in the patients included. Chapter 8 comprises a discussion of the statistical methodology developed in this thesis, including an exploration of how this could influence further work.

1.1 Disclaimer regarding collaborative work

Following the requirements for applied research using data from the CASCADE cohort, the analysis in Chapters 6 and 7 was planned and interpreted in collaboration with external investigators. A consensus decision was made regarding the inclusion criteria for the analysis and other specific contributions are noted where relevant. The modelling framework for the analysis was developed entirely by myself, and I also carried out all programming tasks and processing and presentation of results.

2 Computational issues for mixed effects models

In this chapter we explore potential problems and solutions in the maximum likelihood estimation of extensions to the standard mixed effects models for longitudinal data. This is done in order to allow subsequent chapters to provide a more focused discussion of methodological developments and novel applications. This chapter therefore mostly comprises a review of existing computational techniques, although the development and implementation of a novel computational strategy for the maximum likelihood estimation of linear mixed models that incorporate stochastic process components is described in Section 2.2.2 and we also explain how Brownian motion can be incorporated using existing software for iterative generalised least squares (IGLS) estimation in Section 2.2.3. We do not consider restricted maximum likelihood estimation (REML) in detail, this is because it would present considerable additional technical challenges for the non-linear models developed in later chapters and also because we aim to apply the modelling framework developed to relatively large datasets containing thousands of individuals, for which the differences between maximum likelihood and REML estimates tend to be small.

We begin by reviewing the properties of some Gaussian processes in Section 2.1, and discuss their incorporation into the linear mixed model framework in Section 2.2. We then review the properties and use of the multivariate-t distribution in Section 2.3. We discuss the available options for maximum likelihood estimation of non-linear mixed effects models and other latent variable models that lack a closed form for the marginal likelihood in Section 2.4, and consider estimation for models involving a combination of multivariate normal and multivariate-t distributions in Section 2.5. A brief summary discussion is presented in Section 2.6.

2.1 Properties of Brownian motion and related processes

2.1.1 Scaled Brownian motion

Scaled Brownian motion has been incorporated into linear mixed effects models by a number of researchers[8;17]. When considered in terms of a given set of observation points, a scaled Brownian motion process W_t is defined by the properties[18]:

$$W_0 = 0$$
$$W_t - W_s \sim N(0, \kappa(t-s)) \text{ for } 0 \leq s < t.$$

The process starts at zero at time (t) zero, and increments of the process are stationary, independent (for disjoint periods of time) and normally distributed with mean zero and variance equal to the difference in time between observation points

scaled by a constant factor κ ($\kappa > 0$). The following characteristics arise from these conditions:

$$\begin{aligned} \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \kappa t \\ \mathrm{Cov}[W_s, W_t] &= \kappa * \min(s, t). \end{aligned}$$

The distribution of a set of n observations relating to a given series of time points therefore follows a multivariate normal distribution with a mean vector of n zeros and covariance matrix defined by the formulae given above. As such, scaled Brownian motion meets the definition of a Gaussian process, and can be readily incorporated into the theoretical framework of linear mixed models. For brevity, we refer to 'scaled Brownian motion' as just 'Brownian motion' in later chapters.

2.1.2 Scaled fractional Brownian motion

Fractional Brownian motion represents a generalisation of a Brownian motion process in which increments for disjoint time periods are not constrained to be independent, although they do remain stationary. The process was introduced by Mandelbrot and van Ness[19]. The characteristics of a fractional Brownian motion process are determined by an additional parameter, referred to as H or 'the Hurst index', that may take a value in the range (0,1). Standard Brownian motion represents a special case of fractional Brownian motion, corresponding to $H = \frac{1}{2}$. As for standard Brownian motion, the expectation of the value of the process is zero for all points in time.

When $H < \frac{1}{2}$, successive increments of the process are negatively correlated. This has the consequence, firstly, that the path of the trajectory appears 'jagged' and, secondly, that realisations of the process tend to revert towards the mean of zero. For $H > \frac{1}{2}$, successive increments of the process are positively correlated. This means that the path of the process has a relatively 'smooth' appearance, and also that realisations of the process tend to diverge away from zero. Illustrative simulated realisations of fractional Brownian motion processes generated with varying values of H are shown in Figure 2.1.

As for Brownian motion, a scale parameter (κ, $\kappa > 0$) can be added to the standard definition of fractional Brownian motion[20], corresponding to the variance of the process at $t = 1$. We may then characterise the properties of the process as follows:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \end{aligned}$$

$$\mathrm{Var}[W_t] = \kappa\,|t|^{2H}$$
$$\mathrm{Cov}[W_s, W_t] = \frac{\kappa}{2}\left(|s|^{2H} + |t|^{2H} - |t-s|^{2H}\right).$$

As for Brownian motion, fractional Brownian motion is defined as a continuous-time stochastic process. However, as we are concerned with modelling biomedical measurements obtained at specific time points, we focus here on the properties of the process relating to a finite set of observations. Fractional Brownian motion has been used for mathematical modelling in fields including hydrology[21], computer network traffic[22] and finance[23]. However, this Gaussian process has not previously been incorporated into the linear mixed effects model framework. For brevity, we refer to 'scaled fractional Brownian motion' as just 'fractional Brownian motion' in later chapters.

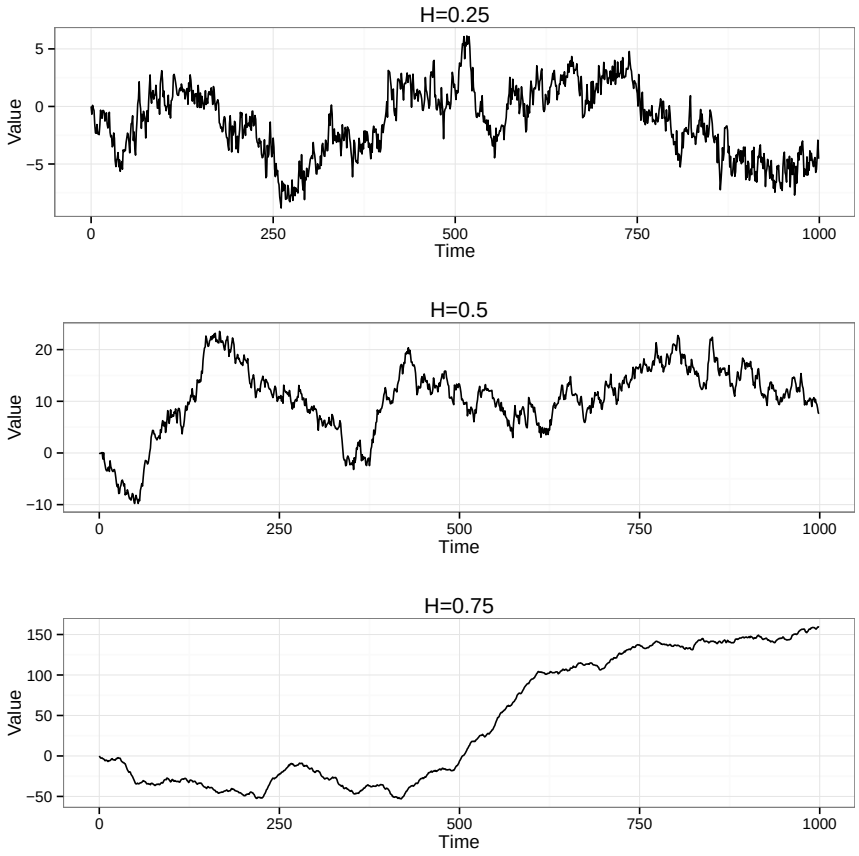

Figure 2.1. Simulated realisations of fractional Brownian motion processes with varying values of H and scale parameter fixed at 1. A finite set of 1000 observations was generated in each case.

2.1.3 Integrated Ornstein–Uhlenbeck process

The integrated Ornstein–Uhlenbeck process (IOU) process is another non-stationary Gaussian stochastic process that has also been used to model CD4 counts in HIV-

positive patients, a full description is provided by Taylor *et al.*[8]. The process has the following characteristics:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \frac{\kappa}{\alpha^3}\left(\alpha t + e^{-\alpha t} - 1\right) \\ \mathrm{Cov}[W_s, W_t] &= \frac{\kappa}{2\alpha^3}\left(2\alpha * \min(s,t) + e^{-\alpha t} + e^{-\alpha s} - 1 - e^{-\alpha|t-s|}\right). \end{aligned}$$

We have used the symbol κ ($\kappa > 0$) to denote the variance scaling parameter (σ^2 was used by Taylor *et al.*[8]). The α ($\alpha > 0$) parameter determines the extent to which the process reverts towards its mean value. For values of α approaching infinity, the process is equivalent to scaled Brownian motion, whereas for values of α approaching zero the process is equivalent to a random slopes model (without a random intercept)[8].

2.2 Maximum likelihood estimation

2.2.1 Marginal distribution and likelihood function

For models incorporating Gaussian processes such as Brownian motion, the fact that the marginal distribution of the full vector of observations of the outcome variable is multivariate normal (MVN) means that parameter estimation can be achieved through adjustment of the methods used for standard linear mixed models. Ignoring the potential for grouping factors beyond that of each individual in a dataset (e.g. those of hospitals or geographical regions), the linear mixed model for longitudinal data can be expressed in the form[1]:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\ \mathbf{e}_i &\sim MVN(\mathbf{0}, \mathbf{R}_i). \end{aligned} \tag{1}$$

Here, $\mathbf{y}_i$ represents the vector of n_i observations for the i^{th} individual, $\mathbf{X}_i$ represents their design matrix for the 'fixed effects' parameters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_i$ is the vector of residual errors for each measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2 ... \mathbf{b}_N$ and residual errors $\mathbf{e}_1, \mathbf{e}_2 ... \mathbf{e}_N$ for each of the N individuals are independent of one another. It can be easily shown that this

formulation leads to the following marginal distribution for $\mathbf{y}_i$:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{R}_i).$$

Using the notation $\mathbf{V}_i = \mathbf{Z}_i\Psi\mathbf{Z}_i^{\mathrm{T}} + \mathbf{R}_i$, $\boldsymbol{\alpha}$ is defined as the parameters that are used to calculate $\mathbf{V}_i$, with $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ representing the vector of all parameters in the marginal model. If Θ_β and Θ_α are defined as the parameter spaces for the fixed effects and random effects/variance parameters, then $\Theta_\beta = \mathbb{R}^p$ and Θ_α is equal to the set of values for $\boldsymbol{\alpha}$ for which $\boldsymbol{\Psi}$ and all $\mathbf{R}_i$ are positive definite[3]. Maximum likelihood estimation of the model parameters is achieved through maximisation of the marginal likelihood function:

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{N}\left\{(2\pi)^{-\frac{n_i}{2}}|\mathbf{V}_i|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right)\right\}. \tag{2}$$

In practice, this is usually achieved through maximisation of the log-likelihood function:

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{N}\left\{-\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_i| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}. \tag{3}$$

When linear mixed models are fitted to longitudinal data, it is common to assume that the residual errors for each observation within each individual, $\mathbf{e}_i$, are independent and with constant variance, σ^2, i.e. $\mathbf{R}_i$ as defined in (1) is equal to $\sigma^2\mathbf{I}_{n_i}$. However, other forms for $\mathbf{R}_i$ are widely used, particularly for the analysis of longitudinal or spatial data. An example is provided by the exponential correlation structure[6], for which the elements (r_{jk}) of $\mathbf{R}_i$ are calculated as a function of the 'distance' s between each pair of observations (in the context of longitudinal data this would be the time difference) and a 'range' parameter η, which is constrained to be greater than zero:

$$r_{jk} = \sigma^2\, exp\left(-\frac{s_{jk}}{\eta}\right).$$

The remaining variability in the model, once the random effects have been accounted for, can also be subdivided into a component relating to a Gaussian process (independent of other model components) with expectation zero for all time points and an independent residual error for each observation (here assumed to have constant variance); this effectively just creates a class of parameterisations for $\mathbf{R}_i$. Defining $\boldsymbol{\Sigma_i}$ as the covariance matrix resulting from the chosen Gaussian process and set of time points $\mathbf{t}_i$ for the i^{th} individual, the linear mixed model can then be expressed

as:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + W_i[\mathbf{t}_i] + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\ W_i[\mathbf{t}_i] &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_i) \\ \mathbf{e}_i &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{4}$$

with marginal distribution:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_i + \sigma^2\mathbf{I}_{n_i}).$$

Diggle[24] observed that "...the correlation between measurements on the same unit usually depends on their separation in time, typically as a monotone decreasing function", and therefore proposed the use of a stationary Gaussian process to account for serial correlation. However, there is no mathematical barrier to the use of a non-stationary Gaussian process for this formulation of the linear mixed model. Indeed, this is what has been implemented by researchers who have incorporated Brownian motion and IOU processes into linear mixed effects models[8;17;25–27]. A naïve approach to parameter estimation for a linear mixed model incorporating any Gaussian process is therefore to use a general-purpose optimising program to maximise the log-likelihood function as expressed in (3) with respect to the full set of parameters $\boldsymbol{\theta}$, with a choice of parameterisation for $\boldsymbol{\alpha}$ that constrains $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma_i}$ (for all individuals) to be positive definite. Whether or not this approach is effective is dependent on the specific model and dataset considered, and on the characteristics of the optimisation algorithm employed. At present, the fitting of such models is not available through easily implemented default routines for most statistical software, although an R package 'lmenssp' that can be used to fit linear mixed effects models that include Brownian motion or IOU processes using this approach has been recently developed[28;29].

A number of different methods of parameterisation for random effect variance–covariance matrices have been proposed that ensure that they remain positive definite during estimation procedures, including use of Cholesky decompositions and matrix logarithms. A review of these methods is provided by Pinheiro and Bates[30]. If a scaled Brownian motion process is added to a random effects model being fitted to a dataset of longitudinal measurements in a set of individuals for which $t \geq 0$ for all observations, then $\boldsymbol{\Sigma_i}$ can be constrained to be positive semi-definite by constraining the scale parameter κ to be ≥ 0; hence, assuming an appropriate parameterisation of $\boldsymbol{\Psi}$, optimisation can be performed in terms of $\log(\kappa)$.

General-purpose optimising programs, which often form the core of maximum

likelihood estimation routines in statistical software, usually make use of a Newton– Raphson-type approach in which a set of working values for each parameter ($\boldsymbol{\theta}_i$) in a model is updated at each iteration (to $\boldsymbol{\theta}_{i+1}$) using a function of the gradient vector $\mathbf{g}(\boldsymbol{\theta}_i)$ and the matrix of second derivatives $\mathbf{H}(\boldsymbol{\theta}_i)$ (or some approximation of these) of the objective function, evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. Iterations are repeated until some specified convergence criteria are achieved. A concise review of the use of Newton–Raphson type algorithms for maximum likelihood estimation is provided in Chapter 1 of Gould *et al.*[31]. The vectors of first and second derivatives required for optimisation can either be computed exactly or approximated by finite differencing. The latter technique is still widely employed, but can lead to problems in the speed and stability of optimisation.

Expectation–maximisation (EM)-type approaches, as outlined by Dempster *et al.*[32], have also been widely used for maximum likelihood parameter estimation of linear mixed models. Indeed, an EM-type algorithm treating the random effects $\mathbf{b}_i$ as a missing data problem was proposed by Laird and Ware[1] in the paper that formalised this class of models. Liu and Rubin further extended the EM-type approach for linear mixed models, terming their new algorithm ECME, for 'expectation/conditional maximisation either'[33], for which they reported improvements in the speed of convergence. They note that the 'EM' approach suggested by Laird and Ware is in fact not a true EM-algorithm, but could be described as an ECME algorithm. Although the ECME approach provides a flexible framework for obtaining maximum likelihood estimates of parameters in the linear mixed model, existing implementations do not allow for arbitrary structures of the residual error covariance matrix, i.e. $\mathbf{R}_i$ in (1).

2.2.2 Profile likelihood methods

A number of specialised computational techniques have been developed for linear mixed models in order to optimise the speed and stability of maximum likelihood estimation. These methods include profiled likelihood techniques, in which expressions for the conditional estimates of a subset of the model parameters (obtained in terms of and conditional on the remaining model parameters) are substituted into the expression used to calculate the log-likelihood. This has the effect of reducing the dimensionality of the optimisation problem, as the log-likelihood is subsequently calculated as a function of a reduced set of unknown parameters that need to be included in the iterative optimization process. For linear mixed effects models with independent residual errors with constant variance, e.g. following the form of (1) with $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$, Pinheiro and Bates[6] demonstrate that the log-likelihood can be efficiently calculated as a function of only the parameters that define a form of the

variance–covariance matrix $\boldsymbol{\Psi}$ for the random effects scaled by the residual variance σ^2:

$$\mathbf{D} = \frac{\boldsymbol{\Psi}}{\sigma^2}.$$

As such, optimisation can be performed only in terms of the parameters that define $\mathbf{D}$, with the fixed effects parameters $\boldsymbol{\beta}$ and residual variance σ^2 calculated at their estimates conditional on $\widehat{\mathbf{D}}$ once convergence has been achieved. The methods of Pinheiro and Bates are also valid for more complex models in which there is more than one grouping factor, but we shall concentrate on single-level models. The profile likelihood method is implemented in the 'nlme' package[34] for the R statistical computing environment (R Foundation, Vienna, Austria). It is worth noting that the fact that optimisation is carried out with the covariance terms parameterised relative to σ^2 facilitates the use of heuristic algorithms to choose starting values for the parameters being entered into the iterative optimisation procedure.

Pinheiro and Bates[6] point out that any linear mixed effects model of the form of (1), with $\mathbf{R}_i = \sigma^2 \mathbf{R}'_i$, can be re-expressed as a transformed model that has independent residual errors with constant variance. $\mathbf{R}_i$ and hence $\mathbf{R}'_i$ is positive-definite for all cases, $\mathbf{R}'_i$ is calculated in each case as a function of the individual's covariates (typically the time variable when considering longitudinal data) and parameter vector $\boldsymbol{\lambda}$, and σ^2 is factored out in order to allow this parameter to be eliminated from the expression for calculation of the profiled log-likelihood. Following from the properties of being positive definite, an invertible symmetric square root can be calculated for $\mathbf{R}'_i$. However, an alternative and computationally efficient transformation is provided by the Cholesky decomposition:

$$\begin{aligned} \mathbf{R}'_i &= \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^{\mathrm{T}} \\ \mathbf{R}'^{-1}_i &= \left(\boldsymbol{\Lambda}_i^{\mathrm{T}}\right)^{-1} \boldsymbol{\Lambda}_i^{-1}. \end{aligned}$$

Where $\boldsymbol{\Lambda}_i$ is an invertible lower triangular matrix with positive diagonal elements (we have chosen to depart from the notation used by Pinheiro and Bates[6] in this context). Applying the inverse Cholesky root transformation to each term of the linear mixed model:

$$\begin{aligned} \mathbf{y}^*_i &= \boldsymbol{\Lambda}_i^{-1} \mathbf{y}_i & \qquad \mathbf{e}^*_i &= \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i \\ \mathbf{X}^*_i &= \boldsymbol{\Lambda}_i^{-1} \mathbf{X}_i & \qquad \mathbf{Z}^*_i &= \boldsymbol{\Lambda}_i^{-1} \mathbf{Z}, \end{aligned}$$

the form of the model can be rewritten as:

$$\mathbf{y}^*_i = \mathbf{X}^*_i \boldsymbol{\beta} + \mathbf{Z}^*_i \mathbf{b}_i + \mathbf{e}^*_i$$

$$\mathbf{b}_i \sim MVN(\mathbf{0}, \boldsymbol{\Psi})$$
$$\mathbf{e}_i^* \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}).$$

The transformed model therefore constitutes a linear mixed model that has independent errors with constant variance, allowing computational methods developed for such models to be applied. The likelihood for the full model being fitted can be evaluated using the standard rules for a transformation of variables, incorporating the Jacobian determinant:

$$\frac{d\mathbf{y}_i^*}{d\mathbf{y}_i} = \boldsymbol{\Lambda}_i^{-1} \qquad\qquad \left|\frac{d\mathbf{y}_i^*}{d\mathbf{y}_i}\right| = \left|\boldsymbol{\Lambda}_i^{-1}\right|$$

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{i=1}^{N} p\left(\mathbf{y}_i|\boldsymbol{\theta}\right) \\ &= \prod_{i=1}^{N} p\left(\mathbf{y}_i^*|\boldsymbol{\theta}\right)\left|\boldsymbol{\Lambda}_i^{-1}\right| = L(\boldsymbol{\theta}|\mathbf{y}^*)\prod_{i=1}^{N}\left|\boldsymbol{\Lambda}_i^{-1}\right|. \end{aligned}$$

Although we have written the likelihood function here in terms of the full parameter vector $\boldsymbol{\theta}$, in practice $\boldsymbol{\beta}$ and σ^2 are profiled out of the likelihood calculations for $\mathbf{y}_i^*$ and hence the dimensionality of the optimisation problem is only increased by the number of parameters in $\boldsymbol{\lambda}$. The same form of transformation can be used when considering maximisation of the REML function, meaning that REML parameter estimates can also be obtained using this strategy.

Pinheiro and Bates[6] propose the use of this computational technique in combination with a decomposition of the within-group covariance structure into a diagonal matrix $\mathbf{S}_i$ that determines the residual variance for each observation and a correlation matrix $\mathbf{C}_i$:

$$\mathbf{R}_i' = \mathbf{S}_i\mathbf{C}_i\mathbf{S}_i,$$

such that:

$$\mathrm{Var}[\mathrm{e}_{ij}] = \sigma^2(\mathbf{S}_i)_{jj}^2 \qquad\qquad \mathrm{Cor}[\mathrm{e}_{ij}, \mathrm{e}_{ik}] = (\mathbf{C}_i)_{jk}.$$

This decomposition is flexible in allowing a wide range of structures for $\mathbf{R}_i'$, but does not allow for models that incorporate a scaled Brownian motion component with a residual error term or those that include an IOU or fractional Brownian motion component, as in such models the parameters that determine the variance and correlation of $\mathbf{e}_i$ cannot be separated into distinct sets. However, we propose an al-

ternative structure for $\mathbf{R}_i'$ in such cases:

$$\begin{aligned}\mathbf{R}_i' &= \mathbf{I}_{n_i} + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_i \\ &= \mathbf{I}_{n_i} + \boldsymbol{\Sigma}_i'.\end{aligned}$$

This formulation can be used to fit linear mixed models that incorporate any Gaussian process in addition to a constant residual error term. $\boldsymbol{\Sigma}_i'$ is constructed at each stage of the iterative optimisation process, following the forms described in Section 2.1, using the current values of $\boldsymbol{\lambda}$ and each individual's covariates (in this context their set of times for each observation). Owing to the factoring out of σ^2, when using scaled Brownian motion or fractional Brownian motion processes the scale parameter ξ used in the optimisation represents the variance of the Gaussian process at $t = 1$ relative to the residual error variance, i.e. $\xi = \kappa/\sigma^2$ where κ is the natural scale parameter representing variance at $t = 1$. The scale parameter relative to the residual error variance is similarly used for optimisation for an IOU process, although this does not correspond to the variance at $t = 1$. Once convergence of the optimisation procedure has been achieved, the estimate of the scale parameter on the natural scale can be calculated as $\hat{\kappa} = \hat{\sigma^2} * \hat{\xi}$.

We have implemented the incorporation of a scaled Brownian motion, scaled fractional Brownian motion or IOU process component into linear mixed models based on the functionality provided by the 'nlme' package[34] for R. Parameter estimation for such models in terms of ξ can be achieved using the package's existing framework by loading functions to create 'user-defined correlation structures' that generate $\mathbf{R}_i'$ for each subject as described. However, the package's default functions for estimating the approximate distributions of parameter estimates need to be adjusted in order to create confidence intervals for estimates of the natural scale parameter κ. We provide functions to achieve this in the new R package 'covBM', which is available for download from the Comprehensive R Archive Network (CRAN)[35]. The package vignette is provided in Appendix A.

Pinheiro[36] demonstrated that, under certain regularity conditions, maximum likelihood parameter estimates for the general linear mixed model are asymptotically consistent and normally distributed. In addition, the estimates of the fixed effects parameters ($\boldsymbol{\beta}$) and those of the parameters that define the covariance structure ($\boldsymbol{\alpha}$) are asymptotically independent. The approximate covariance matrix of the parameter estimates is given by the inverse of the information matrix of the log-likelihood function.

Given the asymptotic independence between estimates of the fixed effects and covariance parameters, one option for creating confidence intervals for parameter estimates when fitting a linear mixed model is to calculate (or approximate) the ob-

served information matrix only for the estimates of the covariance parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\lambda}}$, and to approximate the sampling distribution of the fixed effects parameters conditional on $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\lambda}}$. Such a strategy is suggested by Pinheiro and Bates[6], and fits naturally with the use of the profile log-likelihood expressed solely in terms of covariance parameters. However, if the form of the log-likelihood is used in which the residual variance parameter σ^2 is also profiled, then an alternative form of the log-likelihood function is required at convergence following optimisation; the parameterisation must be altered so that σ^2 is no longer factored out of the other covariance parameters before estimation of the observed information matrix for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\lambda}}$. In the 'nlme' package, the parameterisation used to generate confidence intervals for the parameters of the variance components of the model is in terms of the $\log(SD)$ of each random effect and $\log(\sigma)$. However, for any user-defined correlation structure the variance-covariance matrix for parameter estimates is approximated in terms of the parameters used for optimisation.

If a scaled Brownian motion, fractional Brownian motion or IOU process is included in the model then the estimate and standard error of the log of the scale parameter used in the optimisation procedure ($\log(\xi)$) need to be converted to values for the log of the natural scale parameter ($\log(\kappa)$). In the 'covBM' package, this is done by the wrapper functions `lmeBM` and `nlmeBM` after the default approximate variance-covariance matrix for the parameter estimates has been generated by the original 'nlme' functions. The default 'nlme' behaviour is to provide $\mathrm{Var}\left[\log\left(\hat{\xi}\right)\right]$, but the 'covBM' wrapper functions convert this to $\mathrm{Var}[\log(\hat{\kappa})]$ as follows:

$$\begin{aligned}\mathrm{Var}[\log(\hat{\kappa})] &= \mathrm{Var}[\log\left(\hat{\sigma^2}\hat{\xi}\right)] \\ &= \mathrm{Var}[2\log(\hat{\sigma}) + \log\left(\hat{\xi}\right)] \\ &= 4\mathrm{Var}[\log(\hat{\sigma})] + 4\mathrm{Cov}[\log(\hat{\sigma}), \log\left(\hat{\xi}\right)] + \mathrm{Var}[\log\left(\hat{\xi}\right)].\end{aligned}$$

Similarly, the covariance of $\log(\hat{\kappa})$ with each other variance parameter $\hat{\alpha_i}$ is calculated as:

$$\begin{aligned}\mathrm{Cov}[\log(\hat{\kappa}), \hat{\alpha_i}] &= \mathrm{Cov}[2\log(\hat{\sigma}) + \log\left(\hat{\xi}\right), \hat{\alpha_i}] \\ &= 2\mathrm{Cov}[\log(\hat{\sigma}), \hat{\alpha_i}] + \mathrm{Cov}[\log\left(\hat{\xi}\right), \hat{\alpha_i}].\end{aligned}$$

2.2.3 Iterative generalised least squares

The use of an IGLS algorithm for the estimation of hierarchical linear mixed models was described by Goldstein[37], this technique is very efficient for models in which grouping factors for random effects are hierarchically nested. In brief, the algorithm consists of two steps at each stage. First the estimates of the fixed effects parameters

$\hat{\boldsymbol{\beta}}_{j+1}$ are updated by a generalised least squares calculation conditional on the overall marginal covariance matrix $\hat{\mathbf{V}}_j$ as defined by the current estimates of the parameters that define the random effects and variance structure $\hat{\boldsymbol{\alpha}}_j$:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{j+1}} = \left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}_j^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}_j^{-1}\mathbf{Y},$$

where $\mathbf{X}$ is the entire design matrix for the mean response, and $\mathbf{Y}$ is the full vector of the outcome variable. The second step is to carry out a generalised least squares calculation to update the values of $\hat{\boldsymbol{\alpha}}_j$ conditional on the current estimates for the fixed effects parameters $\hat{\boldsymbol{\beta}}_{j+1}$:

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{j+1}} = \left(\mathbf{X}^{*\mathrm{T}}\hat{\mathbf{V}}_j^{*-1}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\hat{\mathbf{V}}_j^{*-1}\mathbf{Y}^{*},$$

where $\mathbf{Y}^*$ is the vector of the stacked upper triangular elements of:

$$(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{j+1})(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{j+1})^{\mathrm{T}},$$

$\mathbf{X}^*$ is the design matrix relating the parameters of the covariance structure $\boldsymbol{\alpha}$ to $\mathbf{Y}^*$, and $\mathbf{V}^*$ is the covariance matrix of $\mathbf{Y}^*$.

These two steps are repeated until convergence of both the fixed effects and variance parameter vectors is achieved. Goldstein showed that this algorithm provides maximum likelihood estimates of the parameters for models in which the marginal distribution of the response vector is multivariate normal[37]. Goldstein also described a modified version of this algorithm, termed restricted iterative generalised least squares (RIGLS), that produces REML estimates of the model parameters[38].

One limitation of this estimation procedure is that the parameterisation of each element of the covariance structure $\mathbf{V}$ needs to be linear in terms of the unknown parameters in order for the generalised least squares calculation to be constructed for the step in which variance parameters are updated. For example, this means that it is not possible to use the technique to fit linear mixed models in which there is an exponential correlation structure for the observation-specific residuals. In the context of modelling non-stationary stochastic processes, the technique cannot be used for incorporating fractional Brownian motion or IOU processes into models, as the construction of each term of the resulting covariance matrix is not linear in terms of the parameters to be estimated.

However, it is possible to use the IGLS or RIGLS algorithm for linear mixed models that incorporate a scaled Brownian motion process. In this case, the form of each element of the covariance structure *is* linear in terms of the scale parameter to be estimated κ. The additional contribution to element r_{jk} of an individual's covariance matrix $\mathbf{V}_i$ is given by $\kappa * \min(s,t)$, where s is the time at the j^{th} observation and t is

the time at the k^{th} observation. This use of IGLS to fit a linear mixed model including a Brownian motion component has not been previously described in the literature. A macro to add a Brownian motion component to a linear mixed model in the MLwiN software package (Version 2.28; Centre for Multilevel Modelling, University of Bristol, Bristol, UK), which implements IGLS and RIGLS, is given in Appendix B. This method of estimation is very fast compared to others available, fitting a model to around 90 000 observations in less than 30 seconds using a standard PC (2.4 GHz processor, 2 GB RAM).

2.3 Multivariate-t distribution for longitudinal data

A common finding when assessing the goodness of fit of a statistical model based on the normal distribution is the observation of heavier tails than expected on diagnostic plots of residuals. This remains the case in the context of using linear mixed models for the analysis of longitudinal data. Verbeke and Lesaffre found that estimation of fixed effects parameters using linear mixed models is robust to non-normal distributions of the random effects, although they suggested a correction to the estimated covariance matrix for the parameter estimates when non-normality of random effects is suspected[39]. Jacqmin-Gadda *et al.* used simulations to show that inference for fixed effects is robust to misspecification of the error distribution when using linear mixed models in some situations[40]. However, these analyses did not take into account the potential for missing or unbalanced data where this is dependent on the observed values of the outcome variable.

Furthermore, in some situations the structure of the covariance of observations within individuals is of direct interest to the investigator, rather than just a factor in ensuring correct inference regarding the fixed effects parameters in a model. This is the case when constructing models relating to growth curves, or when attempting to make predictions regarding sequential observations of biomedical variables in individual patients. When standard linear mixed models do not appear to adequately describe the variation observed in the data in such cases, the options for fitting extended models that account for this are limited using the currently available statistical software.

As later discussed in Chapter 3, it is not straightforward to separate the variability due to the random effects included in a linear mixed model from that associated with the residual error terms (whether or not these are independent and with constant variance) when evaluating diagnostic plots to assess the adequacy of fit to the data. Nonetheless, when data appear over-dispersed with respect to normality, one potential first step would be to specify that either the distribution of the random effects or that of the residual error terms follows a heavy-tailed distribution. However,

the combination of normally distributed and non-normally distributed components leads to a model for which the likelihood cannot be expressed in closed form, making the challenge of maximum likelihood estimation substantially harder. Numerical techniques, as discussed in Section 2.4, can be applied to approximate the likelihood and allow optimisation in such cases. However, we first consider a model in which the set of observations for each individual as a whole follows a multivariate-t distribution, as maximum likelihood estimation for such models is computationally simpler. The use of such a model for multivariate regression analysis was proposed by Lange *et al.*[41], and was further developed as an extension of the linear mixed model by Welsh and Richardson[42] and Pinheiro *et al.*[43]. We review here the characteristics of the multivariate-t distribution, and in Chapter 4 we present an analysis of pre-treatment CD4 counts that applies the multivariate-t distribution to linear mixed effects models that also include stochastic process components, a combination that had not previously been reported. In Chapters 5, 6 and 7 we also consider novel extensions to mixed effects models that involve combinations of multivariate-t and multivariate normal distributions, and estimation for such models is discussed in Section 2.5.

2.3.1 Characteristics of multivariate-t distribution

There are a number of multivariate generalisations of the univariate t-distribution, and a thorough review of this topic is provided by Kotz and Nadarajah[44]. However, we shall refer to the *multivariate-t distribution* as that with the probability density function:

$$\frac{\boldsymbol{\Gamma}\left(\left(\nu+n_i\right)/2\right)}{\boldsymbol{\Gamma}\left(\nu/2\right)\nu^{n_i/2}\pi^{n_i/2}\left|\mathbf{V}_i\right|^{1/2}\left(1+\frac{1}{\nu}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)^{\mathrm{T}}\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)\right)^{(\nu+n_i)/2}},$$

where n_i represents the length of the random vector $\mathbf{y}_i$ ($\in \mathbb{R}^{n_i}$), $\mathbf{V}_i$ is a $n_i \times n_i$ positive-definite scale matrix, $\boldsymbol{\mu}_i$ is a $n_i \times 1$ location vector and ν is a degrees of freedom parameter. The mean of the distribution is $\boldsymbol{\mu}_i$ if $\nu > 1$ and otherwise undefined, and the variance of the distribution is $\frac{\nu}{\nu-2}\mathbf{V}_i$ if $\nu > 2$ and otherwise undefined. This is the most commonly used definition of the multivariate-t distribution.

In the present context, the mean vector $\boldsymbol{\mu}_i$ will be represented as $\mathbf{X}_i\boldsymbol{\beta}$, i.e. a function of a design matrix $\mathbf{X}_i$ and vector of parameters $\boldsymbol{\beta}$. As for linear mixed models based on the normal distribution, the scale matrix $\mathbf{V}_i$ can be divided into components relating to a random effects structure and a residual error structure, i.e. $\mathbf{Z}_i\Psi\mathbf{Z}_i^{\mathrm{T}}$ and $\mathbf{R}_i$, respectively. Pinheiro *et al.* consider the situation in which the degrees of freedom parameter may vary between subgroups of individuals, but we shall assume that this is a single constant[43].

If a vector of observations $\mathbf{y}_i$ follows a multivariate-t distribution:

$$\mathbf{y}_i \sim t_{n_i}\left(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i, \nu\right),$$

then this can alternatively be represented as a hierarchical model in which $\mathbf{y}_i$ follows a multivariate normal distribution conditional on a gamma-distributed variable γ_i (with parameters given for 'shape' and 'rate', respectively)[43]:

$$\begin{aligned} \mathbf{y}_i|\gamma_i &\sim MVN\left(\mathbf{X}_i\boldsymbol{\beta}, \frac{1}{\gamma_i}\mathbf{V}_i\right) \\ \gamma_i &\sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned} \tag{5}$$

As noted by Pinheiro *et al.*[43], it directly follows that:

$$\gamma_i|\mathbf{y}_i \sim \text{gamma}\left(\frac{\nu + n_i}{2}, \frac{\nu + \delta_i^2(\boldsymbol{\theta})}{2}\right)$$

where, $$\delta_i^2(\boldsymbol{\theta}) = \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right).$$

Here, $\boldsymbol{\theta}$ represents the parameter vector that includes $\boldsymbol{\beta}$ and determines the construction of $\mathbf{V}_i$. From the standard properties of a gamma distribution, it can be seen that:

$$\mathrm{E}\left(\gamma_i|\mathbf{y}_i\right) = \frac{\nu + n_i}{\nu + \delta_i^2(\boldsymbol{\theta})}.$$

The hierarchical form as shown in (5) also provides a route by which a single variance component of a mixed effects model can be specified as following a multivariate-t distribution, through conditioning of only the relevant covariance parameters on a latent variable that follows a gamma distribution.

2.3.2 Maximum likelihood estimation for the multivariate-t distribution

A series of EM-type algorithms for maximum likelihood estimation of parameters for the multivariate-t distribution generalisation of the linear mixed effects model are provided by Pinheiro *et al.*[43]. Most of the algorithms provided require that the structure of the residual error covariance matrix $\mathbf{R}_i$ be known up to a scalar factor σ^2, but one ECME algorithm is reported for which only the γ_i are treated as missing data, with the $\mathbf{b}_i$ integrated out of the complete data likelihood. Using this approach, parameter estimation can be achieved for multivariate-t distribution models that include non-stationary stochastic components such as Brownian motion. However, the algorithm requires maximum likelihood estimation of the parameter vector $\boldsymbol{\theta}$ conditional on $\hat{\boldsymbol{\gamma}}$ for the conditional maximisation step of every iteration, meaning

that this approach is computationally very time-consuming. This issue is particularly problematic when attempting to analyse large datasets. An R package 'tlmec'[45] exists for fitting models generalised from a normal linear mixed effects model with independent error terms of constant variance using an ECME approach[46].

As the likelihood function for the multivariate-t linear mixed effects model has a closed form, whatever the structure of $\mathbf{V}_i$, it is possible to directly apply Newton–Raphson-type optimisation procedures. Lange *et al.* provide some useful results for the derivation of gradient functions[41]. However, the only currently available software package to implement linear mixed models that incorporate stochastic process components and allow a marginal multivariate-t distribution is 'lmenssp' for R[28] (this feature was added to 'lmenssp' after the submission of our paper on the topic to *Statistics in Medicine*[47]), which uses finite differencing to obtain the gradients required for optimisation. Lin reports a hybrid maximisation algorithm, combining an initial ECME approach and subsequent Fisher scoring method, to fit a multivariate-t linear mixed effects model including autoregressive correlation for the residual error terms[48]. However, for the analysis in Chapter 4 we will use direct application of Newton–Raphson-type optimisation implemented in the ADMB software, using the parameter estimates from the equivalent normal linear mixed model and a moderate value for the degrees of freedom parameter (i.e. $\nu = 10$) as starting values for the iterative procedure. An advantage of using a Newton–Raphson-type procedure for parameter estimation is that the asymptotic multivariate normal estimate of the sampling distribution of the parameters can be readily obtained upon convergence.

Although finite differencing can be employed, the use of exactly calculated gradients (with respect to the model parameters) in Newton–Raphson-type procedures can greatly improve stability and speed of convergence. However, in some situations, such as incorporating stochastic process components into the multivariate-t linear mixed effects model, the analytic derivation of the gradients is not trivial. In addition, once an analytic form for each of the gradient terms has been derived, it is required that this be programmed into the computational procedure for the optimisation in an efficient manner. An alternative method is provided by automatic differentiation, whereby a computer program is structured in such a way that it can automatically calculate the derivatives of a mathematical function to the same degree of accuracy as analytical derivatives (to machine precision)[49]. In essence, this is achieved through application of the chain rule to each of the elementary operations that comprise the calculation of the objective function (i.e. the log-likelihood function). The open source Automatic Differentiation Model Builder (ADMB) software (ADMB Foundation, Honolulu, HI, USA) allows optimisation for any statistical model in which a differentiable log-likelihood function can be written in the C++ language[50]; additional statistical and mathematical functions (including matrix and

vector functions and operations) are provided by the software to facilitate this. This software is used for the analyses presented throughout this thesis.

2.4 Estimation for non-linear latent variable models

Up until now in this chapter we have only considered models for which the marginal log-likelihood can be expressed in closed form. However, there is often a motivation to consider models for which this is not the case, and which therefore require use of an approximation to the marginal log-likelihood at each iteration of an algorithm for maximum likelihood estimation. Certain classes of such models are well supported by current statistical software packages, but the potential for flexible model development outside of the standardised options provided is usually limited.

The use of 'non-linear mixed effects models' is well established in biomedical research[4–7]. In these models, the distribution of the outcome variable is normal conditional on the random effect terms, but the function for the expectation of the outcome variable may be non-linear in both the fixed effect parameters and the random effect terms. Ignoring the potential for grouping factors beyond that of each individual in a dataset, as previously in this chapter, the non-linear mixed model for longitudinal data can be expressed in the form[4]:

$$\begin{aligned}\mathbf{y}_i &= \mathbf{g}\left(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{b}_i\right) + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\ \mathbf{e}_i &\sim MVN(\mathbf{0}, \mathbf{R}_i),\end{aligned}$$

where $\mathbf{y}_i$ is the outcome vector for the i^{th} individual, $\mathbf{X}_i$ is a matrix of covariate data relating to the n_i observations for that individual, $\boldsymbol{\beta}$ is a vector of parameters relating to the expectation for each observations — some of which are associated with subject-specific random effects denoted by the vector $\mathbf{b}_i$, $\mathbf{e}_i$ is a vector of residuals and $\mathbf{g}()$ is a vectorised non-linear function. As for the linear mixed effects model, $\mathbf{b}_i$ and $\mathbf{e}_i$ each follow independent multivariate normal distributions with covariance matrices $\boldsymbol{\Psi}$ and $\mathbf{R}_i$, respectively. The matrix $\mathbf{R}_i$ is often assumed to comprise independent error terms with constant variance (i.e. $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$), but as for the linear mixed effects model the structure of $\mathbf{R}_i$ can be specified to reflect correlated error terms or the incorporation of Gaussian process components into the model. The marginal likelihood function for each individual can be expressed as:

$$f\left(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \int_{-\infty}^{\infty} f_{cond}\left(\mathbf{y}_i|\mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) f_{\mathbf{b}}\left(\mathbf{b}|\boldsymbol{\alpha}\right) d\mathbf{b}, \tag{6}$$

for which the right-hand side is a multidimensional integral if there is more than one

random effect term per individual. We use $\boldsymbol{\alpha}$ here to denote a vector containing all variance parameters, relating to both $\boldsymbol{\Psi}$ and $\mathbf{R}_i$. Unless $\mathbf{g}()$ is linear in the random effect terms, this integral does not have a closed form solution.

An in-depth review of methods to approximate integrals of this form was published by Pinheiro and Bates[5], and we do not provide a full account of the history of this topic here but rather outline key features of the algorithms that are currently in use. The simplest method available to approximate an integral as that in (6) is to use a first-order Taylor series expansion of the model function $\mathbf{g}()$ around the expected value of the random effect terms[51], termed the 'first order' (FO) method in the NONMEM software[52]. An alternative option proposed by Lindstrom and Bates[4] is to use a first-order Taylor expansion of the model function around the current parameter estimates and the conditional (on current parameter values) modes of the random effects; maximum likelihood estimation using this approach requires alternation between a penalised least squares step to calculate the conditional modes of the random effects and a 'linear mixed effects' step, in which updated parameter estimates are obtained using the resulting linear approximation to the non-linear model. This approach is implemented in the 'nlme' package[34] for R, and so our 'covBM' package can be used to incorporate stochastic process components in non-linear mixed effects models using this methodology. However, there exist other methods that provide a more accurate approximation to the marginal likelihood, at the cost of higher computational complexity.

The next available method, in terms of increasing accuracy, is the Laplace approximation. Before formalisation of non-linear mixed effects models, the Laplace approximation had been used to obtain approximate marginal posterior densities and predictive distributions in the context of Bayesian analyses (e.g. Tierney and Kadane[53]), and Wolfinger[54] noted that the Laplace approximation of the integral could be used to derive the linear expansion of the model for the REML form of the Lindstrom and Bates algorithm, assuming flat prior distributions for the fixed effect parameters. It was subsequently observed by Pinheiro and Bates[5] and Vonesh[55] that the Laplace approximation of the integral only with respect to the random effect terms could be used to carry out maximum likelihood estimation. The Laplace approximation is based on a second-order Taylor expansion with respect to the variables that are being integrated out of the expression, but both Pinheiro and Bates[5] and Vonesh[55] proposed a first-order approximation to the required matrix of second-order derivatives in order to simplify calculations; this is also the basis for the first-order conditional estimation method (FOCE) implemented in NONMEM[52]. However, in Chapters 5, 6 and 7 we will make use of the full Laplace approximation to integrals as shown in (6), and to integrals of a more complex form, and so we provide an outline of its derivation here[56].

We consider the need to integrate a joint probability density function of vectors of outcome and latent variables conditional on a full parameter vector ($f_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)$, with joint penalized log-likelihood $l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)$) in order to obtain the marginal probability density for the outcome variable alone ($f\left(\mathbf{y}_i|\boldsymbol{\theta}\right)$):

$$\begin{aligned} f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) &= \int_{-\infty}^{\infty} f_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right) d\mathbf{b} \\ &= \int_{-\infty}^{\infty} \exp\left(l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)\right) d\mathbf{b}. \end{aligned} \tag{7}$$

The joint probability density functions for the outcome and latent variables often takes the factorised form as shown in (6) (i.e. $f_{cond}\left(\mathbf{y}_i|\mathbf{b},\boldsymbol{\theta}\right) f_{\mathbf{b}}\left(\mathbf{b}|\boldsymbol{\theta}\right)$), and some software packages place this restriction on the user in their dedicated non-linear mixed model functions with the condition that the random effects follow a normal distribution (e.g. PROC NLMIXED in SAS or the 'nlme' package in R). However, this particular factorisation is not required in order for the Laplace approximation to the marginal likelihood to be used, and indeed we develop models that do not follow this form in Chapters 5, 6 and 7. The Laplace approximation to (7) results from a second-order Taylor expansion of $l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)$ about the conditional modes ($\hat{\mathbf{b}}$) of the latent variables, i.e. the values of the latent variables that maximise the penalised log-likelihood conditional on the data and the current parameter values:

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmax}}\, l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right).$$

Because the Taylor series is evaluated at the conditional maximum of the function, the first-order derivatives are zero and as such the joint penalised log-likelihood can be approximated as:

$$l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right) \approx l_j\left(\mathbf{y}_i,\hat{\mathbf{b}}|\boldsymbol{\theta}\right) + \frac{1}{2}\left(\mathbf{b}-\hat{\mathbf{b}}\right)^{\mathrm{T}} \mathbf{H}\left(\boldsymbol{\theta}\right)\left(\mathbf{b}-\hat{\mathbf{b}}\right),$$

where $\mathbf{H}\left(\boldsymbol{\theta}\right)$ is the matrix of second-order derivatives, i.e. the Hessian:

$$\mathbf{H}\left(\boldsymbol{\theta}\right) = \left.\frac{\delta^2}{\delta\mathbf{b}^2} l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)\right|_{\mathbf{b}=\hat{\mathbf{b}}}.$$

$\mathbf{H}\left(\boldsymbol{\theta}\right)$ is negative-definite, again because it is evaluated at the conditional maximum of the latent variable terms, and so the integral in (7) can be approximated as:

$$\begin{aligned} \int_{-\infty}^{\infty} \exp\left(l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)\right) d\mathbf{b} &\approx \exp\left(l_j\left(\mathbf{y}_i,\hat{\mathbf{b}}|\boldsymbol{\theta}\right)\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\mathbf{b}-\hat{\mathbf{b}}\right)^{\mathrm{T}}\left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)\left(\mathbf{b}-\hat{\mathbf{b}}\right)\right) d\mathbf{b} \\ &= \exp\left(l_j\left(\mathbf{y}_i,\hat{\mathbf{b}}|\boldsymbol{\theta}\right)\right)(2\pi)^{\frac{d}{2}}\left|\det\left(\mathbf{H}\left(\boldsymbol{\theta}\right)\right)\right|^{-\frac{1}{2}}, \end{aligned}$$

where d is the dimension of the integral, i.e. the number of latent variables terms.

This approach effectively approximates the posterior distribution of the latent variable terms, conditional on the model parameters and data, as a multivariate normal distribution. As for the Lindstrom and Bates[4] approach, obtaining maximum likelihood estimates of the parameters for a given model requires an algorithm that alternates between finding the modes of the latent variables conditional on the current parameter vector and then updating the parameter estimates based on the resulting Laplace approximation to the marginal (log)-likelihood. An automated computational approach to achieve this is reported by Skaug and Fournier[57]; this is implemented in the ADMB software[50], using which the user needs only to specify the function $l_j\left(\mathbf{y}_i, \mathbf{b}|\boldsymbol{\theta}\right)$ with respect to the relevant data, parameters and latent variables.

The use of Gauss–Hermite quadrature to approximate the marginal likelihood for non-linear models was proposed by Davidian and Gallant[58] in conjunction with a smooth non-parametric distribution for the random effect terms. Pinheiro and Bates[5] described the use of Gauss–Hermite quadrature for models in which the random effects follow a multivariate normal distribution, involving evaluation of the joint likelihood function at a grid of values for the random effect terms determined by their modelled marginal distribution. Pinheiro and Bates[5] also proposed an improved technique in this setting that they termed 'adaptive Gaussian quadrature', in which the grid of evaluation points is determined by the approximate multivariate normal posterior distribution of the random effect terms. We make use of this technique in Chapters 5 and 6 and so provide an outline of the method here, although we use the term 'adaptive Gauss–Hermite quadrature' to refer to this method throughout as this nomenclature is more specific.

Both Davidian and Gallant[58] and Pinheiro and Bates[5] demonstrated that Gauss–Hermite quadrature for d-dimensional integrals could be simplified by transformation into a series of 1-dimensional integrals. As we make use of adaptive Gauss–Hermite quadrature, this transformation takes the form $\mathbf{b} = \hat{\mathbf{b}} + \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-\frac{1}{2}}\mathbf{z}$, and leads to the following approximation:

$$
\begin{aligned}
f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) &= \int_{-\infty}^{\infty} \phi\left(\mathbf{b}; \hat{\mathbf{b}}, \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-1}\right) \frac{\exp\left(l_j\left(\mathbf{y}_i, \mathbf{b}|\boldsymbol{\theta}\right)\right)}{\phi\left(\mathbf{b}; \hat{\mathbf{b}}, \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-1}\right)} d\mathbf{b} \\
&= \int_{-\infty}^{\infty} \phi\left(\mathbf{z}; \mathbf{0}, \mathbf{I}\right) \left|\det\left(\left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-\frac{1}{2}}\right)\right| \frac{\exp\left(l_j\left(\mathbf{y}_i, \hat{\mathbf{b}} + \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-\frac{1}{2}}\mathbf{z}|\boldsymbol{\theta}\right)\right)}{\phi\left(\mathbf{z}; \mathbf{0}, \mathbf{I}\right)} d\mathbf{z} \\
&= \int_{-\infty}^{\infty} \phi\left(\mathbf{z}; \mathbf{0}, \mathbf{I}\right) \left|\det\left(\mathbf{H}\left(\boldsymbol{\theta}\right)\right)\right|^{-\frac{1}{2}} \left(2\pi\right)^{\frac{d}{2}} \exp\left(\frac{\mathbf{z}^{\mathrm{T}}\mathbf{z}}{2}\right) \exp\left(l_j\left(\mathbf{y}_i, \hat{\mathbf{b}} + \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-\frac{1}{2}}\mathbf{z}|\boldsymbol{\theta}\right)\right) d\mathbf{z} \\
&\approx \sum_{j_1=1}^{N_{GQ}} \ldots \sum_{j_d=1}^{N_{GQ}} \left|\det\left(\mathbf{H}\left(\boldsymbol{\theta}\right)\right)\right|^{-\frac{1}{2}} \left(2\pi\right)^{\frac{d}{2}} \exp\left(\frac{\mathbf{z}_{\mathbf{j}}^{\mathrm{T}}\mathbf{z}_{\mathbf{j}}}{2}\right) \exp\left(l_j\left(\mathbf{y}_i, \hat{\mathbf{b}} + \left(-\mathbf{H}\left(\boldsymbol{\theta}\right)\right)^{-\frac{1}{2}}\mathbf{z}_{\mathbf{j}}|\boldsymbol{\theta}\right)\right) \prod_{k=1}^{d} w_{j_k},
\end{aligned}
$$

where N_{GQ} is the number of quadrature points in each dimension (d) and $\mathbf{z_j}$ ($=\left(z_{j_1},\ldots,z_{j_d}\right)^{\mathrm{T}}$) and w_j ($j=1,\ldots,N_{GQ}$) denote the abscissa and weight values for the (one-dimensional) Gauss–Hermite quadrature rule for N_{GQ} points based on the standard normal kernel[5]. Adaptive Gauss–Hermite quadrature requires function evaluation at N_{GQ}^d points to approximate the marginal likelihood, and so is often not computationally feasible when the number of latent variable terms per individual cluster is more than one or two. However, when it can be applied, the number of quadrature points can be gradually increased until a stable value for the marginal likelihood is achieved, ensuring that it has been calculated accurately. This method is also implemented in the ADMB software, with the user only required to define the function $l_j\left(\mathbf{y}_i,\mathbf{b}|\boldsymbol{\theta}\right)$, as for the Laplace approximation, and the number of quadrature points.

Pinheiro and Bates[5] note that one-point adaptive quadrature (i.e. $N_{GQ}=1$) is equal to the Laplace approximation, as $z_j=0$ and $w_j=1$ for all j. They also note that the method is equivalent to the use of importance sampling to estimate the marginal likelihood, with sampling performed using the approximate normal posterior of the latent variables, but with a pre-specified set of sample points and weights. Importance sampling provides a potential alternative technique for maximum likelihood estimation when adaptive Gauss–Hermite quadrature is not feasible and is implemented in the ADMB software using a randomly generated (but then fixed) set of standard normal sample points for each optimisation, which are transformed to the approximate posterior distribution of the latent variables at each iteration of the algorithm. However, computational problems can arise when the total number of latent variables in the model is high.

An alternative approach to maximum likelihood estimation in non-linear mixed effects models is the stochastic approximation expectation-maximization (SAEM) algorithm developed by Kuhn and Lavielle[59] and implemented in the MONOLIX software[60]. This algorithm involves stochastic approximation of the latent variable terms in a model (updated at each iteration based on their posterior distribution) followed by updating of the conditional parameter estimates. It has only been proven to provide convergence to the maximum likelihood estimates of parameters when the joint model for the data and latent variable terms follows a distribution from the exponential family[59], which would make interpretation of parameter estimates difficult for models in which this condition is not met. Additionally, current implementations are restricted to the factorised form for the marginal likelihood as shown in (6).

2.5 Estimation for combinations of multivariate normal and multivariate-t distributions

In some settings, it may be desirable to fit statistical models that involve a mixture of components that follow multivariate normal and multivariate-t distributions. For example, Wakefield proposed the use of bivariate-t distributed subject-specific random effects in combination with normally distributed measurement error terms in the development of pharmacokinetic models[61]. Wakefield employed a Bayesian model fitting approach, and made use of the hierarchical form of the multivariate-t distribution in order to structure the desired model. Song *et al.* proposed the use of linear mixed effects models in which either the subject-specific random effects follow a multivariate-t distribution and the residual error terms follow a multivariate normal distribution or *vice versa*[62]; maximum likelihood estimation of these models was performed using a 'maximization by parts' algorithm[63], which has not been incorporated into software for general use.

In Chapters 5, 6 and 7 we consider novel extensions to mixed effects models in which only the stochastic process component follows a multivariate-t distribution whilst the random effect and residual error terms remain multivariate normal. Following the model structure as in (4) (page 22) and employing the hierarchical form of the multivariate-t distribution as in (5) (page 31), this gives a model of the form:

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_i + \mathbf{e}_i \\
\mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\
\mathbf{W}_i|_{\gamma_i} &\sim MVN(\mathbf{0}, \frac{1}{\gamma_i}\boldsymbol{\Sigma}_i) \\
\mathbf{e}_i &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}) \\
\gamma_i &\sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).
\end{aligned}$$

The marginal likelihood function for the model can be found by integrating out the latent variables on the standard normal scale, for which the Laplace approximation is optimally accurate[64], and so we fit such models using the following form for the marginal likelihood:

$$f\left(\mathbf{y}_i\right) = \int_{-\infty}^{\infty} f\left(\boldsymbol{y}_i|\gamma_i = F^{-1}\left(\Phi\left(a\right)\right)\right) f_\phi\left(a\right) da,$$

where f_ϕ and Φ are the probability density and cumulative probability functions for a standard normal distribution and F^{-1} is the inverse of the cumulative distribution function for a gamma distribution with 'shape' and 'rate' parameters both equal to $\frac{\nu}{2}$.

2.6 Discussion

For linear mixed effects models, even with extensions to include Gaussian stochastic process components or generalisation of the framework to the multivariate-t distribution, the marginal likelihood function is available in closed form and this allows direct application of optimisation algorithms to obtain maximum likelihood estimates of the model parameters. For conventional linear mixed effects models, methodological researchers have exploited the properties of the multivariate normal distribution to improve the computational efficiency of the optimisation, such as by profiling out a subset of the models parameters from the likelihood function. We have demonstrated a novel strategy for implementing the profile likelihood methods of Pinheiro and Bates[6] for mixed effects models that also include Gaussian process components. We have also noted that a Brownian motion component can be added to linear mixed effect models using existing software for IGLS estimation.

For non-linear mixed effects models, and for other latent variable models that do not fall within standard categories, the marginal likelihood is not in general available in closed form, and so approximations of varying degrees of accuracy and computational complexity are required in order to perform maximum likelihood estimation of model parameters. We note that although many software packages are restrictive in the structure of non-linear mixed effects models that can be fitted, open source software is available that can be used to fit latent variable models for which the structure can be flexibly defined by the user. This functionality allows the development and implementation of models for longitudinal data that are tailored to both the structure of the data under investigation and to the key questions of interest that we hope to answer. In Chapters 5, 6 and 7 we make use of this flexibility to develop and apply novel statistical models for the combined analysis of pre- and post-treatment CD4 counts in HIV patients.

3 Residual diagnostics for mixed effects models

In this chapter we provide a brief review of the methods available for generating and evaluating residual diagnostic plots for fitted mixed effects models. We then review the methods that have been described for mixed effects models that make use of the multivariate-t distribution, and propose additional plots that could be used to assess model fit in this setting.

Several different approaches have been proposed for assessing the goodness of fit of linear mixed models with respect to their implicit assumptions, including the use of diagnostic plots for model residuals that have been generalised from those used in the standard linear regression framework. It is worth noting, as pointed out by Haslett and Haslett[65], that inconsistent terminology has been used in the literature regarding the different types of 'residuals' that can be obtained for linear mixed models; as such we have italicised the definitions that we have chosen to use in this discussion. The set of *marginal residuals*[2] for a fitted linear mixed effects model is given by:

$$\hat{\boldsymbol{r}}_i^{marg} = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}.$$

We use mathematical notation as defined in Chapter 2.

As the linear mixed model does not assume that residual error terms are independent and identically distributed, the marginal residuals alone cannot be used to assess the adequacy of model fit. However, their distribution conditional on known parameter values will always be multivariate normal. For non-linear mixed effects models, the marginal residuals can be calculated as:

$$\hat{\boldsymbol{r}}_i^{marg} = \mathbf{y}_i - \mathbf{g}\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}, \mathbf{b}_i = \mathbf{0}\right).$$

However, the distribution of these marginal residuals conditional on the model parameters will not follow a multivariate normal distribution unless the function $\mathbf{g}$ is linear in any random effect terms[4].

3.1 Division of random effects and residual error for linear mixed effects models

When the fitted model contains a constant residual error term and no stochastic process component, one option is to calculate *subject-specific residuals*[2] based on the predicted values of the random effects for each individual (or other grouping of fac-

tors):

$$\hat{\boldsymbol{r}}_i^{ss} = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i$$

where, $$\hat{\mathbf{b}}_i = \hat{\boldsymbol{\Psi}}\mathbf{Z}_i^{\mathrm{T}}\hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

The random effects in a linear mixed model are assumed to be unobserved variables arising from a normal distribution. This method of prediction for the random effects corresponds to the mean of the posterior distribution of $\mathbf{b}_i$ given the observed data $\mathbf{y}_i$ and taking the parameter estimates for the linear mixed model as fixed. As such, the resulting predictions for $\mathbf{b}_i$ are referred to as 'empirical Bayes' estimates[1].

This approach is appealing as it may initially seem, given that a standard linear mixed model has been correctly specified, that the calculated subject-specific residuals should be approximately independent and normally distributed with constant variance; this would be the case if all of the realisations of random effects in the model could be predicted very accurately, and would allow standard techniques for residual diagnostics such as quantile–quantile (Q–Q) plots to be employed. Such an approach is described by Pinheiro and Bates[6], who also propose a further generalisation of this technique for use with models that include a non-constant residual variance by transforming the subject-specific residuals using the inverse Cholesky root of the estimated covariance structure additional to that induced by the random effects (i.e. $\mathbf{R}_i$ as in (1), page 20, with decomposition: $\hat{\mathbf{R}}_i = \hat{\sigma}^2\hat{\mathbf{R}}_i' = \hat{\sigma}^2\hat{\boldsymbol{\Lambda}}_i\hat{\boldsymbol{\Lambda}}_i^T$):

$$\hat{\boldsymbol{r}}_i^{ss*} = \hat{\sigma}^{-1}\hat{\boldsymbol{\Lambda}}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i\right).$$

However, as reported by Jacqmin-Gadda *et al.*[40], it can be shown that if $\mathbf{R}_i = \sigma^2\mathbf{I}$ then the subject-specific residuals can be expressed as:

$$\hat{\boldsymbol{r}}_i^{ss} = \hat{\sigma}^2\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right).$$

Hence, even if all of the parameters in the model were known, then the subject-specific residuals would be distributed as:

$$\boldsymbol{r}_i^{ss} \sim MVN\left(\mathbf{0}, \sigma^4\mathbf{V}_i^{-1}\right),$$

which means that interpretation of such residual plots is not straightforward.

In addition to their use for the calculation of subject-specific residuals, the $\hat{\mathbf{b}}_i$ themselves are often used to detect outlying groups or individuals or to assess the goodness of fit of the model with respect to its implicit assumptions. For example, DeGruttola *et al.* reported the use of 'random slopes' models for the progression of T-cell counts in HIV-positive men and used histograms and Q–Q plots of the pre-

dicted values of the random effects to evaluate whether the assumption of a normal distribution was met[66]. This type of analysis is widely facilitated by currently available software, including the 'nlme' package for R and MLwiN. However, the use of diagnostic methods that rely on the prediction of random effects requires caution, as the predictions $\hat{\mathbf{b}}_i$ of $\mathbf{b}_i$ will always show 'shrinkage' towards zero (i.e. the population average)[67]. A related result, as noted by Verbeke and Molenberghs[2](page 81), is that for any linear combination of the random effects, $\mathbf{a}$, the following inequality holds:

$$\mathrm{Var}\left(\mathbf{a}^{\mathrm{T}}\hat{\mathbf{b}}_i\right) \leq \mathrm{Var}\left(\mathbf{a}^{\mathrm{T}}\mathbf{b}_i\right).$$

This complicates any interpretation of diagnostic procedures for model fit that make use of the predicted values of random effects, whether to create subject-specific residuals or for analysis in themselves. Furthermore, it was demonstrated by Verbeke and Lesaffre that the assumed distribution of the random effects strongly affects the results of such analyses, and that even when data are generated with random intercept terms from a distinctly bimodal normal distribution the predicted values from a fitted model may appear unimodal[68]. Verbeke and Lesaffre also note that the distribution of $\hat{\mathbf{b}}_i$ will only be identical across subjects if $\mathbf{Z}_i$ is the same for every subject; this condition will be met in some situations, but in the biomedical setting there will very often be subjects with missing or irregularly collected data and highly unbalanced datasets are common in the case of observational studies.

Attempts have been made to address these difficulties in assessing the goodness of fit of linear mixed models. For example, Hilden-Minton suggested finding a linear transformation of the subject-specific residuals that minimises the influence of the random effects in the model, terming these the *least confounded residuals*[69]. Nobre and Singer used simulation to show that a standardised form of the least confounded residuals could be used in combination with Q–Q plots to correctly indicate a normally distributed residual error term when an associated random error term was generated from a non-normal distribution[70]. This technique has the inherent limitations that the transformed residuals no longer correspond to individual observations, making interpretation more difficult, and that the distribution of the random effects themselves are not assessed.

Another approach to the problem of confounding between different components of the residuals in linear mixed models is to generate tolerance bands for diagnostic plots of subject-specific residuals and predictions of random effects using a parametric bootstrap-type technique, as proposed by Schützenmeister and Piepho[71]. Using this method, data are resampled using the parameter estimates obtained and the model is refitted and corresponding residual diagnostics calculated for each of the generated datasets; these are then used to calculate tolerance bands under the as-

sumption of a correctly specified model for the residual values obtained from the original data. Schützenmeister and Piepho acknowledge that the simulation-based approach that they present can take a long period of time to run using standard computers, which will limit its use, certainly for models fitted to large datasets.

3.2 Transformation of marginal residuals

Instead of attempting to divide the observed marginal residuals of linear mixed models into components related to the random effects in the model and a residual error term, an alternative method is to transform them directly. In the context of analysis of longitudinal data, Fitzmaurice, Laird and Ware propose the use of a Cholesky decomposition of the estimated marginal covariance matrix for each individual for this purpose[72]. This suggestion followed previous work by Waternaux, Laird and Ware in which an equivalent transformation was generated following the fitting of a model for longitudinal data and used to transform the outcome and predictive variables, leading to an ordinary least squares optimisation in terms of the fixed effects parameters only that could be fitted and evaluated using standard linear regression procedures[73]. If $\mathbf{L}_i$ is a lower triangular matrix such that:

$$\hat{\mathbf{V}}_i = \mathbf{L}_i \mathbf{L}_i^{\mathrm{T}}.$$

Then we will term the *Cholesky-transformed residuals* as follows:

$$\hat{\boldsymbol{r}}_i^{Chol} = \mathbf{L}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right).$$

This transformation creates a standardised residual for the first observation in each individual and has the effect of standardising each subsequent observation for each individual conditional on all preceding observations, for the k^{th} observation in the i^{th} individual providing an estimate of:

$$\frac{\mathrm{y}_{ik} - \mathrm{E}\left(\mathrm{y}_{ik}|\mathrm{y}_{i1},\ldots,\mathrm{y}_{ik-1}\right)}{\sqrt{\mathrm{Var}\left(\mathrm{y}_{ik}|\mathrm{y}_{i1},\ldots,\mathrm{y}_{ik-1}\right)}}.$$

Given a correct model specification, these transformed residuals are asymptotically independent, each following a standard normal distribution. As such, the adequacy of model fit can be assessed by generating Q–Q plots of the transformed residuals, and by plotting them against Cholesky-transformed predicted mean values (i.e. $\mathbf{L}_i^{-1}\mathbf{X}_i\hat{\boldsymbol{\beta}}$) or covariates such as time. Haslett and Haslett[65] note that the Cholesky decomposition of the marginal covariance matrix for each individual is not the only decomposition available for the purpose of standardising the marginal residuals.

Houseman *et al.* propose the use of a transformation resulting from the Cholesky decomposition of the inverse of the marginal covariance matrix for each individual, and present a derivation of the standard error of an empirical cumulative distribution function for the transformed residuals, facilitating interpretation of Q–Q plots[74], while Louis made use of symmetric square-root of the covariance matrix[75].

An advantage of using a direct transformation of the marginal residuals in a linear mixed model is that the values obtained, under a correctly specified model, are not dependent on the number of observations or the covariate values of each individual in the dataset. This is particularly beneficial in the presence of missing data that are thought to be 'missing at random', with the probability of data being missing only dependent on the observed data and covariates. The Cholesky-transformed residuals only depend on the previous observations for that individual, and so are not influenced by the fact that later observations may be missing.

3.3 The semivariogram function

The semivariogram function is used to describe the spatial and/or temporal dependence in random fields and stochastic processes. The technique has its roots in the analysis of geological and geographical data, but was introduced in the context of longitudinal data by Diggle[24]. The function is of the form:

$$\begin{aligned}\zeta(s,t) &= \frac{1}{2}\mathrm{Var}\left(\left(\mathrm{y}_s - \mu(s)\right) - \left(\mathrm{y}_t - \mu(t)\right)\right) \\ &= \frac{1}{2}\mathrm{E}\left(\left|\left(\mathrm{y}_s - \mu(s)\right) - \left(\mathrm{y}_t - \mu(t)\right)\right|^2\right),\end{aligned}$$

where y_s and y_t denote the observed response variables at times s and t and $\mu(\cdot)$ is a function giving the expected value. If the observations under consideration are assumed to arise from a stationary process $\mathrm{Y}(t)$ of deviations from the mean, or from a stochastic process in which the increments are stationary, then the function can be expressed in terms of time lag u:

$$\zeta(u) = \frac{1}{2}\mathrm{E}\left(|\mathrm{Y}(t) - \mathrm{Y}(t-u)|^2\right).$$

If $\mathrm{Y}(t)$ is stationary, then the semivariogram function is directly related to the autocorrelation function, $\rho(u)$, of the process[3]:

$$\zeta(u) = \sigma^2\left(1 - \rho(u)\right).$$

As such, sample semivariograms can be used to suggest plausible structures for the correlation of residual errors under this assumption. These are produced as a smoothed plot of the half-squared-differences between marginal residuals (v_{ijk}) for

each pair of observations in each individual, plotted against the time difference between observations (u_{ijk}) in each pair[3]:

$$v_{ijk} = \frac{1}{2}\left(\hat{r}_{ij} - \hat{r}_{ik}\right)^2$$
$$u_{ijk} = t_{ij} - t_{ik}.$$

Whilst this technique can be useful in some circumstances, the required assumption that the underlying process is stationary can be limiting. One approach to this problem when dealing with more complex linear mixed models is to use the subject-specific residuals (based on the predicted values of the random effects for each individual) for calculation of the empirical semivariogram, with the aim of evaluating the correlation between residual variation once the random effects have been taken into account; this is suggested by Pinheiro and Bates[6], and is implemented in the 'nlme' package for R. However, we demonstrate in Section 3.3.1 that this approach can lead to potentially misleading results even when the statistical model is correctly specified.

Another approach to allow the use of the semivariogram to evaluate residual correlation in linear mixed models was described by Verbeke *et al.*[76]. These authors propose the use of a transformation of the ordinary least squares (OLS) residuals that creates a projection of the residuals that is orthogonal to the random effects design matrix for each subject (i.e. $\mathbf{Z}_i$), and which can be used to obtain a semivariogram plot that does not depend on the random effects part of the model. One problem with this approach is that it is based on the OLS estimates of the fixed effects parameters being consistent, which may not be the case in the presence of missing data. Another limitation is that this approach is not implemented in any statistical software packages.

Alternatively, Fitzmaurice, Laird and Ware[72] propose the use of the sample semivariogram to check the independence of Cholesky-transformed residuals. Under a correctly specified model, these residuals are asymptotically normally and independently distributed with variance 1 and so a sample semivariogram should show random scatter around 1 as a function of time lag between observations. However, a limitation of this approach is that it is not obvious how to interpret systematic deviations from the expected pattern, other than to conclude that the covariance structure of the model does not perfectly describe that observed in the data.

3.3.1 Semivariogram for subject-specific residuals

We consider here the use of a sample semivariogram to assess the distribution of subject-specific residuals in a correctly specified 'random slopes' linear mixed model,

with independent residual errors of constant variance. To simplify matters, we consider the model parameters to be known. We denote σ^2 as the residual variance, with random effects following a bivariate normal distribution:

$$\begin{pmatrix} \mathrm{b}_0 \\ \mathrm{b}_1 \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Psi}\right),$$

where,

$$\boldsymbol{\Psi} = \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix}.$$

In this context, the theoretical semivariogram function for observations of subject-specific residuals at time points s and t, with $t - s = u$, is given by:

$$\begin{aligned} 2\zeta^*(s,t) &= \mathrm{Var}\left(\left(\mathrm{y}_t - \mu(t) - \hat{\mathrm{b}}_0 - t\hat{\mathrm{b}}_1\right) - \left(\mathrm{y}_s - \mu(s) - \hat{\mathrm{b}}_0 - s\hat{\mathrm{b}}_1\right)\right) \\ &= \mathrm{Var}\left(\left(\mathrm{y}_t - \mu(t)\right) - \left(\mathrm{y}_s - \mu(s)\right) - u\hat{\mathrm{b}}_1\right) \\ &= \mathrm{Var}\left(\mathrm{y}_t\right) - 2\mathrm{Cov}\left(\mathrm{y}_t, \mathrm{y}_s\right) + \mathrm{Var}\left(\mathrm{y}_s\right) - 2u\mathrm{Cov}\left(\mathrm{y}_t, \hat{\mathrm{b}}_1\right) + 2u\mathrm{Cov}\left(\mathrm{y}_s, \hat{\mathrm{b}}_1\right) + u^2\mathrm{Var}\left(\hat{\mathrm{b}}_1\right) \\ &= 2\sigma^2 + \left(s^2 + t^2 - 2st\right)\sigma_{11} - 2u\mathrm{Cov}\left(\mathrm{y}_t, \hat{\mathrm{b}}_1\right) + 2u\mathrm{Cov}\left(\mathrm{y}_s, \hat{\mathrm{b}}_1\right) + u^2\mathrm{Var}\left(\hat{\mathrm{b}}_1\right). \end{aligned}$$

If we let $\mathbf{y}$ be the vector of outcome variables, $\mathbf{B}$ be a row vector of zeros with a single '1' corresponding to the position of time point t_i, $\mathbf{Z}$ be the design matrix for the random effects and $\mathbf{V}$ be the marginal covariance matrix of $\mathbf{y}$, then the term $\mathrm{Cov}\left(\mathrm{y}_{t_i}, \hat{\mathrm{b}}_1\right)$ is given by the second element of:

$$\begin{aligned} \mathrm{Cov}\left(\boldsymbol{\Psi}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}, \mathbf{B}\mathbf{y}\right) &= \boldsymbol{\Psi}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}\mathrm{Cov}\left(\mathbf{y}, \mathbf{y}\right)\mathbf{B}^{\mathrm{T}} \\ &= \boldsymbol{\Psi}\begin{pmatrix} 1 & \dots & 1 & \dots & 1 \\ t_1 & \dots & t_i & \dots & t_n \end{pmatrix}\begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \\ &= \boldsymbol{\Psi}\begin{pmatrix} 1 \\ t_i \end{pmatrix}. \end{aligned}$$

Therefore

$$\mathrm{Cov}\left(\mathrm{y}_{t_i}, \hat{\mathrm{b}}_1\right) = \sigma_{01} + t_i\sigma_{11}.$$

Using this expression, we have:

$$\begin{aligned}2\zeta^*(s,t) &= 2\sigma^2 + \left(s^2+t^2-2st\right)\sigma_{11} - 2(t-s)(\sigma_{01}+t\sigma_{11}) + 2(t-s)(\sigma_{01}+s\sigma_{11}) + u^2\mathrm{Var}\left(\hat{\mathrm{b}}_1\right)\\ &= 2\sigma^2 - \left(s^2+t^2-2st\right)\sigma_{11} + u^2\mathrm{Var}\left(\hat{\mathrm{b}}_1\right)\\ &= 2\sigma^2 - u^2\left(\sigma_{11} - \mathrm{Var}\left(\hat{\mathrm{b}}_1\right)\right).\end{aligned}$$

Following from the fact that $\mathrm{Var}\left(\mathbf{a}^{\mathrm{T}}\hat{\mathbf{b}}\right) \leq \mathrm{Var}\left(\mathbf{a}^{\mathrm{T}}\mathbf{b}\right)$ for any linear combination $\mathbf{a}$, and hence $\mathrm{Var}\left(\hat{\mathrm{b}}_1\right) \leq \sigma_{11}$, it can be seen that the semivariogram function in this situation quadratically decreases with respect to the time difference between observations. This characteristic of the semivariogram approach when using subject-specific residuals is not discussed by Pinheiro and Bates[6].

3.4 Residual diagnostics for multivariate-t linear mixed effects models

The evaluation of diagnostic plots of the residuals resulting from fitted statistical models forms an important part of model criticism and development. Such plots can be used to check the adequacy of fitted models to describe the data under investigation and, when problems are observed, to suggest how further improvements might be made. This is particularly important when there is interest in understanding patterns of variability within and between individuals as well as ensuring correct inference for fixed effects parameters.

3.4.1 Subject-level residuals

Much of the focus regarding the use of multivariate-t linear mixed effects models has been with respect to providing robust inference for the fixed effects; this follows from the fact that individuals with observations that are further from the mean in each case are down-weighted in the estimation of the fixed effects parameters. Lange *et al.* were concerned with achieving robust multivariate regression, and suggested the use of diagnostic residual plots that indicated whether the fitted model adequately reflected the presence of outlying sets of measurements (i.e. each set corresponding to the various measurements conducted on a single individual)[41]. They point out that for a normal linear mixed model, the statistic:

$$\hat{\delta}_i^2(\boldsymbol{\theta}) = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)$$

for each individual would asymptotically follow a χ^2 distribution with n_i degrees of freedom. However, under a multivariate-t model, the statistic $\frac{\hat{\delta}_i^2(\boldsymbol{\theta})}{n_i}$ would asymptotically follow an F-distribution with n_i and $\hat{v}$ degrees of freedom. Lange *et al.* trans-

form these statistics to standard normal deviates, and then use Q–Q plots to assess model fit.

Pinheiro *et al.*[43] suggest direct plotting of the standardised sum of squares for each individual $\frac{\hat{\delta}_i^2(\boldsymbol{\theta})}{n_i}$, alongside a decomposition of the standardised sum of squares into that predicted to be due to the random effects terms:

$$\frac{\hat{\delta}^2_{\mathbf{b}_i}(\boldsymbol{\theta})}{q} = \frac{\hat{\mathbf{b}_i}^{\mathrm{T}} \hat{\boldsymbol{\Psi}}^{-1} \hat{\mathbf{b}_i}}{q},$$

where q is the length of the random effects vector, and that predicted to be due to residual error:

$$\frac{\hat{\delta}^2_{\mathbf{e}_i}(\boldsymbol{\theta})}{n_i} = \frac{\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i\right)^{\mathrm{T}} \hat{\mathbf{R}}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\hat{\mathbf{b}}_i\right)}{n_i}.$$

In the example given by Pinheiro *et al.*[43], these statistics for each individual are calculated using the parameter estimates from a multivariate-t model, but are compared to their asymptotic expected value of 1 under the equivalent Gaussian model. This technique may be of use in identifying unusual individuals in a dataset, but does not seem to directly address the adequacy of the multivariate-t model to describe the data. In addition, as described in Section 3.1, any method that relies on the predicted values of the random effects in each individual may result in misleading findings.

Wang and Fan report the use of a hybrid ECME–scoring approach to fit a multivariate-t model for joint modelling of longitudinal observations of CD4 and CD8 cell counts in a sample of 30 HIV-positive individuals[9]. Their model included random intercept and random slope components, with autoregressive correlation for the residual error terms. To assess model fit, Wang and Fan provide Q–Q plots equivalent to those suggested by Lange *et al.*[41], summarizing the distribution of the standardised distance from the expected mean values for the set of observations obtained from each individual as a whole. As demonstrated by Lange *et al.*[41] and Wang and Fan[9], these plots can demonstrate the inadequacy of the normal linear mixed effects model to describe the observed data. However, the plots do not directly show whether the multivariate-t model correctly describes variability between individual measurements.

Wang and Fan[9] also propose plotting quantiles of the empirical conditional distributions of the γ_i for each individual, with those for which the 95 % credible interval does not include the mean of 1 suggested as potential outliers in the population. In their sample dataset this technique flags up 9 out of 30 patients as potential outliers, suggesting that this method may not perform ideally. Indeed, use of the multivariate-t model is making the assumption that the γ_i values differ between individuals, and

the more data is gathered per individual the narrower (and less likely to contain 1) the empirical credible interval for each individual will be; this would be a particular problem for interpretation when analysing unbalanced datasets with varying numbers of observations per individual.

3.4.2 Measurement-level residuals

We propose that the gamma–normal formulation of the multivariate-t linear mixed model, as given in (5) (page 31), can be also used to assess whether the multivariate-t distribution fully describes the patterns of variability observed for all individual measurements in a dataset. This would be important when the motivation for an analysis is to be able to make predictions regarding future individual measurements or to simulate datasets in which the exact pattern of values within each individual is important. As the observations for the i^{th} individual are assumed to follow a multivariate normal distribution conditional on γ_i, one option is to use empirical Bayes estimates (i.e. the mean of the predicted posterior distribution) of the γ_i:

$$\hat{\gamma}_i = \frac{\hat{\upsilon} + n_i}{\hat{\upsilon} + \hat{\delta}_i^2(\boldsymbol{\theta})}$$

to estimate the normal covariance matrix ($\hat{\mathbf{V}}'_i$) for each individual:

$$\begin{aligned}\hat{\mathbf{V}}'_i &= \frac{1}{\hat{\gamma}_i}\hat{\mathbf{V}}_i \\ &= \mathbf{L}'_i\mathbf{L}'^{\mathrm{T}}_i.\end{aligned}$$

This could then be used to transform the marginal residuals for the i^{th} individual as for a normal linear mixed model (i.e. using the inverse of a Cholesky decomposition), with the transformed residuals for all individuals displayed in a Q–Q plot. However, assuming the empirical Bayes estimates of the γ_i to be correct for all individuals might result in misleading conclusions in a similar manner to that which can be observed when evaluating the empirical Bayes estimates of random effects in a normal linear mixed model (e.g. as reported by Verbeke and Lesaffre[68]). An alternative would be to draw a number of repeated samples from the predicted posterior distribution of the full vector of $\boldsymbol{\gamma}$, using each sample to generate a full set of $\hat{\mathbf{V}}'_i$ matrices and corresponding Cholesky-transformed marginal residuals. The sets of transformed marginal residuals could then be used individually to generate multiple Q–Q plots, or used together to derive a single Q–Q plot showing the distribution of 'observed quantiles' over multiple realisations of the $\boldsymbol{\gamma}$. The sets of Cholesky-transformed marginal residuals based on the simulated values of $\boldsymbol{\gamma}$ could also be used to produce other standard diagnostic plots.

We also note that the strategies proposed could be used to generate plots of residuals for models that include a combination of multivariate-t and multivariate normal distributed components, as described in Section 2.5. In such cases, the posterior predictive distribution of the latent scaling variables $\boldsymbol{\gamma}$ is not available in closed form. However, the ADMB software that we use provides the mean and covariance matrix of the multivariate normal approximation to the posterior distribution of any latent variable terms included in a fitted model.

The concept of using multiple simulated samples of missing data and/or latent variables to assess the fit of Bayesian and/or hierarchical models has been particularly promoted by Gelman, for example in Gelman (2004)[77] and Gelman *et al.* (2005)[78]. Gelman *et al.*[78] point out that the concepts of 'missing' and 'latent' variables are closely connected, as illustrated by the way in which they are treated in EM-type approaches, and that the distinction between them can be seen to depend on their structural relation to the observed data included in the model; missing data have the same structure as observed data when imputed whereas latent variables do not.

The examples provided by Gelman[77] and Gelman *et al.*[78] focus on the use of graphical plots to compare observed sets of data with either fully simulated sets of data or 'completed' datasets based on the fitted model, with each imputation plotted as a separate graph or with summary statistics from multiple imputations displayed in a single graph. The examples given do not include the use of multiple imputations of latent variables to evaluate the possible distribution of standardised residuals as we propose, but this is a natural use of the model-checking framework, whereby the use of multiple imputations allows for an intuitive assessment of model fit.

The gamma–normal formulation of the marginal multivariate-t model provides another route to model-checking through the separate evaluation of each individual in the dataset. Assuming that the model parameters are known, then it can be seen that the distribution of the transformed marginal residuals using the inverse of the Cholesky decomposition of the scale matrix for each individual ($\mathbf{V}_i$, not adjusted for $\hat{\gamma}_i$) are normally and independently distributed with mean $\mathbf{0}$ and variance $\frac{1}{\gamma_i}$, conditional on the value of γ_i:

$$\mathbf{V}_i = \mathbf{L}_i \mathbf{L}_i^{\mathrm{T}}$$

$$\mathbf{L}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)|\gamma_i \sim MVN\left(\mathbf{0}, \frac{1}{\gamma_i}\mathbf{I}_{n_i}\right).$$

Hence, for a model that correctly describes the data, separate Q–Q plots (with respect to the standard normal distribution) of these transformed residuals for each individual should each indicate a normal distribution (with differing variance). For small datasets, it may be possible to create faceted graphics that simultaneously display the Q–Q plots for all individuals, but for larger datasets it would be necessary

to select a random sample of individuals for inspection. This approach will be more effective when there are a greater number of observations per individual, as it is difficult to assess the assumption of normality for very small samples. This reflects the fact that the presence of a greater number of observations per individual in a dataset will provide more information as to whether there truly is a difference in underlying variability between individuals, as represented by the values of γ_i.

3.5 Discussion

For linear mixed effects models, a range of techniques for the use of residuals to evaluate model fit have been developed. However, the best practice for the evaluation of model residuals has not been firmly established and the relevant terminology has not been fully standardised[65]. The options available for non-linear mixed effects models are more limited, due to the greater complexity of model structure.

Some techniques have been proposed to evaluate the residuals from models that make use of a marginal multivariate-t distribution, but we argue that the existing methods do not fully assess the fit of the model to the data. We propose two alternative methods for assessing the residuals from such models, one of which could also be generalised to models that combine components relating to both multivariate-t and multivariate normal distributions; these methods are applied in Chapters 4 and 5.

4 Application of the multivariate-t distribution with stochastic processes to pre-treatment CD4 counts

In this chapter we demonstrate a novel extension to the linear mixed effects models available for longitudinal biomedical data, combining the addition of a fractional Brownian motion component with generalisation of the model to follow a multivariate-t distribution. The model developed is applied to pre-treatment CD4 count data in HIV (type 1) patients, and background regarding the disease area and the development of statistical methods in this context is given in Section 4.1. The motivating dataset of pre-treatment CD4 counts from a multinational cohort collaboration used for analysis is introduced in Section 4.2, some further technical details of the model fitting process are given in Section 4.3 and the results of model-fitting and residual diagnostic procedures are presented in Section 4.4. Simulation studies informed by the results are described in Section 4.5, demonstrating differences in predictions made by the more complex models regarding the timing of treatment initiation in population cohorts and showing that the application of simpler models can lead to substantial bias in parameter estimates when there is censoring dependent on observed values of the outcome variable. Practical and methodological implications of the work are discussed in Section 4.6. The contents of this chapter form the basis for a publication in *Statistics in Medicine*[47], which is provided as Appendix C (reproduced under CC BY 4.0 license). We include here some additional residual diagnostic plots that are not included in the published paper.

4.1 Background

4.1.1 Monitoring of CD4 counts in HIV patients

CD4 cells are a type of white blood cell for which counts are monitored over time both before and after treatment initiation in HIV patients in order to evaluate the progress of the disease and state of the immune system[79]. Although the CD4 counts within an individual can vary erratically over time, on average the counts decline steadily from normal levels following HIV infection. A small minority of patients maintain high CD4 counts up to and beyond 10 years from the date of seroconversion[80] without initiating treatment.

Over the last 20 years, effective regimens of HAART have been developed for the treatment of HIV, allowing long-term management of the condition and greatly improving the life expectancy and quality of life of affected individuals, at least for those with the condition diagnosed in a resource-rich country. In most patients CD4 counts recover after the initiation of HAART, reaching levels within the normal range for non-infected people for the majority of patients with a baseline CD4 count of

≥ 350 cells/μL after a number of years on constant therapy[81]. The CD4 cell count is a strong predictor of subsequent progression to AIDS in both untreated HIV patients[82;83] and in patients initiating HAART[84;85].

Until recently, clinical guidelines regarding the initiation of treatment varied between countries. In the USA, the Health and Human Services Panel on Antiretroviral Guidelines for Adults and Adolescents have for a number of years recommended immediate initiation of antiretroviral therapy (ART) for most patients newly diagnosed with HIV[86], whereas in Europe guidelines recommended monitoring of CD4 count in most patients, with treatment initiated once this dropped below 350 cells/μL[87]. However, a recent randomised controlled trial (RCT) has provided definitive evidence of the benefit of immediate initiation of HAART on diagnosis of HIV[88], leading to a shift in clinical guidelines towards early treatment initiation in all well-resourced countries, including the UK[89]. Although CD4 counts will no longer be routinely monitored prior to the initiation of HAART in developed countries, there remains a motivation to better understand their pre-treatment dynamics in HIV-patients as for many patients there is a delay from infection to diagnosis and an improved understanding may also facilitate investigation of post-treatment recovery.

4.1.2 Estimation of seroconversion date in HIV patients

The term 'seroconversion' describes the appearance of HIV antibodies in a patient's blood. After the infection event there is a delay of 1–3 weeks before detectable viral RNA appears in the plasma, and following this it takes approximately another 2 weeks for HIV antibodies to reach a detectable level[90]. However, the interval from infection to a detectable level of antibodies was several weeks longer for the first generation of assays that were developed[91]. Some HIV-infected patients develop 'seroconversion illness', flu-like symptoms that occur during the period of seroconversion[92].

In a minority of patients, the timing of seroconversion can be accurately dated because they either presented with seroconversion illness or they underwent laboratory tests during the seroconversion period that definitively revealed a recent infection. However, in most patients diagnosed with HIV the exact timing of infection and subsequent seroconversion is not known. In order to enable modelling of the natural progression of HIV infection in larger cohorts of patients, many analyses make an assumption that seroconversion occurred at the mid-point between last negative and first positive diagnostic tests among those patients who were undergoing regular testing. The range of potential dates may cover an interval of months or years, depending on the frequency of testing in the population under investigation and on the inclusion criteria specified for any given analysis. As well as providing a larger sample size, the inclusion of patients with 'mid-point' estimation of seroconversion date

means that a less selective group of patients can be analysed, as those presenting with seroconversion illness may show differences in the progression of their disease.

In this chapter, the estimated date of seroconversion is treated as fixed and known for each patient. However, models that do not make this assumption are developed later in the thesis in Chapter 7.

4.1.3 Models for pre-treatment CD4 counts

The study in which Taylor *et al.*[8] first proposed the addition of a Brownian motion component to a 'random slopes' linear mixed model found that this led to a significant improvement in model fit for a dataset of 722 measurements obtained from 87 seroconverters, patients who had been observed to transition from an HIV-negative to HIV-positive state (with the midpoint of the interval between these observations taken as the time of infection, $t = 0$). Taylor *et al.*[8] also investigated the use of an IOU process, of which Brownian motion is a special case, in this context, but did not find a further improvement in model fit. In this paper, CD4 counts were modelled on a fourth-root scale to better match the assumptions of normality made.

Sy *et al.*[25] further developed the Brownian motion and IOU process models reported in Taylor *et al.*[8], using data from a nearly identical set of patients, by fitting a bivariate model incorporating these stochastic process components to both fourth-root transformed CD4 count and beta-2-microglobulin measurements. In another study, again using a similar patient population, Taylor and Law[26] used models fitted to fourth-root transformed CD4 data up to a given calendar cut-off and used these to predict the value of subsequent measurements for individuals in which these were available. They found that a random intercept model incorporating an IOU process gave a better combination of mean squared error and prediction interval coverage than did a linear (random slopes) or quadratic mixed effects model in terms of t. This study also compared the goodness of fit of these models for the log, square-root, cube-root, fourth-root and untransformed CD4 data. Accounting for the Jacobian of the transformation, they found that square-root, cube-root or fourth-root transformations provided similar optimal log-likelihood values for each of the models considered, and these transformations also performed similarly in terms of their mean squared error and coverage of predictions for future measurements.

Wolbers *et al.*[17] fitted a model including a random intercept and a Brownian motion component to a dataset of pre-treatment CD4 counts of 2820 HIV patients from the multinational CASCADE cohort collaboration study[16], and found that this had the optimal AIC of a set of models considered. Again using data from the CASCADE study, Babiker *et al.*[27] subsequently applied a model also including a random slope component to nearly 90 000 CD4 count observations, using a square root scale, in

over 15 000 seroconverters prior to the occurrence of any AIDS-defining events or the initiation of any ART, reporting a substantial improvement in model fit in comparison to a random slopes model. Additionally, Babiker *et al.*[27] analysed post-ART CD4 counts obtained from the same patients using a similar model, but with linear splines with a break at $t = 1 year$ in order to enable drop-off in treatment response to be assessed. The pre-ART fitted model was used to simulate series of CD4 measurements following HIV infection from a large cohort of patients in order to estimate the proportion that would initiate ART as a function of time according to various follow-up and treatment regimens, and the post-ART model was used to simulate the response to treatment that would be observed. The motivation for the analyses of Babiker *et al.*[27] was to enable power and sample size calculations for a RCT to investigate whether immediate or delayed ART treatment of patients newly diagnosed with HIV leads to better overall outcomes.

The multivariate-t distribution was used by Wang and Fan[9] to model CD4 counts in a small sample of 30 HIV-positive patients taken from a historic trial of ART medication. Here observations were recorded on a regular schedule, and Wang and Fan used a random slopes structure with an additional first-order autoregression parameter for the residual error. The same authors have also reported the fitting of a similar multivariate-t model with a second-order autoregressive structure to a sample of 50 patients from the same historic dataset using a Bayesian approach[93]. In the context of HIV, Matos *et al.*[46] have also reported the use of a multivariate-t model for right-censored HIV RNA assays in untreated patients with acute infection using a non-linear random effects model for the mean with independent error terms; their model was fitted to 830 observations in 320 individuals. However, there were no publications prior to Stirrup *et al.*[47] in which multivariate-t models have been fitted to CD4 data with the addition of Brownian motion or other stochastic process components, and multivariate-t models have not been used to analyse large-scale datasets containing tens of thousands of observations.

4.2 Dataset

We present here a reanalysis of the dataset of pre-ART CD4 counts described by Babiker *et al.*[27]. The total dataset includes 89 176 CD4 count observations in 15 274 individuals whose date of HIV seroconversion is well documented; comprising all available measurements prior to the occurrence of AIDS-defining illness or initiation of ART up to December 2007 from 26 cohorts participating in the CASCADE study[16]. The CASCADE cohort includes patients in whom there was a maximum interval between negative and positive HIV antibody tests of 3 years or in whom there was laboratory evidence of seroconversion; for the first of these groups the date of reported

seroconversion illness is used, if it is recorded, as the estimated date of seroconversion, otherwise the 'mid-point' estimate is used.

Only 3955 (4.4 %) measurements from 789 (5.2 %) patients were recorded at a time of more than 10 years, and so we chose to model only those CD4 measurements obtained up to 10 years from the estimated date of seroconversion. This resulted in a dataset of 85 221 measurements in 15 164 individuals. A further 365 observations were excluded for which an identical CD4 measurement was recorded only 1 day after the previous count for that patient, as these were found to cause problems with model estimation and were assumed to result from data-entry errors, resulting in a dataset of 84 856 measurements for analysis.

CD4 cell counts are measured as cells per microlitre, and we followed established practice in modelling the counts on a square root scale[27]. As an illustrative example, the CD4 measurements were modelled only in terms of time from seroconversion, expressed as continuous in years, although it would be possible to include other predictive variables. The median number of CD4 observations per individual in the analysed dataset was 4, with a range of 1–57 and an interquartile range (IQR) of 2–8. There was no rigid pattern to the timing of observations in each patient, with a median interval between measurements of 112 (IQR, 70–182) days. The highly unbalanced nature of the dataset and the irregular observation schedule necessitate the use of flexible modelling strategies that can accommodate such features. Visual inspection of the CD4 data suggests that the trajectories over time for each individual do not follow predictable paths and that there may be between-patient differences in variability over time, motivating the combination of stochastic process components and the multivariate-t distribution, respectively, as presented in this chapter. A total of 9831 (64.8 %) patients were censored from the dataset at initiation of ART, 1111 (7.3 %) because of a recorded AIDS event and 318 (2.1 %) at death, 2444 (16.1 %) patients can be considered lost to follow-up (with no clinic visit recorded for 12 months and no censoring event) and the remaining 1460 (9.6 %) were in follow-up at the time that the data were gathered.

4.3 Model fitting

The initial model fitted was a standard linear mixed effects model including correlated random intercept and slope terms and independent measurement error terms of constant variance. An exponential decay correlation structure was considered for the error terms of this model, and the initial model was then extended to also include either a scaled Brownian motion process or a scaled fractional Brownian motion process. The equivalent set of four models was then fitted using a marginal multivariate-t distribution, i.e. with the scale matrix $\mathbf{V}_i$ structured in the same manner but assum-

ing an unobserved scaling variable for each individual as described in Section 2.3.

For all models, maximum likelihood estimates of the parameters were obtained using the ADMB software (ADMB Foundation)[50]. The 'R2admb' package[94] was used to run analyses and manage results through the R statistical computing environment. Starting values are required for all parameters when using ADMB. These were obtained by using approximate values from a 'nlme' model fit for the initial 'random slopes' linear mixed model, and subsequent models were fitted using parameter estimates from the previous simpler model as the initial value. When fitting models with a Brownian motion component, an initial value of 1 was used for the scale parameter, and for models with fractional Brownian motion, an initial value of 0.5 was used for the H index. For models based on the multivariate-t distribution, an initial value of 10 was used for the degrees of freedom parameter.

The 'fixed effects' for each model included an intercept (β_0) and a slope (β_1) parameter. For the 'random effects' covariance/scale matrix ($\mathbf{\Psi}$) for each model, U_{00} and U_{11} represent the variance of the random intercepts and random slopes, respectively, for each individual, with ρ representing the correlation between them. For the multivariate-t models, this interpretation holds conditional on the vector of unobserved latent variables $\boldsymbol{\gamma}$. Optimisation was performed in terms of log-transformations of U_{00} and U_{11} and a generalised logistic transformation of ρ. For all models, the residual error term was optimised in terms of $\log(\sigma)$ (i.e. the log of the residual standard deviation). The exponential correlation structure was optimised in terms of the log of the range parameter (η), and Brownian motion models (including fractional) used the log of the scale parameter (κ). Fractional Brownian motion was parameterised in terms of the logistic transformation of H. A log-transformation was used for the degrees of freedom parameter (ν) in multivariate-t models. For all model parameters, confidence intervals are reported derived from the estimated asymptotic multivariate normal distribution on the transformed scales.

Nested models are compared using the likelihood ratio test; as only a single parameter is being added to the model in each of the comparisons presented, the critical value for change in 2×log-likelihood ($2\Delta\ell$) at the 5 % significance level is only 3.84. Non-nested models are compared using the BIC statistic, using the total number of observations in the dataset for the calculation of the penalty term; this is supported by the derivation of Cavanaugh and Neath[95]. The AIC statistic is also provided for each model.

4.4 Results and diagnostic checks

Summaries of the linear mixed models, with marginal multivariate normal distribution, fitted to the pre-ART CASCADE data are provided in Table 4.1. The addition to the initial random slopes model of an exponential correlation structure for the residual variance resulted in a significant improvement in model fit, with change in 2×log-likelihood ($2\Delta\ell$) of 460 for 1 parameter (P<0.001). However, the addition of a Brownian motion component to the random slopes model led to a greater increase in log-likelihood ($2\Delta\ell$ 4940 for 1 parameter, P<0.001), with a subsequently lower value of BIC for this model. A further improvement in model fit was observed when the Brownian motion component was generalised to a fractional Brownian motion process ($2\Delta\ell$ 160 for 1 parameter, P<0.001). As such, the 'random slopes + fractional Brownian motion + measurement error' model was found to have the lowest BIC of the fitted linear mixed models. A 'random slopes + IOU process' model was also considered, but was found to return the special case of a Brownian motion process (i.e. with a very large estimate for the α parameter[8]).

Table 4.1. Summaries of linear mixed models (with marginal multivariate normal distribution) fitted to square-root transformed pre-antiretroviral therapy CD4 measurements from the CASCADE dataset

	Random slopes + measurement error	Random slopes + exp. cor. + measurement error	Random slopes + Brownian motion+ measurement error	Random slopes + fBM + measurement error
β_0	24.13 (24.02 to 24.24)	24.12 (24.01 to 24.23)	23.81 (23.70 to 23.92)	23.82 (23.71 to 23.92)
β_1	-1.36 (-1.40 to -1.33)	-1.35 (-1.38 to -1.31)	-1.15 (-1.18 to -1.11)	-1.15 (-1.19 to -1.12)
U_{00}	33.68 (32.65 to 34.73)	33.22 (32.20 to 34.27)	28.69 (27.72 to 29.70)	27.46 (26.46 to 28.51)
ρ	-0.39 (-0.41 to -0.36)	-0.38 (-0.41 to -0.35)	-1 (—)	-0.59 (-0.63 to -0.54)
U_{11}	1.62 (1.54 to 1.71)	1.54 (1.46 to 1.62)	0.20 (0.16 to 0.24)	0.58 (0.49 to 0.68)
σ	2.76 (2.74 to 2.77)	2.79 (2.77 to 2.81)	2.28 (2.26 to 2.29)	2.01 (1.94 to 2.07)
η	—	0.03 (0.03 to 0.03)	—	—
κ	—	—	7.00 (6.78 to 7.22)	9.32 (8.78 to 9.91)
H	—	—	—	0.30 (0.27 to 0.33)
ℓ	-232 579	-232 349	-230 109	-230 029
AIC	465 170	464 712	460 232	460 074
BIC	465 226	464 777	460 297	460 149

Parameter estimates are given with 95 % confidence intervals in parentheses. AIC, Akaike's information criterion; BIC, Bayesian information criterion; exp. cor., exponential correlation structure for residual error term; fBM, fractional Brownian motion.

It is of particular interest that the estimate of the H parameter for the model incorporating a fractional Brownian motion process is below 0.5, indicating that successive increments of the process are negatively correlated and hence that the process will tend to revert towards its mean. The mean in this case would include the subject-specific random effects for the intercept and slope. The correlation between the random intercept and random slope for each individual for the model incorporating a standard Brownian motion process is estimated to be −1.00, which seems

rather unnatural. However, when the process is generalised to a fractional Brownian motion, an estimate of -0.59 (95 % CI, -0.63 to -0.54) is obtained for this correlation.

The Cholesky-transformed residuals of the commonly used random slopes model and of the best-fitting linear mixed model, incorporating a fractional Brownian motion component, were analysed to assess the goodness of fit (as described in Section 3.2). For the 'random slopes' model, a plot of mean Cholesky residuals against time (Figure 4.1a) indicates that for times above 5 years the observed value is on average lower than that expected conditioning on the previous observations for each individual. The Q–Q plot of the Cholesky residuals indicates that their distribution is heavy-tailed in comparison to the expected standard normal under a correctly specified model (Figure 4.1c). A semivariogram plot of the Cholesky residuals also indicates a lack of independence between them (Figure 4.1d), with apparent residual correlation at time lags of less than 2.5 years.

For the 'random slopes + fractional Brownian motion' linear mixed model, the plot of mean Cholesky-transformed residuals against time (Figure 4.2a) shows an acceptable scatter around zero, and the semivariogram plot is relatively close to the expected value (under a correct model) of 1 for all time lags up to 8 years (Figure 4.2d) with little data available beyond this. However, the plot of mean squared Cholesky residuals shows a systematic downward trend with time (Figure 4.2b) and the Q–Q plot still shows heavier tails than expected (Figure 4.2c), indicating that the fit of the model to the data is not perfect.

Summaries of the multivariate-t distribution models fitted to the pre-ART CASCADE data are provided in Table 4.2. The addition to the initial random slopes model of an exponential correlation structure for the residual variance resulted in a significant improvement in model fit ($2\Delta\ell$ 1032 for 1 parameter, P<0.001). However, as for the normal model, the addition of a Brownian motion component to the random slopes model led to a greater increase in log-likelihood ($2\Delta\ell$ 4412 for 1 parameter, P<0.001). A further improvement in model fit was observed when the Brownian motion component was generalised to a fractional Brownian motion process ($2\Delta\ell$ 270 for 1 parameter, P<0.001). As such, the 'random slopes + fractional Brownian motion + measurement error' model was found to have the lowest BIC of the fitted multivariate-t distribution models. Furthermore, all of the multivariate-t models were found to have lower BIC values than all of the normal linear mixed models. The difference in 2ℓ between the normal and the multivariate-t 'random slopes + fractional Brownian motion + measurement error' models is 8298, indicating a significant and substantial improvement in model fit (1 parameter, P<0.001).

The degrees of freedom parameter ($\hat{v}$) was found to be between 5 and 6 for all of the fitted multivariate-t models, in accordance with the heavy tails observed in the Q–Q plots for the normal linear mixed models. However, the heavy tails could

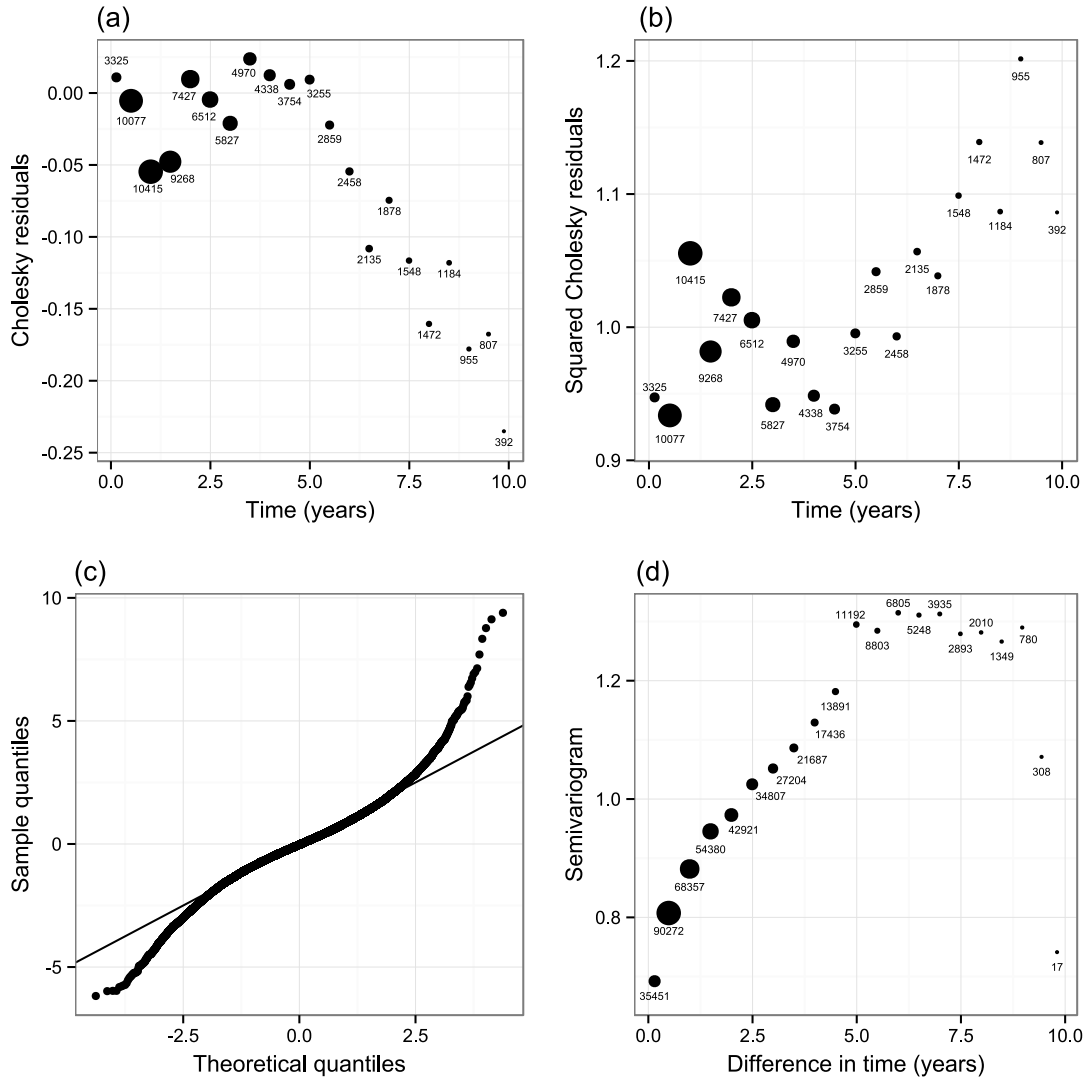

Figure 4.1. Plots of Cholesky-transformed residuals from the 'random slopes + measurement error' linear mixed model fitted to the pre-antiretroviral therapy CD4 counts from the CASCADE dataset. In (a) and (b) mean values are plotted grouped by nearest multiple of 6 months, with size of points approximately proportional to the number of observations in each group and n values shown. (c) Quantile–quantile plot for Cholesky-transformed residuals with respect to a standard normal distribution, with line of equality. (d) Semivariogram of Cholesky-transformed residuals with respect to difference in time between observations, grouped by nearest multiple of 6 months, with size of points approximately proportional to the number of observation pairs in each group and n values shown.

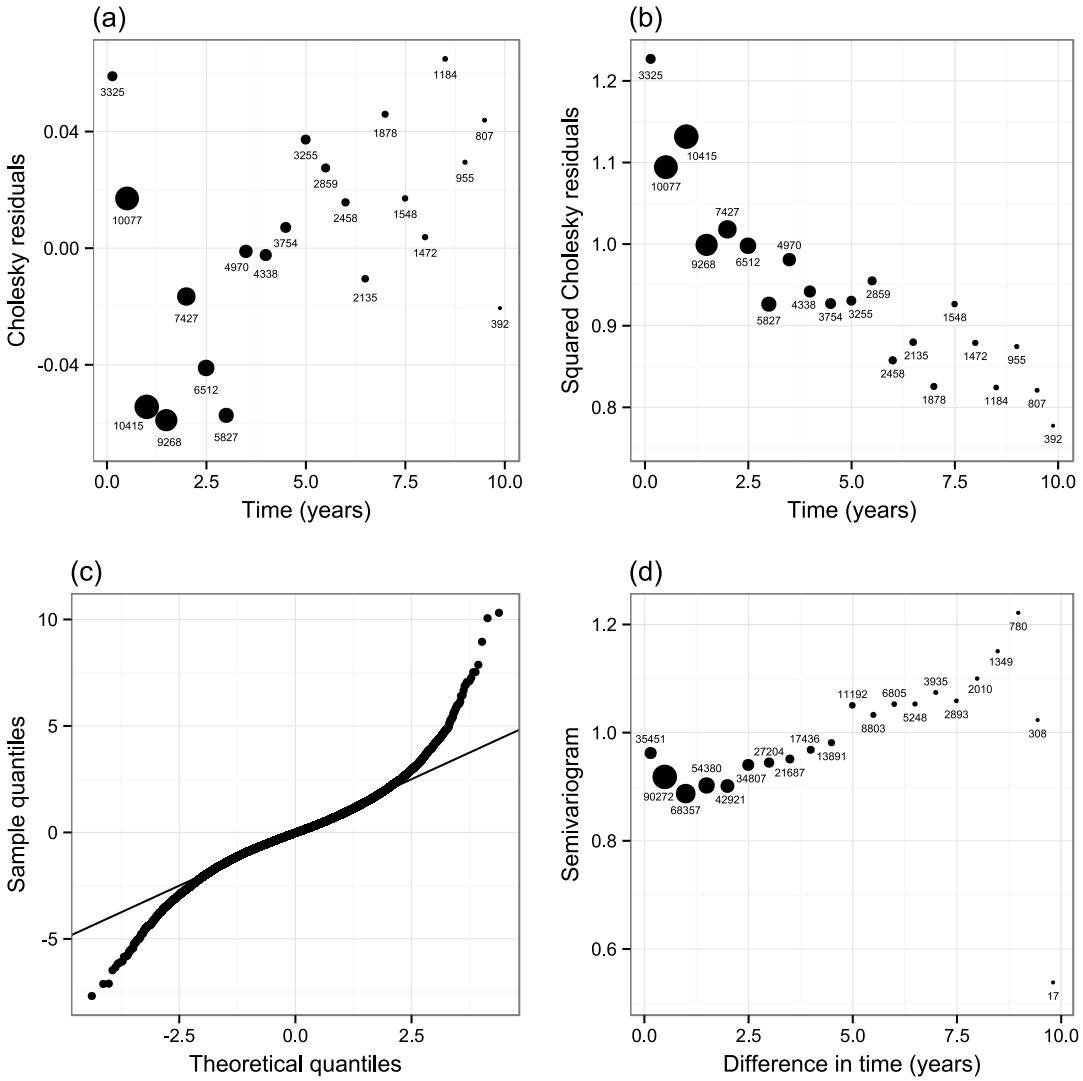

Figure 4.2. Plots of Cholesky-transformed residuals from the 'random slopes + fractional Brownian motion + measurement error' linear mixed model fitted to the pre-antiretroviral therapy CD4 counts from the CASCADE dataset. In (a) and (b) mean values are plotted grouped by nearest multiple of 6 months, with size of points approximately proportional to the number of observations in each group and n values shown. (c) Quantile–quantile plot for Cholesky-transformed residuals with respect to a standard normal distribution, with line of equality. (d) Semivariogram of Cholesky-transformed residuals with respect to difference in time between observations, grouped by nearest multiple of 6 months, with size of points approximately proportional to the number of observation pairs in each group and n values shown.

Table 4.2. Summaries of multivariate-t distribution models fitted to square-root transformed pre-antiretroviral therapy CD4 measurements from the CASCADE dataset

	Random slopes + measurement error	Random slopes + exp. cor. + measurement error	Random slopes + Brownian motion+ measurement error	Random slopes + fBM + measurement error
β_0	23.77 (23.67 to 23.87)	23.76 (23.66 to 23.86)	23.57 (23.47 to 23.67)	23.59 (23.49 to 23.69)
β_1	-1.27 (-1.31 to -1.24)	-1.23 (-1.27 to -1.20)	-1.10 (-1.13 to -1.07)	-1.11 (-1.14 to -1.07)
U_{00}	23.82 (22.99 to 24.69)	22.83 (22.00 to 23.68)	20.30 (19.5 to 21.14)	18.82 (17.98 to 19.7)
ρ	-0.37 (-0.4 to -0.34)	-0.36 (-0.39 to -0.33)	-1 (—)	-0.51 (-0.55 to -0.47)
U_{11}	1.17 (1.10 to 1.23)	1.01 (0.95 to 1.08)	0.12 (0.10 to 0.15)	0.49 (0.43 to 0.55)
σ	2.25 (2.23 to 2.27)	2.32 (2.30 to 2.35)	1.88 (1.86 to 1.90)	1.45 (1.35 to 1.55)
η	—	0.07 (0.06 to 0.07)	—	—
κ	—	—	5.17 (4.98 to 5.36)	8.02 (7.44 to 8.64)
H	—	—	—	0.23 (0.21 to 0.26)
ν	5.64 (5.40 to 5.88)	5.34 (5.12 to 5.57)	5.83 (5.58 to 6.09)	5.76 (5.52 to 6.02)
ℓ	-228 221	-227 705	-226 015	-225 880
AIC	456 456	455 426	452 046	451 778
BIC	456 521	455 501	452 121	451 862

Parameter estimates are given with 95 % confidence intervals in parentheses. AIC, Akaike's information criterion; BIC, Bayesian information criterion; exp. cor., exponential correlation structure for residual error term; fBM, fractional Brownian motion.

be due to distributional structures other than the marginal multivariate-t distribution employed, for example the random effects and any Gaussian processes included could follow multivariate normal distributions with the residual error terms following independent-t distributions. As such, there is a need for further investigation to assess the goodness of fit of the chosen multivariate-t model with respect to the data.

For the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model, 1000 simulations of the vector of latent variables $\boldsymbol{\gamma}$ were generated, based on the predicted posterior distribution in each individual, and used to calculate sets of Cholesky- transformed residuals for the model (as described in Section 3.4.2). Summaries of the distributions of these residuals are displayed in Figure 4.3. Although the presence of unobserved latent variables leads to uncertainty in the exact values of the residuals, plots of the distributions of the mean residuals, mean squared residuals and semivariogram calculations show similar patterns to those for the equivalent normal linear mixed model. The Q–Q plot of the Cholesky residuals derived using the empirical Bayes estimate ($\hat{\gamma}_i$) for each individual shows a near perfect fit to the standard normal distribution. However, taking quantiles over multiple simulations of $\boldsymbol{\gamma}$ indicates the presence of slightly heavier tails than expected. A subject-level residual plot for this model, as proposed by Lange *et al.*[41] (described in Section 3.4.1), is also presented in Figure 4.4. This plot does not indicate any major problems with model fit.

The goodness of fit of the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model was further investigated by inspection of Q–Q

plots of residuals for individual patients transformed by the inverse of the Cholesky decomposition of their estimated scale matrix ($\hat{\mathbf{V}}_i$) without any correction for γ_i. As little would be gained by evaluating patients with very few observations, only those with greater than 15 measurements in the dataset were considered; 1044 (6.9 %) of individuals in the dataset met this criterion. Q–Q plots for 25 randomly selected individuals are shown in Figure 4.5. Under a correctly specified model, each of the plots should approximately show a straight line of points, with differing slopes between individuals; for the i^{th} individual the expected slope is a function of their unobserved scale variable: $\gamma_i^{-1/2}$, where $\gamma_i \sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, with ν being the degrees of freedom parameter in the multivariate-t model. From these plots, it seems plausible that there are indeed differences in overall variability between individuals as implied by the marginal multivariate-t model, for example Plot 9 shows a clearly steeper slope than Plot 3. However, some of the plots, for example number 24, appear to show a heavy-tailed rather than a normal distribution for that particular patient, implying that the multivariate-t model (in which each individual follows a multivariate normal distribution conditional on an unobserved scale variable) does not wholly account for the heavy-tailed nature of the residuals observed. This evidence is consistent with that provided by the overall Q–Q plot in Figure 4.3c, the standardised residuals using simulated values of $\boldsymbol{\gamma}$ are much closer to fitting a standard normal distribution in comparison to those from the equivalent normal linear mixed model, but still appear to be slightly heavy-tailed.

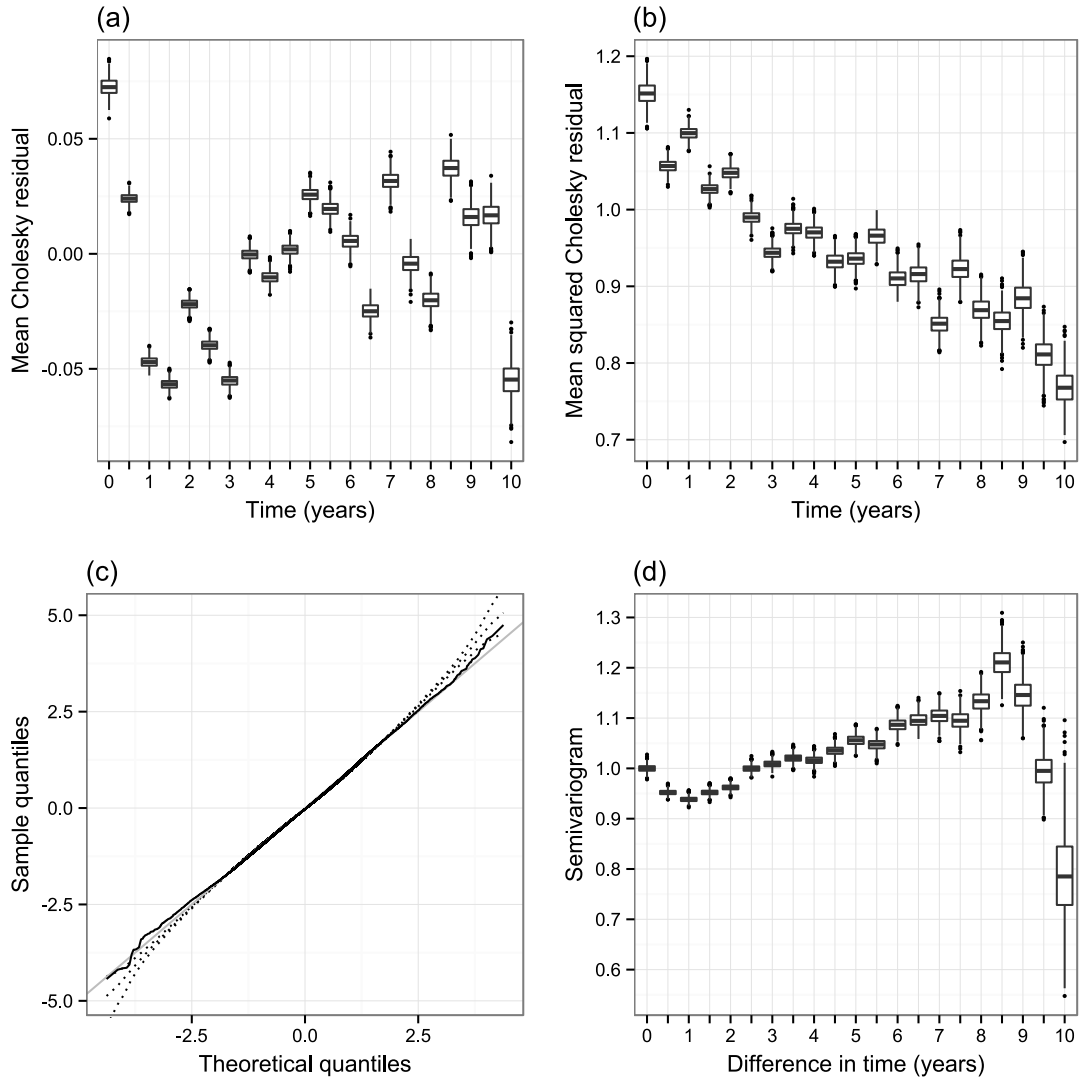

Figure 4.3. Plots of the distribution of Cholesky-transformed residuals from the 'random slopes + fractional Brownian motion + measurement error' multivariate-t distribution model fitted to the pre-antiretroviral therapy CD4 counts from the CASCADE dataset, based on 1000 simulations of the vector of latent variables $\boldsymbol{\gamma}$. In (a) and (b) box plots of mean values for each simulation are plotted grouped by nearest multiple of 6 months. (c) Quantile–quantile plot for Cholesky-transformed residuals with respect to a standard normal distribution; the dotted lines show the 2.5^{th}, 50^{th} and 97.5^{th} percentiles of the sample quantiles for each theoretical quantile corresponding to the total number of observations, the solid black line shows the sample quantiles derived using the empirical Bayes estimate ($\hat{\gamma}_i$) for each individual, with the line of equality also displayed in grey. (d) Box plots of the distribution of mean semivariogram values, over multiple simulations of $\boldsymbol{\gamma}$, of Cholesky-transformed residuals with respect to difference in time between observations, grouped by nearest multiple of 6 months. The numbers of observations contributing to the mean at each time point per simulation of $\boldsymbol{\gamma}$ is not shown, but match those given in Figures 4.1 and 4.2.

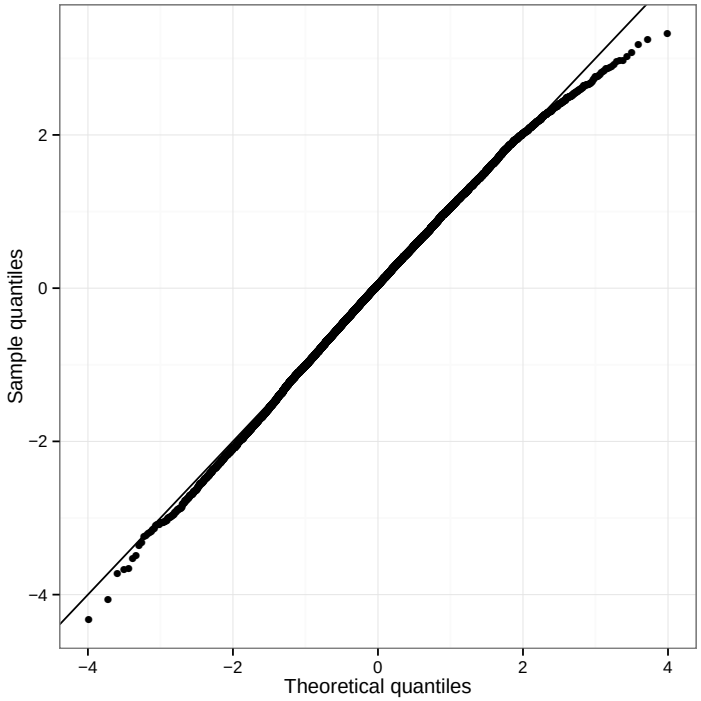

Figure 4.4. Plot of subject-level residuals, as proposed by Lange *et al.*[41], for the 'random slopes + fractional Brownian motion + measurement error' multivariate-t distribution model fitted to the pre-antiretroviral therapy CD4 counts from the CASCADE dataset. As described in Section 3.4.1, $\hat{\delta}_i^2(\boldsymbol{\theta}) = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}} \hat{\mathbf{V}}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)$ was calculated for each patient, and the cumulative probability function of an F-distribution with n_i and $\hat{\upsilon}$ degrees of freedom was applied to $\frac{\hat{\delta}_i^2(\boldsymbol{\theta})}{n_i}$; these values were then converted to quantiles of a standard normal distribution, which are displayed in a Q–Q plot.

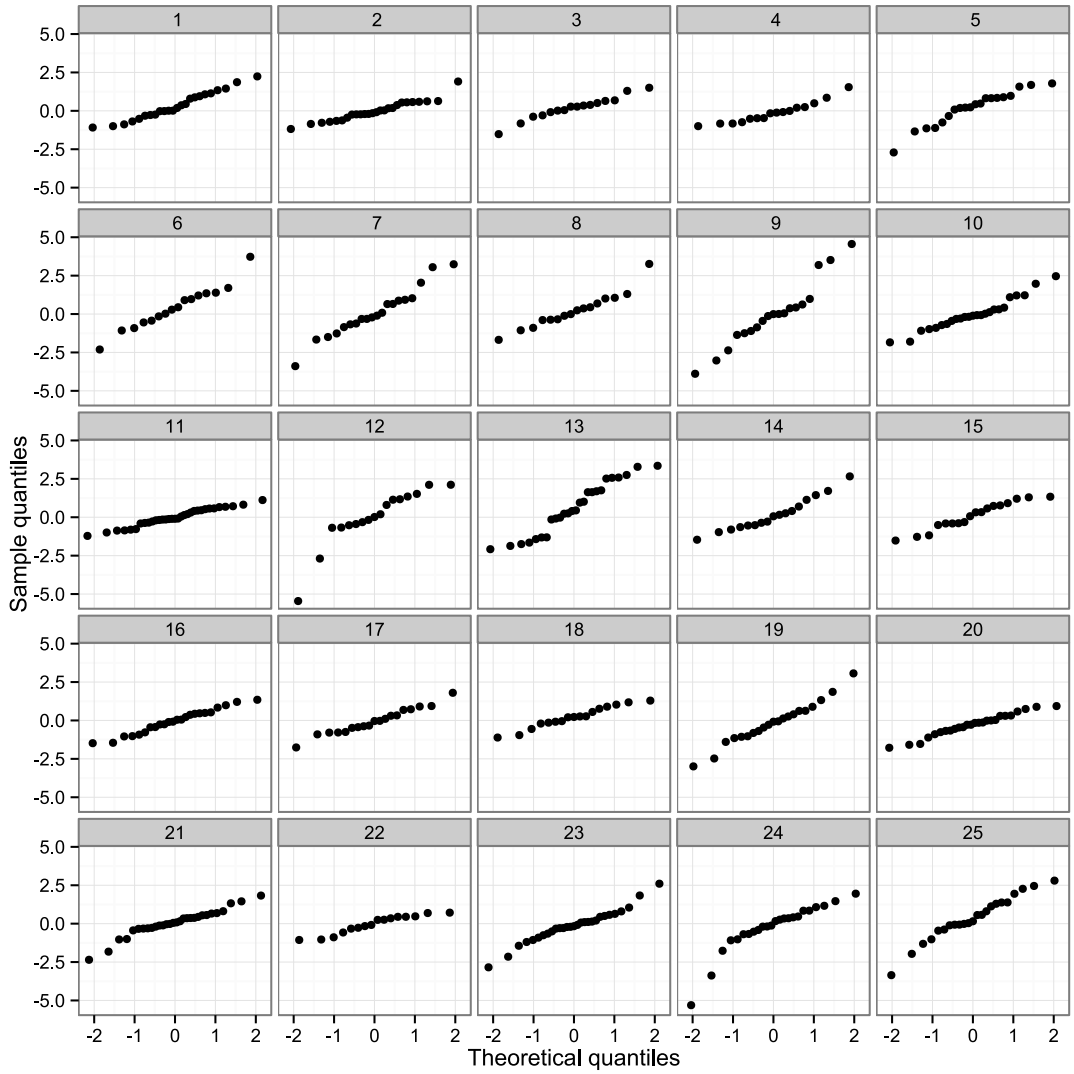

Figure 4.5. Quantile–quantile plots for the residuals under the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model of 25 randomly selected individuals with greater than 15 observations. The residuals for individual patients have been transformed by the inverse of the Cholesky decomposition of their estimated scale matrix ($\hat{\mathbf{V}}_i$) without any correction for the unobserved scale variable γ_i. Theoretical quantiles in each case are those from the standard normal distribution.

4.5 Simulation study

4.5.1 Impact of model choice on treatment initiation predictions

Until recently, the initiation of ART for asymptomatic HIV-positive patients in European countries was commonly based on the observations of a CD4 count below a given threshold, with the most appropriate cut-off (or whether treatment should be given immediately upon diagnosis) for any given setting a matter of evolving debate[96]. As such, there was interest in determining the proportion of patients that will cross any given threshold and initiate ART as a function of time from seroconversion, as this would have had an impact on clinical practice and on the cost of different healthcare strategies. Lodi *et al.*[97] used random slopes linear mixed models fitted to over 175 000 CD4 measurements from the CASCADE cohort (including the data analysed in this chapter) to predict the proportion of untreated patients reaching thresholds of <500, <350 and <200 cells/μL with respect to time from seroconversion, reflecting the cut-offs used in various versions of official guidelines. In their analysis, the distribution of subject-specific slopes was used to estimate the proportion of patients with 'true' CD4 count below each threshold value.

Babiker *et al.*[27] describe simulations performed for the planning of the Strategic Timing of Antiretroviral Treatment (START) trial, which randomised HIV patients with CD4 cell counts $\geq$ 500 cells/μL to either initiate treatment immediately or for this to be delayed until their count had dropped to <350 cells/μL. Using their fitted linear mixed model including a Brownian motion component, Babiker *et al.*[27] investigated the proportion of patients reaching a threshold of <350 cells/μL through simulation of sets of longitudinal measurements for tens of thousands of individuals. This approach has the advantage of allowing realistic assessment of the characteristics of a cohort in practice, and several regimes for the scheduling of measurements and initiation of ART were considered in their simulations. However, the predictions made from the simulations were not directly compared to those that would have been obtained using a normal random slopes model.

The START trial was ended earlier than planned following an interim analysis in May 2015 by the independent data and safety monitoring board, which found that early initiation of treatment led to statistically and clinically significant reductions in both serious AIDS-related and serious non-AIDS-related events[88]. The results of the trial led to changes to the World Health Organisation[98] and UK[89] guidelines on initiation of ART in HIV patients, which now state that ART should be started in all patients regardless of CD4 count. In developed countries, patients will therefore no longer undergo monitoring of CD4 counts before initiation of ART. However, we present here a comparison of predictions regarding treatment initiation patterns following the previous CD4 cut-offs that have been used, based on several of the fitted

models described in Section 4.4. Although this is no longer of direct clinical relevance, it illustrates how the use of more complex models for longitudinal data could have impacted on an important medical problem.

Simulated cohorts of individuals were generated based on three multivariate normal models: the random slopes model, the Brownian motion model and the fractional Brownian motion model (with the latter two also including a random slopes structure and all including measurement error). In addition, a cohort was generated using the fitted multivariate-t fractional Brownian motion model (again, including a random slopes structure and measurement error). For each of these models, data for 5 million individual patients were simulated based on scheduled measurements being obtained every 4 months for up to 10 years. Data were also generated for measurements 1 month after the scheduled observation in each case for use in the analysis, corresponding to a confirmatory test. CD4 thresholds of <500, <350 and <200 cells/μL for ART initiation were investigated. If a scheduled measurement was observed below a given threshold then the value 1 month later was assessed, to mimic the conduct of an additional confirmatory test as commonly performed in clinical practice. The patient was considered to initiate ART if this second value was also below the threshold.

The results of the analysis of the simulated cohorts are presented in Figure 4.6. The differences in predictions made by each of the fitted models are large enough to have had practical implications were CD4 cell cut-offs still used for the initiation of treatment, particularly within a public health or health economics context, for example using the <500 cells/μL threshold the proportion of patients on ART 2 years after seroconversion is predicted to be 57 % by the normal random slopes model and to be 62 % by the multivariate-t model with fractional Brownian motion. The planning of the START trial described by Babiker *et al.*[27] made use of predictions of the proportion of patients initiating ART at the 350 cells/μL threshold, for which we found only small differences between each of the models that included a stochastic process component (i.e. excluding the standard random slopes model). It is interesting to note that for the 500 and 350 cells/μL cut-offs the predictions for the models incorporating stochastic process components converge as time increases towards 10 years, separate to the lower predictions made by the standard random slopes model.

4.5.2 Parameter bias in slope estimates

One interesting feature of the various models fitted to the CASCADE pre-ART CD4 data is that the mean slope (β_1) of CD4 decline is substantially less negative for the linear mixed models that include standard or fractional Brownian motion components (both –1.15) than for the random slopes model (–1.36). The estimated slopes

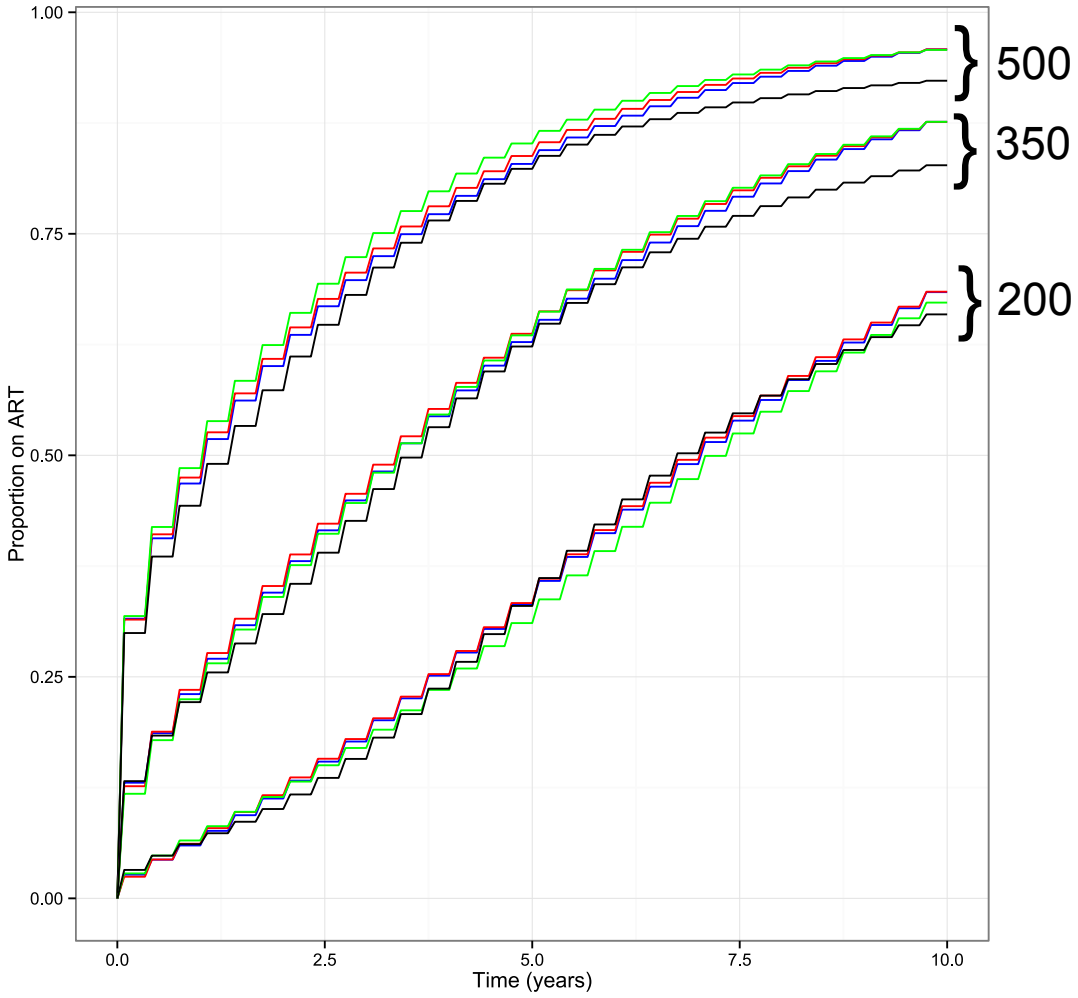

Figure 4.6. The proportion of HIV-positive patients predicted to have initiated antiretroviral therapy (ART) as a function of time since seroconversion, based on simulation from the fitted normal random slopes model (——), Brownian motion model (——) and fractional Brownian motion model (——) and the multivariate-t fractional Brownian motion model (——). Results are presented using CD4 thresholds for ART initiation of <500, <350 and <200 cells/μL, as indicated at top right of the graph. Simulations are based on CD4 measurements being obtained every 4 months, with initiation of ART conditional on an additional observation below the cut-off concerned 1 month after the 'scheduled' measurement.

for the equivalent multivariate-t models were also less steep in each case (see Tables 4.1 and 4.2). We performed a simulation study to assess the impact of model choice and missing data patterns on this difference, which may indicate apparent bias from the use of simpler models.

For a given observation schedule, it follows from Liang and Zeger[99] that a linear mixed model analysis of longitudinal data will give consistent estimates of the fixed effects given that either there is no missing data or that data is 'missing completely at random' (MCAR) (following the terminology of Rubin[100]). This also requires the structure of the fixed effects to be correctly specified in the model, but not the exact distribution of observations or covariance between them. If the observation schedule is instead considered to result from a random process, then maximum likelihood estimation of a model for the outcome variable alone is consistent, without the need for specification of a model for the distribution of follow-up times, on the condition that the timing of observations is dependent only on previously observed outcomes[101]; however, this result requires a correctly specified model for the outcome variable (including covariance structure). Hence it seems that the substantial differences in slope estimates between different models fitted to pre-ART CD4 data are due to the presence of missing data for which the missingness is not MCAR, or to a dependency of the observation schedule on the observed values of the outcome variable.

It is often postulated that the missingness of observations in pre-ART datasets can be treated as 'missing at random' (MAR)[102], i.e. that it is independent of the unobserved outcome variable conditional on the observed values of the outcome variable and other covariates included in the model, and that as such the missingness can be ignored under maximum likelihood estimation such as the use of linear mixed models. Although less often explicitly stated, it is also common to assume that the timing of observations can be ignored in the statistical model[101]. If the missing data in a study can be considered monotone, i.e. if a scheduled observation is missed by a patient then they also miss all further scheduled observations, then the 'ignorable timing of observation' and the MAR assumptions are very similar in form. Furthermore, if observations are only considered to occur at a set of discrete time points (e.g. within a finite set of days or months), then any statistical analysis of longitudinal data can be considered in terms of Rubin's missing data framework[100], without consideration of a model for the timing of observations. Hence, further discussion on this topic is framed in terms of the MAR assumption.

The MAR assumption is plausible if patients are thought to mainly drop out of the dataset upon initiation of ART, and if this is entirely dependent on their observed CD4 counts. However, the beneficial properties of maximum likelihood-based inference (i.e. consistency and asymptotic normality and efficiency of estimates) with

respect to MAR data are dependent on a correctly specified model for the likelihood. The fact that adding stochastic process components and/or generalising to a multivariate-t distribution leads to a very substantial improvement in BIC indicates that the standard random slopes model does not correctly describe the covariance structure or probability model for pre-ART CD4 data.

To further investigate bias in parameter estimates resulting from overly simplistic models, the best-fitting model (i.e. multivariate-t with fractional Brownian motion) was assumed to be 'correct' and cohorts of patient data simulated from it. CD4 cell count observations were generated from 0 to 5 years, for groups of either 100 or 200 patients and with an annual observation frequency of 1 or 3; 500 cohorts were generated for each combination. For each simulated cohort, models were first fitted to the complete uncensored data (although this would include impossible negative values), and subsequently to the data following censoring corresponding to ART initiation at CD4 cut-off values of 200, 350 and 500 cells/μL. The 'correct' multivariate-t model and three normal linear mixed models (the random slopes model, the Brownian motion model and the fractional Brownian motion model) were applied to each simulated cohort under each condition. For the analyses involving censoring, additional confirmatory measurements were generated 1 month after the 'scheduled' observations; these were only considered to be observed when the scheduled measurement was below the cut-off value, and the patient was only censored when the confirmatory value was also below the cut-off. The censored datasets could therefore be considered to correspond to observations being MAR but not MCAR. As the MAR condition holds for any possible realisation, this scenario meets the 'everywhere MAR' definition provided by Seaman *et al.*[103], allowing valid frequentist likelihood inference. Model fitting was considered to have failed when parameter estimates were not returned or when the covariance matrix of parameter estimates was not positive-definite.

Limited bias was observed in the estimation of the intercept term when using simplified models, and so the results of this analysis are only presented for estimation of the slope parameter β_1. Bias in the estimation of β_1 and the coverage of 95 % confidence intervals for this parameter are presented in Table 4.3. As expected, a lack of bias (or only very minimal bias) and appropriate coverage intervals were observed when the correctly specified model was fitted, even in the presence of censoring. Interestingly, no or only minimal bias was observed when the equivalent normal linear mixed model (including a fractional Brownian motion component) was used. Linear mixed models including a Brownian motion component showed some downward bias in the presence of censoring, with this most marked when censoring was applied using the CD4 cut-off of 500 cells/μL. Substantial downward biases and poor coverage of confidence intervals were observed when a standard random slopes lin-

ear mixed model was applied in the presence of censoring, with the degree of bias clearly linked to the extent of censoring.

A summary of the standard deviations of point estimates for the mean slope and the average estimated standard error for this parameter in the simulations is also provided in Table 4.4. There were not large discrepancies between these two measures of the standard error. The mean slope estimates from the correctly defined model showed slightly lower variance than the estimates from the incorrectly defined models in any given situation, but the scale of these differences seems relatively small compared to the large biases observed.

The differences in slope estimates observed between models under the censoring conditions in this simulation study correspond to the differences observed between the models when applied to the real dataset. This provides supporting evidence that special attention should be given to the probability model used, and in particular the covariance structure, when analysing a dataset for which there is substantial missing data that is not MCAR. These simulations imply that an analysis using a wrongly specified model might incorrectly indicate differences between two groups in their average rate of decline if they have been subject to different censoring mechanisms. We carried out an additional investigation in which two groups of either 100 or 200 patients each were simulated with three observations per year, with the first group subject to censoring at the '200 cut-off' while the '500 cut-off' was applied for the second group. Other details of the simulation and model-fitting were as previously described, but two additional 'fixed effects' parameters were added to the models to allow the mean intercept (δ_0) and slope (δ_1) of the second group to differ from the first group (with the true value of these parameters being zero). These simulations confirmed that bias could occur in the estimation of between-group differences in slope within a single model (estimated bias for random slopes model with 200 patients per group: –0.163, see Table 4.5).

Table 4.3. Summary of the results of simulation analyses to assess bias in the estimate of mean slope (β_1) when models that are simpler than the data-generating process are applied in the presence of 'missing at random' censoring. Bias is calculated as the mean estimate of β_1 minus the true value, and is presented with coverage of nominal 95 % confidence intervals in parentheses. For each combination of number of simulated patients (N) and annual frequency of observation (freq), 500 cohorts were generated and analysed under different censoring regimes, corresponding to treatment initiation at CD4 cut-offs of 200 (ART200), 350 (ART350) or 500 (ART500). All cohorts were simulated with a follow-up of 5 years, including an observation at time zero for each patient. Data were generated according to a multivariate-t distribution (MVT) incorporating a fractional Brownian motion (fBM) process and measurement error (ME) and, alongside a model of the correct form, normal linear mixed models were fit with a random slopes (RS) structure alone and with RS in combination with Brownian motion (BM) and fBM processes.

			RS+ME		RS+BM+ME		RS+fBM+ME		MVT: RS+fBM+ME	
	Prop. cens. (median (IQR))	*n* obs. (median (IQR))	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)
N=100, freq=1										
Uncensored	0 (0–0)	600 (600–600)	0	0.013 (93.8)	0.0	0.012 (94.6)	2.6	0.011 (94.0)	1.2	0.012 (96.0)
ART200	21 (18–24)	573 (567–579)	0	-0.100 (88.2)	0.4	-0.021 (96.8)	1.4	0.000 (95.3)	0.2	0.008 (95.4)
ART350	51 (48–55)	505 (493–515)	0	-0.244 (73.0)	1.2	-0.071 (95.1)	2.6	-0.001 (96.5)	1.4	0.000 (95.9)
ART500	77 (75–80)	400 (388–413)	0	-0.384 (72.0)	2.4	-0.146 (94.9)	5.2	-0.022 (94.3)	1.0	-0.019 (93.5)
N=100, freq=3										
Uncensored	0 (0–0)	1600 (1600–1600)	0	0.000 (95.4)	0.0	-0.001 (97.0)	4.0	-0.002 (96.0)	2.0	-0.001 (96.7)
ART200	30 (27–34)	1414 (1386–1441)	0	-0.161 (84.4)	0.8	-0.054 (95.8)	4.8	-0.009 (96.0)	3.4	-0.004 (96.1)
ART350	63 (60–66)	1095 (1060–1131)	0	-0.289 (71.8)	0.4	-0.125 (95.0)	7.0	-0.002 (97.8)	2.6	-0.002 (97.9)
ART500	85 (82–87)	724 (687–761)	0	-0.322 (84.6)	2.2	-0.270 (92.0)	9.8	-0.025 (94.5)	3.0	-0.027 (94.2)
N=200, freq=1										
Uncensored	0 (0–0)	1200 (1200–1200)	0	0.004 (94.6)	0.0	0.005 (94.4)	1.0	0.004 (93.7)	0.0	0.001 (94.4)
ART200	22 (20–24)	1144 (1136–1152)	0	-0.106 (82.4)	0.2	-0.019 (95.4)	1.2	0.000 (94.1)	0.6	0.000 (95.0)
ART350	52 (49–54)	1008 (993–1023)	0	-0.234 (56.0)	0.4	-0.057 (95.8)	2.4	0.007 (95.3)	0.8	0.005 (95.2)
ART500	78 (76–80)	799 (780–816)	0	-0.360 (56.2)	0.4	-0.118 (92.6)	6.8	0.007 (94.6)	1.6	-0.001 (94.5)
N=200, freq=3										
Uncensored	0 (0–0)	3200 (3200–3200)	0	-0.002 (94.6)	0.0	-0.001 (95.6)	2.4	-0.001 (94.7)	0.4	0.000 (92.2)
ART200	30 (28–32)	2840 (2804–2874)	0	-0.161 (66.6)	0.2	-0.054 (93.0)	2.2	-0.012 (94.5)	0.8	-0.005 (92.1)
ART350	62 (60–65)	2197 (2142–2244)	0	-0.300 (44.6)	0.4	-0.127 (88.0)	4.4	-0.014 (96.2)	2.4	-0.007 (96.1)
ART500	84 (83–86)	1454 (1404–1508)	0	-0.337 (70.2)	1.0	-0.243 (86.3)	12.4	-0.004 (93.8)	5.8	-0.017 (93.8)

Model fitting was considered to have failed when parameter estimates were not returned or when the covariance matrix of parameter estimates was not positive-definite. IQR, interquartile range; *n* obs., total observations included in analysis per simulated cohort; Prop. cens., proportion of patients in simulated cohort subject to censoring before 5 years.

Table 4.4. Summary of the standard deviation of slope estimates obtained ($SD(\hat{\beta}_1)$) and mean of standard error estimates ($\mu(\widehat{SE_{\beta_1}})$) from the simulation analyses to assess bias in the estimate of mean slope (β_1) when models that are simpler than the data-generating process are applied in the presence of 'missing at random' censoring. For each combination of number of simulated patients (N) and annual frequency of observation (freq), 500 cohorts were generated and analysed under different censoring regimes, corresponding to treatment initiation at CD4 cut-offs of 200 (ART200), 350 (ART350) or 500 (ART500). All cohorts were simulated with a follow-up of 5 years, including an observation at time zero for each patient. Data were generated according to a multivariate-t distribution (MVT) incorporating random slopes (RS), a fractional Brownian motion (fBM) process and measurement error (ME) and, alongside a model of the correct form, normal linear mixed models were fit with a RS structure alone and with RS in combination with Brownian motion (BM) and fBM processes (all with ME).

	RS+ME		RS+BM+ME		RS+fBM+ME		MVT: RS+fBM+ME	
	$SD(\hat{\beta}_1)$	$\mu(\widehat{SE_{\beta_1}})$	$SD(\hat{\beta}_1)$	$\mu(\widehat{SE_{\beta_1}})$	$SD(\hat{\beta}_1)$	$\mu(\widehat{SE_{\beta_1}})$	$SD(\hat{\beta}_1)$	$\mu(\widehat{SE_{\beta_1}})$
N=100, freq=1								
Uncensored	0.131	0.131	0.131	0.132	0.131	0.130	0.110	0.115
ART200	0.147	0.146	0.137	0.144	0.135	0.139	0.116	0.122
ART350	0.201	0.184	0.177	0.188	0.179	0.181	0.156	0.158
ART500	0.347	0.292	0.297	0.317	0.311	0.301	0.273	0.268
N=100, freq=3								
Uncensored	0.133	0.132	0.129	0.138	0.130	0.129	0.107	0.110
ART200	0.165	0.153	0.140	0.153	0.137	0.143	0.118	0.121
ART350	0.205	0.203	0.187	0.213	0.178	0.200	0.154	0.168
ART500	0.383	0.340	0.372	0.391	0.363	0.353	0.310	0.304
N=200, freq=1								
Uncensored	0.096	0.093	0.095	0.093	0.094	0.092	0.083	0.081
ART200	0.111	0.103	0.099	0.101	0.100	0.099	0.088	0.086
ART350	0.142	0.130	0.121	0.131	0.126	0.127	0.113	0.111
ART500	0.228	0.207	0.213	0.220	0.211	0.212	0.189	0.189
N=200, freq=3								
Uncensored	0.094	0.093	0.095	0.097	0.091	0.092	0.078	0.077
ART200	0.106	0.103	0.101	0.106	0.094	0.099	0.084	0.084
ART350	0.145	0.130	0.135	0.146	0.131	0.127	0.110	0.118
ART500	0.270	0.207	0.263	0.268	0.268	0.212	0.213	0.215

Table 4.5. Summary of simulation study to assess bias in the estimation of between-group differences in intercept (δ_0) and slope (δ_1) within a single model when different patterns of censoring are applied to the two groups. Using the probability distribution of the 'random slopes (RS) + fractional Brownian motion (fBM) + measurement error (ME)' multivariate-t (MVT) model fitted to the real dataset, two groups of either 100 or 200 patients each were simulated with three observations per year up to 5 years, with the first group subject to censoring at the '200 cut-off' while the '500 cut-off' was applied for the second group (with corresponding 'confirmatory' observations in each case). Alongside a model of the correct form, normal linear mixed models were fit with a RS structure alone and with RS in combination with Brownian motion (BM) and fBM processes (all with ME). 500 iterations of the simulation were performed for each group size. The true value of both δ_0 and δ_1 was set to zero, and were estimated as the difference in mean intercept and slope parameter for the second group relative to the first.

	RS+ME	RS+BM+ME	RS+fBM+ME	MVT:RS+fBM+ME
Group size=100 each				
Bias: $\hat{\delta_0}$	0.091	0.027	0.022	0.020
$SD(\hat{\delta_0})$	0.806	0.771	0.758	0.651
$Mean(\widehat{SE_{\delta_0}})$	0.805	0.765	0.754	0.655
Coverage: δ_0	95.0	94.8	94.5	94.6
Bias: $\hat{\delta_1}$	-0.143	-0.108	0.004	0.004
$SD(\hat{\delta_1})$	0.275	0.270	0.271	0.230
$Mean(\widehat{SE_{\delta_1}})$	0.278	0.294	0.268	0.237
Coverage: δ_1	92.0	95.6	94.5	95.6
Failed (%)	0.0	0.6	2.2	3.8
Group size=200 each				
Bias: $\hat{\delta_0}$	0.04	-0.020	-0.014	-0.007
$SD(\hat{\delta_0})$	0.576	0.554	0.547	0.471
$Mean(\widehat{SE_{\delta_0}})$	0.572	0.544	0.536	0.465
Coverage: δ_0	95.0	95.0	94.6	94.6
Bias: $\hat{\delta_1}$	-0.163	-0.128	-0.016	-0.013
$SD(\hat{\delta_1})$	0.197	0.192	0.189	0.164
$Mean(\widehat{SE_{\delta_1}})$	0.197	0.208	0.190	0.168
Coverage: δ_1	87.0	92.4	95.0	94.8
Failed (%)	0.0	0.0	0.8	0.2

4.6 Discussion

In this chapter we have further developed the statistical modelling of longitudinal biomarker data, through application to pre-treatment CD4 counts in patients with HIV, in which we have shown that the combination of a fractional Brownian motion component and generalisation of the normal linear mixed model to a multivariate-t distribution leads to substantial improvements in model fit. This novel combination of model features provides additional information regarding the between- and within-patient variability in observations over time. Evidence is provided for the appropriateness of using a multivariate-t distribution in the studied dataset through evaluation of novel diagnostic plots. Furthermore, simulation studies are presented to demonstrate the impact of model choice on cohort-level predictions and on bias in mean slope estimates when data are MAR.

The presence of non-stationary stochastic process components in models for longitudinal data imply that the progress of the state of the underlying biological system for each individual does not follow a deterministic relationship with time, but rather follows an unpredictable path. This finding seems intuitive in the context of the extremely complex interactions between viral replication and immune system response that influence the CD4 count series that are observed in HIV-positive patients. When using a fractional Brownian motion component the H values obtained were less than 0.5, indicating that the process is erratic but displays some reversion towards an underlying mean. The estimates of the degrees of freedom parameter for the multivariate-t models of between 5 and 6 indicate substantial between-patient differences in variability over time.

Through simulations based on generating data from the more complex fitted model, we have demonstrated that the use of a normal random slopes model is associated with substantial bias in the estimation of the mean slope parameter in the presence of censoring, with the degree of bias strongly dependent on the choice of censoring regime. This is important, as estimates of this parameter are often used as a proxy for rate of decline in health and compared between groups. As initiation of ART was historically dependent on observed CD4 values, the MAR condition has been invoked to argue that likelihood-based model estimation will lead to valid inferences (e.g. Lodi *et al.*[97]), but this only holds conditional on the correct specification of the likelihood-model. It can therefore be argued that in this context greater effort should be made to make use of statistical models that adequately describe the distributional and covariance patterns present in the data.

Diagnostic Q–Q plots of Cholesky-transformed marginal residuals from multivariate normal models fitted to square-root CD4 counts show very heavy tails, indicating clear violation of the modelling assumptions. We have demonstrated that the

use of a multivariate-t distribution in combination with a non-stationary stochastic process component leads to a very substantial improvement in BIC, with diagnostic Q–Q plots that only indicate relatively mild violation of the model's assumptions. Such models can be fit efficiently and to large datasets using the open-source ADMB software[50], with this task made easier by the fact that the log-likelihood of the multivariate-t distribution is available in closed form. It is of interest to investigate whether models comprised of different combinations of multivariate-t and normal distributions could provide a better fit to the data, such models have been previously discussed by Song *et al.*[62]. For example, it may be considered more biologically plausible to fit a statistical model in which the variability of the stochastic process component differs between individuals (i.e. follows a multivariate-t distribution) but the random effects and measurement error terms do not (i.e. they follow normal distributions); we did attempt to fit this model to the dataset analysed in this chapter, but maximum likelihood estimation failed when Gauss–Hermite quadrature was used — for such models the likelihood function is not available in closed form, making the computations required for parameter estimation substantially more complex. However, we do make use of models combining multivariate-t and multivariate normal distributions in Chapters 5, 6 and 7.

Our research has been focused on CD4 cell counts in HIV-positive patients, but the modelling framework developed may be of use for the analysis of longitudinal data in other biomedical applications. For example, Diggle *et al.* recently described the use of an extended linear mixed model including another non-stationary stochastic process, integrated Brownian motion, for the analysis of estimated glomerular filtration rates in patients at risk for renal failure[29]. The authors provide plots of 'Cholesky-standardised' residuals produced from the application of the model, which show very heavy-tails. The multivariate-t distribution implies differences in the volatility of observations between patients, which may by useful in planning and interpreting the monitoring of biomarkers in HIV and other disease areas.

Whilst it is arguably impossible to claim that any statistical model exactly represents the data-generating mechanism under investigation, it seems that both the addition of stochastic process components to the standard linear mixed model and the use of a multivariate-t distribution can be used to gain a greater understanding of longitudinal biomedical data. Such models provide greater flexibility, but require only a small number of additional parameters and follow a model specification that can be interpreted in terms of the underlying biological process; as such, the potential gains in inference and understanding through their use are likely to greatly outweigh any drawbacks of increased model complexity. There is therefore a motivation to continue to develop more efficient methods of fitting such models and to make these more widely available.

5 Development of a combined model for pre- and post-treatment data

In this chapter, we develop a novel modelling framework for the combined analysis of pre- and post-treatment data. The work is motivated by the ambition to better understand the factors that predict recovery in CD4 counts after the initiation of HAART in HIV patients, and we illustrate the methodology through application to data from the UK Register of HIV Seroconverters cohort[104]. This chapter serves to explain the framework developed and to discuss the methodology primarily in the context of the relevant statistics literature. Following the development of models for pre-treatment CD4 counts as described in Chapter 4, we also incorporate stochastic process components and between-patient differences in variability over time. The approach developed is applied to a larger dataset in Chapters 6 and 7, in which patient and drug regimen characteristics are also included in the analysis and in which the results are discussed in more depth with reference to previous research on this topic within the field of HIV. The contents of this chapter form the basis for a publication in *BMC Medical Research Methodology*[105], which is provided as Appendix D (reproduced under CC BY 4.0 license). We include here some additional residual diagnostic plots and simulations that are not included in the published paper. The simulation analyses in Section 5.13 serve to firstly provide a check that the statistical methodology is functioning as intended, and secondly to further explore the potential problems with existing approaches.

5.1 Background

In medical research, there is often interest in evaluating response to treatment conditional on the baseline value at initiation of the biomarker under investigation. In the setting of RCTs, designed primarily to assess the difference between treatment conditions, some authors have argued that optimal efficiency is gained by treating the baseline measurement as an outcome variable within a parametric model[10;11], whilst Senn has argued that conditioning estimation of treatment effect on the baseline observation through the use of ANCOVA is preferable in most trial situations[12] and Kenward *et al.* demonstrated that with correct adjustments for sample size the two approaches have nearly identical properties[106]. However, both of these approaches can be problematic when applied to the estimation of response to treatment using longitudinal observational datasets, in which the timing and choice of treatment have not been randomised and in which baseline observations immediately prior to treatment may not be available for all patients. Furthermore, there is often substantial interest in the influence of the baseline value of the biomarker itself in de-

termining the level of response to treatment, rather than just using this to provide a better estimate of the differences between treatment choices. In this chapter we describe the development of flexible parametric models for this situation, providing a combined analysis of pre- and post-treatment data in which the response of the biomarker to treatment is dependent on a 'true' baseline value that is not directly observed; this combines elements of previous approaches in that the pre-treatment data are modelled as 'response variables', but the trajectory of the biomarker after treatment initiation can also be modelled using flexible functions of the baseline value.

Tango recently proposed a format for RCTs in which multiple observations for both the pre- and post-treatment periods are included in the analysis[107], demonstrating that this would lead to a reduction in the required sample size in some circumstances. Tango considered models in which a random intercept term is shared by pre- and post-treatment observations, or in which an additional random effect allows response to treatment to be correlated with the pre-treatment random effect term. However, given the RCT setting, Tango did not consider models for a progressive disease in which the timing of treatment could vary between individuals; in this case specifying the combined model only in terms of correlated random effects provides limited flexibility in modelling the dependency of response to treatment conditional on the state of the patient at initiation.

The models that we develop in this chapter are applied to CD4 cell counts in HIV-positive patients who initiate HAART. Although the CD4 counts within an individual can vary erratically over time, on average the counts decline steadily from normal levels following HIV infection[97] and then in most cases recover towards normal levels following initiation of HAART[81;108–111]. In observational datasets, the timing of recorded CD4 measurements can be highly variable between patients. In much of the existing literature about the long-term response of CD4 counts to HAART, the investigators have avoided any associated complications in their analyses by converting the available data into a set of discrete time points, typically corresponding to annual or 6-monthly observations. This has been done by linear interpolation (Kaufmann *et al.*)[108], selecting only the observation closest to the chosen time point (Moore and Keruly)[109] or taking the mean measurement within intervals (Lok *et al.*)[110]. Each of these studies included an analysis stratified by intervals of baseline CD4 count and, although the statistical methodology varied between studies, each found that higher baseline CD4 counts were associated with higher values after several years of ART.

An alternative approach, which can make use of all post-treatment data on its original time-scale, is to fit linear mixed effects models to the post-treatment data with stratification based on pre-treatment observations. This has been done using

linear splines (Gras *et al.*[81]) or fractional polynomial functions (Hughes *et al.*)[111] for the post-treatment model. These modelling approaches find similar conclusions but again discard most pre-treatment data and involve conditioning of the model on pre-treatment observations that are subject to measurement error, which can lead to substantial biases in the estimation of treatment effects in some situations[112].

A study by Le *et al.* suggested that the long-term response to ART in HIV-positive patients is improved if it is initiated within the first few months after infection, with this effect independent of the CD4 count at baseline[113]. Le *et al.*[113] also relied on stratification of patients into groups, and used a generalised estimating equation analysis with exchangeable correlation structure (equivalent to a random intercept linear mixed model) and splines for the mean (the exact methodology used is not fully specified in the paper, but visual inspection of the Figures indicates the use of cubic splines). Ding *et al.* also found a more rapid initial increase in CD4 count, assessed by absolute increase at 6 months, and a higher probability of attaining a CD4 count ≥ 600 or ≥ 900 cells/μL within 3 years of treatment, for patients in whom treatment was initiated within 2 months of diagnosis of a recent infection[114].

We now also know that early treatment of HIV leads to a substantial reduction in the occurrence of both AIDS-defining conditions and serious non-AIDS events[88], but there nonetheless remains clinical interest in understanding the factors that are predictive of the recovery in CD4 counts upon HAART initiation as for many patients there is a substantial delay between infection and diagnosis and suboptimal CD4 recovery remains a concern for patients and clinicians[115]. The principal aim of the research described in this chapter is the development of a flexible parametric framework for the combined modelling of pre- and post-treatment CD4 data in HIV positive individuals. This is motivated by the clinical interest in investigating the factors that determine the characteristics of long-term response to HAART, in particular the influences of baseline CD4 count and the time elapsed from infection to treatment initiation. However, the modelling strategy developed could also be used in other settings in which a biomarker is monitored prior to some treatment initiation or clinical intervention.

The modelling strategy described in this chapter represents a flexible extension of established non-linear mixed effects models, fitted through maximum likelihood estimation based on all observed data using time as a continuous variable. As well as allowing inclusion of all available data in its original format (other than global transformations for normalisation) and the combined assessment of multiple predictive factors, the approach will have the advantage that the characteristics of CD4 trajectories of individual patients over time will be quantified, creating a complete framework for epidemiological simulations or patient-specific predictions, whereas previously this has been done using separate models for pre- and post-treatment

data[27]. We also incorporate stochastic process components and between-patient differences in variability over time into the models developed, with the aim of defining models that are as realistic as possible in representing the structure of the biological measurements under investigation. This is particularly important when considering analyses for datasets in which missing data and irregular follow-up times are a substantial concern.

5.2 Dataset

The UK Register of HIV Seroconverters is an observational cohort study of patients whose date of infection can be reliably estimated[104], which contributes to the CASCADE collaboration. Recruitment to the cohort began in 1994, but, as we are interested in modelling the response to modern HAART regimens, we restrict our analysis to patients with an estimated date of HIV (type 1) seroconversion during or after 2003. Data are included up until January 2014. Patients who started a suboptimal regimen of antiretroviral drugs prior to HAART were excluded, as were patients without at least one post-treatment CD4 count recorded. Patients without any pre-treatment CD4 counts were, however, included in the analysis. HAART is defined by a regimen of at least three antiretroviral drugs from at least two different classes (unless abacavir or tenofovir is used in a regimen with three nucleoside/nucleotide analog reverse-transcriptase inhibitors (NRTIs)).

Application of these conditions resulted in a study population of 852 patients, with a total of 5805 pre-HAART and 7302 post-HAART CD4 observations recorded. The median (IQR) number of pre-HAART CD4 counts was 5 (3–10), whilst that for post-HAART observations was 6 (3–12). There were a total of 39 patients without any pre-HAART CD4 counts recorded. The median (IQR) time from estimated date of seroconversion to initiation of HAART was 1.3 (0.6–2.8) years, with 192 patients starting HAART within 6 months and 149 starting between 6 months and 1 year from seroconversion.

As in Chapter 4, CD4 cell counts are measured as cells/μL and we followed established practice in modelling the counts on a square-root scale[27;47]. For the pre-treatment part of the model, time is measured in years from date of HIV seroconversion (the estimate of which is treated as fixed in each patient), whilst for the post-treatment part of the model it is measured in years from HAART initiation. We have censored patients at recorded interruption of HAART (including switch to suboptimal treatment) for more than 1 week, but have not censored according to viral load (VL) status or change to HAART regimen. Treatment interruption was recorded in 124 (14.6 %) patients, and there were a total of seven deaths recorded (three of which occurred after censoring due to interruption of HAART). Data from a random subset

of 100 of the patients analysed are shown in Figure 5.1.

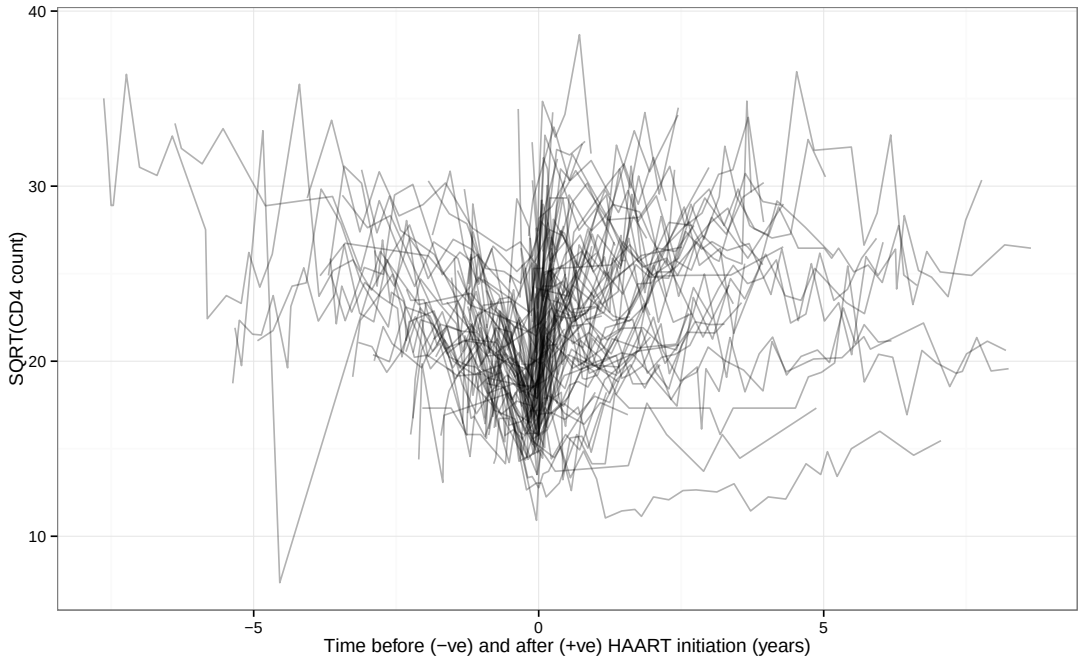

Figure 5.1. 'Spaghetti plot' of the square root of CD4 counts from a random sample of 100 patients. Patients are from the UK Register of HIV Seroconverters dataset. Lines are semi-translucent to aid visualisation. Time has been centred at the date of highly active antiretroviral therapy (HAART) initiation for each patient.

5.3 Baseline state as a latent variable

It can be shown that in situations in which the initiation of treatment is conditional on a biomarker that is monitored over time, and which is measured with error, the observed value of the biomarker at the start of treatment provides a biased estimate of the 'true' underlying value[27]. This presents a problem when attempting to model treatment response conditional on the baseline value. We propose that one option in this situation is to build a combined model for both the pre- and post-treatment data, allowing the response to treatment to be conditional on all available pre-treatment data rather than on just a single baseline value. Such an approach would also have the advantage that patients could be included for whom no measurement close to the start of treatment had been obtained. Additionally, fewer assumptions regarding the marginal distribution of 'true' baseline values of any given population would be required. For example, such an approach could appropriately deal with a set of distinct treatment initiation guidelines applied across different periods of time or subpopulations, which might lead to a multimodal distribution of baseline values in the total study population, whereas a standard mixed model approach would generally assume the observed baseline values to follow a normal distribution for the popu-

lation as a whole. The methodology recently proposed by Tango[107] allows response to treatment to be modelled in combination with multiple pre-treatment observations, but nonetheless assumes a fixed distributional form for the baseline prior to treatment (represented by the population mean plus a random intercept term).

As described in Chapter 2, any linear mixed effects model implies a marginal multivariate normal distribution[1], for which the log-likelihood function can be expressed in closed form. However, this is not true (except for some special cases) for non-linear mixed effects models[4] and for such models some approximation of the log-likelihood is required. Among the available options, adaptive Gauss–Hermite quadrature is particularly attractive as an increasing number of quadrature points can be used for each random effect to ensure that the log-likelihood is evaluated to an adequate degree of accuracy. However, adaptive Gauss–Hermite quadrature is not generally used when there are more than one or two random effects terms per individual defined in a model, and the computational requirements to attain high accuracy in calculation of the log-likelihood function are lowest when there is only one random effect term per individual.

Because of these computational issues, to undertake the combined modelling of pre- and post-treatment CD4 data we focus on the use of non-linear latent variable models that require numerical integration only over the unobserved 'true' CD4 count at treatment initiation (which we will term u). The rationale of this approach is that it will allow adequate flexibility in model structure without increasing the computational requirements to a level that will prevent application to the dataset available. In order to achieve this, we will specify linear mixed models for the pre-treatment data ($\mathbf{y}_{pre}$) and non-linear models for the post-treatment data ($\mathbf{y}_{post}$), conditioned on the 'true' baseline CD4 count, that *are* linear in any other random effects terms (allowing a closed form expression for each of these two parts of the model). Under such a scheme, the likelihood function for the combined pre- and post-treatment data for each individual can therefore be expressed as:

$$\begin{aligned} f\left(\mathbf{y}_{pre},\mathbf{y}_{post}\right) &= \int_{-\infty}^{\infty} f_{pre,post,u}\left(\mathbf{y}_{pre},\mathbf{y}_{post},u\right) du \\ &= \int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}\right) f_{post,u}\left(\mathbf{y}_{post},u|\mathbf{y}_{pre}\right) du \\ &= \int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}\right) f_{post}\left(\mathbf{y}_{post}|\mathbf{y}_{pre},u\right) f_{u}\left(u|\mathbf{y}_{pre}\right) du. \end{aligned}$$

For simplicity above, we suppress notation to indicate that each element of the likelihood function is dependent on model parameters. However, we now consider calculation of the likelihood function dependent on the values of a parameter vector relating to the pre-treatment part of the model '$\boldsymbol{\theta}_{pre}$', a parameter vector relating to

the post-treatment part of the model '$\boldsymbol{\theta}_{post}$' and a shared measurement error variance parameter 'σ^2'. If we assume that the post-treatment response depends on the pre-treatment data only though the true baseline value at treatment initiation, i.e. that $\mathbf{y}_{post}$ is independent of $\mathbf{y}_{pre}$ given u, then we may write:

$$f\left(\mathbf{y}_{pre}, \mathbf{y}_{post}\right) = \int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}|\boldsymbol{\theta}_{pre}, \sigma^2\right) f_{post}\left(\mathbf{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right) f_u\left(u|\mathbf{y}_{pre}, \boldsymbol{\theta}_{pre}, \sigma^2\right) du.$$

This follows a similar form to the likelihood expression for standard random effects models but here the distribution of the latent variable u, which is integrated out to obtain the marginal likelihood, is conditioned on the pre-treatment data for each individual rather than following a pre-specified distribution across the population. For those patients in whom no pre-treatment observations were obtained, the likelihood contribution can be calculated solely for the post-treatment observations:

$$f\left(\mathbf{y}_{post}\right) = \int_{-\infty}^{\infty} f_{post}\left(\mathbf{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right) f_u\left(u|\boldsymbol{\theta}_{pre}, \sigma^2\right) du.$$

In Section 5.7, we describe the addition of two further latent variables to the model for each individual in order to allow for between-patient differences in variability over time.

5.4 Pre-treatment model structure

At present we consider only linear mixed model formulations for the likelihood of $\mathbf{y}_{pre:i}$, representing the observed vector of $n_{pre:i}$ pre-treatment observations for the i^{th} individual. However, this is inclusive of stochastic Gaussian process components, such as Brownian motion[8;17] or fractional Brownian motion[19], as these do not prevent the use of a (multivariate normal) closed form for the pre-treatment likelihood function f_{pre} (as described in Chapter 2). Denoting the vector of values of the stochastic process $\mathbf{W}_{pre:i}$ at times $\mathbf{t}_{pre:i}$, and defining $\boldsymbol{\Sigma}_{pre:i}$ as the covariance matrix resulting from the chosen Gaussian process for the i^{th} individual, the linear mixed model can then be expressed as:

$$\begin{aligned}
\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\
\mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\
\mathbf{W}_{pre:i} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{pre:i}) \\
\mathbf{e}_{pre:i} &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{pre:i}}).
\end{aligned}$$

Here, $\mathbf{X}_i$ represents the pre-treatment design matrix for the 'fixed effects' param-

eters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the pre-treatment 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_{pre:i}$ is the vector of residual errors for each pre-treatment measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2 ... \mathbf{b}_N$, residual errors $\mathbf{e}_{pre:1}, \mathbf{e}_{pre:2} ... \mathbf{e}_{pre:N}$ and stochastic process realisations $\mathbf{W}_{pre:1}, \mathbf{W}_{pre:2} ... \mathbf{W}_{pre:N}$ for each of the N individuals are independent of one another. It can be easily shown that this formulation leads to the following marginal distribution for $\mathbf{y}_{pre:i}$:

$$\mathbf{y}_{pre:i} \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_{pre:i} + \sigma^2\mathbf{I}_{n_{pre:i}}).$$

We shall use $\mathbf{V}_{pre:i}$ to denote the marginal covariance matrix for $\mathbf{y}_{pre:i}$.

In this analysis, we shall consider only a 'random intercepts and slopes' structure for the fixed and random effects parts of the pre-treatment model. We shall also include fractional Brownian motion as a Gaussian process component, along with an independent residual error term, following the optimal multivariate normal model for pre-treatment data as found in Chapter 4.

5.5 Conditional distribution of 'true' baseline

The use of a pre-treatment model with marginal multivariate normal distribution means that the conditional distribution of the 'true' baseline value (u_i) at treatment initiation for each individual given their observed pre-treatment data can be readily obtained. We denote the time of treatment initiation from the start of observation (HIV seroconversion in this case) as $t_{trt:i}$. We shall assume that u_i is formed by the sum of the fixed effects parameter vector ($\boldsymbol{\beta}$) multiplied by a row vector ($\mathbf{X}_{trt:i}$) corresponding to an extension of the design matrix ($\mathbf{X}_i$) for that individual relating to variable values (e.g. time) at $t_{trt:i}$, the equivalent term for the subject-specific random effects (i.e. $\mathbf{Z}_{trt:i}\mathbf{b}_i$) and the realisation of the subject's stochastic process at $t_{trt:i}$:

$$u_i = \mathbf{X}_{trt:i}\boldsymbol{\beta} + \mathbf{Z}_{trt:i}\mathbf{b}_i + W_{trt:i}.$$

As such, the joint distribution $\mathbf{y}_{pre:i}$ and u_i is multivariate normal:

$$\begin{pmatrix}\mathbf{y}_{pre:i}\\ u_i\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mathbf{X}_i\boldsymbol{\beta}\\ \mathbf{X}_{trt:i}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{pre:i} & \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_{trt:i}^{\mathrm{T}} + \mathrm{Cov}\left[\mathbf{W}_{pre:i}, W_{trt:i}\right]\\ \mathbf{Z}_{trt:i}\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathrm{Cov}\left[W_{trt:i}, \mathbf{W}_{pre:i}\right] & \mathbf{Z}_{trt:i}\boldsymbol{\Psi}\mathbf{Z}_{trt:i}^{\mathrm{T}} + \mathrm{Var}\left[W_{trt:i}\right]\end{pmatrix}\right).$$

The variance and covariance terms for the stochastic component of the model can be calculated for any given Gaussian process based on $\mathbf{t}_{pre:i}$, $t_{trt:i}$ and any pre-treatment model parameters relating to the process. The conditional probability

density function of u_i given $\mathbf{y}_{pre:i}$, $f_u\left(u_i|\mathbf{y}_{pre:i},\boldsymbol{\theta}_{pre},\sigma^2\right)$, can therefore be obtained using the standard result for a partitioned multivariate normal distribution. Using a simplified notation:

$$\begin{pmatrix}\mathbf{y}_{pre:i}\\ u_i\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mathbf{X}_i\boldsymbol{\beta}\\ \mathbf{X}_{trt:i}\boldsymbol{\beta}\end{pmatrix},\begin{pmatrix}\mathbf{V}_{pre:i} & \mathbf{v}_{12:i}\\ \mathbf{v}_{21:i} & v_{22:i}\end{pmatrix}\right),$$

it is known that:

$$\begin{aligned} &u_i|\mathbf{y}_{pre:i} \sim N\left(\mu\prime, v\prime\right),\\ &\text{where } \mu\prime = \mathbf{X}_{trt:i}\boldsymbol{\beta} + \mathbf{v}_{21:i}\mathbf{V}_{pre:i}^{-1}\left(\mathbf{y}_{pre:i} - \mathbf{X}_i\boldsymbol{\beta}\right)\\ &\quad\text{and } v\prime = v_{22:i} - \mathbf{v}_{21:i}\mathbf{V}_{pre:i}^{-1}\mathbf{v}_{12:i}. \end{aligned}$$

If a patient has no pre-treatment observations, then the probability density function for the baseline value is simply that for a normal distribution with mean $\mathbf{X}_{trt:i}\boldsymbol{\beta}$ and variance $v_{22:i}$. We note that in forming the conditional distribution for u, we assume that it is independent of the decision to initiate treatment at t_{trt} given the value of t_{trt} and the observed $\mathbf{y}_{pre}$ for each patient.

The conditional distribution of each u_i is normal and so will include potential negative realisations, even if the probability of this is vanishingly small for most individuals. As such, we use the notation u_i^+ to indicate a latent variable for which all probability mass for values $u_i < 0$ is assigned instead to $u_i = 0$, i.e. $u_i^+ = Max(0, u_i)$.

5.6 Post-treatment model structure

5.6.1 Mean response to treatment

Although a range of models could be considered for the post-treatment observations, we focus on the use of an asymptotic regression model for the underlying mean structure. Such models have been used to describe CD4 recovery over several years from treatment initiation in children[116;117]. In our definition of this model, the mean value for the i^{th} individual at time after initiation of treatment t_{post}, conditional on the 'true' baseline value u_i^+, is given by the function:

$$g\left(t_{post}, u_i^+\right) = \phi_{1:i} + \left(u_i^+ - \phi_{1:i}\right)\exp\left(-\exp\left(\phi_{2:i}\right)t_{post}\right). \tag{8}$$

This function takes the value u_i^+ when $t_{post} = 0$ (i.e. at the exact time of treatment initiation), and it has a horizontal asymptote at $\phi_{1:i}$ as $t_{post} \to \infty$. The value of $\phi_{2:i}$ determines the speed of transition from u_i^+ to $\phi_{1:i}$, i.e. from the value of the response variable at baseline to its long-term mean, as t_{post} increases. The shape of the function is illustrated in Figure 5.2. It is useful to note that, as this function involves a

change from a baseline value to a long-term maximum that follows an 'exponential decay'-type curve, the 'half life' of this transition can be calculated as $\frac{\log(2)}{\exp(\phi_{2:i})}$; this facilitates interpretation of the estimated values of parameters that define $\phi_{2:i}$.

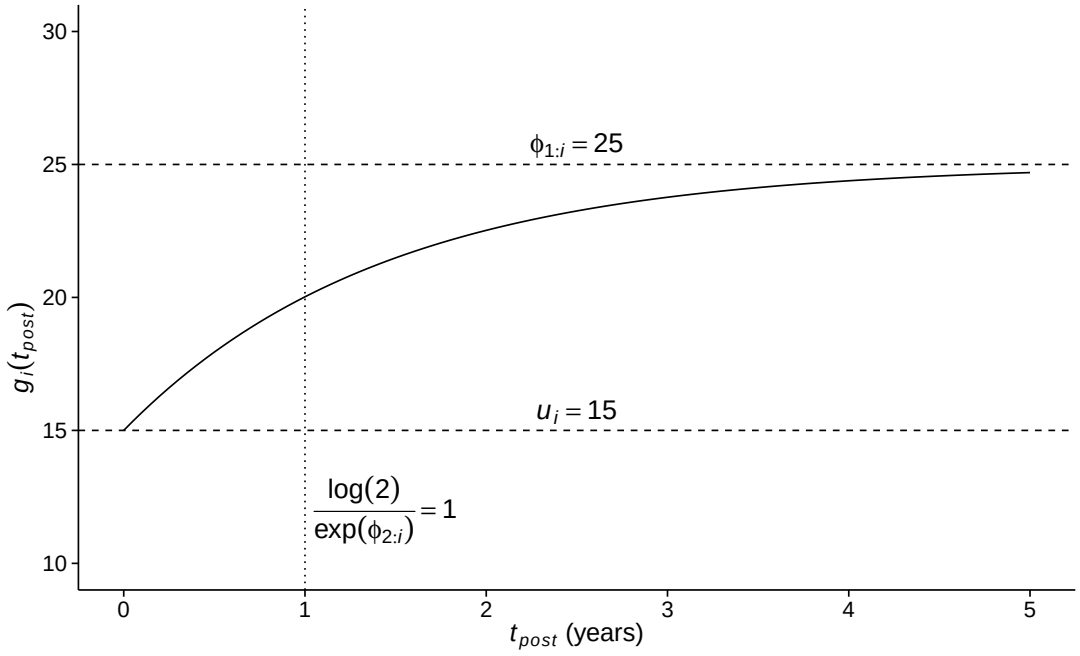

Figure 5.2. Illustrative plot of an asymptotic regression curve. Here the baseline (u_i) is set to 15, the asymptotic maximum ($\phi_{1:i}$) is set to 25 and the rate of recovery parameter ($\phi_{2:i}$) is set to log(log(2)), leading to a 'half-life' of 1.

In models of this type, the place of u_i^+ in this function is usually taken by a single parameter (or a linear function of a set of parameters) to be estimated, potentially with an associated subject-specific random effect term. However, we instead make use of the fact that a subject-specific distribution for u_i^+ can be included in the model conditioned on the observed pre-treatment data for that individual. Similarly, we will consider $\phi_{1:i}$ and $\phi_{2:i}$ as potentially being determined as a function of u_i^+, alongside other variables, i.e. we will investigate whether the long-term average value of the response variable and the speed at which this is attained are predicted by the 'true' value of the variable at treatment initiation.

5.6.2 Long-term maximum response to treatment

The simplest potential model for the long-term maximum response to treatment in each individual, i.e. the horizontal asymptote $\phi_{1:i}$, is to assume that this is equal to a single constant for the entire population:

$$\phi_{1:i} = A_1, \text{ for all } i.$$

The implication of this model is that the long-term response to treatment does not depend on the value of the variable in any given patient at treatment initiation, or on any other factors. This formulation also assumes that there is no random variation in the long-term maximum response between patients, but we will include a subject-specific random-effect term 'τ_i', alongside any deterministic function ($\phi_1(...)$), throughout:

$$\phi_{1:i} = \phi_1(...) + \tau_i, \text{ where } \tau_i \sim N(0, \Omega),$$

with the variance parameter Ω to be estimated. Although the post-treatment model defined in Equation (8) is non-linear in terms of the parameters, using this formulation it *is* linear in terms of the subject-specific random effect. As such $f_{post}\left(\mathbf{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right)$ can be expressed in closed form as a multivariate normal distribution (assuming no further random effect terms are added to the model), even though it does not constitute a linear mixed effects model conditioned on the unobserved baseline variable. Further details are given in Section 5.6.5.

The next model considered is that the expected long-term maximum (working on the square-root scale for CD4 counts) for any given patient follows a linear dependence on their 'true' value at treatment initiation:

$$\phi_1\left(u_i^+\right) = A_1 + A_2 u_i^+,$$

where A_1 and A_2 are parameters to be estimated.

We then wish to investigate whether ϕ_1 is a more complex, non-linear, function of u_i^+. One option would be to specify that ϕ_1 is some specific non-linear function of u_i^+. However, the fact that the relationship between $\phi_{1:i}$ and u_i^+ cannot be directly visualised using the raw data means that there is no obvious way to go about selecting the functional form. Another option is the use of cubic splines defined in terms of u_i^+, this approach has the advantage of allowing consideration of a wide variety of possible relationships between the predictive and outcome variable. In order to restrict the total number of model parameters and improve stability of optimisation, we make use of natural cubic splines derived from a truncated power series basis as described by Hastie, Tibshirani and Friedman [118]. We use knots at 15.5, 17.5, 19.5 and 22 in terms of square-root CD4, corresponding to approximately the 20$^{\text{th}}$, 40$^{\text{th}}$, 60$^{\text{th}}$ and 80$^{\text{th}}$ centiles of the last observed CD4 count before treatment initiation, when available, in the UK Register of HIV Seroconverters dataset. The use of natural cubic splines in this context was suggested by Ronald Geskus (*pers. com.*) following circulation of an early draft report on this work to investigators from the CASCADE collaboration.

We also consider models in which the relationship between the long-term maximum response and the baseline value (u_i^+) can vary according to the time elapsed between seroconversion and treatment initiation for each patient ($t_{trt:i}$). Although ideally this would be done using a smooth function of u_i^+ and $t_{trt:i}$, in this chapter for computational stability and simplicity of exposition we fit separate functions of u_i^+ stratified by $t_{trt:i}$ (in years) as follows: $0 \leq t_{trt:i} \leq 0.5$, $0.5 < t_{trt:i} \leq 1.0$ and $1.0 < t_{trt:i}$. These grouping were chosen based on a combination of findings reported previously in the literature, the level of uncertainty in terms of estimated dates of seroconversion in our study population and the need to ensure that an adequate number of patients were included in each group to allow parameter estimates to be obtained for the model.

Were patient characteristics (i.e. age, gender *etc.*) to be included in the model for $\phi_{1:i}$, and assuming a linear function in terms of u_i^+ for simplicity of exposition, we would have an extended function for ϕ_1 of the form:

$$\phi_1\left(u_i^+, \mathbf{x}_i\right) = A_1 + A_2 u_i^+ + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{\phi_1},$$

where $\mathbf{x}_i$ is the patient-specific vector of data specifying relevant characteristics and $\boldsymbol{\beta}_{\phi_1}$ is the associated vector of parameters that determines their effects.

5.6.3 Speed of response to treatment

As for the function for the long-term maximum value, we consider first a constant value for $\phi_{2:i}$ across the population ($\phi_{2:i} = B_1$) and secondly a linear dependence on u_i^+:

$$\phi_{2:i} = B_1 + B_2 u_i^+,$$

where B_1 and B_2 are parameters to be estimated. We then consider a natural cubic spline function of u_i^+, including an analysis with stratification according to groups defined by the time elapsed from seroconversion to treatment. The addition of a subject-specific random effect to this function was also considered, this required integration of the log-likelihood function over an additional latent variable for each patient and so the Laplace approximation was used.

5.6.4 Residual variance structure

We propose the following model for the vector of post-treatment observations ($\mathbf{y}_{post:i}$) for the i^{th} individual, conditioned on their 'true' baseline value at treatment initia-

tion (u_i^+):

$$\begin{aligned}
\mathbf{y}_{post:i}|_{u_i^+} &= \mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
\tau_i &\sim N(0, \Omega) \\
\mathbf{W}_{post:i} &\sim MVN(\mathbf{0}, \mathbf{\Sigma}_{post:i}) \\
\mathbf{e}_{post:i} &\sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{post:i}}).
\end{aligned}$$

The vector of observation times $\mathbf{t}_{post:i}$ relates to time since treatment initiation, with $n_{post:i}$ post-treatment observations for the i^{th} subject. The function $\mathbf{g}$ here represents a vectorised version of g in equation (8), i.e.:

$$\mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) = \begin{pmatrix} g\left(t_{post:i1}, u_i^+, \tau_i\right) \\ g\left(t_{post:i2}, u_i^+, \tau_i\right) \\ \vdots \\ g\left(t_{post:in_{post:i}}, u_i^+, \tau_i\right) \end{pmatrix}.$$

For the stochastic process component $\mathbf{W}_{post:i}$, we include a 'new' fractional Brownian motion process with value zero at time of treatment initiation and separate parameters to the pre-treatment process. The vector $\mathbf{e}_{post:i}$ represents independent residual measurement errors (or very short-term physiological variation), with a variance parameter (σ^2) that is shared with the pre-treatment model.

5.6.5 Marginal distribution for post-treatment model

Although the models that we have defined for the post-treatment data are non-linear in their parameters, they are all linear in their random terms conditional on the value of u_i^+:

$$\begin{aligned}
\mathbf{y}_{post:i}|_{u_i^+} =& \mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
=& \begin{pmatrix} \phi_{1:i} + \left(u_i^+ - \phi_{1:i}\right) \exp\left(-\exp\left(\phi_{2:i}\right) t_{post:i1}\right) \\ \vdots \\ \phi_{1:i} + \left(u_i^+ - \phi_{1:i}\right) \exp\left(-\exp\left(\phi_{2:i}\right) t_{post:in_{post:i}}\right) \end{pmatrix} + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
=& \begin{pmatrix} \phi_1\left(u_i^+\right) + \tau_i + \left(u_i^+ - \left(\phi_1\left(u_i^+\right) + \tau_i\right)\right) \exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right) t_{post:i1}\right) \\ \vdots \\ \phi_1\left(u_i^+\right) + \tau_i + \left(u_i^+ - \left(\phi_1\left(u_i^+\right) + \tau_i\right)\right) \exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right) t_{post:in_{post:i}}\right) \end{pmatrix} \\
&+ \mathbf{W}_{post:i} + \mathbf{e}_{post:i}
\end{aligned}$$

$$\begin{aligned}\tau_i &\sim N(0, \Omega)\\ \mathbf{W}_{post:i} &\sim MVN(\mathbf{0}, \mathbf{\Sigma}_{post:i})\\ \mathbf{e}_{post:i} &\sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{post:i}}).\end{aligned}$$

As such, for the models defined, the post-treatment data follows a marginal multivariate normal distribution conditional on the value of u_i^+, with mean vector given by:

$$\mathrm{E}\left[\mathbf{y}_{post:i}\right]_{|u_i^+} = \begin{pmatrix} \phi_1\left(u_i^+\right) + \left(u_i^+ - \phi_1\left(u_i^+\right)\right)\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:i1}\right) \\ \vdots \\ \phi_1\left(u_i^+\right) + \left(u_i^+ - \phi_1\left(u_i^+\right)\right)\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:in_{post:i}}\right) \end{pmatrix},$$

and covariance matrix given by:

$$\mathrm{Var}\left[\mathbf{y}_{post:i}\right]_{|u_i^+} = \mathbf{Q}_i + \mathbf{\Sigma}_{post:i} + \sigma^2 \mathbf{I}_{n_{post:i}},$$

where the jk^{th} element of $\mathbf{Q}_i$, $q_{i_{jk}}$, is given by:

$$\begin{aligned} q_{i_{jk}} &= \mathrm{Cov}\left[\left(1-\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:ij}\right)\right)\tau_i,\ \left(1-\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:ik}\right)\right)\tau_i\right] \\ &= \left(1-\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:ij}\right)\right)\times\left(1-\exp\left(-\exp\left(\phi_2\left(u_i^+\right)\right)t_{post:ik}\right)\right)\times\Omega. \end{aligned}$$

5.7 Differences in variability between patients

In Chapter 4 we demonstrated that generalisation of the model structure for pre-treatment CD4 counts as described in Section 5.4 to a multivariate-t distribution leads to a substantial improvement in model fit in terms of the log-likelihood and residual diagnostic plots. However, the application of a marginal multivariate-t distribution is not possible in the current setting, in which a combined model is defined for pre- and post-treatment data. We instead consider models in which the stochastic process components before and after treatment each follow a marginal multivariate-t distribution, with correlated scaling variables. We checked that this was the optimal model structure using only the pre-treatment CD4 data from the UK Register of HIV Seroconverters cohort, as this allowed the use of high-dimensional (15-point) adaptive Gauss–Hermite quadrature for maximum likelihood estimation (with integration only required for one latent scaling variable per patient). For a marginal multivariate-t model the log-likelihood was −14221.5, whilst for models in which the stochastic process component (i.e. fractional Brownian motion) alone was multivariate-t it was −14220.1 and for the measurement error alone it was −14229.8;

for all of these models the number of parameters was nine, and so the model with between-patient differences in variability of the stochastic process component was indeed optimal in terms of AIC or BIC. We note that maximum likelihood estimation of this model did not converge when tested on the dataset analysed in Chapter 4 and we speculate that the later cut-off relating to date of seroconversion for inclusion and the use of data from only a single cohort in the present analysis may be associated with higher data quality (in terms of data-entry errors *etc.*). This is supported by the fact that for the analysis in Chapter 4 we deleted observations when an identical CD4 count was recorded 1 day after an initial observation as likely data-entry errors, but there were no such records for the data included in the present analysis.

The desired model structure for a combined analysis of pre- and post-treatment data requires the use of a bivariate gamma distribution, of which a number are available (as reviewed by Balakrishna and Lai[119]). Such models will include three latent variables per patient, and as such a Laplace approximation to the log-likelihood[5;50;57] rather than adaptive Gauss–Hermite quadrature will be used. Because of this, Moran's bivariate gamma distribution[119;120] makes a natural choice. This distribution is defined by first transforming random variables (A and B) from the standard normal bivariate distribution with correlation ρ_{Moran} into a copula $C(\Phi(a),\Phi(b))$, where Φ is the standard normal cumulative distribution function, and secondly using the inverse cumulative distribution functions of univariate gamma distributions ($\Gamma_1 = F^{-1}(\Phi(A))$, $\Gamma_2 = G^{-1}(\Phi(B))$) to find the joint distribution function of Γ_1 and Γ_2 (each of which has a marginal univariate gamma distribution). F is here defined as the cumulative distribution function for gamma distribution with 'shape' and 'rate' parameters both equal to $\frac{\nu_1}{2}$, whilst G is that for the gamma distribution with parameters both equal to $\frac{\nu_2}{2}$.

The model for pre-treatment CD4 counts is then defined as:

$$\begin{aligned}
\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\
\mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\
\mathbf{W}_{pre:i}|_{\gamma_{1:i}} &\sim MVN(\mathbf{0}, \frac{1}{\gamma_{1:i}}\boldsymbol{\Sigma}_{pre:i}) \\
\mathbf{e}_{pre:i} &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{pre:i}}),
\end{aligned}$$

whilst, the model for post-treatment data is:

$$\begin{aligned}
\mathbf{y}_{post:i}|_{u_i^+} &= \mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
\tau_i &\sim N(0, \Omega) \\
\mathbf{W}_{post:i}|_{\gamma_{2:i}} &\sim MVN(\mathbf{0}, \frac{1}{\gamma_{2:i}}\boldsymbol{\Sigma}_{post:i})
\end{aligned}$$

$$\mathbf{e}_{post:i} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{post:i}}),$$

with the scaling factors jointly following Moran's bivariate gamma distribution:

$$\begin{pmatrix} \gamma_{1:i} \\ \gamma_{2:i} \end{pmatrix} \sim Moran\left(\rho_{Moran}; \frac{\nu_1}{2}, \frac{\nu_1}{2}; \frac{\nu_2}{2}, \frac{\nu_2}{2}\right).$$

The ν_1 and ν_2 parameters equate to the degrees of freedom for the pre- and post-treatment parts of the model, respectively.

This specific bivariate gamma distribution is a natural choice because the marginal log-likelihood function for the model can be found by integrating out the latent variables on the standard normal scale, for which the Laplace approximation is optimally accurate[64], as follows (omitting indexing for each individual and dependence on model parameters):

$$f\left(\mathbf{y}_{pre}, \mathbf{y}_{post}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}|\gamma_1 = F^{-1}\left(\Phi\left(a\right)\right)\right) f_{post}\left(\mathbf{y}_{post}|u, \gamma_2 = G^{-1}\left(\Phi\left(b\right)\right)\right)$$
$$f_u\left(u|\mathbf{y}_{pre}, \gamma_1 = F^{-1}\left(\Phi\left(a\right)\right)\right) f_{ab}\left(a, b\right) du\, da\, db,$$

where f_{ab} is the probability density function for a standard bivariate normal distribution with correlation ρ_{Moran}. The ρ_{Moran} parameter can be estimated from the data through maximum likelihood estimation as for other model parameters.

5.8 Overall model structure and interpretation

A directed acyclic graph depicting the proposed model structure is shown in Figure 5.3. For simplicity, we omit here the extension to the basic model in which further latent variables are added to the model to allow between-patient differences in variability over time as described in Section 5.7. This diagram illustrates the fact that in the model, response to treatment is linked to pre-treatment data only through the 'true' baseline value u and the time from seroconversion to treatment initiation. These links are mediated through variables representing the long-term maximum response to treatment (ϕ_1) and the speed at which this is attained (ϕ_2) in each patient. When fitted to the dataset under investigation, this structure should allow estimates of individual parameters of the model to be interpreted in a meaningful way. It is be relatively straightforward to extend the model to assess whether patient characteristics such as age and gender or drug regimen choice are independently predictive of response to treatment, and such extensions are considered in Chapters 6 and 7.

The primary interpretation of our models as presented is the prediction of the response to HAART in terms of prior CD4 counts and time from seroconversion. It

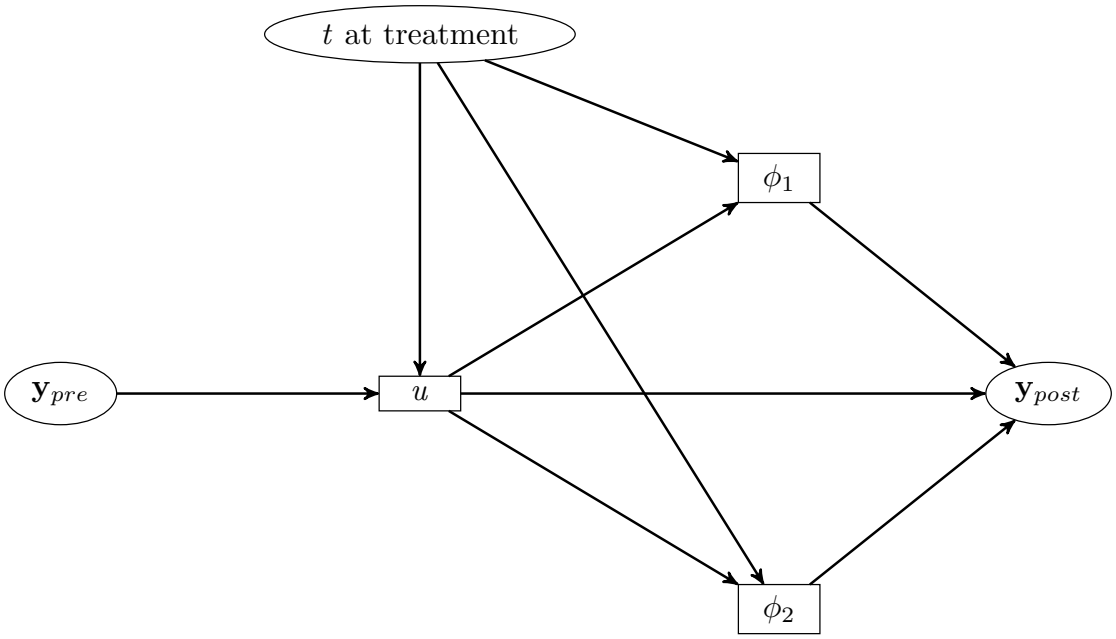

Figure 5.3. Directed acyclic graph depicting the proposed model structure for each patient. Observed variables are shown within ellipses, whilst unobserved latent variables are shown within rectangles.

has been argued that causal effects can only be estimated from observational studies with respect to clearly defined interventions[121]. Whilst interventions with regard to the monitoring of CD4 counts and guidelines for treatment initiation can be defined within the present context, it is not possible to begin treatment conditional on the 'true' value of a patient's CD4 count, as this cannot be observed directly. Furthermore it is not possible to define a treatment policy in terms of a specific simultaneous combination of 'time from seroconversion' and 'true CD4 count', when in a certain period a patient may only experience a limited range of CD4 counts. Hence, the links in Figure 5.3 show dependencies in the fitted probability model rather than direct causal effects.

As we have censored patients at recorded interruption of HAART but not according to VL status, the fitted models can be taken to represent treatment response for all patients were they all to remain on HAART (regardless of success or failure of virological suppression). All included patients had at least one post-HAART CD4 observation, but beyond this the number and timing of CD4 cell counts recorded for each individual were highly variable. We have assumed that the missingness of observations can be treated as MAR (following the terminology of Rubin[100]), i.e. that the 'missingness' of any observation is independent of the unobserved data conditional on the observed values of the outcome variable and any other covariates included in the model. Similarly we assume that the timing of observations is dependent only on previously observed outcomes, under which condition maximum likelihood estimation of a model for the outcome variable alone is consistent, without the need

for specification of a model for the distribution of follow-up times[101]. As noted in Chapter 4, a correct model specification is required to guarantee consistency of parameter estimates under both of these assumptions, and the 'timing of observations assumption' can be subsumed into the MAR assumption given that observations are recorded at discrete days.

5.9 Maximum likelihood estimation

All models presented have been fitted by direct maximum likelihood estimation using the open source AD Model Builder software (Version 11.2; ADMB Foundation)[50]. The 'random effects' mode was used for ADMB, allowing optimisation of a log-likelihood function with automated integration over latent variables[57], as described in Section 2.4. The log-likelihood function for each individual (for their complete pre- and post-treatment data) was defined using the 'separable function' utility, allowing computational efficiency to be gained from the modelled independence of each individual. 15-point adaptive Gauss–Hermite quadrature was used to obtain the maximum likelihood estimates for all models described in this chapter for which only one latent variable was included per individual (i.e. the 'true' baseline). However, for the models including additional latent variables associated with between-patient differences in variability over time, and for those tested with an additional random effect for the speed of post-treatment recovery, Gauss–Hermite quadrature was not feasible and the Laplace approximation was used.

Models were parameterised using logarithmic, logistic and generalised logistic transformations where appropriate such that parameter estimates could be obtained using unrestricted optimisation (e.g. maximum likelihood estimation was carried out using log-transformed variance parameters, with a parameter space of $(-\infty, +\infty)$ rather than $[0, +\infty)$). For all model parameters, confidence intervals are reported derived from the estimated asymptotic multivariate normal sampling distribution based on the observed information on the transformed scales. The 'R2admb' package[94] was used to output data files in the necessary format through the R statistical computing environment (R Foundation, Vienna, Austria).The ggplot2 package for R[122] was used for statistical graphics. All maximum likelihood estimates reported in this chapter were obtained using a computer cluster running with Linux operating systems (UCL Legion High Performance Computing Facility). Fitting each of the models presented to the UK Register of HIV Seroconverters dataset took between 1 and $2\frac{1}{2}$ hours (using a core with 4GB RAM), whereas fitting one of the models using a mid–low specification personal laptop (4GB RAM, Celeron Dual-Core CPU T3500 @ 2.1 GHz) required around 10 hours.

When considering only a single latent variable per patient, nested models are

compared using the generalised likelihood ratio test, comparing the change in 2×log-likelihood ($\Delta 2\ell$) to a χ^2 distribution. Non-nested models are compared using the BIC statistic, using the total number of observations in the dataset for the calculation of the penalty term. It is worth noting that these methods are only valid because adaptive Gauss–Hermite quadrature can be used to calculate the log-likelihood of the fitted models to a high degree of accuracy[5]; this is not the case for less computationally intensive approximations of the log-likelihood.

5.10 Model fitting

Summaries of the set of models fitted to the UK Register of HIV Seroconverters dataset are presented in Table 5.1, and to facilitate their interpretation Table 5.2 provides a description of each model parameter. The most basic model considered included constant parameters for the mean long-term maximum CD4 count (on square-root scale) and the rate of recovery from baseline at treatment initiation, without division of patients according to time from seroconversion to initiation of HAART ($Model_1$ in Table 5.1). Modelling the long-term maximum (ϕ_1) and speed of response to treatment (ϕ_2) as linear functions of the baseline value in each individual (u_i^+) led to a significant improvement in model fit ($Model_2$ *vs* $Model_1$, $\Delta 2\ell$ 460.4 for 2 parameters; $P < 0.0001$). A model equivalent to $Model_2$ but without pre- and post-treatment stochastic process components was also fitted for comparison and was found to have a much higher BIC value (64 398); correspondingly the model including stochastic processes showed a significant improvement in fit ($\Delta 2\ell$ 844.8 for 4 parameters; $P < 0.0001$). The extension of $Model_2$ to allow natural cubic spline functions to define the relationships between u_i^+ and ϕ_1 and ϕ_2 led to a further significant improvement in model fit ($Model_3$ *vs* $Model_2$, $\Delta 2\ell$ 31.4 for 4 parameters; $P < 0.0001$).

Table 5.1. Summary of the results of combined models for pre- and post- highly active antiretroviral therapy (HAART) CD4 cell count data, after square root transformation, for patients from the UK Register of HIV Seroconverters dataset. The interpretation of each model parameter is listed in Table 5.2.

	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$
β_0	22.44 (22.13 to 22.74)	22.45 (22.16 to 22.74)	22.44 (22.15 to 22.73)	22.26 (21.96 to 22.56)	22.26 (21.96 to 22.56)	22.23 (21.94 to 22.53)
β_1	-1.36 (-1.52 to -1.2)	-1.39 (-1.55 to -1.23)	-1.39 (-1.55 to -1.23)	-1.3 (-1.46 to -1.14)	-1.32 (-1.47 to -1.16)	-1.36 (-1.5 to -1.21)
U_{00}	12.37 (10.64 to 14.37)	13.39 (11.77 to 15.23)	13.42 (11.79 to 15.28)	14.43 (12.68 to 16.43)	14.53 (12.77 to 16.54)	12.92 (11.29 to 14.8)
ρ	-0.65 (-0.79 to -0.44)	-0.86 (-0.99 to 0.18)	-0.84 (-0.98 to -0.1)	-0.95 (-1 to 1)	-0.92 (-1 to 0.91)	-0.63 (-0.76 to -0.44)
U_{11}	0.55 (0.33 to 0.93)	0.25 (0.08 to 0.75)	0.28 (0.1 to 0.75)	0.2 (0.05 to 0.74)	0.21 (0.06 to 0.74)	0.49 (0.31 to 0.77)
κ_{pre}	9.68 (8.77 to 10.68)	5.91 (5.23 to 6.67)	5.9 (5.22 to 6.68)	5.99 (5.29 to 6.8)	5.92 (5.21 to 6.72)	5.37 (4.37 to 6.6)
H_{pre}	0.11 (0.09 to 0.14)	0.3 (0.25 to 0.37)	0.3 (0.24 to 0.36)	0.31 (0.25 to 0.37)	0.31 (0.25 to 0.38)	0.16 (0.13 to 0.19)
σ	1.25 (1.09 to 1.42)	1.95 (1.89 to 2.01)	1.94 (1.87 to 2)	1.92 (1.85 to 1.99)	1.92 (1.86 to 1.99)	1.32 (1.19 to 1.46)
ϕ_1 model: long-term maximum	Constant for all patients	Linear for all patients	NCS for all patients	Linear for all patients stratified by ART_t	Linear for early treatment groups, NCS for late group	Linear for all patients stratified by ART_t
$At1_1$	—	—	—	7.04 (4.75 to 9.33)	7.06 (4.77 to 9.35)	8.44 (6.05 to 10.83)
$At1_2$	—	—	—	0.9 (0.79 to 1.01)	0.9 (0.79 to 1)	0.84 (0.72 to 0.95)
$At2_1$	—	—	—	10.73 (7.93 to 13.53)	10.68 (7.85 to 13.51)	12.32 (9.28 to 15.35)
$At2_2$	—	—	—	0.67 (0.54 to 0.81)	0.67 (0.53 to 0.81)	0.64 (0.47 to 0.8)
A_1	25.93 (25.49 to 26.36)	11.42 (9.74 to 13.09)	5.1 (0.3 to 9.9)	14.58 (12.3 to 16.86)	3.76 (-1.99 to 9.51)	14.35 (12.32 to 16.38)
A_2	—	0.69 (0.62 to 0.77)	1.14 (0.84 to 1.44)	0.55 (0.44 to 0.66)	1.23 (0.86 to 1.6)	0.57 (0.46 to 0.67)
A_3	—	—	-0.43 (-0.64 to -0.22)	—	-0.32 (-0.63 to -0.01)	—
A_4	—	—	0.82 (0.43 to 1.2)	—	0.52 (-0.07 to 1.11)	—
ϕ_2 model: recovery speed	As ϕ_1	As ϕ_1	As ϕ_1	As ϕ_1	As ϕ_1	As ϕ_1
$Bt1_1$	—	—	—	2.66 (0.52 to 4.79)	2.8 (0.76 to 4.84)	5.68 (2.94 to 8.43)
$Bt1_2$	—	—	—	0.02 (-0.08 to 0.11)	0.01 (-0.08 to 0.1)	-0.14 (-0.29 to -1.98e-03)
$Bt2_1$	—	—	—	-0.99 (-3 to 1.02)	-0.92 (-2.97 to 1.13)	0.23 (-1.39 to 1.86)
$Bt2_2$	—	—	—	0.15 (0.05 to 0.26)	0.15 (0.04 to 0.26)	0.01 (-0.1 to 0.12)
B_1	-0.16 (-0.3 to -0.02)	-3.34 (-4.19 to -2.48)	1.82 (-0.23 to 3.87)	-3.64 (-4.7 to -2.59)	2.42 (0.26 to 4.58)	-2.25 (-3.3 to -1.21)
B_2	—	0.24 (0.2 to 0.28)	-0.11 (-0.24 to 0.02)	0.23 (0.17 to 0.29)	-0.15 (-0.29 to -0.02)	0.13 (0.07 to 0.19)
B_3	—	—	0.28 (0.19 to 0.38)	—	0.19 (0.04 to 0.33)	—
B_4	—	—	-0.52 (-0.71 to -0.33)	—	-0.28 (-0.58 to 0.02)	—
Ω	11.09 (8.76 to 14.03)	2.97 (2.09 to 4.23)	3.05 (2.13 to 4.38)	3.07 (2.19 to 4.31)	3.31 (2.39 to 4.59)	2.72 (1.71 to 4.31)
κ_{post}	7.59 (6.79 to 8.49)	3.09 (2.46 to 3.89)	3.17 (2.53 to 3.98)	3.36 (2.7 to 4.18)	3.3 (2.66 to 4.11)	4.33 (3.5 to 5.36)
H_{post}	0.08 (0.07 to 0.1)	0.42 (0.32 to 0.52)	0.4 (0.3 to 0.5)	0.38 (0.29 to 0.48)	0.39 (0.3 to 0.5)	0.13 (0.11 to 0.16)
Dif in var	No	No	No	No	No	Yes
v_1	—	—	—	—	—	3.84 (3.06 to 4.82)
v_2	—	—	—	—	—	4.28 (3.4 to 5.38)
ρ_{Moran}	—	—	—	—	—	0.37 (0.19 to 0.52)
n_{pars}	13	15	19	23	27	26
ℓ	-31954.8	-31724.6	-31708.9	-31664.5	-31656.5	-31299.7*
BIC	64032.85	63591.41	63597.94	63547.06	63568.98	62845.9*

Parameter estimates are given with 95 % confidence intervals in parentheses. *Not comparable to other values in Table, as calculated using Laplace approximation. ART_t, time from seroconversion to treatment initiation; BIC, Bayesian information criterion; Dif in var, differences in variability between patients; ℓ, log-likelihood; NCS, natural cubic spline; n_{pars}, number of parameters estimated in model.

Fitting a model with separate linear relationships between u_i^+ and ϕ_1 and ϕ_2 according to timing of HAART subgroup ($Model_4$) led to a reduction in BIC relative to the single-group natural cubic spines model. It was not possible to obtain a model fit for natural cubic spline functions defined separately for each subgroup (due to lack of convergence), but allowing linear functions in the early start subgroups in combination with natural cubic spline functions for the remaining patients led to a further improvement in model fit ($Model_5$ *vs* $Model_4$, $\Delta 2\ell$ 16.0 for 4 parameters; $P = 0.003$). However, $Model_4$, with linear link functions for all subgroups, retained the lowest BIC value and so we have focused on interpretation of this model.

It is harder to make a direct comparison for $Model_6$, which matches $Model_4$ with the addition of jointly distributed latent scaling variables for the pre- and post-treatment fractional Brownian motion processes. Because of the need to integrate the log-likelihood function over multiple latent variables, parameter estimates for $Model_6$ were obtained using the Laplace approximation, meaning that generalised likelihood ratio tests or comparisons of the BIC statistic are not appropriate. However, the low values obtained for the estimates of the pre- and post-treatment degrees of freedom parameters (which are effectively fixed at $+ve\infty$ for the other models considered) indicate that this model may better reflect the structure of the observed data. Convergence of parameter estimates was not achieved when the same extension was made to $Model_5$.

Convergence of parameter estimates also failed when a subject-specific random effect was added to the speed of response to treatment function (ϕ_2) for $Model_4$, $Model_5$ or $Model_6$. We also attempted to extend each of these models to allow an independent linear effect of the patient-specific slope of pre-HAART decline (requiring an additional two latent variable per patient for their random intercept and slope terms), but convergence of parameter estimates was not achieved in each case. Using $Model_4$, we checked the assumption that the pre- and post-HAART measurement error variance can be treated as constant, and no significant improvement in model fit was observed when separate parameters were fitted for the two periods ($\Delta 2\ell$ 0.6 for 1 parameter; $P = 0.44$).

5.11 Model interpretation

All models fitted (other than $Model_1$ by definition) showed a positive association between baseline CD4 count at HAART and the long-term maximum; this finding was consistent across subgroups of patients defined by timing of treatment initiation with only relatively small differences in the fitted functions for each group in models 4–6 (Figures 5.4, 5.5 and 5.6). When modelled as a linear function across all patients (i.e. $Model_2$), the speed of response to treatment also showed a positive as-

Table 5.2. Description of parameters for combined models of pre- and post-treatment data. Some of the parameters relate to the link functions between the 'true' value of the response variable at treatment initiation, u_i^+, and the post-treatment model.

Model parameter	*Description*
β_0	Pre-treatment mean intercept.
β_1	Pre-treatment mean slope.
U_{00}	Pre-treatment intercept subject-specific random effect variance.
ρ	Correlation between pre-treatment intercept and slope subject-specific random effects.
U_{11}	Pre-treatment slope subject-specific random effect variance.
σ	Standard deviation of residual error term for each measurement, shared by pre- and post-treatment parts of model.
κ_{pre}	Scale parameter for pre-treatment fBM process.
H_{pre}	Hurst index for pre-treatment fBM process.
ϕ_1 model	These parameters relate to the long-term maximum value of the response variable after treatment initiation.
$At1_1, At1_2$	Intercept and slope terms in relationship with u_i^+ for patients treated within 6 months of seroconversion.
$At2_1, At2_2$	Intercept and slope terms in relationship with u_i^+ for patients treated beyond 6 months but within 1 year of seroconversion.
A_1, A_2	Intercept and slope terms in relationship with u_i^+ for linear or NCS models*.
A_3, A_4	Third and fourth coefficients for NCS models*.
ϕ_2 model	These parameters relate to the rate of recovery of the response variable after treatment initiation.
$Bt1_1, Bt1_2$	Intercept and slope terms in relationship with u_i^+ for patients treated within 6 months of seroconversion.
$Bt2_1, Bt2_2$	Intercept and slope terms in relationship with u_i^+ for patients treated beyond 6 months but within 1 year of seroconversion.
B_1, B_2	Intercept and slope terms in relationship with u_i^+ for linear or NCS models*.
B_3, B_4	Third and fourth coefficients for NCS models*.
Ω	Residual variance for long-term maximum ($\phi_{1:i}$) not explained by u_i^+.
κ_{post}	Scale parameter for post-treatment fBM process.
H_{post}	Hurst index for post-treatment fBM process.
ν_1	Degrees of freedom parameter for pre-treatment stochastic process.
ν_2	Degrees of freedom parameter for post-treatment stochastic process.
ρ_{Moran}	Correlation parameter for latent scaling variables of pre- and post-treatment stochastic processes.

*Only applicable to patients with treatment initiation more than 1 year after seroconversion when separate terms are included for earlier groups. fBM, fractional Brownian motion; NCS, natural cubic spline.

sociation with baseline CD4 count at HAART. However, when the link function was defined by HAART-timing subgroup, the speed of response to treatment was found to be substantially higher at moderate and lower baseline CD4 counts (below around 25 on the square-root scale) in those patients who started treatment within 6 months of seroconversion, with an intermediate difference observed for the subgroup who started treatment after 6 months but within 1 year. This overall pattern of findings was consistent across models 4–6, although the exact shape of the link functions showed some differences.

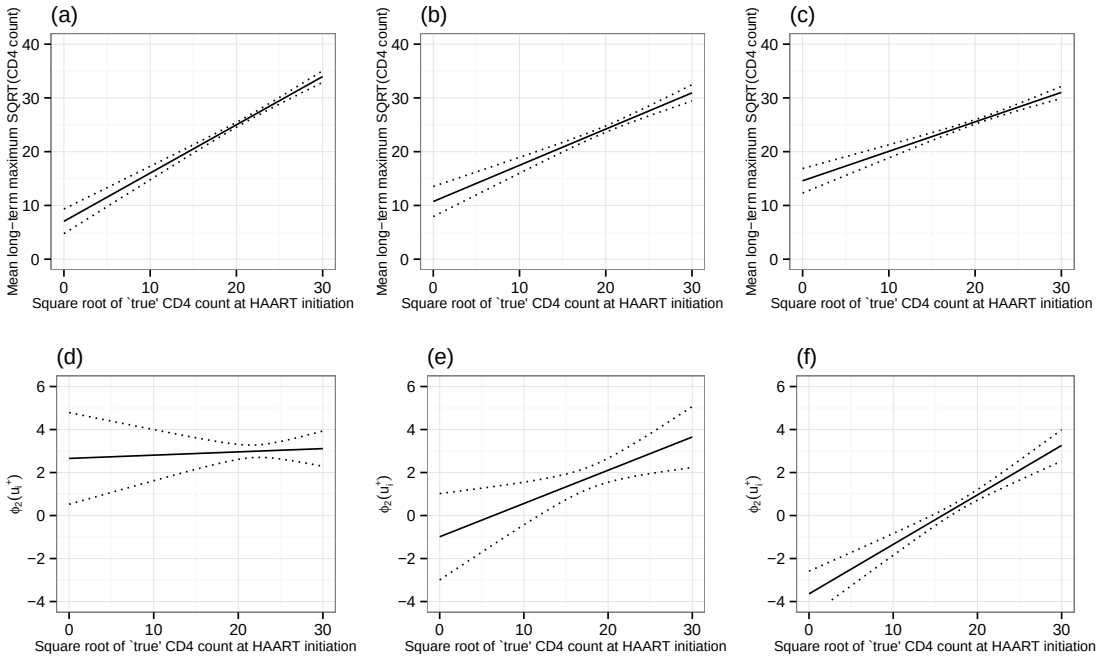

Figure 5.4. Plots of $\phi_1\left(u_i^+\right)$ (a–c, relating to long-term maximum) and $\phi_2\left(u_i^+\right)$ (d–f, relating to speed of response) for $Model_4$. Graphs on the left of each row (a,d) show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre (b,e) show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right (c,f) show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown (·····).

As the full vector of pre- and post-treatment data and u_i for each individual do not jointly follow a multivariate normal distribution, it is not possible to derive a closed form for the posterior predictive distribution of the u_i conditioned on the observed data in the way that would be done for the realizations of the random effects in a linear mixed model. However, the values of u_i for each individual that maximise $f\left(\mathbf{y}_{pre:i}, \mathbf{y}_{post:i}, u_i\right)$, $\hat{u}_i$, conditional on the current values of the model parameters, are calculated at each iteration of the adaptive Gauss–Hermite quadrature algorithm. The values of $\hat{u}_i$ corresponding to the final parameter estimates for each model are returned by ADMB, and these correspond to the posterior mode of $f_{u|\mathbf{Y}_{pre}=\mathbf{y}_{pre},\mathbf{Y}_{post}=\mathbf{y}_{post}}(u)$ for each individual. Kernel density plots for the u_i values for

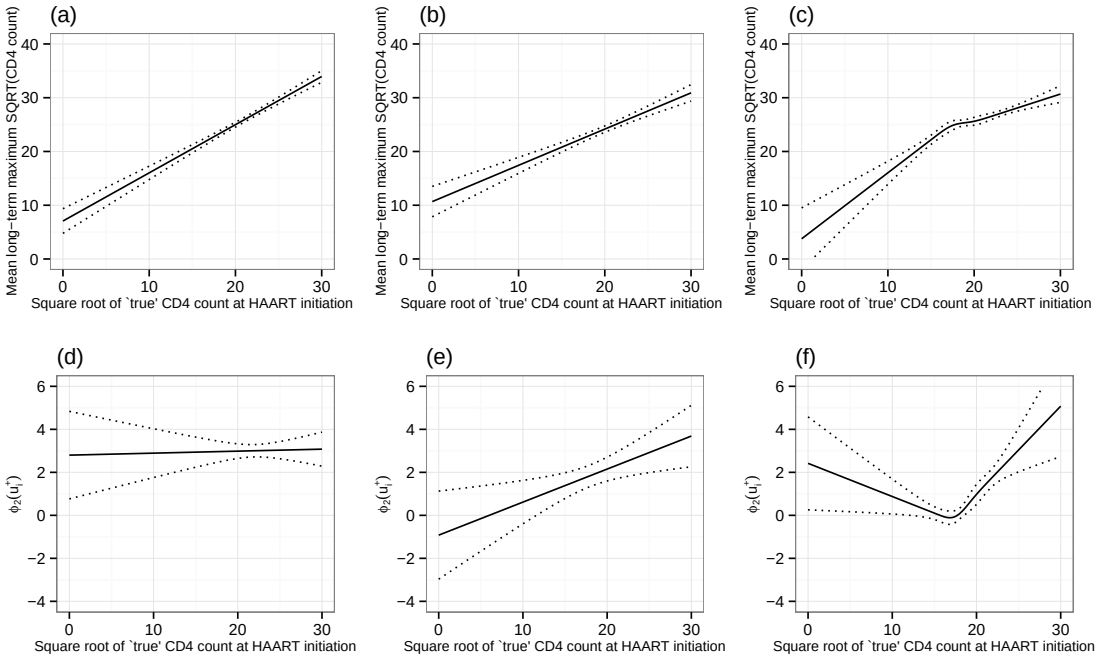

Figure 5.5. Plots of $\phi_1(u_i^+)$ (a–c, relating to long-term maximum) and $\phi_2(u_i^+)$ (d–f, relating to speed of response) for $Model_5$. Graphs on the left of each row (a,d) show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre (b,e) show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right (c,f) show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown (·····).

each subgroup in $Model_4$ are presented in Figure 5.7, approximating the distribution for $f_{u|\mathbf{Y}_{pre}=\mathbf{y}_{pre},\mathbf{Y}_{post}=\mathbf{y}_{post}}(u)$ as normal and making use of subject-specific standard deviation estimates also resulting from the adaptive Gauss–Hermite quadrature algorithm. Equivalent plots for $Model_5$ and $Model_6$ did not show substantial differences. Histograms of the last observed square-root CD4 count before treatment for those individual in whom this was recorded within 6 months of treatment initiation are also presented in Figure 5.7 for comparison, showing a similar shaped distribution in each subgroup. As expected given the results of previous simulations regarding treatment initiation based on observed CD4 cell counts[27], for more than half of patients (63 %) the mode of the posterior predictive distribution ($\hat{u}_i$) was greater than the last observed CD4 count (where available within 6 months); the median difference for $CD4_{last\_obs} - \hat{u}_i$ was −18 cells/μL when transformed back to the original measurement scale.

Predicted post-treatment ranges for CD4 cell counts based on $Model_4$ are shown in Figure 5.8 for patients with a 'true' CD4 counts at initiation of HAART of 200, 350 and 500 cells/μL. These charts further illustrate the model predictions that, in general, patients with a higher CD4 cell count at treatment initiation will go on to show a higher long-term maximum and will attain higher values more quickly after the

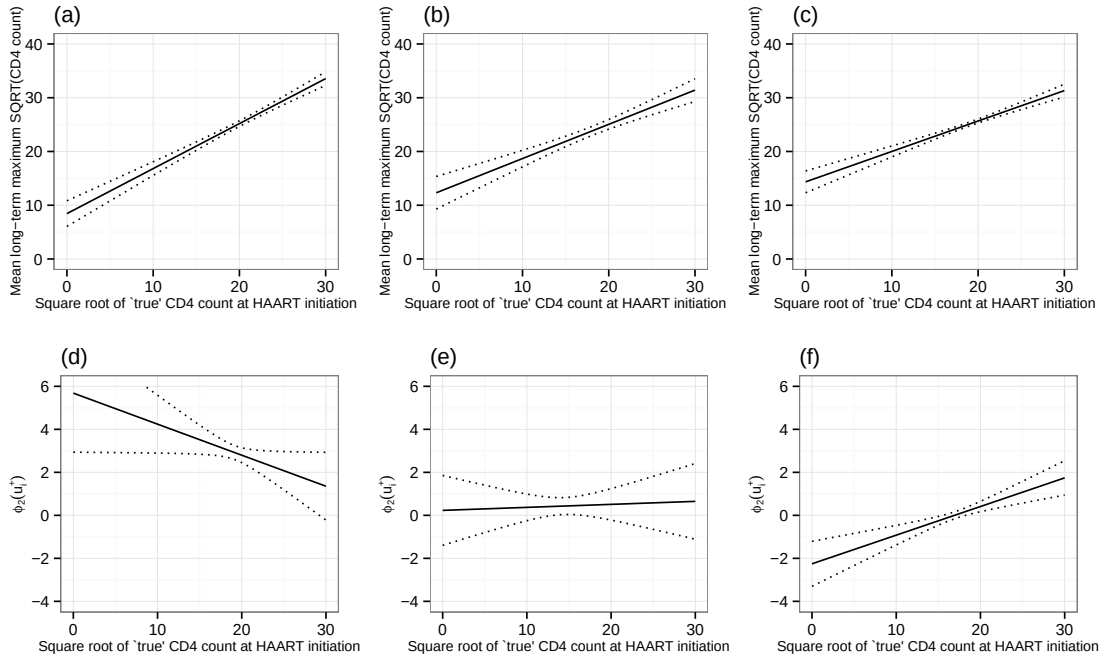

Figure 5.6. Plots of $\phi_1(u_i^+)$ (a–c, relating to long-term maximum) and $\phi_2(u_i^+)$ (d–f, relating to speed of response) for $Model_6$. Graphs on the left of each row (a,d) show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre (b,e) show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right (c,f) show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown (·····).

start of treatment, but that response to treatment is rapid if it is initiated within 6 months of seroconversion regardless of baseline CD4. These charts also illustrate that the model predicts considerable variability in response to treatment between patients at any given baseline CD4 value. However, in the models presented in this chapter we have not included variables such as patient age, gender and mode of infection that may also be predictive of response to treatment, and so it is possible that more fully developed models would include less unexplained variance in the long-term response to treatment. The inclusion of such potential confounding variables may also affect estimates of the influence of baseline value of CD4 at treatment initiation on each patient's response to treatment. Equivalent plots for $Model_6$ are presented in Figure 5.9, showing a very similar overall pattern of predictions. One interesting difference is that the inclusion of between-patient differences in variability in $Model_6$ leads to a more stable overall variance beyond around 2 years after initiation of HAART, this is due to the much lower estimate of the H index (0.13 for $Model_6$ *vs* 0.38 for $Model_4$) for the fractional Brownian motion process, which indicates stronger reversion towards the mean level for each patient over time.

For $Model_6$, estimates of the pre- and post-treatment degrees of freedom parameters (3.84 (95 % CI, 3.06–4.82) and 4.28 (3.4–5.38), respectively) indicate that there

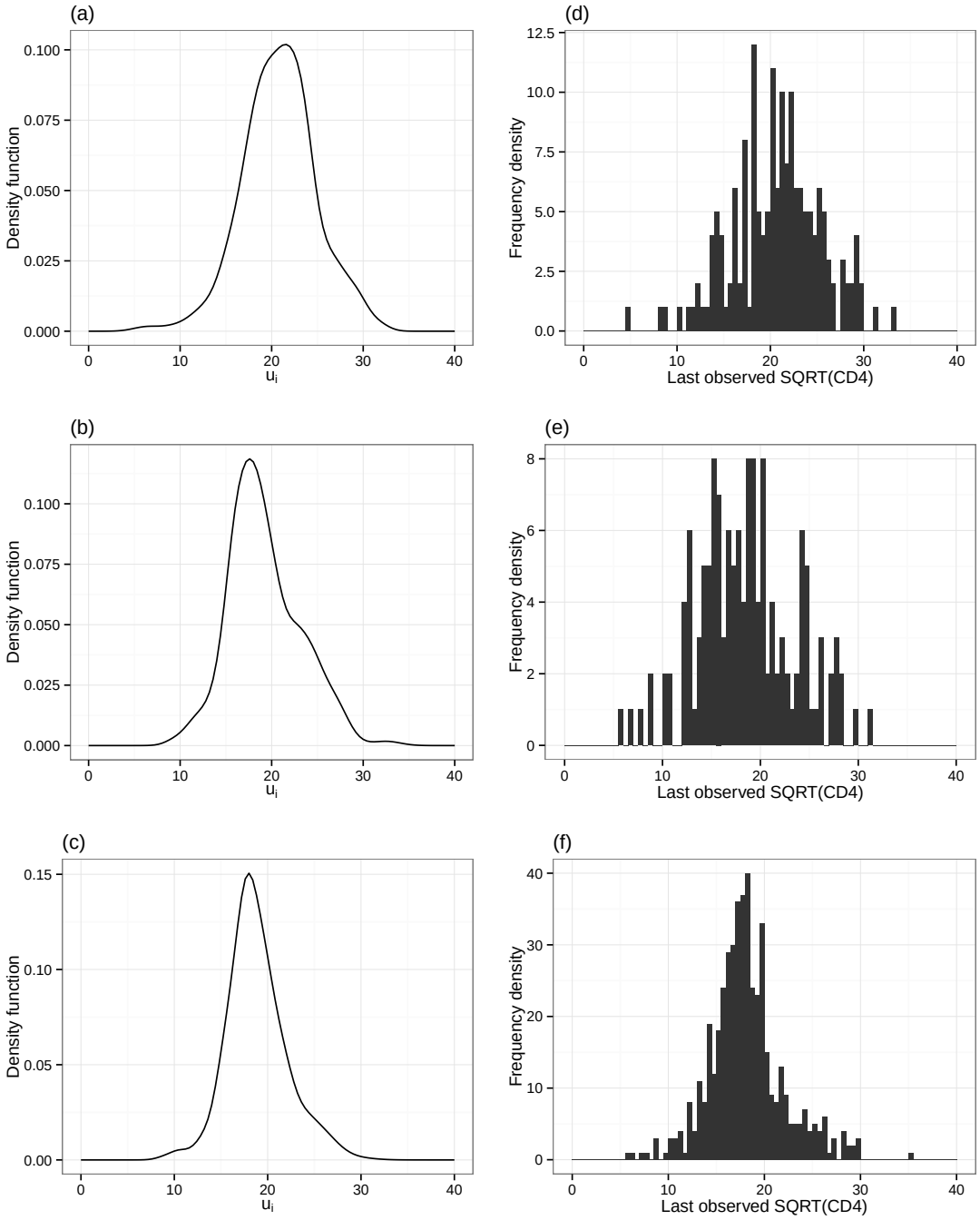

Figure 5.7. Kernel density plots (a–c) for the 'true' baseline square root CD4 counts based on $Model_4$ and (d–f) histograms of the last observed square-root CD4 count before treatment. (a–c) Kernel density plots for the 'true' baseline square root CD4 counts for each individual (u_i), approximating the posterior distribution of each as normal (with subject-specific standard deviation as estimated during model fitting), and (d–f) histograms of the last observed square-root CD4 count before treatment for those individual in whom this was recorded within 6 months of treatment initiation ($n = 170$, $n = 141$ and $n = 486$, respectively). Graphs in the top row (a,d) relate to patients initiating treatment within 6 months of seroconversion, those in the centre row (b,e) relate to patients initiating treatment beyond 6 months but within 1 year and those on the lower row (c,f) are for patients who started treatment beyond 1 year.

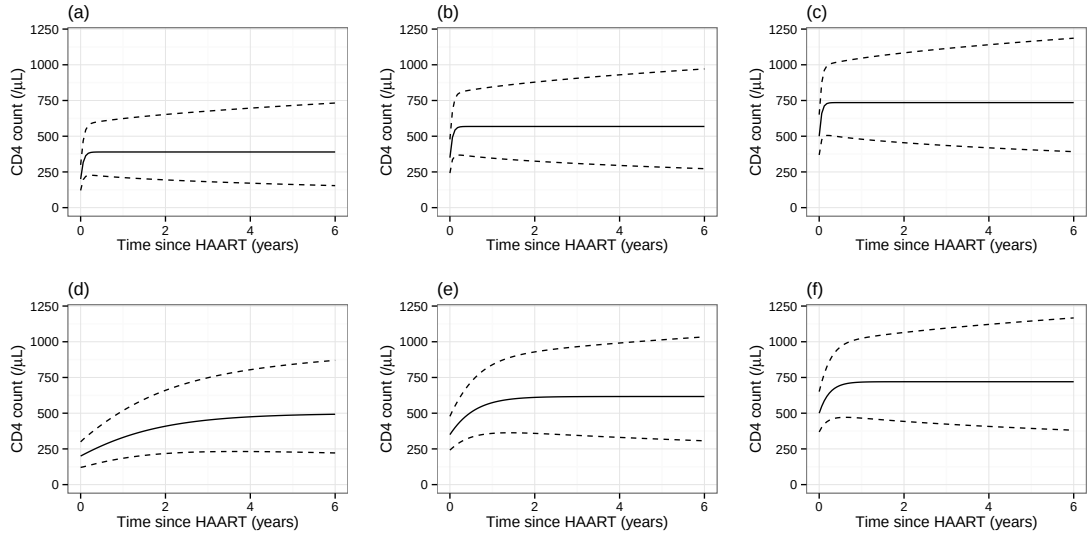

Figure 5.8. Predictions for hypothetical patients made from fitted model. Plots of the median (——) and 5th and 95th centiles (- - -) of CD4 counts predicted by $Model_4$ for a population of patients initiating highly active antiretroviral therapy (HAART) either within 6 months (a–c) or more than 1 year (d–f) from seroconversion, with 'true' CD4 counts at treatment initiation of: (a,d) 200, (b,e) 350 and (c,f) 500 cells/μL. The predicted ranges include measurement error (alongside the stochastic process component and variance in the subject-specific long-term maximum), explaining the variance present at time zero. The ranges shown have been back-transformed from the model predictions generated on the square-root scale.

are considerable between-patient differences in the variability of observations over time. It is interesting to note that the correlation parameter between the pre- and post-treatment latent scaling variables was positive, but only of moderate magnitude ($\hat{\rho}_{Moran}$ = 0.37 (0.19–0.52)), i.e. the degree of variability over time before and after treatment for each patient shows a moderate positive correlation. It is also of interest that the estimated H-index for the post-treatment fractional Brownian motion process in this model was much lower than that for the equivalent model without the latent scaling variables (0.13 (0.11–0.16) *vs* 0.38 (0.29–0.48)), indicating that although some patients show high variability in CD4 observations over time, successive increments of the stochastic process are strongly negatively correlated and there is an associated reversion of the process towards the underlying mean in each patient. It is possible to use the modes of the posterior predictive distributions of the latent scaling variables for each patient to identify those individuals with particularly smooth or erratic patterns of CD4 counts over time; observations for the two patients with the most extreme values obtained for the post-treatment latent scaling variable are plotted in Figure 5.10.

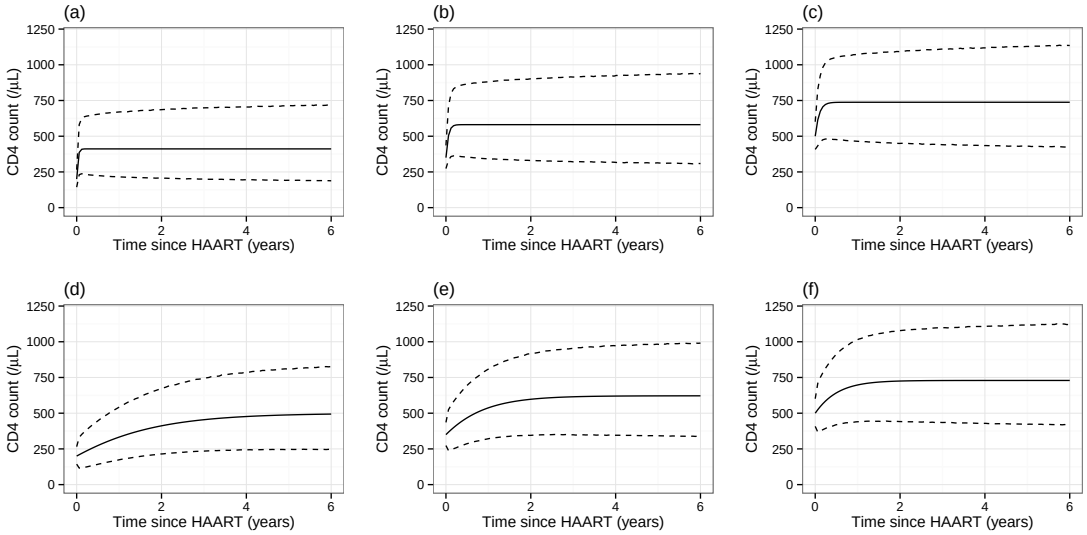

Figure 5.9. Predictions for hypothetical patients made from fitted model. Plots of the median (——) and 5th and 95th centiles (- - -) of CD4 counts predicted by $Model_6$ for a population of patients initiating highly active antiretroviral therapy (HAART) either within 6 months (a–c) or more than 1 year (d–f) from seroconversion, with 'true' CD4 counts at treatment initiation of: (a,d) 200, (b,e) 350 and (c,f) 500 cells/μL. The predicted ranges include measurement error (alongside the stochastic process component and variance in the subject-specific long-term maximum), explaining the variance present at time zero. The ranges shown have been back-transformed from the model predictions generated on the square-root scale. The marginal distribution is assumed for the latent scaling variable for the fractional Brownian motion process, i.e. without conditioning on any potential pre-treatment information, and the combination of multivariate normal and t distributions is approximated through averaging over 1000 draws from the relevant gamma distribution.

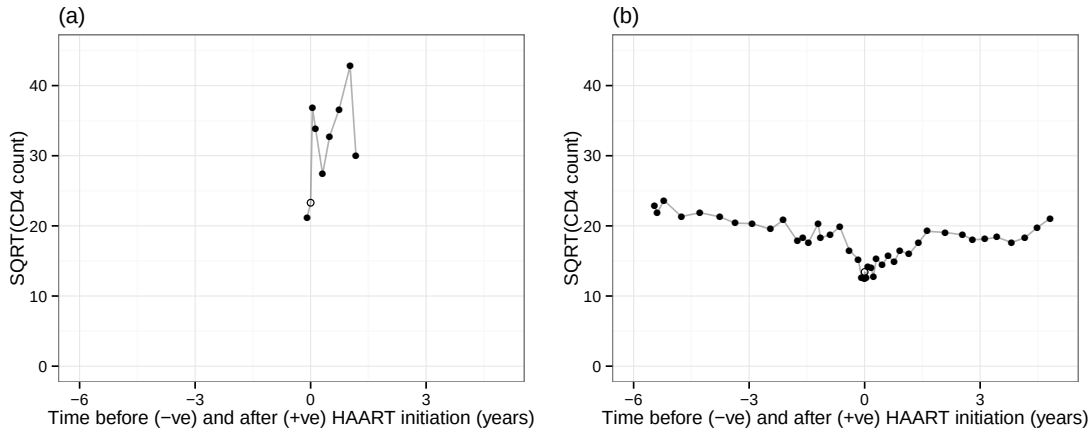

Figure 5.10. Plots of CD4 counts (•) observed in the two patients with the most (a) and least (b) erratic response to highly active antiretroviral therapy (HAART). Variability of response was assessed as indicated by the modes of the posterior predictive distributions of the post-treatment latent scaling variables ($\widehat{\gamma_{2:i}}$) obtained from $Model_6$. The mode of the posterior predictive distribution for the 'true' baseline value ($\widehat{u_i}$) is also shown in each case (∘).

5.12 Residual diagnostics and model checks

We present plots of residuals obtained from the fit of $Model_6$ to the UK Register of Seroconverters dataset. The approach taken is as described in Section 3.4.2. Firstly we note that, for pre-treatment CD4 cell counts, the distribution for the full set of observations for each patient is multivariate normal conditional on the value of the latent scaling variable associated with the pre-treatment fractional Brownian motion process:

$$\begin{aligned}
\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\
\mathbf{b}_i &\sim MVN(\mathbf{0}, \boldsymbol{\Psi}) \\
\mathbf{W}_{pre:i}|_{\gamma_{1:i}} &\sim MVN(\mathbf{0}, \frac{1}{\gamma_{1:i}}\boldsymbol{\Sigma}_{pre:i}) \\
\mathbf{e}_{pre:i} &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{pre:i}}).
\end{aligned}$$

We can therefore obtain an estimate of the pre-treatment marginal covariance matrix specific to each patient based on the posterior predictive mode of their latent scaling variable, $\hat{\gamma}_{1:i}$:

$$\hat{\mathbf{V}}_{pre:i} = \mathbf{Z}_i\hat{\boldsymbol{\Psi}}\mathbf{Z}_i^{\mathrm{T}} + \frac{1}{\hat{\gamma}_{1:i}}\hat{\boldsymbol{\Sigma}}_{pre:i} + \hat{\sigma}^2\mathbf{I}_{n_{pre:i}},$$

or alternatively generate covariance matrices based on samples from the posterior predictive distribution of $\boldsymbol{\gamma}_1$.

As discussed in Section 3.4.2, if the model parameters and the value of the scaling variable were known, then the distribution of the transformed marginal residuals using the inverse of the Cholesky decomposition of the covariance matrix for each individual, $\mathbf{V}_{pre:i}$, would be normally and independently distributed with mean $\mathbf{0}$ and variance $\mathbf{1}$:

$$\begin{aligned}
\mathbf{V}_{pre:i} &= \mathbf{L}_i\mathbf{L}_i^{\mathrm{T}} \\
\mathbf{L}_i^{-1}\left(\mathbf{y}_{pre:i} - \mathbf{X}_i\boldsymbol{\beta}\right) &\sim MVN\left(\mathbf{0}, \mathbf{I}_{n_i}\right).
\end{aligned}$$

For post-treatment observations, the distribution for the full set of observations for each patient is multivariate normal conditional on the value of both the true baseline CD4 value and the latent scaling variable associated with the post-treatment fractional Brownian motion process:

$$\begin{aligned}
\mathbf{y}_{post:i}|_{u_i^+} &= \mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
\tau_i &\sim N(0, \Omega)
\end{aligned}$$

$$\mathbf{W}_{post:i}|_{\gamma_{2:i}} \sim MVN(\mathbf{0}, \frac{1}{\gamma_{2:i}}\mathbf{\Sigma}_{post:i})$$
$$\mathbf{e}_{post:i} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{post:i}}).$$

As noted in Section 5.6.5 this forms a multivariate normal distribution conditional on u_i^+ and $\gamma_{2:i}$, given that $\mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right)$ is linear in τ_i. As such, a covariance matrix can be constructed for the post-treatment observations of each patient based on the posterior predictive modes of u_i^+ and $\gamma_{2:i}$, or from samples from their joint posterior predictive distribution, and Cholesky-transformed residuals can be calculated as for the pre-treatment data.

We present plots summarising Cholesky-residuals conditional on multiple samples from the joint posterior predictive distribution of the latent variables using the approximate multivariate normal distribution, as returned by the ADMB software. For the latent scaling variables relating to the pre- and post-treatment stochastic process components of the model, sampling was based on the bivariate normal a and b variables as used for the Laplace approximation of the integral, with transformation to the necessary gamma variates as described in Section 5.7. Plots based on 1000 sets of samples are shown in Figures 5.11 and 5.12. These plots of residuals derived from $Model_6$ do not indicate substantial problems with the fitted model. For the pre-treatment data (Figure 5.11), the appearance of the plots is very similar to those obtained for the marginal multivariate-t distribution model fitted in Chapter 4. For the post-treatment data (Figure 5.12) the plot of mean squared residuals and the Q–Q and semivariogram plots indicate near perfect fit of the model to the data, whilst the mean residual plot (Figure 5.12a) does show a regular pattern in relation to post-treatment time of minor deviations from the expected value of zero, suggesting that some further fine-tuning of the shape of the post-treatment response curve might be possible.

As a further check of the model structure developed, the fitted $Model_6$ was used to simulate pre- and post-treatment CD4 counts of a cohort of 100 patients. As we have not developed a probabilistic model for the timing of initiation of treatment, and in order to generate a range of different conditions, these patients were randomised to initiate treatment either: (1) immediately at the time of seroconversion, (2) 1 year after seroconversion or at the first observation below (3) 500, (4) 350 or (5) 200 cells/μL. Data were generated on the square-root(CD4) scale, and cut-off points for initiation of treatment were accordingly transformed to this scale. Each patient, up until the point of treatment initiation, was scheduled to be observed at 4-month intervals from seroconversion; treatment was initiated at 8 years if the threshold for a specific patient had not been triggered before this point. Following treatment initiation, observations were simulated after 1, 2, 3 and 4 months, and at 4-month in-

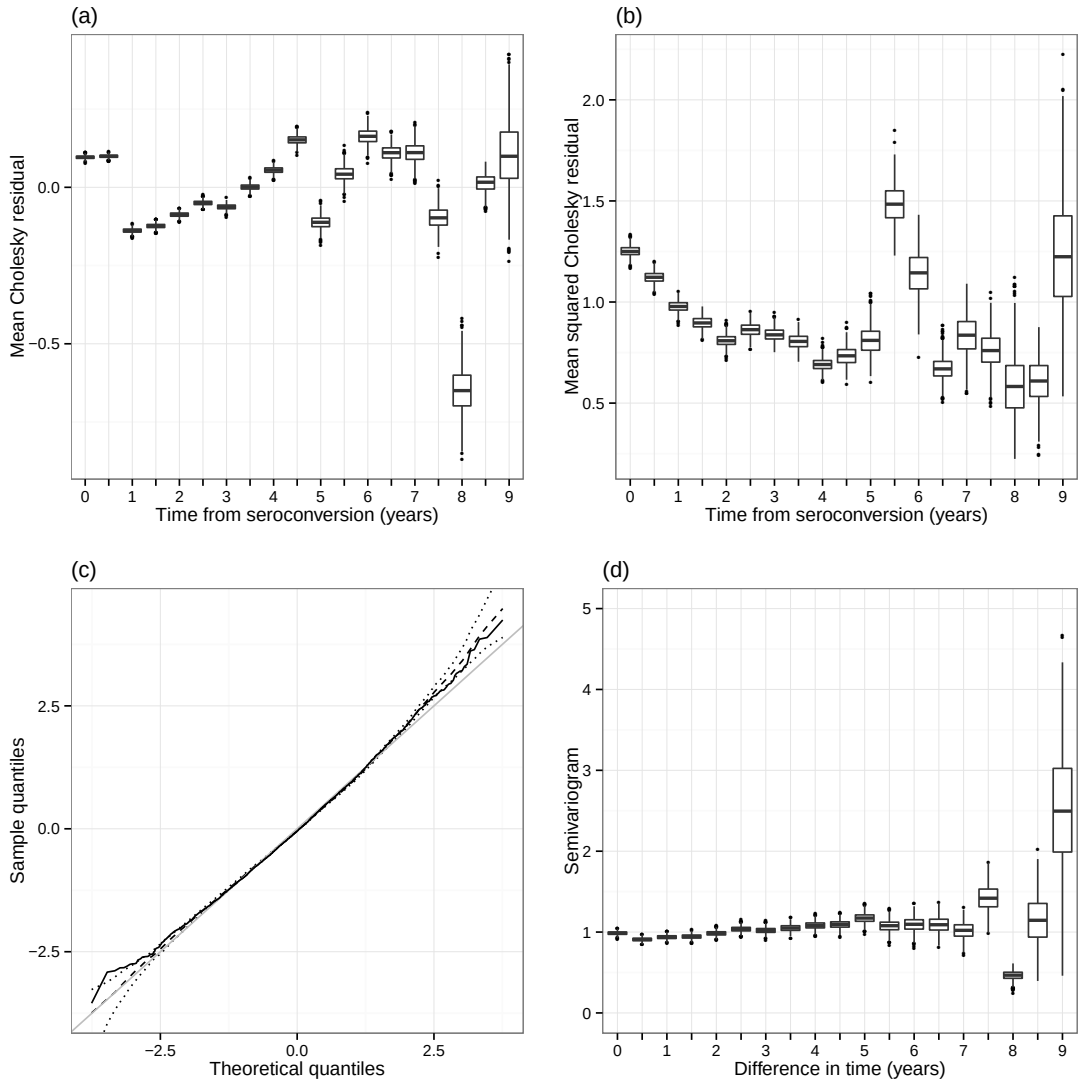

Figure 5.11. Plots of the distribution of Cholesky-transformed residuals for pre-treatment CD4 counts derived from $Model_6$, based on 1000 simulations of the vector of latent variables $\boldsymbol{\gamma}_1$. In (a) and (b) box plots of mean values for each simulation are plotted grouped by nearest multiple of 6 months. (c) Quantile–quantile plot for Cholesky-transformed residuals with respect to a standard normal distribution; the dotted lines show the 2.5^{th}, 50^{th} and 97.5^{th} percentiles of the sample quantiles for each theoretical quantile corresponding to the total number of observations, the solid black line shows the sample quantiles derived using the empirical Bayes estimate ($\hat{\gamma}_{1:i}$) for each individual, with the line of equality also displayed in grey. (d) Box plots of the distribution of mean semivariogram values, over multiple simulations of $\boldsymbol{\gamma}_1$, of Cholesky-transformed residuals with respect to difference in time between observations, grouped by nearest multiple of 6 months.

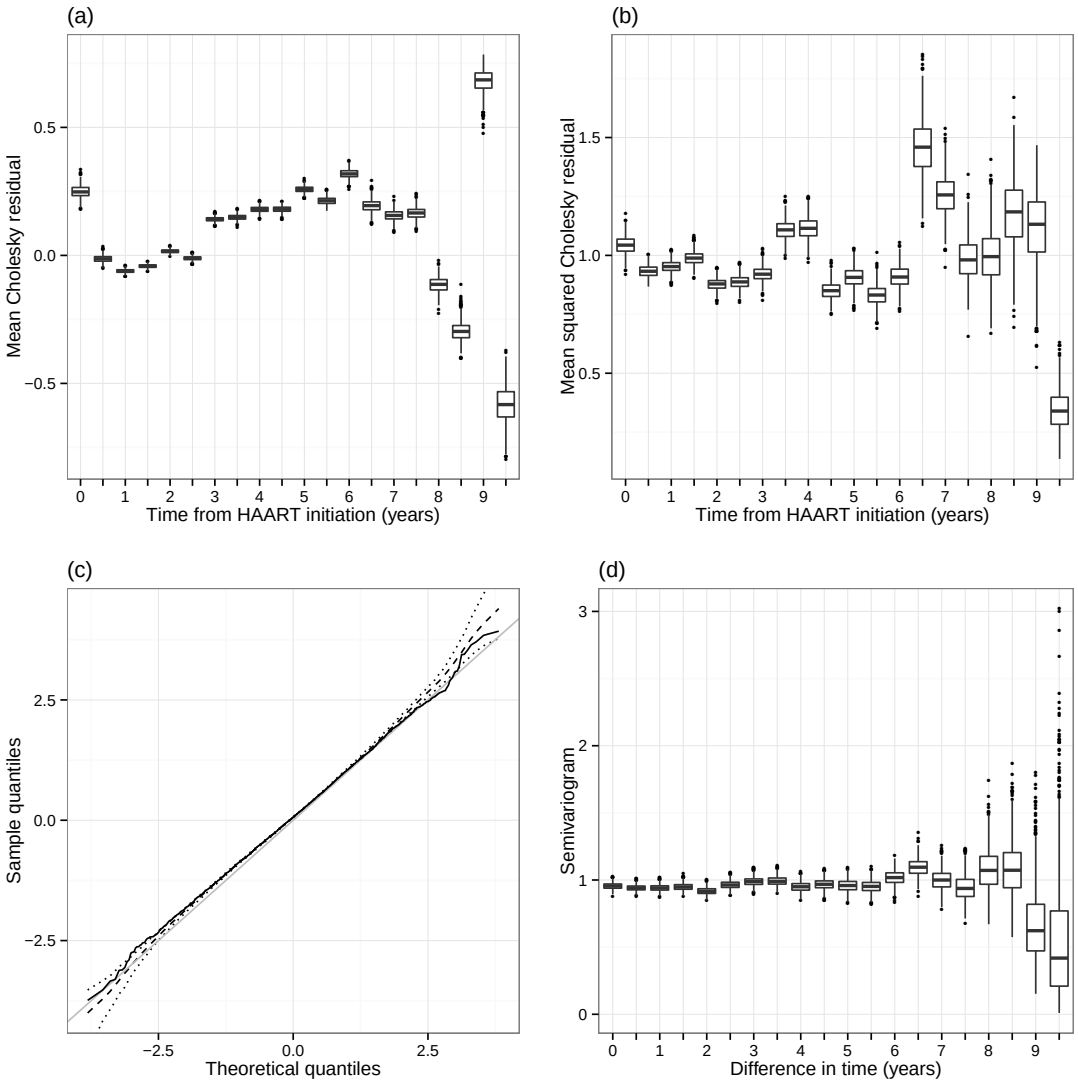

Figure 5.12. Plots of the distribution of Cholesky-transformed residuals resulting for post-treatment CD4 counts derived from $Model_6$, based on 1000 simulations of the vector of latent variables $\boldsymbol{u}^+$ and $\boldsymbol{\gamma}_2$. In (a) and (b) box plots of mean values for each simulation are plotted grouped by nearest multiple of 6 months. (c) Quantile–quantile plot for Cholesky-transformed residuals with respect to a standard normal distribution; the dotted lines show the 2.5^{th}, 50^{th} and 97.5^{th} percentiles of the sample quantiles for each theoretical quantile corresponding to the total number of observations, the solid black line shows the sample quantiles derived using the empirical Bayes estimates ($\hat{u}_i^+$ and $\hat{\gamma}_{2;i}$) for each individual, with the line of equality also displayed in grey. (d) Box plots of the distribution of mean semivariogram values, over multiple simulations of $\boldsymbol{u}^+$ and $\boldsymbol{\gamma}_2$, of Cholesky-transformed residuals with respect to difference in time between observations, grouped by nearest multiple of 6 months.

tervals thereafter up until a maximum of 5 years. A plot of CD4 counts from the simulated cohort is provided in Figure 5.13. This plot is visually consistent with the equivalent plot of 100 randomly selected patients from the real dataset (Figure 5.1, page 82), although in the artificial dataset no allowance has been made for irregular timing of observations or of loss-to follow-up or administrative censoring of patients. This comparison could be described as a posterior predictive check[77].

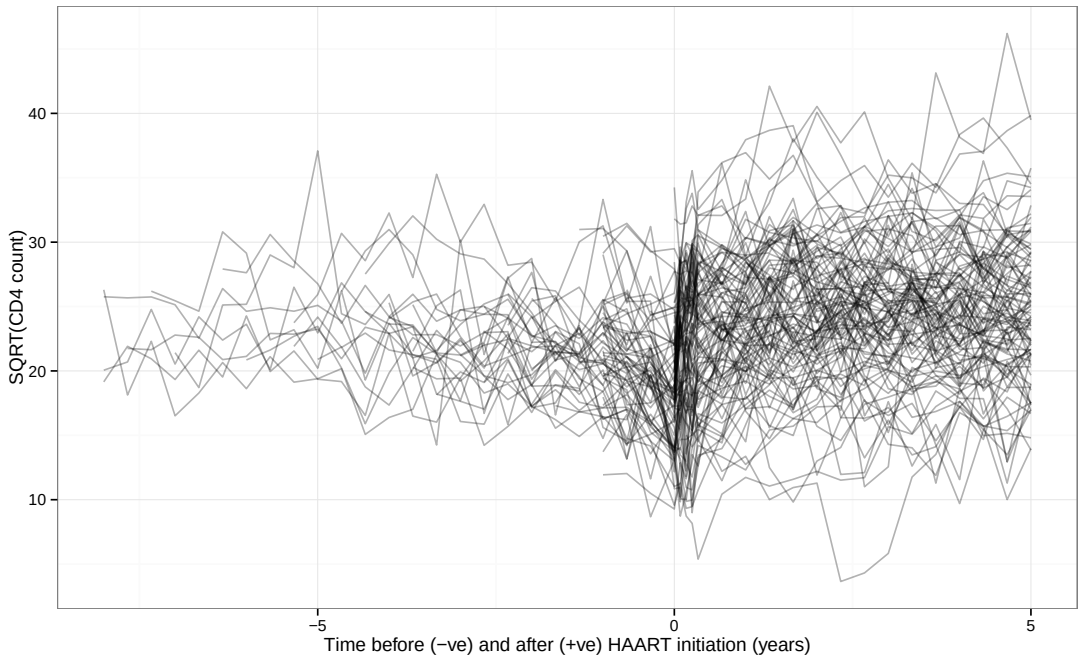

Figure 5.13. Plot of CD4 counts relative to the initiation of HAART for a simulated cohort of 100 patients based on $Model_6$.

5.13 Simulation study

In this section we present simulation analyses based on the fitted models. In Subsection 5.13.1 we refit combined models to simulated pre- and post-treatment data, providing a check that the statistical methodology proposed can be used to draw appropriate inferences regarding the association between the true baseline value of a biomarker and its trajectory after initiation of treatment. In Subsection 5.13.2 we use simulated data to further explore potential problems in analysing response to treatment conditional on an observed baseline value. The results obtained are later discussed in Chapter 6 in relation to inconsistent findings reported in the applied literature regarding the association between observed baseline CD4 count and the absolute increases observed after initiation of treatment.

5.13.1 Model fitting to simulated data

In order to check that the use of natural cubic splines would be able to recover non-linear functions for $\phi_1\left(u_i^+\right)$ and $\phi_2\left(u_i^+\right)$, we simulated cohorts of patients based on a modified version of $Model_6$. The point estimates of parameters were used as obtained from the UK Register of Seroconverters dataset, but to simplify the analysis the recovery of CD4 counts after initiation of treatment was assumed to depend only on the 'true' CD4 value at baseline and *not* on the time elapsed from seroconversion to initiation. Furthermore, $\phi_1\left(u_i^+\right)$ and $\phi_2\left(u_i^+\right)$ were modified to follow non-linear sigmoidal functions:

$$\phi_1\left(u_i^+\right) = 15 + \frac{15}{1+\exp(-0.5(u_i^+ - 15))}$$
$$\phi_2\left(u_i^+\right) = \frac{2}{1+\exp(-0.5(u_i^+ - 20))}.$$

Twelve cohorts of 250 patients were generated, using the observation and treatment initiation schedule as described in Section 5.12 with reference to Figure 5.13, and $Model_3$ was fitted to each cohort — i.e. with a natural cubic spline function to approximate $\phi_1\left(u_i^+\right)$ and $\phi_2\left(u_i^+\right)$, without any dependence on the time from seroconversion to treatment initiation and without accounting for between-patient differences in variability over time. The latter discrepancy with the model used to generate the data was decided because we were not able to fit link functions using natural cubic splines and account for between-patient differences in variability within the same model when analysing the real data, likely because of the need to rely on the less accurate Laplace approximation to the marginal likelihood for such models, but we wanted the simulated data to reflect the characteristics of the observed data. The results of this simulation are relevant to the analyses presented in Chapters 6 and 7, in which we largely focus on inferences from statistical models for CD4 counts in which between-patient differences in variability were not taken into account. The number of simulated cohorts was chosen for convenience as the maximum number of separate processes that could be initiated from R using the cluster system available, and also as a number that would allow simultaneous visual inspection of the fitted models.

Convergence of maximum likelihood estimates of the model parameters was achieved for 10/12 of these simulated cohorts. The fitted functions for $\phi_1\left(u_i^+\right)$ and $\phi_2\left(u_i^+\right)$ in each case are shown in Figures 5.14 and 5.15, respectively. A histogram of the 'true' CD4 values at treatment initiation for each patient in the first cohort is shown in Figure 5.16.

The plots of the fitted functions for $\phi_1\left(u_i^+\right)$ and $\phi_2\left(u_i^+\right)$ indicate that natural cubic splines can be used to approximate non-linear relationships between latent vari-

ables, even if the probability model as a whole is not completely correctly specified. However, the natural cubic splines are constrained to a linear function beyond the upper and lower boundary knots, and this clearly affects the ability of the approach to model response to treatment in patients with very high or very low baseline CD4 at treatment initiation. Adding more knots to the natural cubic spline basis would allow more flexibility in the fitted function, but at the cost of reduced computational stability. Hence these plots indicate that caution is required when interpreting predictions or attempting to draw inferences regarding patients with unusually high or low CD4 values at treatment initiation, and reinforce the general principle that fitted relationships should not be extrapolated beyond the range of values observed in the dataset under analysis.

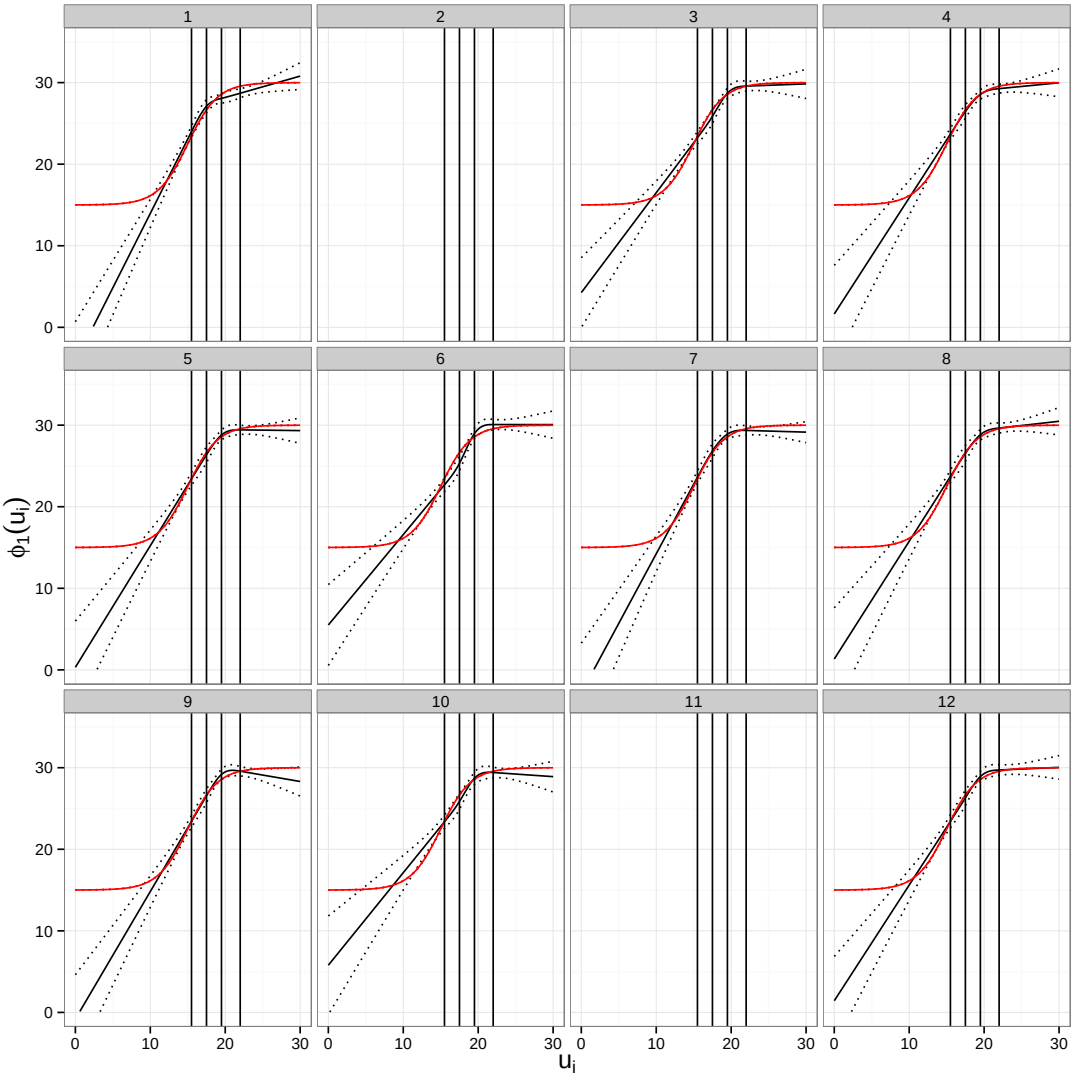

Figure 5.14. Plots of estimates of $\phi_1(u_i^+)$ (black curved line, with dotted 95 % CI) obtained by fitting $Model_3$ to 12 simulated cohorts of 250 patients. The function specified as used to generate the data is shown in red. The vertical black lines show the positions of knots for restricted natural cubic spline basis.

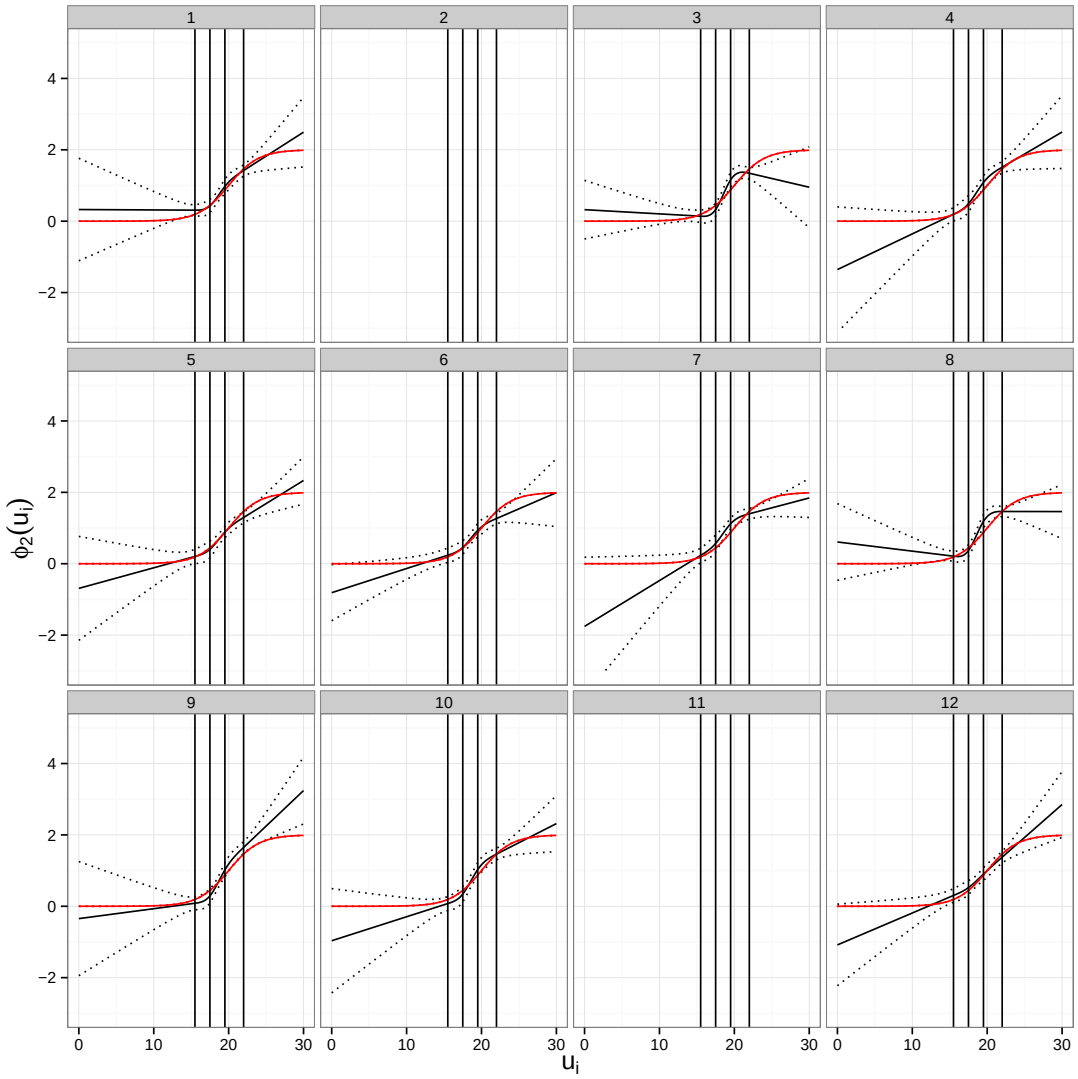

Figure 5.15. Plots of estimates of $\phi_2(u_i^+)$ (black curved line, with dotted 95 % CI) obtained by fitting $Model_3$ to 12 simulated cohorts of 250 patients. The function specified as used to generate the data is shown in red. The vertical black lines show the positions of knots for restricted natural cubic spline basis.

5.13.2 Measurement errors at treatment initiation

As noted in Section 5.3, Babiker *et al.*[27] have demonstrated that if treatment is initiated conditionally on the observed value of a biomarker that is monitored over time, then the observation at treatment initiation can provide a biased estimate of the 'true' underlying value. In this section, we use simulations based on the structure and fitted parameters of $Model_6$ to explore how this could affect inferences regarding the associations between the baseline value of the biomarker, time from infection to treatment initiation and the characteristics of post-treatment recovery. To simplify this investigation, we assume that time from seroconversion to treatment initiation does not affect the characteristics of response to treatment and use the fitted link

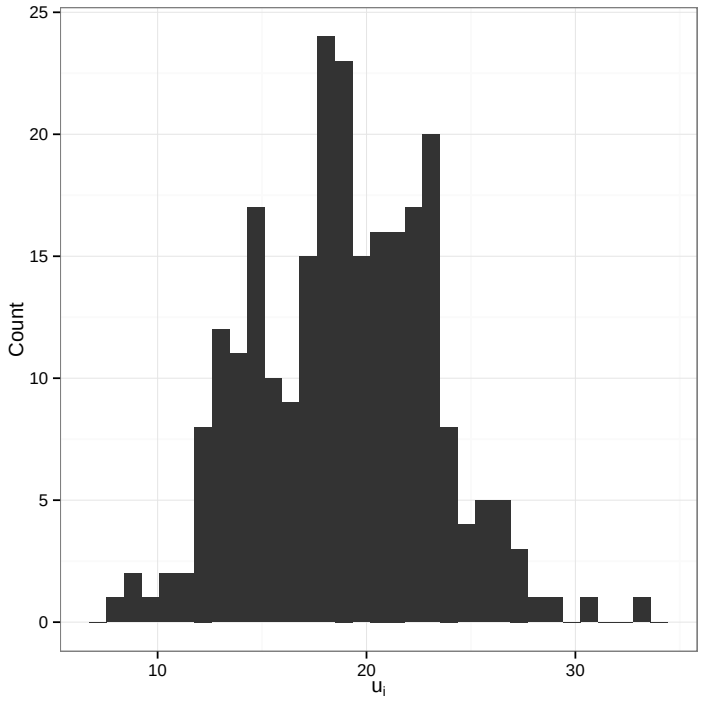

Figure 5.16. Histogram of 'true' CD4 values at treatment initiation for each patient in the first cohort of the simulation study described in Section 5.13.1.

functions for treatment initiation beyond 1 year from seroconversion for all patients. We use the combination of observation and treatment initiation rules as specified in Section 5.12. For convenience, we repeat here that simulated patients were randomised to initiate treatment either: (1) immediately at the time of seroconversion, (2) 1 year after seroconversion or at the first observation below (3) 500, (4) 350 or (5) 200 cells/μL.

For those patients that initiate treatment without reference to their CD4 observations (i.e. following rule (1) or (2)), the observed CD4 count will provide an unbiased estimate of the true underlying value (without measurement error). However, for patients in which treatment is initiated based on their CD4 count observations (i.e. following rule (3), (4) or (5)), the observed CD4 cell count will show a negative bias. Under the simulation set-up described here, all patients who initiate treatment at more than 1 year from the date of seroconversion will therefore be expected to exhibit a negative bias in their baseline observed CD4 count, whereas for the majority of patients initiating treatment within 1 year from seroconversion, the observed baseline CD4 count will be unbiased with respect to the true underlying value. Although this simulation represents a simplification of the complex combination of clinical and historical factors that have influenced the timing of treatment initiation for the patients in real observational datasets, it provides an illustration of how measurement errors could affect inferences in this setting.

We created a simulated dataset of 240 000 patients, and compared the last observed CD4 count before treatment initiation to the 'true' value (without any measurement error) in each individual (on the square root scale on which the model was fitted). When simulated patients were stratified by their time from seroconversion to treatment initiation, those with an interval of ≤ 6 months showed a small negative bias in their observed baseline CD4 count (median difference=-0.26 , IQR: -1.15 to 0.63, n=96 324) as did those with an interval to treatment of >6 months but ≤ 1 year (median difference=-0.22 , IQR: -1.13 to 0.69, n=64 621) whereas those that initiated treatment beyond 1 year showed a more substantial negative bias (median difference=-0.85 , IQR: -1.70 to -0.02, n=79 055). Boxplots of the differences between observed baseline CD4 counts and the underlying true value (according to the model used for simulation) stratified by timing of treatment initiation are shown in Figure 5.17. These simulated data can also be used to demonstrate a general 'regression to the mean' effect, regardless of treatment initiation rule, in that those patients with a particularly low observed CD4 count at treatment initiation (i.e. around <15 on the square-root scale) are likely to in fact have a higher true baseline value, whilst the opposite is true for patients with a particularly high observed baseline CD4 count (i.e. around >25 on the square-root scale), this is demonstrated in Figure 5.18.

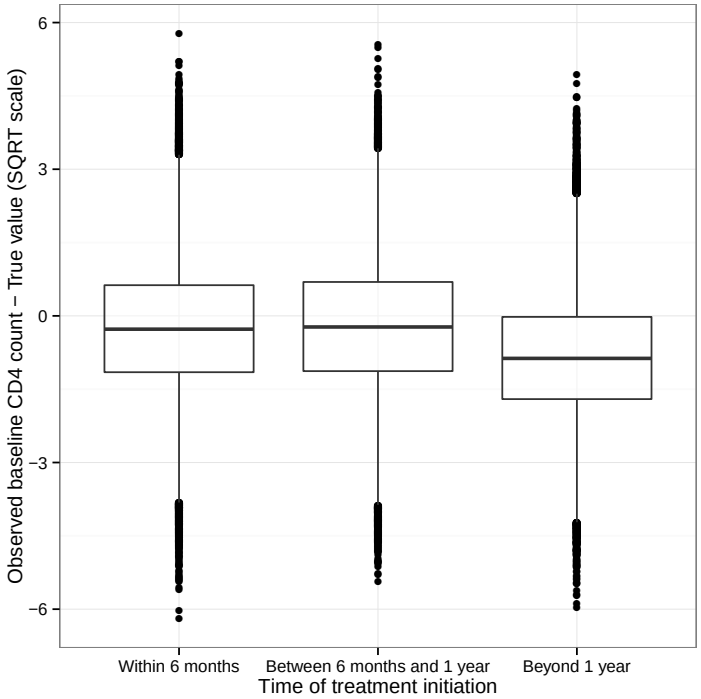

Figure 5.17. Boxplots of differences between observed baseline CD4 counts and the underlying true value stratified by timing of treatment initiation for the simulation study described in Section 5.13.2.

In Figure 5.19, average post-treatment CD4 counts (on the square-root scale) from this simulation are shown with stratification by baseline CD4 observation (150 –

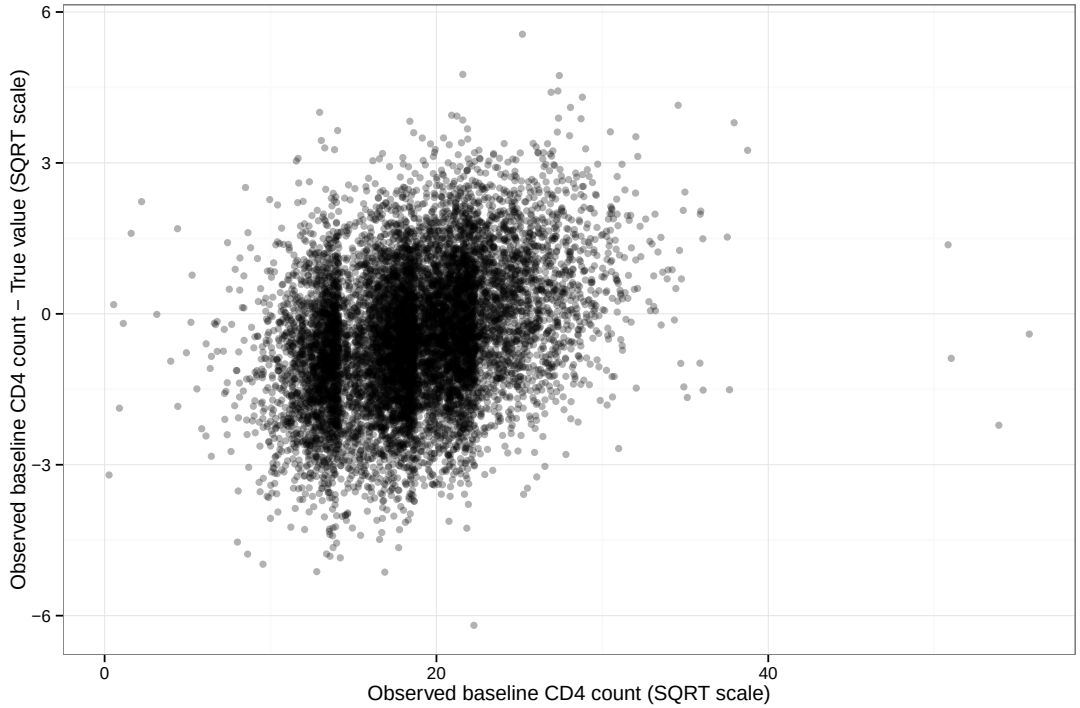

Figure 5.18. Scatter plot of differences between observed baseline CD4 count and the underlying true value against the observed baseline CD4 count for the simulation study described in Section 5.13.2. Datapoints are shown for a random selection of 10 000 of the simulated patients.

250 *vs* 450–550 cells/μL) and by time from seroconversion to treatment initiation. In the simulation, the characteristics of response to treatment were dependent on the 'true' baseline CD4 count but not on timing of initiation itself. However, stratification by timing of treatment leads to differences in the distributions of both true and observed baseline CD4 counts between the groups defined within each stratum of baseline CD4. For the lower CD4 stratum considered (i.e. 150 – 250 cells/μL), there are differences in the mean observed CD4 count at baseline between the groups defined by timing of initiation that persist throughout the post-treatment period, whereas for the higher CD4 stratum considered (i.e. 450 – 550 cells/μL) there are differences at baseline that disappear after treatment initiation. The scale of the observed differences between groups in this simulation (in which no effect is assigned to the timing of treatment initiation, conditional on 'true' CD4) are modest, but the simulation nonetheless illustrates some of the problems faced in unpicking factors associated with treatment response conditional on the baseline value of the variable under investigation.

A number of studies in the literature investigating response to HAART have analysed the change from the observed CD4 count at baseline to the value observed after a given number of months, for example Smith *et al.*[123] report factors associated with the observed change at 3 months. However, as we have discussed, the ob-

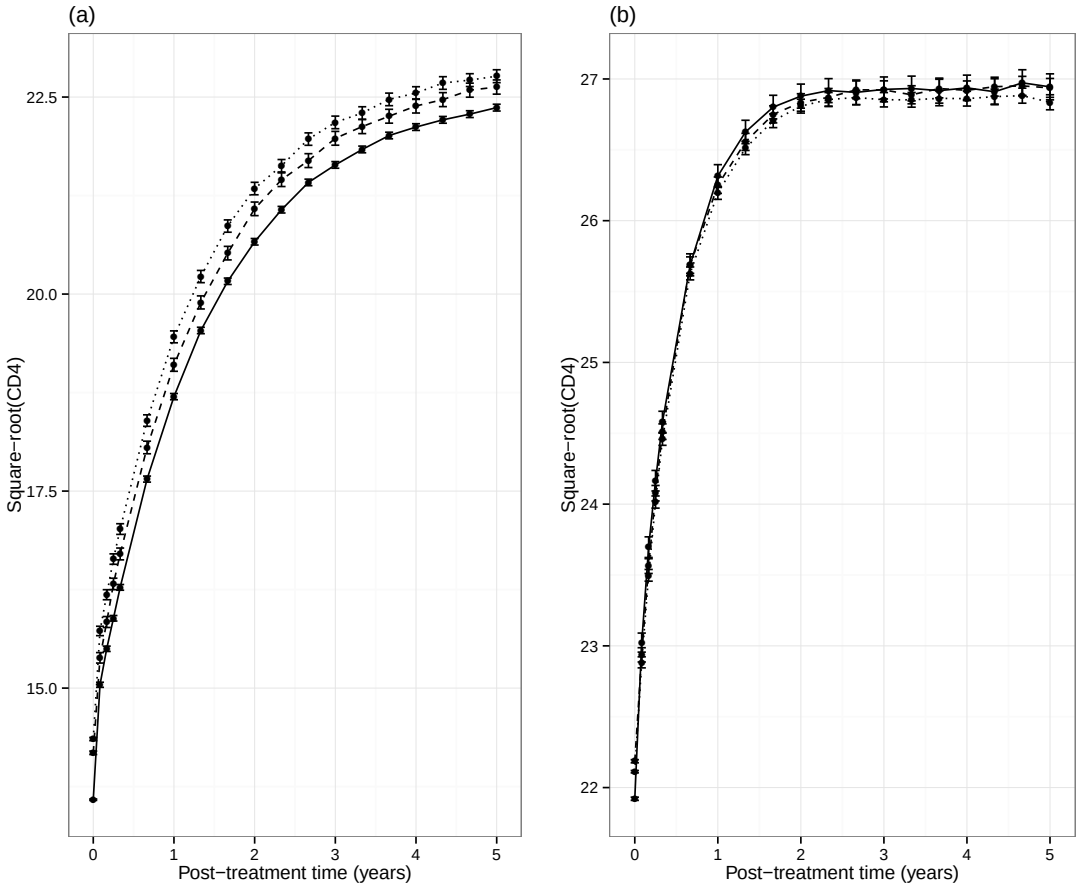

Figure 5.19. Mean observed CD4 counts following initiation of treatment resulting from the simulation study described in Section 5.13.2. Plots are shown for patients with a last observed CD4 count of (a) 150 – 250 or (b) 450–550 cells/μL, and each plot is stratified by treatment initiation: within 6 months of seroconversion (·····), beyond 6 months but within 1 year (- - -) and beyond 1 year (——). Whiskers show the 95% CI of the mean for each group at each point in time (reflecting sampling variability in the simulation as defined).

served baseline CD4 count can provide a biased estimate of the true underlying value because of the combined influence of selective treatment initiation and regression to the mean, and the extent of the bias for any given observed baseline value will depend on the pre-treatment observation and treatment initiation schedules for a given population. For the simulated cohort described in this section, we present in Figure 5.20 plots of observed changes in CD4 count at 3 months and the equivalent change that would have been observed had the true baseline value been known. For this model and set of treatment initiation rules, a maximum median change of around 115 cells/μL is observed for an observed baseline value of 425 cells/μL (Figure 5.20a). However, were it possible to assess the observed change from the true baseline value, a maximum increase would be observed from a baseline value of 625 cells/μL (Figure 5.20b). This shows that caution is required in interpreting changes from observed baseline values in such a situation.

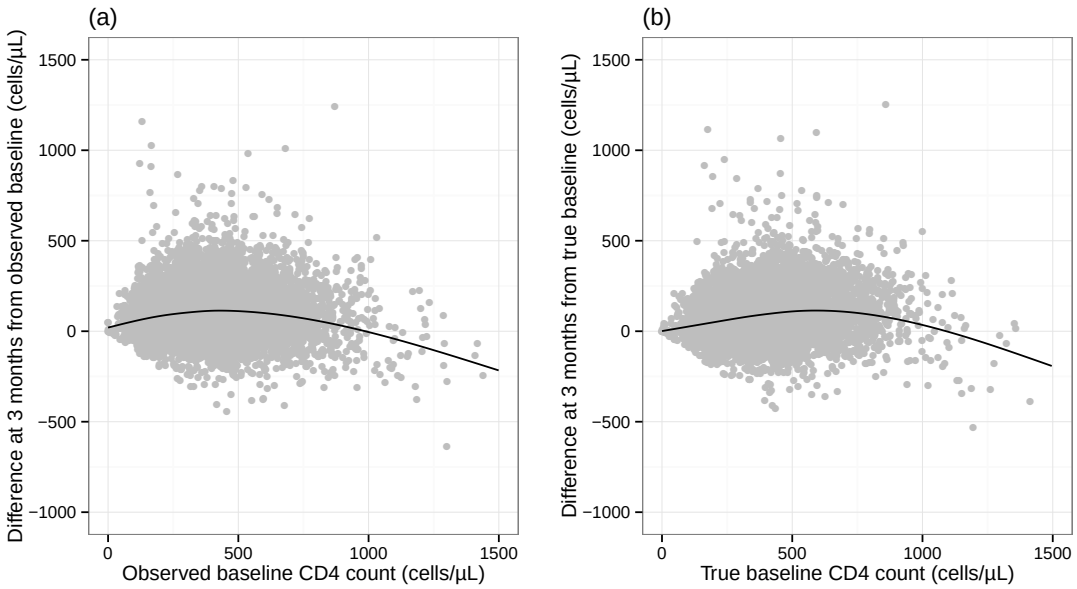

Figure 5.20. Plots of the change from (a) observed baseline CD4 count and (b) true baseline CD4 count at 3 months after initiation of treatment resulting from the simulation study described in Section 5.13.2. Individual datapoints are shown for a random selection of 10 000 of the simulated patients (•), and a smoothed graph of the median change (based on all simulated patients) is also shown (——).

5.14 Discussion

The statistical methodology developed in this chapter provides a novel framework for the combined analysis of pre- and post-treatment longitudinal biomarker data. The approach proposed has the advantage of making use of all available data, does not require an *a priori* assumption regarding the distribution of baseline values at treatment across the studied population as a whole and allows a flexible choice of functions to link the pre- and post-treatment trajectories of the biomarker under investigation for each patient. When applied to CD4 data from the UK Register of Seroconverters cohort, the resulting fitted models provide evidence of a positive association between baseline CD4 count at initiation of HAART and the long-term maximum achieved by each patient, which is consistent with previous published literature on this topic[81;108–111]. In addition the fitted models suggest that initiation of HAART closer to the date of HIV seroconversion is associated with a more rapid response to treatment, regardless of the baseline CD4 value. This finding warrants further investigation with inclusion of additional factors that are thought to be associated with response to treatment into the modelling framework; this extension is straightforward using the methodology developed and is implemented using a larger dataset in Chapter 6.

The standard non-linear mixed effects model approach in this situation, ignoring observations before the start of treatment, would require rigid assumptions regard-

ing the distribution of the biomarker variable at treatment initiation and its relationship to subsequent post-treatment observations, i.e. typically that baseline values and the long-term maximum value for each patient follow a bivariate normal distribution. The modelling strategy that we have developed allows greater flexibility in the link between baseline and post-treatment maximum values of the biomarker, and does not restrict the shape of the overall marginal distribution of baseline values in the studied population.

Alternatively, the standard use of baseline observations as a predictive variable also discards any information from measurements obtained prior to this point in time and would require a separate imputation model for missing values of the baseline measurement, which would not be straightforward to define for observational data with highly irregular number and timing of measurements for each patient. Furthermore, it is not obvious how the primary model for multiple post-treatment observations should be structured in this context, as it would be overly restrictive to assume a constant fixed effect coefficient for the baseline observation for all time points after the initiation of treatment. One option is to stratify the modelled mean according to intervals of baseline observations[81;111], but this discards some of the information provided by the baseline value as a continuous variable. Geng *et al.*[124] addressed this issue by fitting a linear mixed model for post-treatment CD4 counts that included a combination of linear splines with knots at 4 and 12 months for post-treatment time, a restricted cubic spline basis for the effect of pre-treatment CD4 count and an interaction between these parts of the model. However, baseline observations at treatment initiation are subject to measurement error, with the distribution of these errors dependent on both the treatment initiation rules that have been implemented and on the value of the observation relative to the underlying distribution, which further complicates model fitting and interpretation.

The proposed model for the analysis of pre- and post-treatment CD4 data has been structured so that the estimated parameters of the different components of the model each have a clear practical interpretation, i.e. it is of direct interest to clinicians to know how baseline CD4 and time from seroconversion at initiation of HAART are associated with the speed and maximal level of treatment response that can be expected. If further patient variables were added to the functions that determine the characteristics of response to treatment then the modelled effects would be independent of the influence of the true baseline value of the biomarker, making interpretation of estimated coefficients relatively simple. If a mixed effects model is fitted to only baseline and post-treatment measurements, then assessment of the influence of a covariable on treatment response conditional on a baseline observation requires an additional stage of statistical adjustment[125].

The cost of using a combined model for pre- and post-treatment data is that we

are required to assume that the proposed model structure provides an adequate description of the data under analysis. The requirement for strong assumptions regarding the correctness of model structure has been used as an argument against the use of integrated models for baseline and treatment response data[12]. In the present analysis, the motivation for the inclusion of pre- and post-treatment stochastic process components in the models and for the use of natural cubic spline functions to link baseline CD4 and characteristics of the treatment response trajectory was to maximise model flexibility and therefore provide an optimal fit to the data. The residual plots produced indicate a good fit of the final model to the data.

An advantage of the extension of the non-linear mixed effects modelling approach as developed in this chapter is that the nature of the variability in biomarker observations over time within each patient can be investigated, whereas this is often lost when using approaches that only consider population mean values or the marginal distribution of observations across the population at each point in time. A focus on realistic modelling of the patterns of variation in the data is also required in order to provide valid inference under the 'missing at random' assumption for missing data and when the timing of observations is dependent on previous outcomes[101]. A limitation of the present analysis is that we have not considered the possibility of censoring being related to underlying latent variable terms rather than just the observed CD4 counts. Such joint modelling of longitudinal and event time data[126;127] would provide useful information regarding the patterns of drop-out from the cohort, but would add further to the computational complexity of estimation.

The fitted models in the present analysis show that there is considerable unexplained variance in the long-term asymptotic maximal response to treatment for each patient, even after accounting for baseline CD4 and time from seroconversion to initiation of HAART, although this might be reduced by the inclusion of additional patient and drug regimen variables into the model. There is also considerable erratic post-treatment variability over time, represented by the fractional Brownian motion process as introduced for the analysis of pre-treatment CD4 data in Chapter 4. The parameter estimates for the model in which the stochastic process components were generalised to follow marginal multivariate-t distributions indicate substantial between-patient differences in their variability over time, with a moderate positive association between the degree of pre- and post-treatment variability within each patient, which are novel findings in this context. The fact that the models fitted follow a structure that can accommodate any combination of number and timing of observations in each patient means that they can be readily used for simulation studies of patient cohorts.

The methodology developed in this chapter could also be applied to other medical settings in which an intervention is triggered following monitoring of a biomarker

of interest, and in which the response to treatment may be conditional on the state of the patient (as indicated by the value of the biomarker) at the time of treatment initiation. Seroconverter cohorts have a special status in HIV research, and in other disease settings the 'zero time' for pre-treatment observations might be time of diagnosis or another clinically significant event. The framework proposed could be applied with different choice of pre- and post-treatment model components, but those demonstrated may be a natural choice in many settings.

6 Application of combined model to CASCADE dataset

6.1 Disclaimer regarding collaborative work

Following the requirements for applied research using data from the CASCADE cohort, the analyses in this chapter and in Chapter 7 were planned and interpreted in collaboration with the following external investigators: Andrew Phillips, M. John Gill, Ronald Geskus, Giota Touloumi, James Young and Heiner Bucher. A consensus decision was made regarding the inclusion criteria for the analysis and for the specification and coding of potential predictive factors. However, I completed all of the work relating to the development of the modelling framework, programming and processing and presentation of the results. The collaborators suggested relevant papers from the literature for consideration, but I alone have written the discussion of the results obtained presented in this thesis.

6.2 Dataset and estimation

We now apply the modelling framework developed in Chapter 5 to the full CASCADE dataset of seroconverters. As in Chapter 5 we restrict our analysis to patients with an estimated date of HIV seroconversion during or after 2003, and patients who started a suboptimal regimen of antiretroviral drugs prior to HAART were excluded as were patients without at least one post-treatment CD4 count recorded. Data were included up to March 2014 (with the analysis using a more recently updated CASCADE dataset than that in Chapter 4).

HAART is defined by a regimen of at least three drugs, with at least two different classes (unless abacavir or tenofovir is used in a '3N' regimen with three NRTIs). The non-nucleoside reverse-transcriptase inhibitors (NNRTI) regimen includes at least one NNRTI and at least one NRTI. The 'PI' regimen includes at least one ritonavir-boosted protease inhibitor (PI) with at least one NRTI. The integrase strand transfer inhibitor (INSTI) category includes at least one integrase inhibitor with any combination of NRTI, NNRTI and PI.

Applying these conditions results in a cohort of 8175 patients. We are interested in modelling response to HAART as a function of multiple patient characteristics and so we also exclude patients in whom no VL measurement was recorded within 6 months before the start of HAART ($n = 818$) and those for whom the mode of infection is unknown ($n = 326$). Because there is some overlap between these groups, the resulting cohort for analysis includes a total of 7065 patients, with 37 728 pre-treatment and 55 961 post-treatment CD4 count observations. Ideally the missing data would be imputed rather than these patients being excluded, but doing this appropriately in the present setting would be very challenging and the number of

exclusions represents a reasonably small proportion of the total number of patients (13.6 %). However, a further analysis in which the patients with missing pre-treatment VL are included is presented in Chapter 7. Where the recorded pre-treatment VL measurement is below the lower limit of detection for the assay used we impute the value as a half of the lower limit, assuming this limit to be 50 copies/mL (the most common value) if this is itself not recorded.

We are also interested in whether a diagnosis of hepatitis C virus (HCV) prior to initiation of HAART is predictive of recovery in CD4 counts. However, there was no HCV test recorded prior to HAART in a substantial proportion of patients in the cohort (1202/7065) and as such it was decided not to exclude these patients but rather to treat them as a separate grouping in the analysis along with those with a positive ($n = 387$) or negative ($n = 5476$) HCV test. Patient characteristics are summarised in Table 6.1.

As in Section 5.2, we have censored patients at recorded interruption of HAART (including switch to suboptimal treatment) for more than 1 week, but have not censored at change to HAART regimen. Analyses are first conducted without censoring due to virological failure (as in Chapter 5). However, we also conducted analyses in which post-treatment observations are censored at the observation of detectable VL beyond 6 months after the initiation of HAART. The rationale for this was to provide an estimate of CD4 recovery conditional on perfect adherence to treatment, under the assumption that most occurrences of virological failure in patients who have initiated HAART are due to imperfect adherence to their regimen. Due to the requirement for at least one post-treatment CD4 count, this censoring led to a slightly smaller total number of 7015 patients for analysis, with 37 526 pre-treatment and 40 921 post-treatment CD4 count observations. Of these patients, 2275 (32.4 %) had virological failure observed at some point in time, at a median of 0.90 years (IQR, 0.65–1.57 years).

The primary analysis relates to models fitted with a latent variable for each patient only for the 'true' baseline CD4 value (on the square-root scale) at initiation of HAART, with maximum likelihood estimation carried out using 10-point adaptive Gauss–Hermite quadrature; this is used rather than the 15-point adaptive Gauss–Hermite quadrature as in Chapter 5 in order to make the analysis feasible using the larger dataset (refitting of models in Chapter 5 using 10-point quadrature showed only negligible differences). Model fitting was also attempted with inclusion of between-patient differences in variability over time as described in Section 5.7, with the Laplace approximation used in such cases. As in Chapter 5, maximum likelihood estimation was carried out using the random effects mode of the ADMB software, run on the UCL Legion High Performance Computing Facility.

Table 6.1. Demographic and treatment characteristics of patients included in the primary analysis (n=7065)

Characteristic	n (%) or median (IQR)
Calendar date of SC	15 Sep 2007 (3 Jul 2005 – 17 Jan 2010)
SC date estimated by:	
SC illness	232 (3.3)
lab evidence	1371 (19.4)
mid-point	5462 (77.3)
Interval between HIV-1 tests (years)*	0.84 (0.44–1.5)
Infection group:	
Male homosexual	5443 (77.0)
Male heterosexual	667 (9.4)
Male IDU	151 (2.1)
Female heterosexual	758 (10.7)
Female IDU	46 (0.7)
Pre-HAART VL (log10(copies/mL))	4.82 (4.25–5.32)
Age at HAART initiation (years)	36.1 (29.9–43.5)
Pre-HAART AIDS Dx	204 (2.9)
Pre-HAART HCV test:	
+ve	387 (5.5)
–ve	5476 (77.5)
not available	1202 (17.0)
Time from SC to HAART (years)	1.45 (0.67–2.82)
$0 \leq t_{trt} \leq 0.5$	1366 (19.3)
$0.5 < t_{trt} \leq 1.0$	1170 (16.6)
$1.0 < t_{trt}$	4529 (64.1)
HAART regimen:	
NNRTI	2998 (42.4)
rb-PI	2485 (35.2)
INSTI	438 (6.2)
other	1144 (16.2)
3N	862 (75.3)
other PI	226 (19.8)
fusion inhibitor	43 (3.8)
other classification	13 (1.1)
n pre-HAART CD4 counts	4 (2–7)
n post-HAART CD4 counts	6 (3–11)
Time to last recorded post-HAART CD4 count (years)†	1.79 (0.74–3.41)

Mid-point estimates of seroconversion date are used for data shown in this table. *Of those used for mid-point estimates of SC date. †From date of HAART initiation, of those observations included in the analysis. 3N, triple nucleoside analog reverse-transcriptase inhibitors; AIDS, acquired immune deficiency syndrome; Dx, diagnosis; HAART, highly active antiretroviral therapy; HCV, hepatitis C virus; IDU, injecting drug user; INSTI, integrase strand transfer inhibitor; IQR, interquartile range; NNRTI, non-nucleoside reverse-transcriptase inhibitor; PI, protease inhibitor; rb-PI, ritonavir-boosted PI; SC, seroconversion; t_{trt}, time from SC to HAART (years).

6.3 Model structure and hypothesis tests

Given the much larger cohort of patients in this analysis in comparison to that presented in Chapter 5, we allow greater flexibility in the shape of the asymptotic recovery curve. This adjustment was made in response to feedback from a collaborator on the analysis (Andrew Phillips) who felt that CD4 counts continue to slowly improve, on average, many years after the initiation of HAART, rather than levelling off as implied by the recovery curve used in Chapter 5. We use an extension to the asymptotic growth curve attributed to Janoshek and Sager[128–130]:

$$g\left(t_{post}, u_i^+\right) = \phi_{1:i} + \left(u_i^+ - \phi_{1:i}\right) \exp\left(-\exp\left(\phi_{2:i}\right) t_{post}^D\right).$$

This function matches that given in Section 5.6.1 with an additional power transformation of the post-treatment time variable t_{post} by exponent D, a parameter to be estimated with value $D > 0$. For values of $D > 1$ the growth curve is sigmoidal, for $D = 1$ growth follows a standard asymptotic curve and for $D < 1$ growth is more rapid at time points closer to zero. The $\phi_{1:i}$ and $\phi_{2:i}$ terms still reflect long-term maximum and speed of recovery, respectively, but parameter estimates may not be straightforward to interpret directly if recovery does not reach the modelled asymptotic maximum within the time-frame under consideration (i.e. if substantial recovery in CD4 counts is still ongoing beyond around 5 years after initiation of HAART).

Given the accurate approximation to the marginal log-likelihood for each fitted model in the primary analysis (with a latent variable term only included for the 'true' baseline), statistical hypothesis tests for comparison of nested models are carried out using generalised likelihood-ratio tests. We initially fit a model in which the post-treatment recovery in CD4 count follows a standard asymptotic curve, for which the long-term maximum and speed of response are each linearly dependent on the 'true' baseline CD4 value (u^+) but not on any other patient or treatment characteristics. We then test whether the Janoshek–Sager curve provides a better fit to the data. Following Chapter 5, we subsequently stratify the functions that specify the dependence of $\phi_{1:i}$ and $\phi_{2:i}$ on u_i^+ according to whether treatment was initiated within 6 months, beyond 6 months and within 1 year or beyond 1 year from the estimated date of seroconversion.

We consider whether VL before treatment initiation is predictive of the speed of recovery in CD4 counts or long-term maximum; VL (in copies/mL) is transformed to the $\log_{10}$ scale and used to generate a natural cubic spline basis with boundary and internal knots at $(3, 4, 4.7, 5, 6)$, no intercept is included (as this would not be identifiable) and the basis is centred at 4.7 (i.e. $\log_{10}(50\,000)$), which results in four model parameters relating to $\phi_{1:i}$ and an additional four relating to $\phi_{2:i}$. Sets of parameters relating to groupings of patients determined by gender and mode of infection

are then added to the model; we have combined these characteristics into a single step in model development because of the inherent dependence between gender and mode of infection (i.e. the majority of the cohort are homosexual men, but there is not an equivalent group of women) and the potential for gender differences to vary according to whether the patient is an injecting drug user (IDU). Male homosexual patients were treated as the reference group ($n = 5443$), with parameters added to the models for $\phi_{1:i}$ and $\phi_{2:i}$ for heterosexual men and women ($n = 667$ and $n = 758$, respectively) and male and female IDUs ($n = 151$ and $n = 46$, respectively).

Patient age at treatment (in years) is then added to the models for $\phi_{1:i}$ and $\phi_{2:i}$ using a natural cubic spline basis and knots at (25,31,36,41,51), approximately equivalent to the 10$^{\text{th}}$, 30$^{\text{th}}$, 50$^{\text{th}}$, 70$^{\text{th}}$ and 90$^{\text{th}}$ centiles. As for VL there is no intercept, and so four parameters are added to the model for both $\phi_{1:i}$ and $\phi_{2:i}$, and the basis is centred at 36 years. Parameters relating to the diagnosis of an AIDS-defining illness prior to initiation of HAART ($n = 204$) are then added, followed by parameters linked to either a positive test for HCV ($n = 387$) or no record of a test for HCV ($n = 1202$) prior to HAART. The predictive value of HAART regimen classification at initiation was then assessed, with patients grouped with NNRTI regimen as reference ($n = 2998$), and parameters added relating to ritonavir-boosted PI ($n = 2485$), INSTI ($n = 438$) or other treatment regimens ($n = 1144$).

After the addition of the specified patient and treatment characteristics to the model, the functions linking (square-root) baseline CD4 value and the speed and long-term maximum of recovery were generalised (from a linear relationship) using a natural cubic spline basis with knots at 15.5, 17.5, 19.5 and 22 (stratified by time from estimated date of seroconversion to treatment initiation, as in Chapter 5). This was done after the addition of the patient characteristics to the model because the combination of a large number of additional parameters (12) and the use of natural cubic splines applied to a latent variable term increased the required computation time to fit the model up to a level that was close to the maximum available (72 hours).

6.4 Results without censoring due to virological failure

The models fitted to the full CASCADE dataset (without censoring related to VL) are summarised in Table 6.2. Generalising the baseline model (Mod_1) so that CD4 recovery followed a Janoshek–Sager curve (Mod_2) led to a highly significant improvement in model fit ($2\Delta\ell 2\,422$ for 1 parameter, $P < 0.0001$), and so this extension to the model was maintained. As was found in Chapter 5, stratifying the functions linking baseline CD4 to recovery by the time elapsed from estimated date of seroconversion to treatment initiation also led to a highly significant improvement in model fit (Mod_3; $2\Delta\ell 608$ for 8 parameters; $P < 0.0001$). A further highly significant improvement in

model fit was found when VL prior to treatment initiation was added as a predictor (Mod_4; $2\Delta\ell 578$ for 8 parameters; $P < 0.0001$). Adding each of the remaining patient and drug regimen characteristics to the model as predictors (Mod_5–Mod_9) led to statistically significant improvements in model fit (with $P < 0.01$ in all cases, and corresponding reductions in AIC); however, the improvement in log-likelihood for each of these models was modest relative to the size of the dataset under investigation, and no further improvements in BIC were seen. Similarly, the use of natural cubic splines to create more flexible link functions between baseline CD4 and the nature of post-treatment recovery led to a statistically significant improvement in model fit ($Mod_{10}\, vs\, Mod_9$; $2\Delta\ell 68$ for 12 parameters; $P < 0.0001$), but not a reduction in BIC.

Table 6.2. Summary of fitted combined models for CD4 cell counts before and after the initiation of highly active antiretroviral therapy (HAART) in patients from the CASCADE cohort. All models shown are nested within that described in the row below.

Model	Predictors	Curve	n_{pars}	ℓ	AIC	BIC	$2\Delta\ell$
Mod_1	Linear-u	Asym.	15	-229390	458810	458952	NA
Mod_2	Linear-u	JS	16	-228179	456390	456541	2422
Mod_3	As above + trt-time grp	JS	24	-227875	455798	456025	608
Mod_4	As above + baseline VL	JS	32	-227586	455236	455538*	578
Mod_5	As above + gender/inf grp	JS	40	-227571	455222	455600	30
Mod_6	As above + age	JS	48	-227556	455208	455661	30
Mod_7	As above + AIDS Dx	JS	50	-227543	455186	455658	26
Mod_8	As above + HCV Dx	JS	54	-227534	455176	455686	18
Mod_9	As above + trt regimen	JS	60	-227505	455130	455697	58
Mod_{10}	As above + NCS-u	JS	72	-227471	455086*	455766	68

The 'Predictors' field lists variables included in the functions to determine both long-term maximum (ϕ_1) and speed of recovery (ϕ_2), and 'Curve' gives shape of expected recovery following HAART. 'trt-time grp' denotes stratification of functions for long-term maximum and speed of recovery in terms of baseline CD4 at treatment initiation according to time elapsed from seroconversion to treatment. *Lowest value of AIC/BIC for set of models. '$2\Delta\ell$' denotes differences in 2×log-likelihood in comparison to model described in the row above in each case. AIC, Akaike information criterion; AIDS, acquired immune deficiency syndrome; Asym., asymptotic; BIC, Bayesian information criterion; Dx, diagnosis prior to HAART; grp, group; HCV, hepatitis C virus; inf, mode of infection; JS, Janoshek–Sager; ℓ, log-likelihood of model fit; NA, not applicable; NCS, natural cubic spline; n_{pars}, number of parameters in model; trt, treatment; VL, viral load.

Although it seems therefore that Mod_4 might provide a more parsimonious model for response to treatment, we further investigate the implications of the fitted Mod_{10} to evaluate the role of patient and drug regimen characteristics to predict response to HAART. Parameter estimates for Mod_{10} are given in Table 6.3. Direct interpretation of the parameter estimates is complicated for many of the patient characteristics by the fact that the sign (i.e. +/−) of the effect on long-term maximum CD4 is opposite for that on speed of recovery, and so it is not immediately obvious whether

or not an associated benefit is predicted. This problem is compounded by the fact that the D parameter for the Janoshek–Sager curve was estimated to be less than one ($\hat{D}$=0.42 for Mod_{10}), indicating a rapid early response to treatment but with a very gradual later response; this has the effect that the modelled long-term maximum for any given patient is not attained within the time-frame for the available data for this analysis, and so the parameters relating to 'long-term maximum' and 'speed of recovery' cannot be interpreted in isolation. As such, evaluation of the fitted model is focused on generating and comparing predicted recovery curves for hypothetical patients. However, we start by inspecting the fitted natural cubic spline functions for baseline CD4, VL prior to treatment initiation and patient age.

The link functions for ϕ_1 and ϕ_2 in terms of baseline CD4 (u_i^+), and stratified by elapsed time from estimated date of seroconversion to initiation of HAART fitted in Mod_{10} are shown in Figure 6.1. As found in the more limited analysis in Chapter 5, both the long-term maximum and speed of recovery were positively associated with the baseline 'true' CD4 count, and those patients that initiated treatment within 6 months of seroconversion were found to show a more rapid recovery for a given baseline CD4 count. The effect of pre-treatment VL on ϕ_1 and ϕ_2 is plotted in Figure 6.2, VL does not appear to predict the long-term maximum CD4 count after treatment initiation, but higher than average VL values do seem to predict a substantially higher speed of recovery. Patient age at treatment initiation was estimated to have little effect on long-term maximum CD4, and greater age was found to be associated with a small reduction in the speed of recovery (Figure 6.3).

The predicted median recovery in CD4 counts following initiation of HAART for a series of hypothetical patients is presented in Figures 6.4 and 6.5. In Figure 6.4 predictions are shown according to 'true' baseline CD4 and time elapsed from estimated date of seroconversion to treatment initiation, again demonstrating the link between baseline CD4 and long-term maximum. The plots also demonstrate that the use of the Janoshek–Sager curve results in a model that predicts (on average) gradual increases in CD4 beyond 5 years from the initiation of HAART, even amongst those patients with a high baseline value. The gain in speed of recovery associated with early initiation of HAART appears to be only moderate, but high VL prior to treatment is also strongly predictive of a rapid response as shown in Figure 6.5a. VL shows a peak close to the date of seroconversion (e.g. Pantazis *et al.*[127]), and so high VL measurements might be acting as a marker that any given patient is close to their date of seroconversion. The CASCADE dataset includes patients with up to 3 years between their last negative and first positive test for HIV, and so this might be the case even for those patients assigned to the group with greater than 1 year between estimated date of seroconversion and treatment initiation. This is further investigated in Chapter 7.

For the remaining patient and treatment characteristics the estimated effect sizes

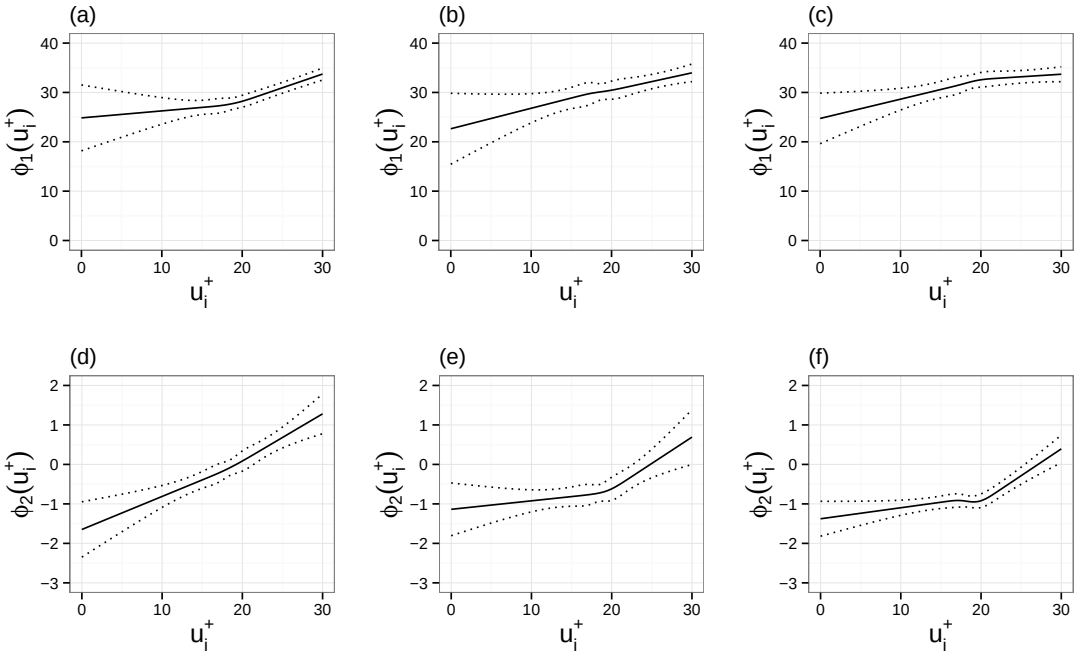

Figure 6.1. Plots of $\phi_1(u_i^+)$ (a–c, relating to long-term maximum) and $\phi_2(u_i^+)$ (d–f, relating to speed of response) for Mod_{10}. Graphs on the left of each row (a,d) show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre (b,e) show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right (c,f) show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown (·····).

were only moderate (Figure 6.5), which makes interpretation difficult given the potential for residual unmeasured confounding factors. Recovery is predicted to be slightly worse for male heterosexuals or female IDUs, but the sample size in the latter group was very small and the 95 CIs of parameter estimates for the effect on ϕ_1 and ϕ_2 both included zero. As also demonstrated in Figure 6.3, recovery is predicted to be better on average in younger patients. A surprising finding is that an AIDS diagnosis prior to treatment initiation was associated with slightly better recovery, although the sample size of such patients was small. A positive HCV test prior to treatment initiation was associated with slightly worse recovery. Of the HAART regimen regimens, the INSTI category was associated with improved recovery, with the 'other' category showing the next best performance. However, it is possible that the use of newer drugs is associated with confounding factors such as earlier treatment initiation (as this is only partially controlled for in the current model), and so caution is required in the interpretation of this finding.

The estimates of variance parameters relating to unexplained variation in post-treatment CD4 recovery were large, representing clinically meaningful differences in response to treatment that could not be attributed to the combination of patient and drug characteristics included in the model. This can be seen in both the estimated

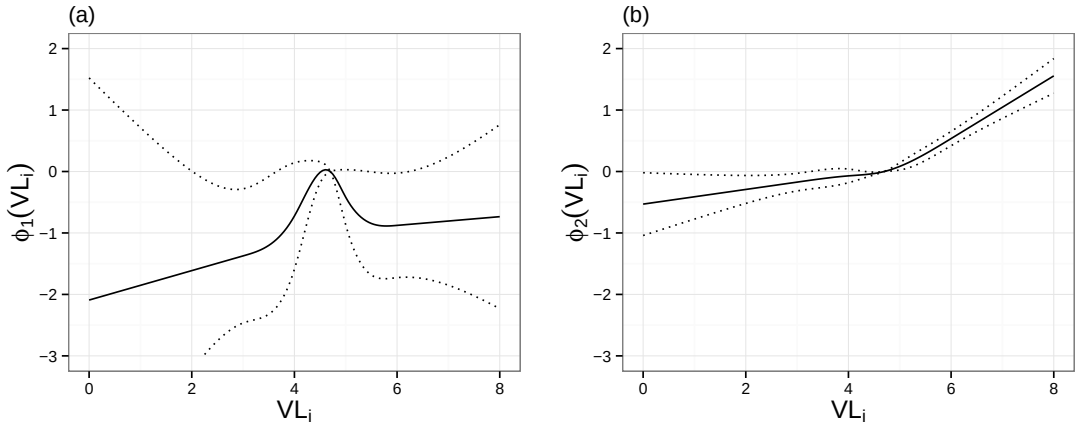

Figure 6.2. Plots of effect on ϕ_1 (a, relating to long-term maximum) and ϕ_2 (b, relating to speed of response) of pre-treatment viral load (VL, expressed using $\log_{10}$ scale on x-axis) as estimated in Mod_{10}. Pointwise 95 % confidence intervals for the functions are shown (⋯⋯). The model is parameterised such that the effect at $\log_{10}$(VL)=4.7 is zero.

variance of the random effect term relating to asymptotic maximum ($\hat{\Omega} = 9.7$) and to a lesser extent in the parameters relating to the post-treatment fractional Brownian motion process ($\hat{\kappa}_{post} = 4.6$ and $\hat{H}_{post} = 0.23$). The residual variation is also illustrated in Figure 6.6, in which the 5$^{\text{th}}$ and 95$^{\text{th}}$ centiles of post-treatment CD4 counts are plotted for hypothetical patients in addition to the median.

6.5 Results with censoring due to virological failure

When the set of models were fitted to the processed dataset with censoring of post-treatment CD4 counts at any occurrence of detectable VL beyond 6 months after treatment initiation, the same pattern was observed of statistically significant improvements in model fit but with optimal BIC for the inclusion of only VL (Table 6.4). Furthermore, predictions generated from the fitted model including all patient and drug characteristics (i.e. Mod_{10}) are nearly identical to those for the equivalent model without censoring due to detectable VL (Figure 6.7). The parameter estimates for Mod_{10} fitted to the two versions of the dataset were correspondingly very similar, as can be seen in Table 6.3. There are some apparent differences between the parameter estimates relating to natural cubic spline functions, but plotting of the fitted functions reveals them to be similar within the range of values well represented in the data, for example the estimated functions linking baseline CD4 to recovery characteristics resulting from the model fitted to the censored dataset are plotted in Figure 6.8.

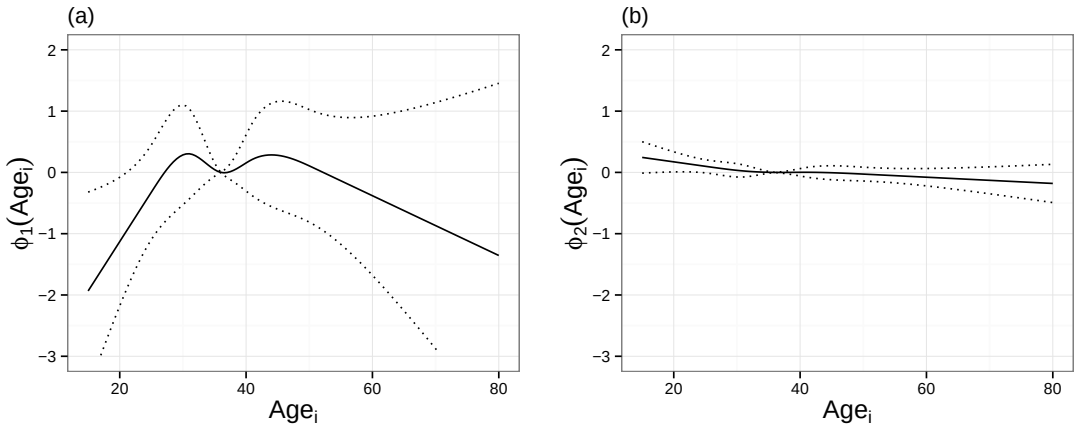

Figure 6.3. Plots of effect on ϕ_1 (a, relating to long-term maximum) and ϕ_2 (b, relating to speed of response) of patient age at treatment initiation as estimated in Mod_{10}. Pointwise 95 % confidence intervals for the functions are shown (·····).The model is parameterised such that the effect at 36 years is zero.

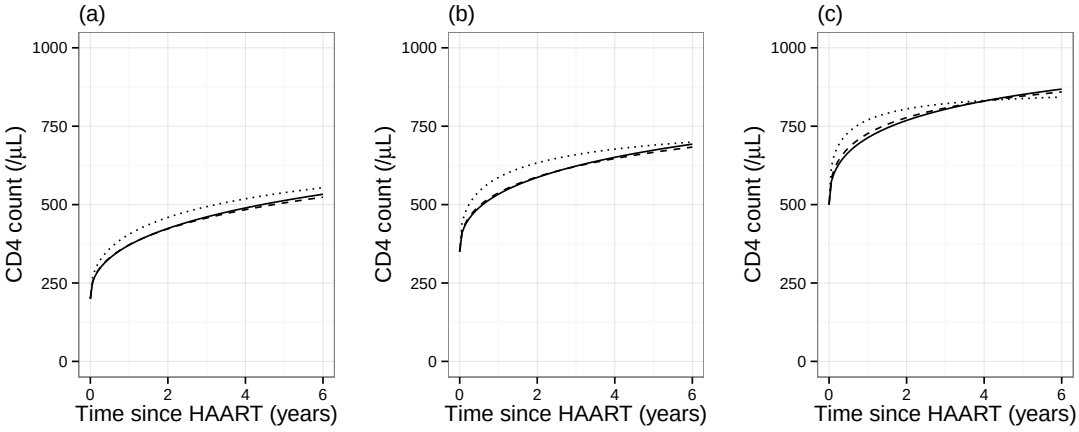

Figure 6.4. Plots of predicted median recovery in CD4 counts, based on Mod_{10}, for patients with a 'true' baseline value of 200 (a), 350 (b) or 500 (c) cells/μL. Predictions are shown for patients initiating treatment within 6 months of seroconversion (·····), patients initiating treatment beyond 6 months but within 1 year (- - -) and for patients who started treatment beyond 1 year (——). For this plot, all patients are assumed to be male homosexual, aged 36 years, with negative test for hepatitis C virus, no prior AIDS diagnosis and starting on a non-nucleoside reverse-transcriptase inhibitor (NNRTI) regimen. Viral load prior to treatment is also fixed at the overall $\log_{10}$ median of 4.825.

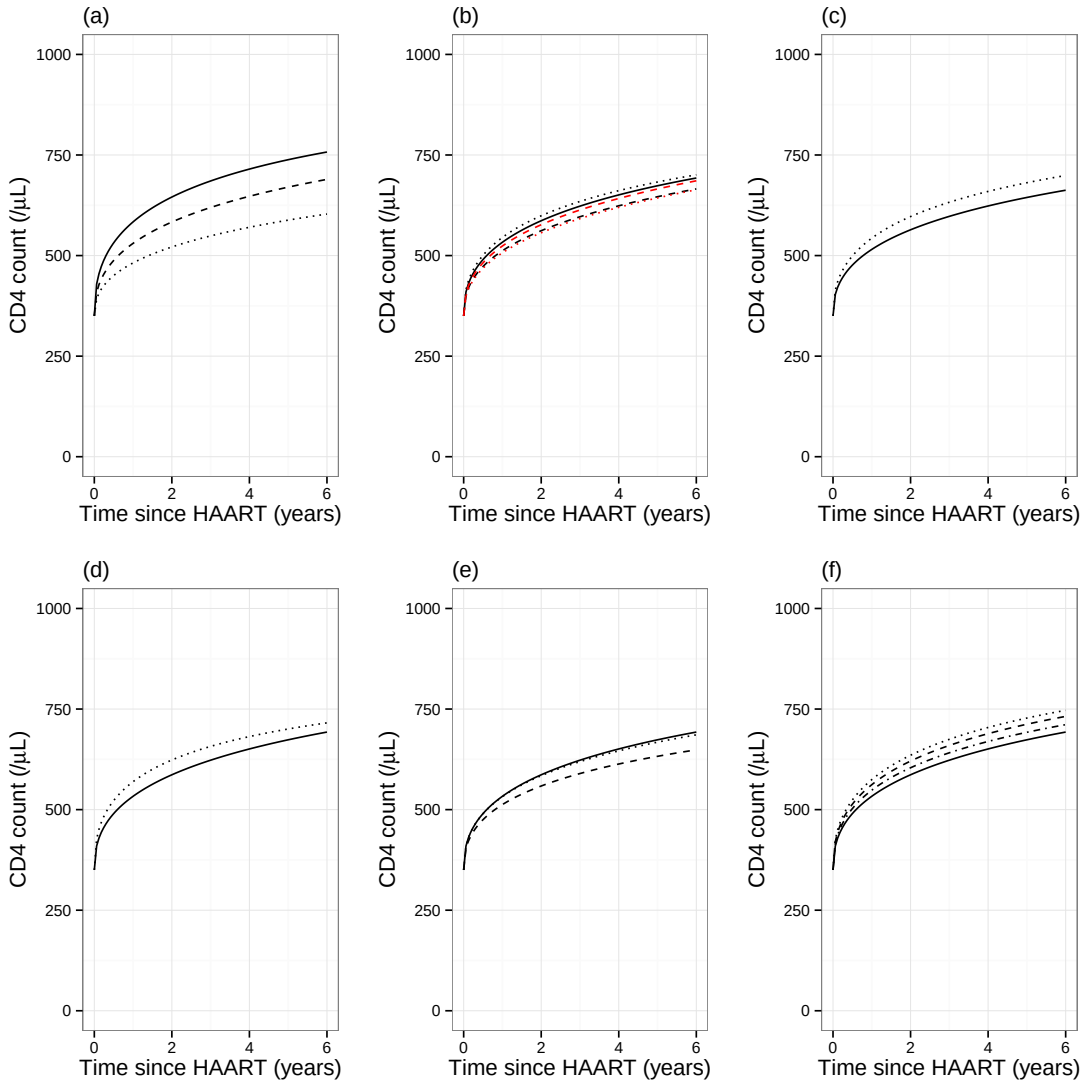

Figure 6.5. Plots of predicted median recovery in CD4 counts, based on Mod_{10}, for patients with a 'true' baseline value of 350 according to: (a) viral load (VL) prior to treatment initiation (·····, $\log_{10}(\text{VL}) = 2.7$; - - - , $\log_{10}(\text{VL}) = 4.7$; ——, $\log_{10}(\text{VL}) = 5.7$); (b) gender and infection groups (——, male homosexual; - - - , male heterosexual; ·····, male injecting drug user; - - - , female heterosexual; ·····, female injecting drug user); (c) patient age at treatment initiation (·····, $age = 20$ years; ——, $age = 60$ years); (d) AIDS diagnosis prior to treatment (·····, yes; ——, no); (e) hepatitis C virus (HCV) status (·····, no test; - - - , +ve test; ——, -ve test); and (f) HAART regimen (·····, integrase strand transfer inhibitor; -·-·-, ritonavir-boosted protease inhibitor; - - - , other; ——, non-nucleoside reverse-transcriptase inhibitor (NNRTI)). All patients are assumed to be male homosexual, aged 36 years, with negative test for HCV, no prior AIDS diagnosis, baseline $\log_{10}$(VL)=4.825 and starting on a NNRTI regimen at more than 1 year since estimated date of seroconversion unless stated otherwise.

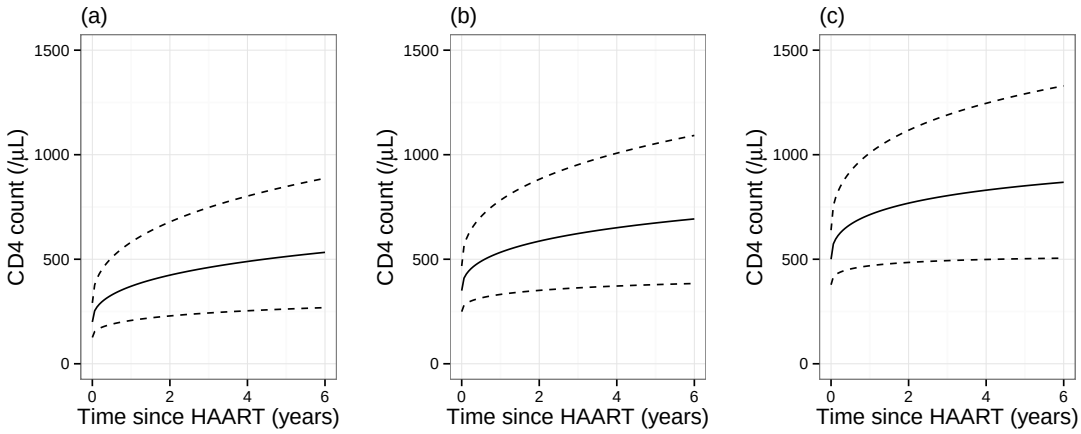

Figure 6.6. Plots of predicted median (——) and 5th and 95th centiles (- - -) for recovery in CD4 counts, based on Mod_{10}, for patients with a 'true' baseline value of 200 (a), 350 (b) or 500 (c) cells/μL. For this plot, all patients are assumed to be male homosexual, aged 36 years, with negative test for hepatitis C virus, no prior AIDS diagnosis and starting on a non-nucleoside reverse-transcriptase inhibitor (NNRTI) regimen beyond 1 year from sero-conversion. Viral load prior to treatment is also fixed at the overall $\log_{10}$ median of 4.825.

Table 6.3. Parameter estimates for Mod_{10} as applied to the full CASCADE dataset, and to a processed dataset with censoring of CD4 counts at occurrence of detectable viral load (VL) beyond 6 months after treatment initiation.

Para.	Full data fit	Cens data fit	Para. (cont. 1)	Full data fit (cont. 1)	Cens data fit (cont. 1)	Para. (cont. 2)	Full data fit (cont. 2)	Cens data fit (cont. 2)
β_0	22.144 (22.015 to 22.274)	22.151 (22.021 to 22.28)	$Bt2_1$	-1.137 (-1.806 to -0.469)	-1.117 (-1.934 to -0.301)	FI_A	3.203 (-0.838 to 7.244)	5.278 (0.519 to 10.038)
β_1	-1.447 (-1.516 to -1.379)	-1.452 (-1.521 to -1.383)	$Bt2_2$	0.021 (-0.026 to 0.069)	0.023 (-0.032 to 0.078)	FI_B	-0.422 (-0.878 to 0.034)	-0.581 (-1.058 to -0.105)
U_{00}	20.927 (19.975 to 21.924)	20.87 (19.918 to 21.868)	$Bt2_3$	0.004 (-0.068 to 0.075)	0.001 (-0.069 to 0.072)	FH_A	1.404 (0.371 to 2.437)	1.994 (0.82 to 3.168)
ρ	-0.649 (-0.724 to -0.573)	-0.652 (-0.73 to -0.573)	$Bt2_4$	0.011 (-0.134 to 0.157)	0.016 (-0.126 to 0.158)	FH_B	-0.179 (-0.296 to -0.062)	-0.236 (-0.366 to -0.107)
U_{11}	0.608 (0.476 to 0.777)	0.592 (0.458 to 0.765)	B_1	-1.377 (-1.819 to -0.935)	-1.057 (-1.531 to -0.584)	$ageA_1$	0.161 (0.02 to 0.302)	0.157 (0.008 to 0.307)
κ_{pre}	6.01 (5.681 to 6.358)	5.951 (5.617 to 6.304)	B_2	0.028 (-0.002 to 0.058)	0.01 (-0.022 to 0.042)	$ageA_2$	-0.041 (-0.091 to 0.01)	-0.044 (-0.097 to 0.01)
H_{pre}	0.313 (0.283 to 0.343)	0.324 (0.292 to 0.356)	B_3	-0.023 (-0.061 to 0.015)	-0.022 (-0.06 to 0.017)	$ageA_3$	0.102 (-0.04 to 0.245)	0.108 (-0.043 to 0.259)
σ	1.775 (1.739 to 1.811)	1.788 (1.751 to 1.826)	B_4	0.064 (-0.011 to 0.138)	0.067 (-0.009 to 0.143)	$ageA_4$	-0.088 (-0.227 to 0.051)	-0.088 (-0.235 to 0.058)
$At1_1$	24.843 (18.168 to 31.518)	24.542 (16.392 to 32.692)	Ω	9.692 (8.384 to 11.205)	10.663 (9.195 to 12.366)	$ageB_1$	-0.014 (-0.036 to 0.007)	-0.011 (-0.034 to 0.012)
$At1_2$	0.142 (-0.282 to 0.565)	0.17 (-0.351 to 0.692)	κ_{post}	4.588 (4.303 to 4.893)	3.87 (3.568 to 4.197)	$ageB_2$	0.001 (-0.006 to 0.008)	0.001 (-0.007 to 0.008)
$At1_3$	0.042 (-0.268 to 0.352)	0.034 (-0.337 to 0.404)	H_{post}	0.228 (0.21 to 0.246)	0.222 (0.199 to 0.244)	$ageB_3$	-0.001 (-0.021 to 0.018)	0.001 (-0.02 to 0.022)
$At1_4$	-0.015 (-0.588 to 0.558)	0.001 (-0.68 to 0.681)	D	0.421 (0.406 to 0.437)	0.427 (0.41 to 0.444)	$ageB_4$	-0.001 (-0.02 to 0.018)	-0.003 (-0.023 to 0.017)
$At2_1$	22.642 (15.452 to 29.832)	21.179 (12.745 to 29.614)	VA_1	0.239 (-0.884 to 1.363)	-0.024 (-1.155 to 1.106)	$AIDS_A$	-2.626 (-3.931 to -1.322)	-3.308 (-4.584 to -2.032)
$At2_2$	0.415 (-0.067 to 0.897)	0.485 (-0.065 to 1.036)	VA_2	1.229 (-0.779 to 3.238)	1.572 (-0.477 to 3.622)	$AIDS_B$	0.502 (0.327 to 0.677)	0.615 (0.428 to 0.802)
$At2_3$	-0.064 (-0.591 to 0.464)	-0.083 (-0.603 to 0.437)	VA_3	-6.108 (-13.28 to 1.065)	-7.661 (-15.039 to -0.282)	HCV_pre_A	-1.901 (-3.341 to -0.461)	-2.561 (-3.898 to -1.224)
$At2_4$	0.117 (-0.905 to 1.14)	0.147 (-0.848 to 1.142)	VA_4	11.937 (-0.486 to 24.36)	14.845 (1.999 to 27.691)	HCV_pre_B	0.057 (-0.153 to 0.268)	0.15 (-0.055 to 0.355)
A_1	24.731 (19.615 to 29.847)	19.655 (14.843 to 24.466)	VB_1	0.119 (-0.037 to 0.276)	0.144 (-0.009 to 0.297)	$no\_HCV\_test_A$	-0.791 (-1.424 to -0.158)	-0.782 (-1.436 to -0.128)
A_2	0.391 (0.053 to 0.728)	0.677 (0.357 to 0.996)	VB_2	-0.053 (-0.317 to 0.211)	-0.076 (-0.341 to 0.19)	$no\_HCV\_test_B$	0.071 (-0.021 to 0.162)	0.084 (-0.012 to 0.181)
A_3	0.026 (-0.334 to 0.385)	-0.033 (-0.387 to 0.321)	VB_3	0.329 (-0.606 to 1.264)	0.437 (-0.518 to 1.392)	$INSTI_A$	-0.632 (-1.616 to 0.351)	-0.705 (-1.713 to 0.304)
A_4	-0.099 (-0.791 to 0.593)	-0.039 (-0.723 to 0.645)	VB_4	-0.311 (-1.948 to 1.327)	-0.424 (-2.116 to 1.268)	$INSTI_B$	0.296 (0.144 to 0.448)	0.309 (0.153 to 0.465)
$Bt1_1$	-1.648 (-2.348 to -0.947)	-1.73 (-2.632 to -0.828)	MI_A	-1.243 (-3.499 to 1.013)	0.336 (-2.219 to 2.891)	PI_A	-0.551 (-1.213 to 0.11)	-0.588 (-1.283 to 0.107)
$Bt1_2$	0.084 (0.035 to 0.132)	0.087 (0.025 to 0.15)	MI_B	0.196 (-0.138 to 0.529)	-0.009 (-0.364 to 0.346)	PI_B	0.142 (0.056 to 0.229)	0.147 (0.055 to 0.24)
$Bt1_3$	0.006 (-0.056 to 0.068)	0 (-0.071 to 0.072)	MH_A	1.356 (0.233 to 2.478)	1.169 (0.004 to 2.333)	$other_A$	-0.426 (-1.202 to 0.35)	-0.257 (-1.076 to 0.561)
$Bt1_4$	-0.005 (-0.13 to 0.12)	0.004 (-0.139 to 0.146)	MH_B	-0.25 (-0.383 to -0.118)	-0.24 (-0.384 to -0.095)	$other_B$	0.207 (0.106 to 0.307)	0.204 (0.097 to 0.311)

95% CIs are given in parentheses. Parameters (Para.) up to H_{post} are as described in Table 5.2. D is the shape parameter for the Janoshek–Sager curve. VA_i parameters relate to the natural cubic spline (NCS) function for the effect of pre-treatment viral load on long-term maximum recovery. MH_A, MI_A, FH_A and FI_A denote the effect of being a male heterosexual, male injecting drug user (IDU), female heterosexual or female IDU, respectively. $ageA_i$ are NCS function parameters for patient age at treatment initiation. Estimated effects are also shown for occurrence of an AIDS diagnosis prior to treatment ($AIDS_A$), a positive hepatitis C virus test (HCV_pre_A) or lack of test before treatment ($no\_HCV\_test_A$), and for HAART initiation on an integrase strand transfer inhibitor regimen ($INSTI_A$), ritonavir-boosted protease inhibitor regimen (PI_A) or 'other' regimen ($other_A$). Equivalent 'B' parameters denote the corresponding estimated effect of the speed of recovery following treatment initiation.

Table 6.4. Summary of fitted combined models for CD4 cell counts before and after the initiation of highly active antiretroviral therapy (HAART) in patients from the CASCADE cohort, with censoring of post-treatment CD4 counts at the observation of detectable viral load beyond 6 months after treatment initiation. All models shown are nested within that described in the row below.

Model	Predictors	Curve	n_{pars}	ℓ	AIC	BIC	$2\Delta\ell$
Mod_1	Linear-u	Asym.	15	-192644	385318	385457	NA
Mod_2	Linear-u	JS	16	-191634	383300	383448	2020
Mod_3	As above + trt-time grp	JS	24	-191336	382720	382942	596
Mod_4	As above + baseline VL	JS	32	-191037	382138	382435*	598
Mod_5	As above + gender/inf grp	JS	40	-191023	382126	382497	28
Mod_6	As above + age	JS	48	-191008	382112	382557	30
Mod_7	As above + AIDS Dx	JS	50	-190992	382084	382548	32
Mod_8	As above + HCV Dx	JS	54	-190982	382072	382573	20
Mod_9	As above + trt regimen	JS	60	-190951	382022	382578	62
Mod_{10}	As above + NCS-u	JS	72	-190915	381974*	382641	72

The 'Predictors' field lists variables included in the functions to determine both long-term maximum (ϕ_1) and speed of recovery (ϕ_2), and 'Curve' gives shape of expected recovery following HAART. 'trt-time grp' denotes stratification of functions for long-term maximum and speed of recovery in terms of baseline CD4 at treatment initiation according to time elapsed from seroconversion to treatment. *Lowest value of AIC/BIC for set of models. '$2\Delta\ell$' denotes differences in 2×log-likelihood in comparison to model described in the row above in each case. AIC, Akaike information criterion; AIDS, acquired immune deficiency syndrome; Asym., asymptotic; BIC, Bayesian information criterion; Dx, diagnosis prior to HAART; HCV, hepatitis C virus; grp, group; inf, mode of infection; JS, Janoshek–Sager; ℓ, log-likelihood of model fit; NA, not applicable; NCS, natural cubic spline; n_{pars}, number of parameters in model; trt, treatment; VL, viral load.

6.6 Results including between-patient differences in variability

We also attempted to fit models in which between-patient differences in variability were accounted for, using linked multivariate-t distributions for the pre- and post-treatment fractional Brownian motion components of the model as described in Section 5.7. This was done using the full dataset without censoring due to virological failure, given that such censoring had little effect on the results obtained. Convergence was not achieved when the extension was applied to the model including all patient and drug regimen characteristic, i.e. Mod_{10}. However, maximum likelihood estimates were obtained when the extension was applied to the model with optimal BIC, i.e. Mod_4 including linear effects for baseline CD4 count stratified by time to treatment initiation and additional effects of baseline VL, using a natural cubic spline basis, on the characteristics of response to treatment. The relationships between these predictive factors and response to treatment indicated by the resulting model were nearly identical to those described for Mod_{10}; this is illustrated by Fig-

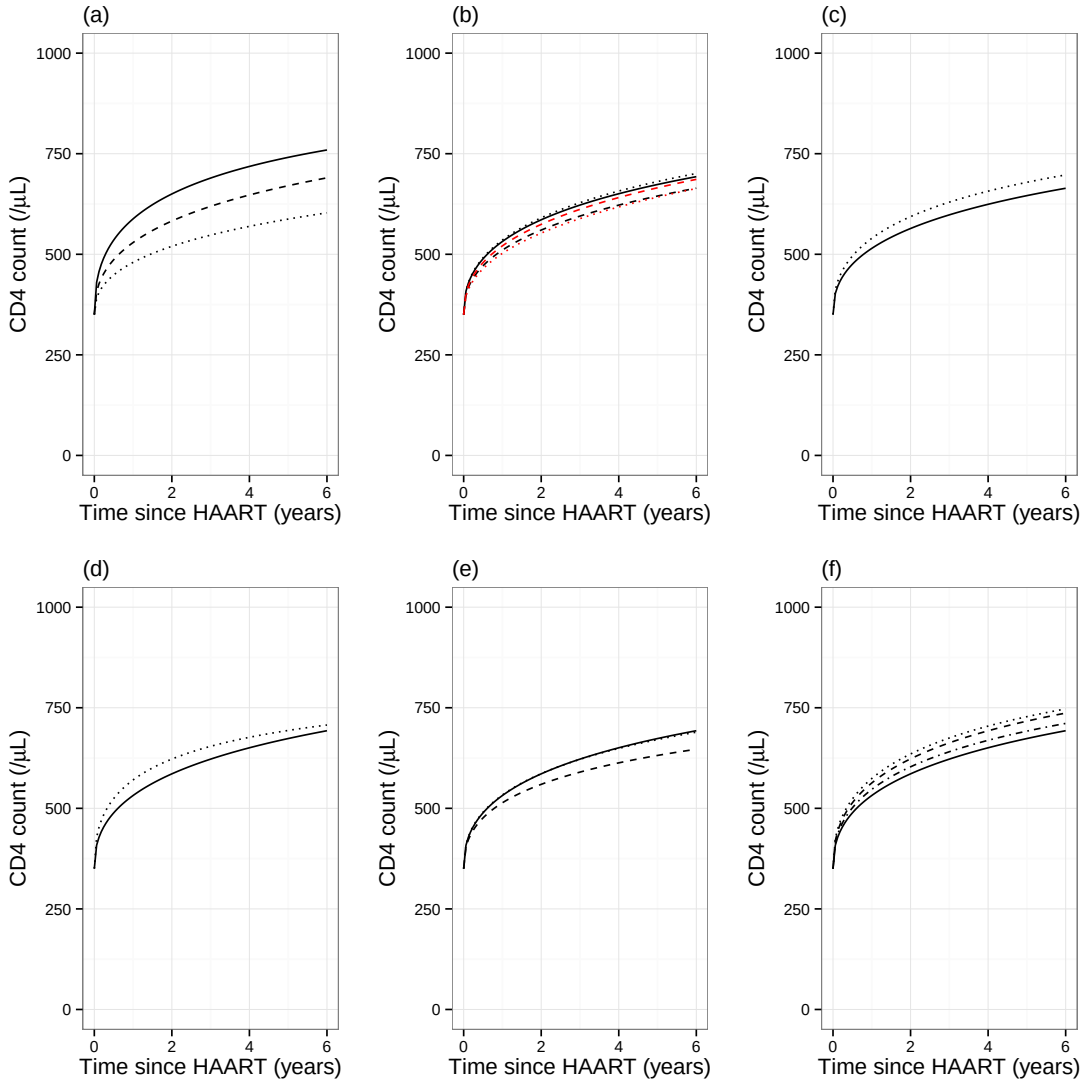

Figure 6.7. Plots of predicted median recovery in CD4 counts, based on Mod_{10} fitted to the dataset with censoring at detectable viral load (VL) after 6 months of treatment, for patients with a 'true' baseline value of 350 according to: (a) VL prior to treatment initiation (·····, $\log_{10}(\text{VL}) = 2.7$; - - - , $\log_{10}(\text{VL}) = 4.7$; ——, $\log_{10}(\text{VL}) = 5.7$); (b) gender and infection groups (——, male homosexual; - - - , male heterosexual; ·····, male injecting drug user; - - - , female heterosexual; ·····, female injecting drug user); (c) patient age at treatment initiation (·····, $age = 20$ years; ——, $age = 60$ years); (d) AIDS diagnosis prior to treatment (·····, yes; ——, no); (e) hepatitis C virus (HCV) status (·····, no test; - - - , +ve test; ——, -ve test); and (f) HAART regimen (·····, integrase strand transfer inhibitor; -·-·-, ritonavir-boosted protease inhibitor; - - - , other; ——, non-nucleoside reverse-transcriptase inhibitor (NNRTI)). All patients are assumed to be male homosexual, aged 36 years, with negative test for HCV, no prior AIDS diagnosis, baseline $\log_{10}$(VL)=4.825 and starting on a NNRTI regimen at more than 1 year since estimated date of seroconversion unless stated otherwise.

ures 6.9 and 6.10. As found in Chapter 5, the spread of the 5^{th} and 95^{th} centiles of predictions was also similar to the models that did not take between-patient differences in variability into account.

The parameter estimates for the extended Mod_4 indicated a substantial level of

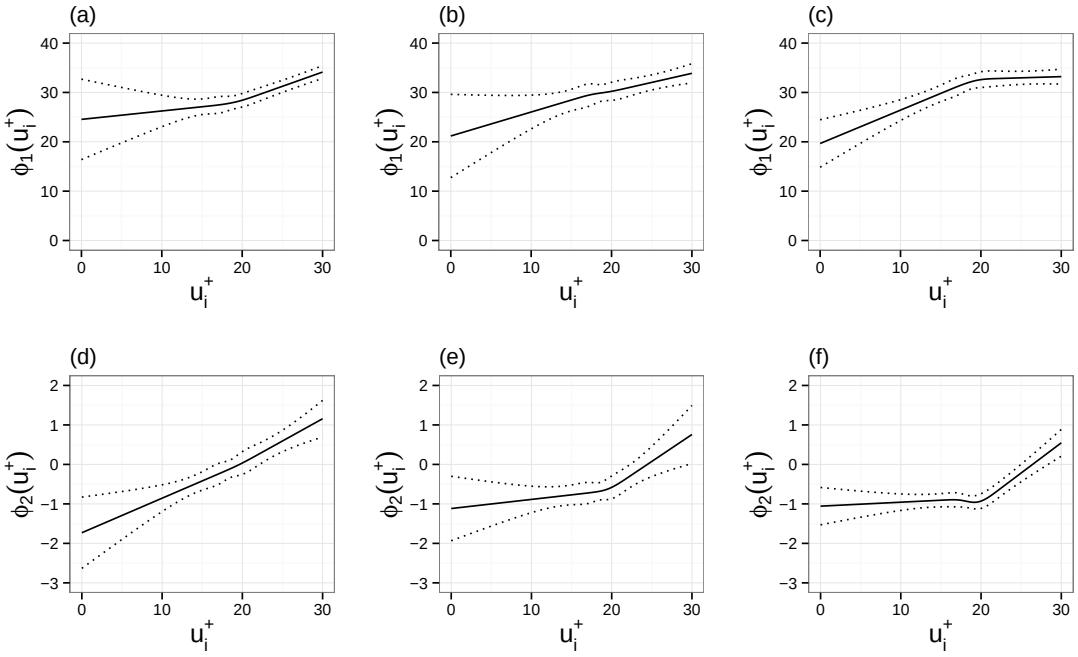

Figure 6.8. Plots of $\phi_1(u_i^+)$ (a–c, relating to long-term maximum) and $\phi_2(u_i^+)$ (d–f, relating to speed of response) for Mod_{10} fitted to the dataset with censoring at detectable viral load (VL) after 6 months of treatment. Graphs on the left of each row (a,d) show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre (b,e) show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right (c,f) show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown (·····).

between-patient differences in variability, with low values obtained for the pre- and post-treatment degrees of freedom parameters for the multivariate-t distributions fitted to the stochastic process components of the model ($\hat{\nu}_1$ = 3.59 (95% CI 3.26–3.96); $\hat{\nu}_2$ = 3.80 (95% CI 3.50–4.12)). There was a low positive correlation estimated between the levels of pre- and post-treatment variability within each patient ($\hat{\rho}_{Moran}$ = 0.18 (95% CI 0.14–0.21)).

6.7 Discussion

In this chapter, we have applied the framework developed in Chapter 5 for the combined modelling of pre- and post-treatment data to CD4 counts from the CASCADE cohort of HIV seroconverters, with the inclusion of potential predictive factors for the characteristics of post-treatment recovery. We discuss here the findings in relation to those previously reported in the literature. However, as this thesis is focused on the development of statistical methodology, we include only minimal discussion regarding how the results obtained might link to the underlying biology of HIV infection.

The analyses presented indicate that the primary factors that predict recovery

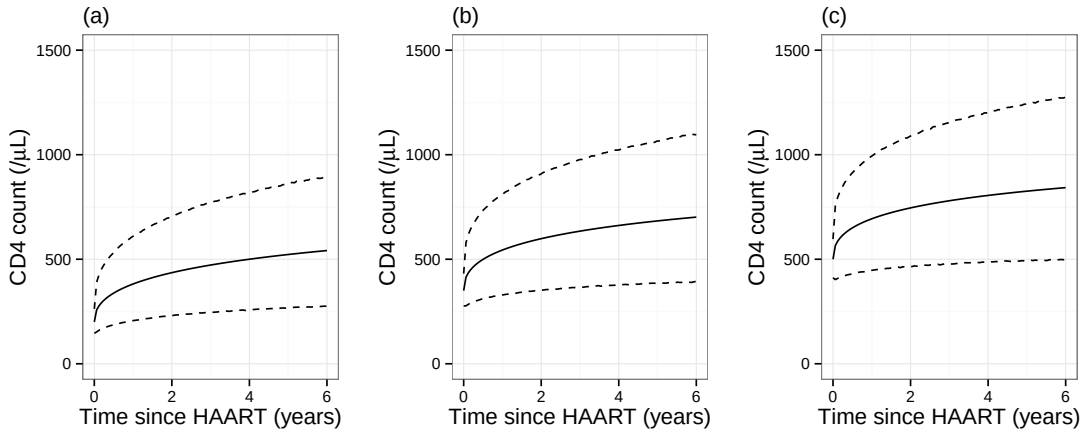

Figure 6.9. Plots of predicted median (——) and 5th and 95th centiles (- - -) for recovery in CD4 counts, based on Mod_4 with the model extended using linked multivariate-t distributions for the pre- and post-treatment stochastic process components, for patients with a 'true' baseline value of 200 (a), 350 (b) or 500 (c) cells/μL. For this plot, viral load prior to treatment is fixed at the overall $\log_{10}$ median of 4.825 and treatment initiation is set to be beyond 1 year from seroconversion. The marginal distribution is assumed for the latent scaling variable for the fractional Brownian motion process, i.e. without conditioning on any potential pre-treatment information, and the combination of multivariate normal and t distributions is approximated through averaging over 1000 draws from the relevant gamma distribution. The pattern of predictions is very close to those resulting from Mod_{10} as displayed in Figure 6.6.

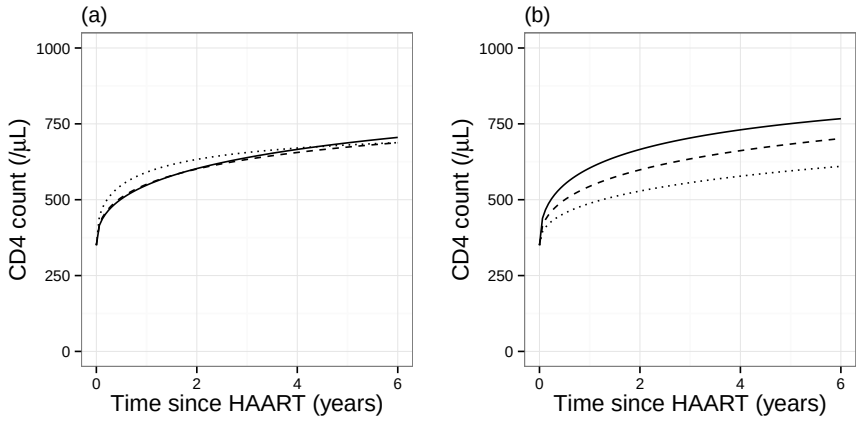

Figure 6.10. Plots of predicted median recovery in CD4 counts, based on Mod_4 with the model extended using linked multivariate-t distributions for the pre- and post-treatment stochastic process components, for patients with a 'true' baseline value of 350 according to: (a) treatment initiation within 6 months of seroconversion (·····), beyond 6 months but within 1 year (- - -) and for patients who started treatment beyond 1 year (——) and (b) with viral load (VL) prior to treatment initiation (·····, $\log_{10}$(VL) = 2.7; - - - , $\log_{10}$(VL) = 4.7; ——, $\log_{10}$(VL) = 5.7). For (a) VL prior to treatment is fixed at the overall $\log_{10}$ median of 4.825 and for (b) treatment initiation is set to be greater than 1 year from seroconversion. The pattern of predictions is very similar to those resulting from Mod_{10} as displayed in Figures 6.4b and 6.5a.

in CD4 counts following the initiation of HAART are the baseline CD4 count before the start of treatment and the pre-treatment VL. The strong positive association of baseline CD4 count with the long-term maximum of post-treatment recovery was expected given the findings of previous research on this topic[81;108–111], as discussed in Chapter 5. There is less of a consensus in the literature regarding the relationship between baseline CD4 count and the initial speed of recovery. Smith *et al.*[123] reported that higher baseline values were associated with both lower observed post-treatment increases at 3 months and a lower rate of increase beyond this point in time, and Hunt *et al.*[131] also reported greater gains in patients with lower baseline values, particularly within the first 2 years after initiation of treatment. However, Florence *et al.*[132] reported that lower baseline CD4 counts were predictive of a poor response within 6–12 months of HAART initiation, and Moore *et al.*[133] also found greater increases in CD4 cell count at 6 months post-treatment for higher baseline values up to 350 cells/μL.

In the framework that we have developed, the model term relating to 'speed of recovery' following treatment initiation represents the speed of transition from the baseline state to the long-term maximum for any given patient, rather than the rate of increase in terms of the CD4 count itself. As such, the models that we have fitted indicate that the absolute rate of increase in CD4 count will be lower for patients with 'true' baseline CD4 counts above around 600 cells/μL, due to the fact that there is less of a difference between the baseline value and the long-term maximum in such cases (this is illustrated by Figure 5.20b in Chapter 5, and equivalent plots for Mod_{10} as described in the present chapter showed a similar pattern). However, this cannot wholly explain the inconsistency in papers in the literature regarding this topic as these have largely described patients with lower CD4 counts at initiation of HAART, for example Smith *et al.*[123] report a median (IQR) pre-treatment CD4 count of 194 (75–314) cells/μL, with very similar values of 195 (118–274) cells/μL reported by Florence *et al.*[132]. For 'true' baseline CD4 counts below 600 cells/μL, our models predict a positive relationship between the baseline value and the rate of post-treatment increase in absolute terms. As such, we suggest that the inconsistencies observed could be due to the fact that lower observed baseline CD4 counts are likely to be more strongly downwardly biased as a result of selective treatment initiation and 'regression to the mean'-type effects as explored in Section 5.13.2, which will result in an increase in the apparent response to treatment, particularly at the first observation for each patient; the degree of bias is dependent on the observation and treatment initiation schedules applied to any given cohort, and so this could explain the observed differences in findings between studies.

Higher VL prior to HAART, conditional on the baseline CD4 count, was found to be associated with faster recovery and a higher long-term maximum. This finding is

consistent with previous reports in the literature, for example Smith *et al.*[123] found positive relationships with both the increase in CD4 count observed at 3 months post-treatment and with the rate of increase thereafter, Florence *et al.*[132] found that higher baseline VL was associated with lower odds of a poor response to HAART and Gras *et al.*[81] found that patients with a baseline VL ≥ 4.5 on the $\log_{10}$ scale demonstrated better long-term response to HAART conditional on their baseline CD4 value. There is some evidence that higher plasma VL levels are associated with sequestration of CD4 cells in lymphoid tissue[134;135], and it has been suggested that this is associated with a more rapid initial increase in circulating CD4 cells following the initiation of HAART[136;137]. However, given the uncertainty in the exact date of seroconversion for many of the patients included in the analysis, we hypothesised that this finding might also be at least partially due to the fact that high VL is a marker that a patient might be close to their true date of seroconversion (as shown by e.g. Pantazis et al.[127]). Investigation of this possibility constitutes a motivating factor for further developments to the modelling framework presented in Chapter 7. An alternative explanation for the fact that low pre-treatment VL predicts lower CD4 recovery is that this may reflect previous exposure to antiretroviral treatment that has not been recorded in the cohort database. However, as only patients with well-estimated date of seroconversion were included in the analysis, hence all patients were under continuous observation and inclusion in the various cohort studies from the time of HIV diagnosis, it could be argued that this is not likely to be the main true cause of this finding.

Initiation of treatment close to the estimated date of seroconversion, within 6 months according to the stratification used, was also associated with a more rapid initial improvement in post-treatment CD4 counts, with the fitted models indicating an additional benefit (beyond that associated with higher baseline CD4) over the first 2 years from the start of treatment and a moderate longer term benefit for those patients with a baseline CD4 count below around 350 cells/μL. This is in line with the findings by Le *et al.*[113]. The potential benefit of early treatment initiation, beyond that associated with a higher baseline CD4 count, and the time interval from seroconversion in which this can be observed is further investigated in Chapter 7.

The other patient and drug regimen characteristics that were included in the models developed only showed small to moderate associations with the characteristics of post-treatment recovery in CD4 counts, adjusting for baseline CD4. Because this is an analysis of an observational dataset, small estimated effect sizes need to be interpreted with caution. Increasing patient age at date of treatment initiation was found to be associated with a moderate reduction in CD4 recovery, which is consistent with previous research on this topic[81;131;132]. Pre-treatment diagnosis of HCV was also found to be associated with a moderate reduction in recovery; although

statistically significant differences in CD4 recovery have not always been found for cases of HCV (for example no differences were reported by Hunt *et al.*[131] or Florence *et al.*[132]), this is a finding that has been observed in other studies, for example Greub *et al.*[138], and a meta-analysis has supported the conclusion that HCV infection is associate with a lower CD4 recovery following initiation of HAART[139].

The finding that a pre-treatment AIDS diagnosis was associated with an improvement in the post-treatment recovery is surprising, but it should be noted that the estimated effect size was small and that this is conditional on baseline CD4 count. It is possible that the difference observed could be explained by greater sequestration of CD4 cells in the lymphoid tissue in such cases, leading to a greater increase on initiation of HAART, as suggested for cases with higher pre-treatment VL[134;135].

Comparison of recovery amongst gender and infection groups in the present analysis is hampered by the fact that sample sizes are very uneven between the groups, and the potential for a wide range of unmeasured confounding factors. Recovery appeared to be slightly worse on average amongst heterosexual men, which is not something that has been reported previously. However, it seems likely that this difference could be due to confounding factors such as ethnicity. Some previous studies have observed better post-treatment CD4 recovery in women, for example those of Hunt *et al.*[131] and Gras *et al.*[81], but we did not find such an association.

Of the classifications of drug regimens included in the analysis, the INSTI regimen at initiation of HAART was found to be associated with a moderate improvement in post-treatment recovery in CD4 relative to the NNRTI regimen, with the mixed 'other' regimen showing a similar performance. These findings warrant further investigation, but are not conclusive on their own given the potential for residual confounding and the moderate effect sizes found. There is evidence that the use of regimens including an integrase inhibitor is associated with more rapid viral suppression than regimens that do not include a drug of this class, but the evidence regarding potential differences in CD4 recovery is less clear. The FLAMINGO trial showed more rapid viral suppression for a regimen including an integrase inhibitor (dolutegravir) in comparison to a ritonavir-boosted protease inhibitor (darunavir) in addition to an NRTI 'backbone'[140], but found no difference in the change in CD4 cell count from baseline at 48 weeks. The NEAT trial randomised patients to an NRTI-sparing regimen (including an integrase inhibitor and ritonavir-boosted protease inhibitor) or a standard NRTI + ritonavir-boosted protease inhibitor regimen, finding that the regimen including an integrase inhibitor was associated with more rapid viral suppression and shorter times to achieve a CD4 count above 500, but with no difference between groups in their change in CD4 count from baseline at 96 weeks[141].

No substantial differences were observed when the models were refitted to a processed dataset with censoring of post-treatment CD4 counts following the observa-

tion of a detectable VL beyond 6 months from initiation of HAART. This is surprising given that the observation of a detectable VL on HAART could be taken to indicate poor adherence and so the exclusion of such data might be expected to lead to a fitted model that provides more optimistic predictions of CD4 count recovery. The fact that no difference was observed following censoring could be explained by high levels of drug adherence within the studied cohort or by effective recording of treatment interruptions (at which patients were censored for all fitted models). Another possible explanation is that the variance terms in the fitted models might already account for between-patient differences in adherence prior to censoring, in which case the censoring event would provide limited information regarding the future CD4 trajectory of any given patient given their prior CD4 count observations.

As noted in Chapter 5, a disadvantage of the approach that has been used for this analysis is that strong assumptions are required regarding the probability model for pre- and post-treatment CD4 cell counts. However, we have aimed to ensure that the model is as flexible as possible in order to allow it to reflect the structure and patterns of variation observed in the data under investigation. A further limitation of this analysis is that we have not included patient characteristics in the pre-treatment part of the model; this would be straightforward to achieve in principle, but the total number of parameters in the most fully developed models that we have fitted were close to the maximum possible within the time and memory constraints of the computing resources available and so the pre-treatment part of the model was not fully developed. We do not believe that this further extension would have a major impact on the conclusions drawn regarding the factors that predict post-treatment recovery in CD4 counts, although it would further refine the distribution of the 'true' baseline value in each patient and so might allow the link functions between these values and the characteristics of response to treatment to be more precisely estimated. It should also be mentioned that there are additional potential predictive factors, such as ethnicity and HIV subtype, that were not available for inclusion in the present analysis.

For all of the models fitted, the residual variation in post-treatment recovery in CD4 cell count not explained by predictive factors was substantial. It is possible that the addition of further potential predictive factors might lead to a reduction in the total residual variance, but it does not seem likely that a substantial reduction could be achieved without an in-depth analysis of immunological and virological factors specific to each patient. As such, although the models developed provide information regarding CD4 response to HAART that will be of interest to researchers and clinicians, they do not offer enough predictive power to confidently identify those patients who are likely to demonstrate a suboptimal immunological response to HAAART — beyond the established finding that the baseline CD4 count is a very important predictor of the level of long-term recovery.

As found in Chapter 5, extending the model to allow between-patient differences in variability over time appeared to provide a better fit to the data, as indicated by low estimates for the degrees of freedom parameters, although statistical comparison of these models is made difficult by the need to use the less accurate Laplace approximation for maximum likelihood estimation. This fitted model provided no substantial differences in inferences regarding the primary predictive factors for CD4 recovery, i.e. baseline CD4 count, baseline VL and time from seroconversion to treatment initiation, and the overall predicted 90 % ranges for post-treatment observations were also similar to the models based only on the multivariate normal distribution. It was not possible to obtain maximum likelihood estimates for such extended models that also included the other patient and drug regimen characteristics as potential predictive factors, and this may be due to the small effect sizes observed for these factors.

For the analysis in this chapter we have retained the assumption that the estimated date of seroconversion for the included patients is fixed and known, even though in some cases it is only known to fall within an interval between last negative and first positive HIV tests of up to 3 years. This is a problem when attempting to investigate the association between early initiation of treatment and recovery, in order to establish whether there is an additional benefit that is not mediated by baseline CD4 cell count, and it also makes it difficult to interpret the results relating to baseline VL measurements as the observation of a high baseline VL might indicate that the patient concerned is in fact closer to their true date of seroconversion than is suggested by their estimated date based on the mid-point approximation. With the aim of addressing these concerns, in Chapter 7 we present further developments to the modelling framework in which a probability model is also included for pre-treatment VL measurements and the uncertainty in exact dates of seroconversion is taken into account.

7 Modelling uncertainty in seroconversion date

7.1 Background

In previous chapters, we have developed models based on the assumption that the estimated date of seroconversion in each patient is correct. However, the uncertainty in exact date of seroconversion for those patients with a 'mid-point' estimate, set at the halfway point between last negative and first positive HIV tests, raises questions regarding the interpretation of any fitted models, particularly when trying to determine whether treatment initiation close to the date of seroconversion might lead to substantial improvements in CD4 recovery as we have done in Chapters 5 and 6.

There has been development of statistical methodology to address the problem of uncertainty in seroconversion dates, both in order to provide more accurate estimation of infection time in 'seroconverters' and to allow modelling of the delay to diagnosis in 'seroprevalent' patients. Early work on this issue was motivated by the need to estimate the survival function for the progression from seroconversion to AIDS (before the availability of effective treatment) [142;143], while more recent research has focused on the need for accurate estimation of infection dates for monitoring of the incidence of new HIV cases in different countries and communities [144;145]. One approach to dealing with the problem is to define a time-to-event regression model for the interval between seroconversion and first observation of a patient, with biomarkers at presentation used as predictive variables; for example, Muñoz *et al.* [142] used a truncated Weibull regression model for the time elapsed since seroconversion among 'seroconverters' with CD4 % as a predictive variable, and then used the results to impute dates for seroprevalent patients. Geskus [143] proposed a non-parametric approach in which the estimated distribution of the timing of seroconversion, conditional on observed CD4 counts, is empirically derived based on data from patients with well estimated date of seroconversion (including patients with an interval of up to 1 year between negative and positive tests).

Taffé *et al.* [144] developed a joint model incorporating time from seroconversion to diagnosis, differences between serial CD4 count measurements following diagnosis and drop-out from the analysis due to either ART or death. Estimation of the parameters for this model requires integration over correlated subject-specific random effect terms for the intercept and slope of CD4 trajectory as well as an independent measurement error term (the latter resulting from the fact that the model for CD4 counts is defined in terms of differences from the first observation) in order to obtain the marginal log-likelihood. In this model, the time from seroconversion to diagnosis is treated as a time-to-event outcome variable, the model for which is defined conditional on the random effect terms. Conditional imputation of the date

of seroconversion for any given patient therefore requires calculations based on the empirical Bayes estimates of their random effects.

A different approach to this problem was proposed by Sommen *et al.*[145], who developed longitudinal models for virological markers of recent infection in which the time elapsed from seroconversion to diagnosis for each patient is itself treated as a latent variable. This has the advantage that the models for the biomarkers under investigation can be defined in terms of the true time since seroconversion, with the marginal log-likelihood obtained by integration over the range of possible seroconversion dates for any given patients as well as any subject-specific random effects. A similar approach was independently described by Drylewicz *et al.*[146], who developed dynamic models for pre-treatment CD4 cell counts and VL measurements in HIV patients with integration of the likelihood function over possible infection dates. The markers under investigation by Sommen *et al.* were antibodies to the immunodominant epitope of gp41 (IDE) and a mixture of five V3 peptides; their model for each comprised an asymptotic increase from zero and independent Brownian motion and measurement error terms, and a uniform prior distribution was assumed for the occurrence of seroconversion between last negative and first positive HIV tests (or over an interval of 70 days prior to signs of symptomatic primary infection or 30 days prior to incomplete Western blot).

In this chapter, we develop a model for pre-treatment CD4 counts and VL measurements and for the recovery in CD4 counts following initiation of HAART conditional on the true date of seroconversion for each patient. We follow the principle proposed by Sommen *et al.*[145] and Drylewicz *et al.*[146] of obtaining the marginal log-likelihood by integration over a prior distribution of the true date for each patient informed by the interval between negative and positive tests, although beyond this the model structure that we develop differs from their work. Our work is novel in that we also model response to HAART in terms of the true time elapsed from seroconversion to initiation of treatment.

7.2 Exact seroconversion date as a latent variable

If a distribution can be assigned for the true seroconversion date t' for each patient, with probability density function $f_{t'}$, then the marginal likelihood for an extension of the combined model for pre- and post-treatment data as described in Chapter 5 can be expressed as follows:

$$f\left(\mathbf{y}_{pre},\mathbf{y}_{post}\right) = \int_{t'_{min}}^{t'_{max}} \int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}|T'=t'\right) f_{post}\left(\mathbf{y}_{post}|u,T'=t'\right) f_u\left(u|\mathbf{y}_{pre},T'=t'\right) f_{t'}\left(t'\right) du\, dt'.$$

This integral is of a form similar to that used by Sommen *et al.*[145].

As noted in Section 5.7, the Laplace approximation to the marginal likelihood is optimally accurate for latent variables integrated out over a normal scale[64], and so for maximum likelihood estimation we instead express this integral as:

$$f\left(\mathbf{y}_{pre}, \mathbf{y}_{post}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}|T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right)\right) f_{post}\left(\mathbf{y}_{post}|u, T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right)\right) \\ f_{u}\left(u|\mathbf{y}_{pre}, T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right)\right) f_{\phi}\left(a\right) du\, da,$$

where f_{ϕ} is the probability density function for a standard normal distribution and $F_{t'}^{-1}$ is the inverse of the cumulative distribution function corresponding to $f_{t'}$.

The expression used by ADMB for the gradient of the Laplace approximation to an integral involves third order partial derivatives of the integrand with respect to the latent variable terms, which means that the response to treatment cannot be modelled according to arbitrary divisions of the time from 'true date of seroconversion' to treatment initiation; for such models the integrand would not be differentiable with respect to t' across its range of potential values. For this extension to the model, we hypothesise that the response in CD4 count to HAART follows distinct relationships with the baseline value 'u^{+}' according to whether treatment is initiated very close to the date of seroconversion or after a long period of time has elapsed. As described in Chapter 5, the response to treatment is modelled as being dependent on the baseline CD4 value through functions that determine the expected long-term maximum and speed of recovery, with separate functions defined for 'early' and 'late' treatment initiation (denoted $\phi_{1:early}(u^{+})$, $\phi_{2:early}(u^{+})$, $\phi_{1:late}(u^{+})$ and $\phi_{2:late}(u^{+})$). However, unlike in previous chapters, we incorporate a smooth transition from the 'early' to the 'late' functions according to the exact value of t', by weighting their respective contributions towards the expected long-term maximum $\phi_{1:i}$ or speed of response $\phi_{2:i}$ for any given patient according to the functions:

$$weight_{early:i} = 2 - 2/\left(1 + \exp\left(-S * t_{trt:i}\right)\right) \\ weight_{late:i} = 2/\left(1 + \exp\left(-S * t_{trt:i}\right)\right) - 1,$$

here S is a parameter to be estimated that determines the balance between 'early' and 'late' treatment response characteristics according to the time elapsed between true date of seroconversion and initiation of HAART '$t_{trt:i}$'. These two weighting function sum to 1 for any value of $t_{trt:i}$, for $t_{trt:i} = 0$ the functions return 1 and 0, and $weight_{early:i} \to 0$ and $weight_{late:i} \to 1$ as $t_{trt:i}$ increases. A plot of these functions with the S parameter estimated from the data is presented later in this chapter in Figure 7.3 (page 154).

The terms for the expected long-term maximum ($\phi_{1:i}$) and speed of response to treatment ($\phi_{2:i}$) conditional on the timing of treatment and true baseline value in each patient are therefore given by the following expressions:

$$\phi_{1:i} = weight_{early:i} \times \phi_{1:early}(u^+) + weight_{late:i} \times \phi_{1:late}(u^+)$$
$$\phi_{2:i} = weight_{early:i} \times \phi_{2:early}(u^+) + weight_{late:i} \times \phi_{2:late}(u^+).$$

7.3 Incorporating viral load into the model

Following our aim of investigating the separate contributions of baseline CD4 count, time from seroconversion to treatment initiation and baseline VL in predicting the characteristics of CD4 recovery on HAART, we also extend the model to include pre-treatment VL as an outcome variable. This development is necessary in order to allow information from VL observations to contribute to the posterior distribution of the true date of seroconversion for each patient, and it also means that patients for whom no VL observations were obtained close to the start of treatment can be included in the analysis. Viral load is analysed on the $\log_{10}$ scale, and we make use of the non-linear model for the mean in terms of time from seroconversion as reported by Pantazis *et al.*[127]:

$$g_{VL}(t_{VL}) = \beta_{0VL} + \beta_{1VL} t_{VL} + \beta_{2VL} \exp\left(-\beta_{3VL} t_{VL}\right), \tag{9}$$

where t_{VL} is the time of VL observation from date of seroconversion and β_{0VL}–β_{3VL} are parameters to be estimated.

However, a patient-specific random effect is only included for the intercept and not for the long-term slope, as we were unable to successfully fit models that also included the latter term (the program crashed or convergence failed). The patient-specific random intercept is modelled as following a joint multivariate normal distribution with the random-intercept and -slope terms of the pre-treatment CD4 part of the model, and there is also an examination-specific independent normal error term for the pre-treatment VL:

$$\begin{aligned}
\mathbf{v}_i &= \mathbf{g}_{VL}(\mathbf{t}_{VL:i}) + \mathbf{1}\mathrm{b}_{VL:i} + \mathbf{e}_{VL:i} \\
\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\
\begin{pmatrix} \mathrm{b}_{VL:i} \\ \mathbf{b}_i \end{pmatrix} &\sim MVN\left(\mathbf{0}, \begin{pmatrix} \psi_{VL} & \mathrm{Cov}(\mathrm{b}_{VL:i}, \mathbf{b}_i) \\ \mathrm{Cov}(\mathbf{b}_i, \mathrm{b}_{VL:i}) & \boldsymbol{\Psi} \end{pmatrix}\right) \\
\mathbf{e}_{VL:i} &\sim MVN(\mathbf{0}, \sigma^2_{VL}\mathbf{I}_{n_{VL:i}}) \\
\mathbf{W}_{pre:i} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{pre:i}) \\
\mathbf{e}_{pre:i} &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{pre:i}}).
\end{aligned}$$

Here, $\mathbf{v}_i$ is the vector of $n_{VL:i}$ pre-treatment VL observations for the i$^{\text{th}}$ patient at times $\mathbf{t}_{VL:i}$, $\mathbf{g}_{VL}$ is a vectorised version of the function in (9), $\mathbf{1}$ is a vector of '1's of length $n_{VL:i}$, $\mathrm{b}_{VL:i}$ is the subject-specific random intercept for VL with variance ψ_{VL}, $\mathbf{e}_{VL:i}$ is a vector of examination-specific residuals for VL with variance σ^2_{VL}, and other terms are as defined in Chapters 5 and 6. The time values in this model are defined with respect to the true date of seroconversion for each patient through conditioning on the corresponding latent variable term.

A further complication is that the VL measurements recorded in the CASCADE dataset are truncated at lower and upper limits of detection, with these limits depending on the equipment used at each examination and ranging from 1–500 copies/mL for the lower limit and 50 000–10^8 copies/mL for the upper limit. Following Thiébaut *et al.*[147;148], we account for this issue by making use of the fact that the likelihood contribution for such an observation below a lower limit of detection, conditional on the subject-specific random intercept, is independent of other observations and can be expressed using the cumulative normal distribution function (Φ)[149] and the lower limit of detection in that case (lim^L_{ij}):

$$L(v_{ij}|\mathrm{b}_{VL:i}) = \Phi\left(\left(lim^L_{ij} - \left(g_{VL}\left(t_{VL:ij}\right) + \mathrm{b}_{VL:i}\right)\right)/\sigma_{VL}\right),$$

while the likelihood contribution for observations above the upper limit of detection can be expressed using the upper limit (lim^U_{ij}) in that case:

$$L(v_{ij}|\mathrm{b}_{VL:i}) = 1 - \Phi\left(\left(lim^U_{ij} - \left(g_{VL}\left(t_{VL:ij}\right) + \mathrm{b}_{VL:i}\right)\right)/\sigma_{VL}\right).$$

This has the consequence that approximation of the marginal log-likelihood requires integration over the VL random intercept term for each patient. If there were no lower limits of detection, then it would be possible to form a joint multivariate normal distribution (with associated closed form probability density function) for both the CD4 count and VL observations in the pre-treatment part of the model. However, we may still express the probability density function for the pre-treatment CD4 count observations in closed form conditional on the VL random intercept term in each patient, making use of standard expressions for conditional normal distributions. If we express the joint distribution for the VL and CD4 random effects as follows:

$$\begin{pmatrix}\mathrm{b}_{VL:i}\\ \mathbf{b}_i\end{pmatrix} \sim MVN\left(\mathbf{0}, \begin{pmatrix}\psi_{VL} & \boldsymbol{\psi}_{12}\\ \boldsymbol{\psi}_{21} & \boldsymbol{\Psi}\end{pmatrix}\right),$$

then the conditional model for the pre-treatment CD4 counts can be expressed as:

$$\begin{aligned}
\mathbf{y}_{pre:i}|\mathrm{b}_{VL:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\
\mathbf{b}_i|\mathrm{b}_{VL:i} &\sim MVN\left(\frac{\boldsymbol{\psi}_{21}\mathrm{b}_{VL:i}}{\psi_{VL}}, \boldsymbol{\Psi} - \frac{\boldsymbol{\psi}_{21}\boldsymbol{\psi}_{12}}{\psi_{VL}}\right) \\
\mathbf{W}_{pre:i} &\sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{pre:i}) \\
\mathbf{e}_{pre:i} &\sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_{n_{pre:i}}).
\end{aligned}$$

We also allow the post-treatment recovery in CD4 cell counts to be dependent on the realisation of b_{VL}, and the marginal log-likelihood for the complete model can therefore be expressed as:

$$\begin{aligned}
f\left(\mathbf{y}_{pre}, \mathbf{v}, \mathbf{y}_{post}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & f_{pre}\left(\mathbf{y}_{pre}|T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right), \mathrm{b}_{VL}\right) \\
& f_{post}\left(\mathbf{y}_{post}|u, T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right), \mathrm{b}_{VL}\right) \\
& f_{VL}\left(\mathbf{v}|T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right), \mathrm{b}_{VL}\right) \\
& f_u\left(u|\mathbf{y}_{pre}, T' = F_{t'}^{-1}\left(\Phi\left(a\right)\right), \mathrm{b}_{VL}\right) \\
& f_{\mathrm{b}_{VL}}\left(\mathrm{b}_{VL}\right) f_\phi\left(a\right) du\, da\, d\mathrm{b}_{VL}.
\end{aligned} \tag{10}$$

A directed acyclic graph to demonstrate the structure of this model is presented in Figure 7.1. As for Figure 5.3 in Chapter 5, links in this graph represent dependencies in the defined probability model rather than direct causal effects.

In the models that we present, the patient-specific random intercept for VL is included as a linear predictor for the long-term maximum (ϕ_1) and speed of recovery (ϕ_2) of post-treatment CD4 counts. As described in Section 7.2, parameters are fitted corresponding to early and late treatment initiation, with the weighting of the two for each patient dependent on the exact time elapsed from seroconversion to treatment initiation (which is itself defined in terms of a latent variable for those patients in whom date of seroconversion is known to fall within an interval between positive and negative tests). The use of the patient-specific VL intercept as a predictor of CD4 recovery (rather than the absolute VL level) means that the parameter estimates can be interpreted in terms of the patient's VL relative to the distribution across the population at any given point in time following seroconversion.

In approximating the marginal likelihood for this model, using the integral as shown in (10), greatest weight is placed on the values for the true date of seroconversion that maximise the joint penalised likelihood function that forms the integrand; this includes the probability density function for the pre-treatment VL observations, as well as the prior distribution for the true date of seroconversion and the probability density functions corresponding to the pre- and post-treatment and baseline

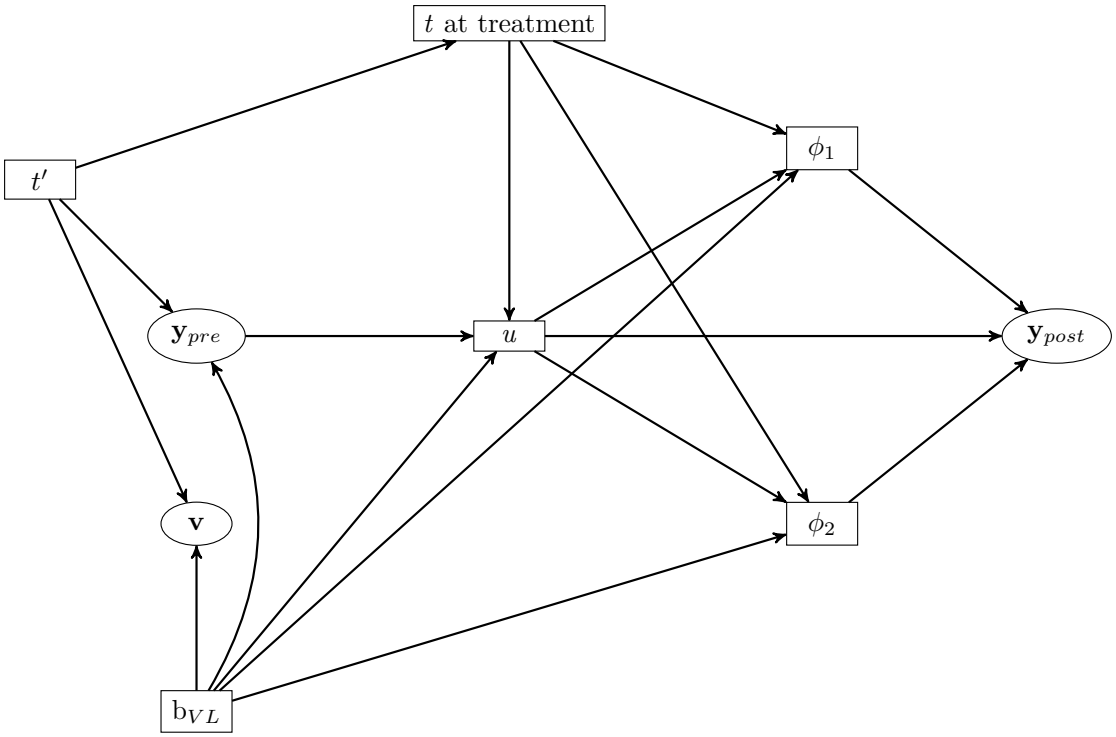

Figure 7.1. Directed acyclic graph depicting the proposed model structure for each patient, accounting for uncertainty in true date of seroconversion (t') and incorporating a probability model for pre-treatment viral load (VL). The distributions for both viral load observations ($\mathbf{v}$) and pre-treatment CD4 counts ($\mathbf{y}_{pre}$) are conditioned on the value of a random intercept variable for VL (b_{VL}), which also influences the long-term maximum (ϕ_1) and speed of recovery (ϕ_2) of post-treatment CD4 counts ($\mathbf{y}_{post}$). The distribution of the 'true' CD4 count at treatment initiation (u) is conditional on the pre-treatment observations and the timing of treatment, and the value of b_{VL} also affects the joint distribution of $\mathbf{y}_{pre}$ and u. Observed variables are shown within ellipses, whilst unobserved latent variables are shown within rectangles.

CD4 counts. All of these aspects of the model as a whole can influence the posterior distribution of the true date of seroconversion for each patient, and so the component of the model relating to pre-treatment VL measurements could affect estimates of how the post-treatment recovery in CD4 counts varies according to other factors such as time elapsed from seroconversion to treatment initiation.

7.4 Prior distribution for true date of seroconversion

We define the 'prior distribution' of true seroconversion dates as that expected before consideration of any CD4 count or VL data. For those patients in whom seroconversion date has been estimated as the midpoint between the last negative and first positive HIV tests, an obvious choice for the prior distribution is a uniform distribution over the interval between tests. However, we found that models using a uniform distribution would not converge, and so instead use a beta distribution scaled to match the duration of the interval between tests with alpha and beta parameters both fixed

at 6. The use of this distribution makes the assumption that seroconversion is most likely to have occurred close to the midpoint between negative and positive tests. This assumption may not be completely justified, but the model nonetheless represents an improvement over the common assumption that seroconversion date is fixed at the midpoint. Plots illustrating these different assumptions for the prior distribution of true seroconversion dates are shown in Figure 7.2. For those patients in whom the date of seroconversion illness or lab evidence of seroconversion (real-time polymerase chain reaction positivity or incomplete Western blot) is recorded, the date of seroconversion was considered to be fixed and known.

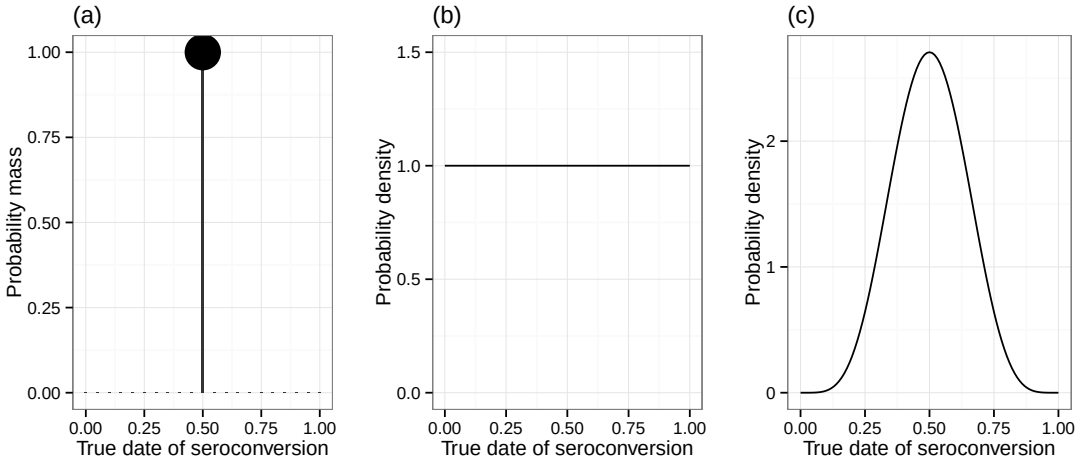

Figure 7.2. Plots illustrating different assumptions for the prior distribution of true seroconversion dates in a patient with their date of seroconversion estimated according to the interval between last negative and first positive HIV tests: (a) mid-point assumption, (b) uniform distribution and (c) beta distribution with alpha=beta=6. An interval between tests of 1 year is shown here, with the x-axis representing the true date of seroconversion within this period (i.e. 0 denotes date of last negative test).

For computational reasons, we also shift the distribution of possible seroconversion dates back in time by 1 day for all patients. This is required because ADMB-generated programs return 'not a number' when asked to return '0^c' for any value of 'c', which causes problems when calculating the covariance terms relating to fractional Brownian motion processes for each patient (involving 't^{2H}' terms). When time is fixed the issue can be avoided by defining a new function that checks that the base is not zero before attempting to calculate a power term, but this is not possible when time is allowed to vary within patients.

7.5 Dataset and estimation

The CASCADE dataset includes patients with a gap between last negative and first positive test of up to 3 years. For this analysis, we apply the same inclusion criteria as specified in Chapter 6, except that patients are not excluded if they lack any VL

observations within 6 months before treatment initiation; this is because VL is being included as a modelled outcome variable and so can be effectively imputed for patients in whom no measurements are available. This leads to a higher number of patients for potential inclusion in the analysis (n = 7849). However, to ensure the coherence of the proposed model we also exclude patients in whom the date of seroconversion was estimated according to the mid-point method but who initiated HAART before their first positive HIV test is recorded in the database (n = 60), leading to a study population of 7789 patients. Similarly, we remove from the analysis any CD4 counts that are recorded before the first positive HIV test. This results in a dataset of 39 854 pre-treatment CD4 counts, 61 057 post-treatment CD4 counts and 36 808 pre-treatment VL measurements.

Of the patients included in this analysis, 6082 (78.1 %) had an estimated date of seroconversion based on the mid-point between last negative and first positive HIV tests, for 1454 (18.7 %) it was based on laboratory evidence of seroconversion and in 253 (3.3 %) it was based on the reported date of seroconversion illness. For those patients in whom a 'mid-point' estimate of seroconversion date was used, the median interval between tests was 308 days and the IQR was 162–548 days.

As in Chapters 5 and 6, maximum likelihood estimation was carried out using the random effects mode of the ADMB software, run on the UCL Legion High Performance Computing Facility. The Laplace approximation to the marginal log-likelihood was used for all models fitted in this chapter. The correlations between random effect terms were parameterised using the Cholesky factor to ensure that the covariance matrix for the joint distribution of $\mathbf{b}_i$ and $\mathrm{b}_{VL:i}$ would remain positive-definite during optimisation, because of this indirect parameterisation confidence intervals are not presented for the estimated correlations.

7.6 Results

The models fitted to the full CASCADE dataset (without censoring related to post-treatment VL) are summarised in Table 7.1. The base-model (Mod'_1) for this section of the analysis includes baseline CD4 and the patient-specific VL random intercept as linear predictors, with 'early' and 'late' parameters weighted as described in Section 7.2, for long-term maximum and speed of recovery of post-treatment CD4 counts; recovery follows a Janoshek–Sager curve with constant D parameter (as defined in Section 6.3). Although it should be noted that use of the Laplace approximation for the marginal log-likelihood means that assessment of AIC and BIC statistics requires caution, the addition of further patient and drug regimen characteristics in Mod'_2–Mod'_6 led to only moderate improvements in model fit, as was found in Chapter 6. Convergence of maximum likelihood estimates of model parameters was not

achieved when natural cubic spline functions were used to provide more flexible link functions between baseline CD4 and VL and the characteristics of post-treatment recovery in CD4, and so all of the results presented in this chapter follow from models in which these variables are treated as linear predictors. In the final model listed in Table 7.1, Mod_7', the D parameter relating to the Janoshek–Sager curve was also allowed to vary according to the time elapsed from seroconversion to treatment initiation, with weighting of 'early' and 'late' parameters on the natural-log scale used for optimisation. We interpret this model in order to provide a comparison with the 'fixed estimate of seroconversion' analysis in Chapter 6.

Table 7.1. Summary of fitted combined models for CD4 cell counts before and after the initiation of highly active antiretroviral therapy (HAART) in patients from the CASCADE cohort, incorporating pre-treatment viral load measurements and uncertainty in the timing of seroconversion. All models shown are nested within that described in the row below.

Model	Predictors	n_{pars}	ℓ †	AIC†	BIC†	$2\Delta\ell$†
Mod_1'	Linear-u + VL by trt-time	33	−284281	568628	568952*	NA
Mod_2'	As above + gender/inf grp	41	−284265	568612	569015	32
Mod_3'	As above + age	49	−284253	568604	569086	24
Mod_4'	As above + AIDS Dx	51	−284247	568596	569097	12
Mod_5'	As above + HCV Dx	55	−284240	568590	569131	14
Mod_6'	As above + trt regimen	61	−284219	568560	569160	42
Mod_7'	As above + D by trt-time	62	−284206	568536*	569146	26

The 'Predictors' field lists variables included in the functions to determine both long-term maximum (ϕ_1) and speed of recovery (ϕ_2). 'trt-time' denotes weighting of functions for long-term maximum and speed of recovery, in terms of baseline CD4 and VL at treatment initiation, according to time elapsed from seroconversion to treatment; the D parameter is treated as a function of time to treatment in Mod_7'. *Lowest value of AIC/BIC for set of models. †All calculations using the log-likelihood for each model are based on the Laplace approximation. '$2\Delta\ell$' denotes differences in 2×log-likelihood in comparison to model described in the row above in each case. AIC, Akaike information criterion; AIDS, acquired immune deficiency syndrome; BIC, Bayesian information criterion; Dx, diagnosis prior to HAART; HCV, hepatitis C virus; grp, group; inf, mode of infection; ℓ, log-likelihood of model fit; NA, not applicable; n_{pars}, number of parameters in model; trt, treatment.

In the models in this chapter, the characteristics of CD4 recovery on HAART follow a smooth function of the time elapsed from seroconversion to initiation of treatment with a transition from an 'early treatment' to a 'late treatment' response conditional on baseline CD4 and VL random intercept. The transition as estimated for Mod_7' is plotted in Figure 7.3, resulting from the estimate of the S parameter as defined in Section 7.2 ($\hat{S} = 6.67$, 95 % CI 5.35 – 8.32). This plot indicates that predictions for CD4 recovery on HAART will depend on the time elapsed from seroconversion up until around 4 months, but that the response conditional on baseline CD4 and VL will be stable beyond this point. The fitted functions linking baseline CD4 and the long-term maximum and speed of recovery for both 'early treatment' and 'late treatment' response are plotted in Figure 7.4, and the corresponding influence of the VL random intercept is plotted in Figure 7.5. When treatment is initiated close to the

date of seroconversion, the predicted long-term maximum CD4 count for a given baseline value is slightly lower, but the speed of recovery is substantially higher. This is further illustrated through the plotting of predicted median recovery for hypothetical patients in Figure 7.6.

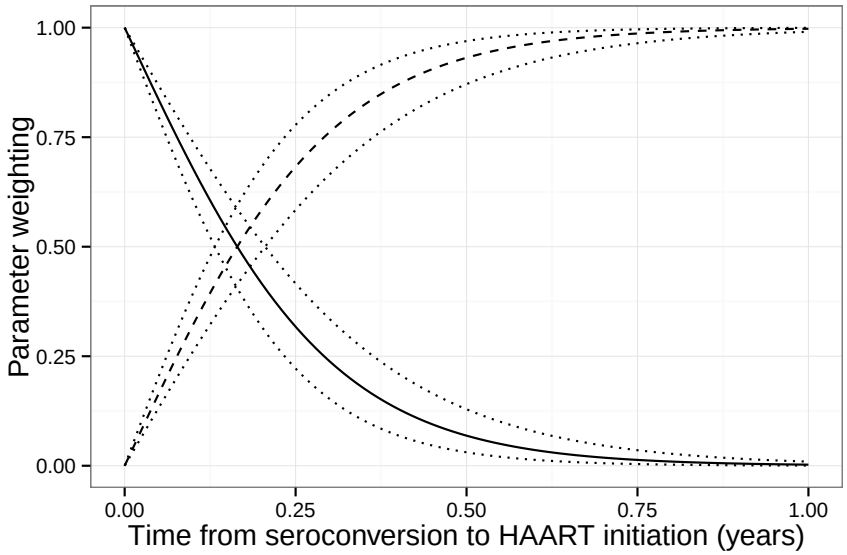

Figure 7.3. Plot of the transition from an 'early treatment' to a 'late treatment' response as estimated for Mod_7'. The weights for the 'early' (——) and 'late' (- - -) parameters linking baseline CD4 and viral load random intercept to CD4 recovery are plotted as a function of the time elapsed from 'true' date of seroconversion to initiation of treatment. 95 % confidence intervals are also plotted (·····).

The patient-specific VL random intercept was positively associated with speed of recovery regardless of the time elapsed from seroconversion to treatment initiation, although its relationship with the predicted long-term maximum did differ according to the time to treatment (Figure 7.5). This is further explored through the plotting of predicted median recovery for hypothetical patients in Figure 7.7. Higher VL, conditional on the baseline CD4 at treatment initiation, consistently predicted a better recovery in CD4 counts following treatment initiation. It should be noted that the VL random intercept term was found to be negatively correlated both with the CD4 count at seroconversion random intercept ($\hat{r} = -0.27$) and with the slope of CD4 change with respect to time from seroconversion ($\hat{r} = -0.48$), indicating that a high VL is associated with a worse prognosis without treatment. For other patient and drug combination characteristics, the predictions from the model (Figure 7.8) were very similar to those from the model in which the estimated date of seroconversion in each patient was treated as fixed (as shown in Figure 6.5).

The parameters relating to pre-treatment VL measurements fitted in Mod_7' were consistent with previous research on this topic (e.g. Pantazis et al.[127]), with a high average VL (on the $\log_{10}$ scale) close to the date of seroconversion that drops down

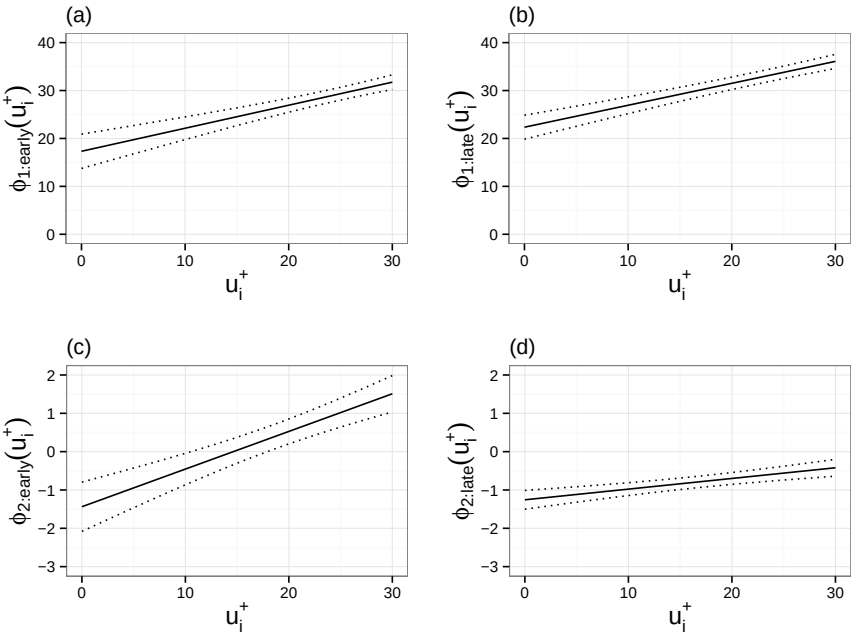

Figure 7.4. Plots of functions linking 'true' baseline CD4 (u_i^+) to post-treatment recovery, $\phi_1(u_i^+)$ (a–b, relating to long-term maximum) and $\phi_2(u_i^+)$ (c–d, relating to speed of response), for Mod_7'. Graphs on the left of each row (a,c) show the fitted functions for 'early treatment' and those on the right (b,d) show the functions 'late treatment'. Pointwise 95 % confidence intervals for the functions are shown (·····).

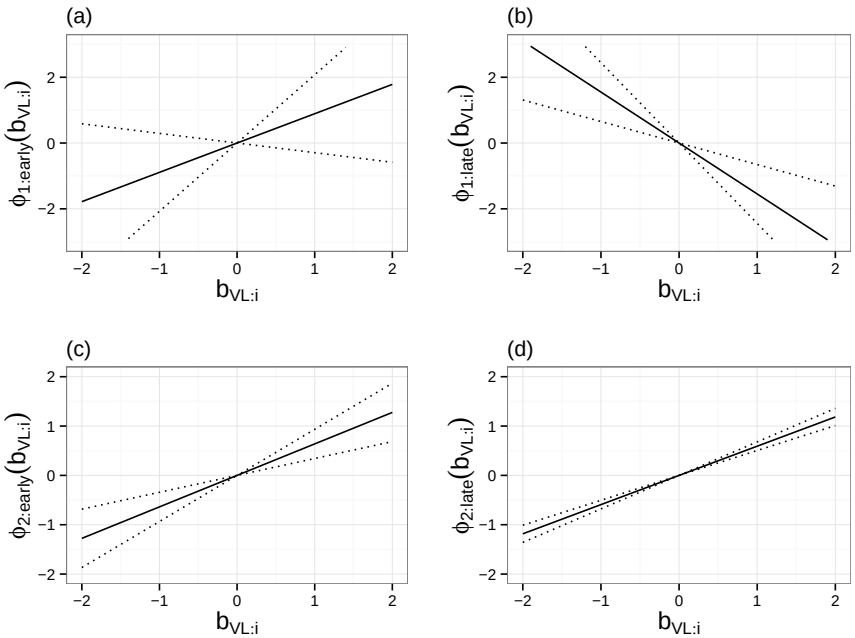

Figure 7.5. Plots of estimated effect of patient-specific viral load random intercept $b_{VL:i}$ on predicted characteristics of post-treatment recovery, $\phi_1(b_{VL:i})$ (a–b, relating to long-term maximum) and $\phi_2(b_{VL:i})$ (c–d, relating to speed of response), for Mod_7'. Graphs on the left of each row (a,c) show the fitted functions for 'early treatment' and those on the right (b,d) show the functions 'late treatment'. Pointwise 95 % confidence intervals for the functions are shown (·····).

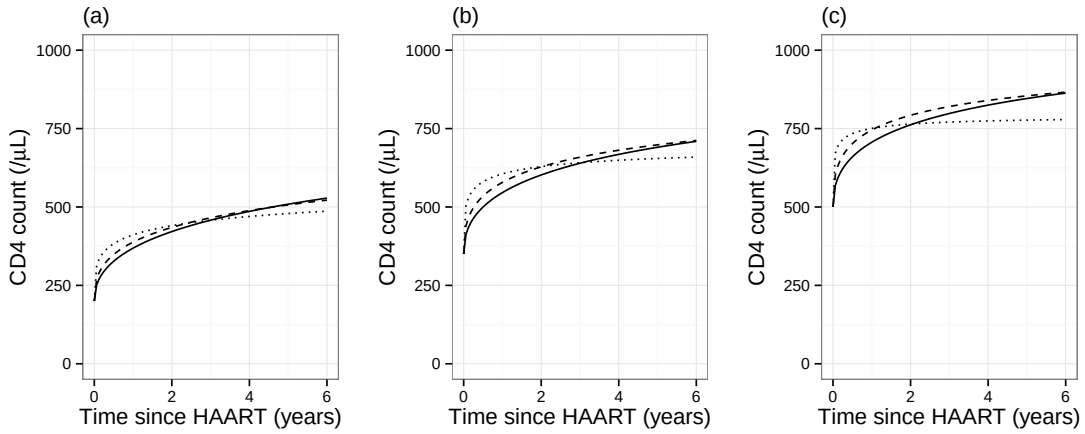

Figure 7.6. Plots of predicted median recovery in CD4 counts, based on Mod_7', for patients with a 'true' baseline value of 200 (a), 350 (b) or 500 (c) cells/μL. Predictions are shown for patients initiating treatment immediately at time of seroconversion (·····), at 3 months (- - -) and at 1 year (——). For this plot, all patients are assumed to be male homosexual, aged 36 years, with negative test for hepatitis C virus, no prior AIDS diagnosis and starting on a non-nucleoside reverse-transcriptase inhibitor (NNRTI) regimen. They are also assumed to have the population median viral load conditional on time from seroconversion.

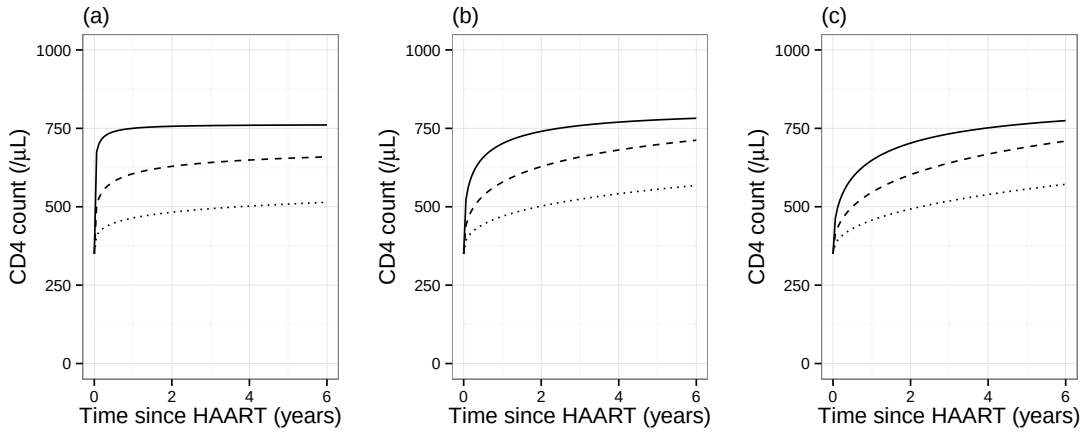

Figure 7.7. Plots of predicted median recovery in CD4 counts, based on Mod_7', for patients with a 'true' baseline value of 350 cells/μL and a patient-specific viral load random intercept (on the $\log_{10}$ scale) corresponding to the 2.5$^{\text{th}}$ centile (·····, -1.44), 50$^{\text{th}}$ centile (- - - , 0) or the 97.5$^{\text{th}}$ centile (——, 1.44). Plots are shown of predictions for patients initiating treatment immediately at time of seroconversion (a), at 3 months (b) and at 1 year (c). For this plot, all patients are assumed to be male homosexual, aged 36 years, with negative test for hepatitis C virus, no prior AIDS diagnosis and starting on a non-nucleoside reverse-transcriptase inhibitor (NNRTI) regimen.

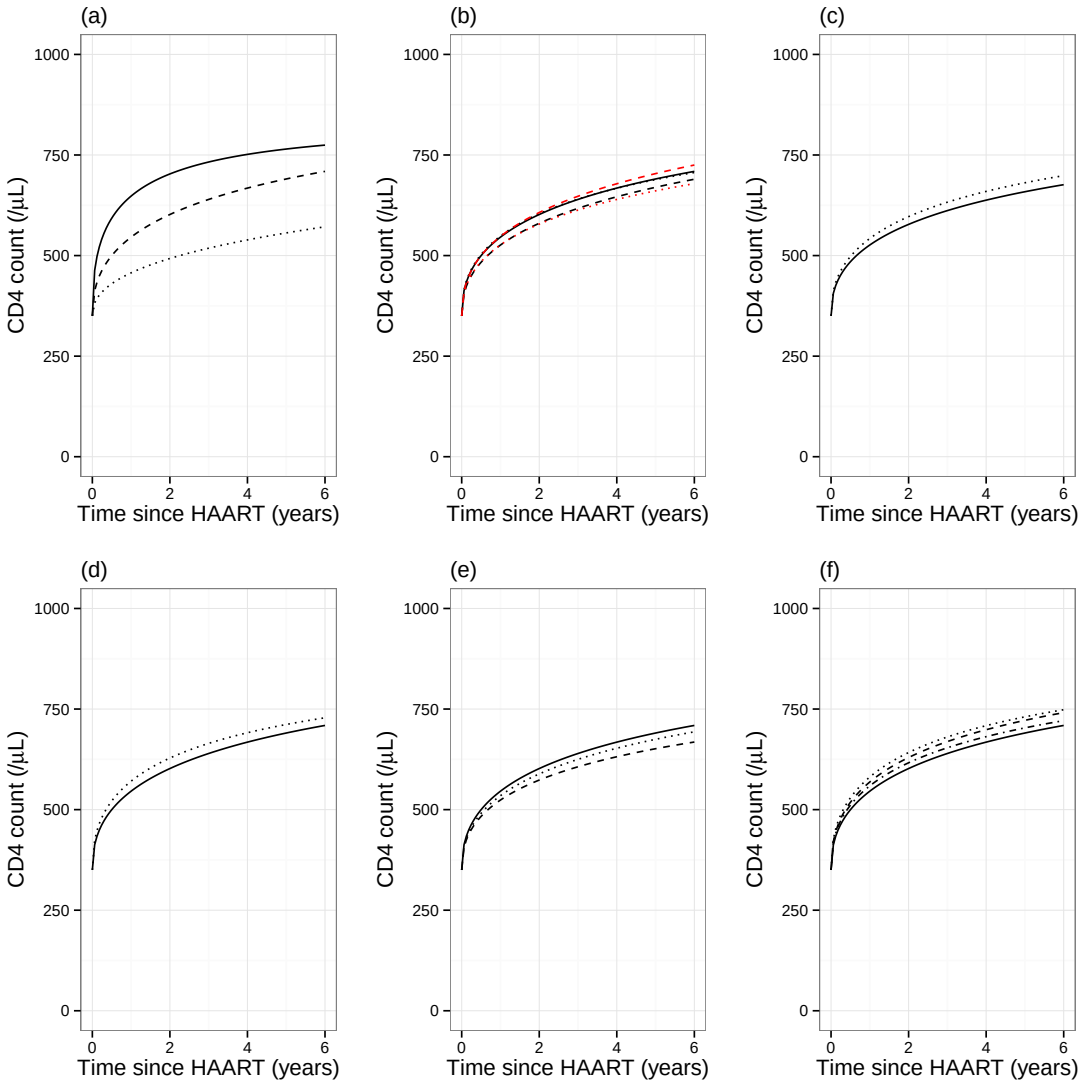

Figure 7.8. Plots of predicted median recovery in CD4 counts, based on Mod'_7, for patients with a 'true' baseline value of 350 according to: (a) patient specific viral load (VL) random intercept (·····, 2.5th centile; - - - , 50th centile; ——, 97.5th centile); (b) gender and infection groups (——, male homosexual; - - - , male heterosexual; ·····, male injecting drug user; - - - , female heterosexual; ·····, female injecting drug user); (c) patient age at treatment initiation (·····, $age = 20$ years; ——, $age = 60$ years); (d) AIDS diagnosis prior to treatment (·····, yes; ——, no); (e) hepatitis C virus (HCV) status (·····, no test; - - - , +ve test; ——, -ve test); and (f) HAART regimen (·····, integrase strand transfer inhibitor; -·-·-, ritonavir-boosted protease inhibitor; - - - , other; ——, non-nucleoside reverse-transcriptase inhibitor (NNRTI)). All patients are assumed to be male homosexual, aged 36 years, with negative test for HCV, no prior AIDS diagnosis, median VL and starting on a NNRTI regimen at 1 year since estimated date of seroconversion unless stated otherwise.

to a more stable level after 2–3 months and a gradual increase over time thereafter (Figure 7.9). The estimate of variance for the patient-specific random intercept term ($\widehat{\phi_{VL}} = 0.542$) was larger than the estimate of variance for the examination-specific measurement error ($\widehat{\sigma^2_{VL}} = 0.333$), indicating the presence of consistent differences between patients that were used to link their pre-treatment VL level to the characteristics of CD4 recovery following treatment initiation.

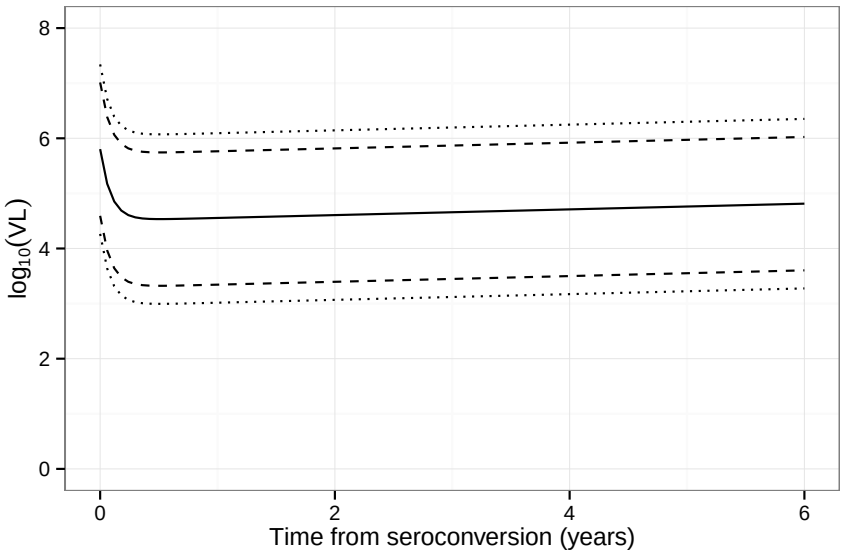

Figure 7.9. Plot of fitted mean viral load (VL, on the $\log_{10}$ scale; ——), as a function of time elapsed since seroconversion in the absence of treatment, resulting from Mod'_7. 5th and 95th percentiles for individual measurements (·····) are plotted, and the equivalent percentiles are also shown for the 90 % range of 'true' VL (- - -; i.e. including between-patient difference relating to the random intercept term in the model, but not examination-specific measurement error).

Mod'_7 was also fitted to a dataset in which CD4 counts were censored if a detectable VL was observed beyond 6 months after the start of treatment. As shown in Figure 7.10, the predictions made were very similar, although a pre-treatment AIDS diagnosis was no longer associated with any substantial relative improvement in post-treatment recovery. The parameter estimates for Mod'_7 as fitted to the full dataset and to the processed dataset with censoring at detectable VL observations are presented in Table 7.2.

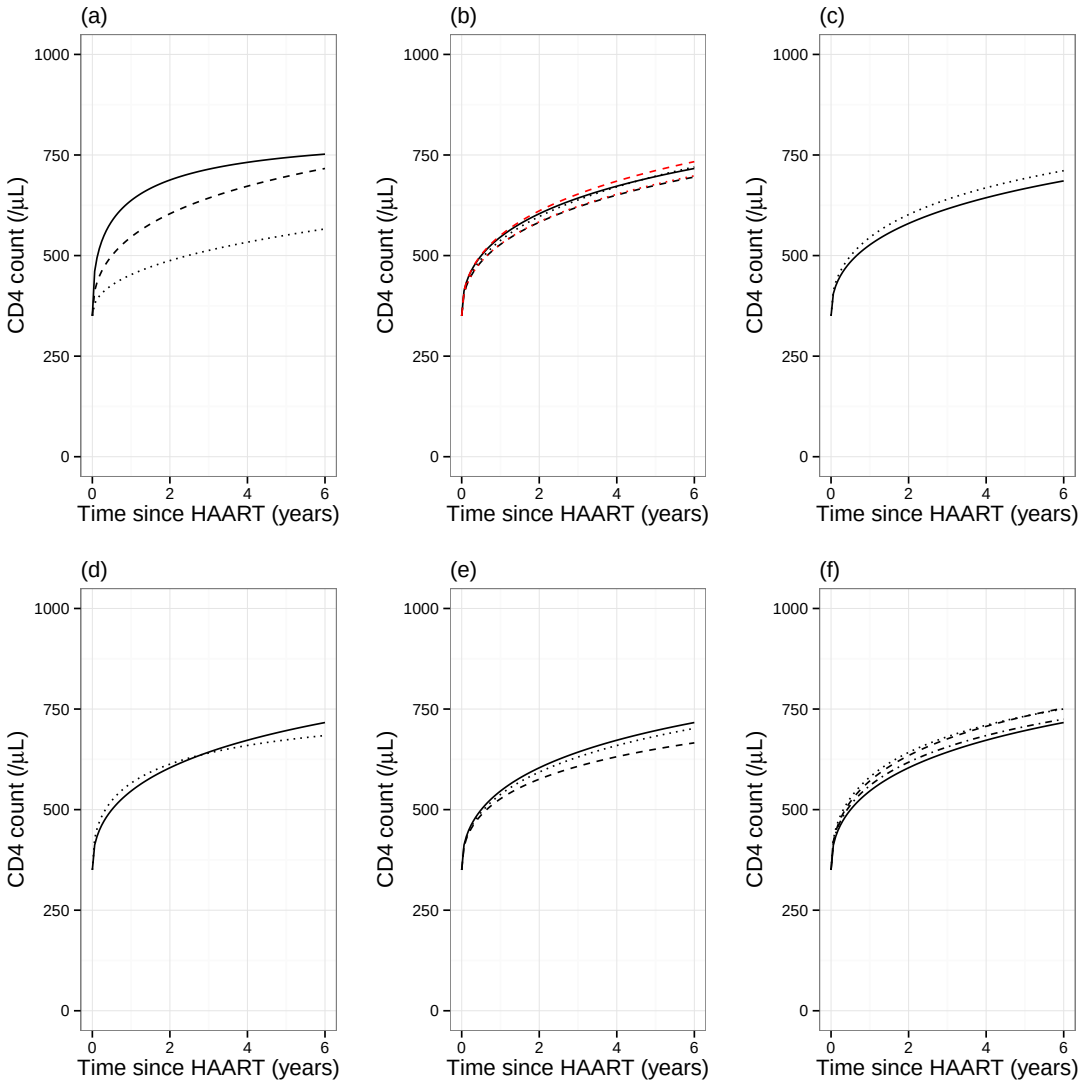

Figure 7.10. Plots of predicted median recovery in CD4 counts, based on Mod_7' fitted to the dataset with censoring at detectable viral load (VL) after 6 months of treatment, for patients with a 'true' baseline value of 350 according to: (a) patient specific viral load (VL) random intercept (·····, 2.5th centile; - - - , 50th centile; ——, 97.5th centile); (b) gender and infection groups (——, male homosexual; - - - , male heterosexual; ·····, male injecting drug user; - - - , female heterosexual; ·····, female injecting drug user); (c) patient age at treatment initiation (·····, $age = 20$ years; ——, $age = 60$ years); (d) AIDS diagnosis prior to treatment (·····, yes; ——, no); (e) hepatitis C virus (HCV) status (·····, no test; - - - , +ve test; ——, -ve test); and (f) HAART regimen (·····, integrase strand transfer inhibitor; -·-·-, ritonavir-boosted protease inhibitor; - - - , other; ——, non-nucleoside reverse-transcriptase inhibitor (NNRTI)). All patients are assumed to be male homosexual, aged 36 years, with negative test for HCV, no prior AIDS diagnosis, median VL and starting on a NNRTI regimen at 1 year since estimated date of seroconversion unless stated otherwise.

Table 7.2. Parameter estimates for Mod_7' as applied to the full CASCADE dataset, and to a processed dataset with censoring of CD4 counts at occurrence of detectable viral load (VL) beyond 6 months after treatment initiation.

Para.	Full data fit	Cens data fit	Para. (cont. 1)	Full data fit (cont. 1)	Cens data fit (cont. 1)	Para. (cont. 2)	Full data fit (cont. 2)	Cens data fit (cont. 2)
β_0	22.228 (22.104 to 22.352)	22.234 (22.11 to 22.358)	VA_{t1}	0.891 (-0.293 to 2.075)	0.121 (-1.321 to 1.564)	$AIDS_B$	0.272 (0.074 to 0.47)	0.66 (0.455 to 0.865)
β_1	-1.591 (-1.655 to -1.527)	-1.597 (-1.661 to -1.532)	VA_{t2}	-1.549 (-2.444 to -0.654)	-2.442 (-3.438 to -1.446)	HCV_pre_A	-1.107 (-2.967 to 0.752)	-2.497 (-4.205 to -0.789)
U_{00}	21.052 (20.132 to 22.014)	21.013 (20.093 to 21.975)	VB_{t1}	0.638 (0.342 to 0.934)	0.747 (0.445 to 1.049)	HCV_pre_B	-0.016 (-0.263 to 0.231)	0.155 (-0.07 to 0.379)
ρ	-0.526	-0.530	VB_{t2}	0.592 (0.505 to 0.679)	0.672 (0.583 to 0.762)	$no\_HCV\_test_A$	0.013 (-0.825 to 0.85)	-0.225 (-1.16 to 0.709)
U_{11}	0.643 (0.525 to 0.786)	0.62 (0.5 to 0.768)	MI_A	-0.656 (-3.574 to 2.262)	2.527 (-1.989 to 7.044)	$no\_HCV\_test_B$	-0.069 (-0.178 to 0.041)	-0.028 (-0.144 to 0.088)
κ_{pre}	5.788 (5.475 to 6.119)	5.684 (5.368 to 6.02)	MI_B	0.088 (-0.293 to 0.47)	-0.268 (-0.725 to 0.19)	$INSTI_A$	-0.578 (-1.753 to 0.597)	-1.176 (-2.533 to 0.181)
H_{pre}	0.299 (0.271 to 0.327)	0.316 (0.285 to 0.346)	MH_A	1.295 (-0.051 to 2.641)	1.041 (-0.519 to 2.601)	$INSTI_B$	0.262 (0.096 to 0.429)	0.317 (0.143 to 0.492)
σ	1.761 (1.726 to 1.798)	1.779 (1.743 to 1.816)	MH_B	-0.245 (-0.398 to -0.092)	-0.205 (-0.377 to -0.033)	PrI_A	-0.562 (-1.364 to 0.24)	-1.058 (-2.021 to -0.095)
$At1_1$	17.323 (13.742 to 20.904)	19.991 (15.42 to 24.562)	FI_A	-0.297 (-4.174 to 3.58)	1.39 (-3.874 to 6.653)	PrI_B	0.138 (0.039 to 0.238)	0.187 (0.079 to 0.295)
$At1_2$	0.481 (0.345 to 0.617)	0.402 (0.241 to 0.564)	FI_B	-0.088 (-0.596 to 0.42)	-0.231 (-0.831 to 0.368)	$other_A$	0.065 (-0.891 to 1.022)	0.006 (-1.101 to 1.114)
$At2_1$	22.349 (19.844 to 24.853)	22.176 (19.092 to 25.26)	FH_A	1.554 (0.165 to 2.943)	1.405 (-0.121 to 2.931)	$other_B$	0.123 (0.004 to 0.241)	0.138 (0.013 to 0.263)
$At2_2$	0.458 (0.362 to 0.553)	0.507 (0.392 to 0.623)	FH_B	-0.141 (-0.287 to 0.004)	-0.106 (-0.259 to 0.046)	S	6.673 (5.35 to 8.323)	6.565 (5.17 to 8.336)
$Bt1_1$	-1.437 (-2.08 to -0.794)	-1.693 (-2.368 to -1.018)	$ageA_1$	0.098 (-0.074 to 0.269)	0.12 (-0.079 to 0.319)	β_{0VL}	4.502 (4.48 to 4.524)	4.503 (4.481 to 4.526)
$Bt1_2$	0.098 (0.068 to 0.128)	0.099 (0.069 to 0.129)	$ageA_2$	-0.028 (-0.088 to 0.032)	-0.043 (-0.113 to 0.026)	β_{1VL}	0.052 (0.046 to 0.058)	0.052 (0.046 to 0.058)
$Bt2_1$	-1.256 (-1.502 to -1.01)	-1.264 (-1.542 to -0.986)	$ageA_3$	0.079 (-0.091 to 0.248)	0.12 (-0.075 to 0.315)	β_{2VL}	1.301 (1.233 to 1.369)	1.302 (1.234 to 1.37)
$Bt2_2$	0.028 (0.016 to 0.04)	0.024 (0.011 to 0.038)	$ageA_4$	-0.078 (-0.243 to 0.087)	-0.114 (-0.303 to 0.076)	β_{3VL}	11.072 (9.888 to 12.397)	11.114 (9.921 to 12.451)
Ω	9.829 (8.184 to 11.803)	12.62 (10.295 to 15.47)	$ageB_1$	-0.002 (-0.027 to 0.022)	-0.003 (-0.029 to 0.022)	ϕ_{VL}	0.542 (0.519 to 0.565)	0.542 (0.52 to 0.565)
κ_{post}	4.809 (4.524 to 5.112)	4.016 (3.718 to 4.338)	$ageB_2$	0 (-0.008 to 0.008)	0 (-0.008 to 0.009)	σ_{VL}	0.577 (0.572 to 0.582)	0.577 (0.572 to 0.582)
H_{post}	0.218 (0.202 to 0.234)	0.208 (0.188 to 0.228)	$ageB_3$	0.001 (-0.021 to 0.023)	-0.001 (-0.024 to 0.023)	$\rho_{VL:CD4_0}$	-0.275	-0.272
D_{t1}	0.28 (0.236 to 0.333)	0.257 (0.213 to 0.311)	$ageB_4$	-0.001 (-0.023 to 0.02)	0 (-0.023 to 0.023)	$\rho_{VL:CD4_1}$	-0.482	-0.498
D_{t2}	0.436 (0.419 to 0.454)	0.436 (0.417 to 0.456)	$AIDS_A$	-1.122 (-2.71 to 0.466)	-4.166 (-5.774 to -2.557)			

95% CIs are given in parentheses. Parameters (Para.) are defined as in Table 6.3 unless stated otherwise. $At1_1$ and $At1_2$ are the intercept and slope of the function linking baseline CD4 to long-term maximum for 'early treatment', while $At2_1$ and $At2_2$ are the equivalent parameters for 'late treatment'. $Bt1_1$ and $Bt1_2$, and $Bt2_1$ and $Bt2_2$ are the corresponding parameters for rate of recovery. D_{t1} and D_{t2} are the shape parameters for the Janoshek–Sager curve corresponding to 'early treatment' and 'late treatment', respectively. VA_{t1} and VB_{t1} represent the effect of the patient specific viral load (VL) random intercept on the long-term maximum and speed of recovery for 'early treatment', while VA_{t2} and VB_{t2} are the equivalent parameters for 'late treatment'. β_{0VL}, β_{1VL}, β_{2VL} and β_{3VL} relate to the population median VL as described in Section 7.3, and ϕ_{VL} and σ_{VL} denote the random intercept variance and VL measurement error SD. $\rho_{VL:CD4_0}$ and $\rho_{VL:CD4_1}$ are the correlations between VL random intercept and CD4 random intercept and slope, respectively.

7.7 Checks of model performance

In this section we investigate the posterior predictive modes of the latent variables relating to the timing of seroconversion for each patient in whom this is known to lie in an interval between last negative and first positive HIV tests (n = 6082). This is done using Mod_7' as fitted to the whole dataset (without censoring linked to post-treatment observations of detectable VL). As described in Section 7.4, the prior distribution of seroconversion times for such patients was assumed to follow a beta-distribution (with both shape parameters set to 6) scaled over the time interval between tests. A histogram of the posterior predictive modes for the timing of seroconversion on a standardised interval between negative and positive tests, along with the specified prior distribution, is shown in Figure 7.11. This plot shows substantial 'shrinkage' to the mean for the posterior predictive modes, with most clustered very close to 0.5, indicating that only a limited amount of information regarding the exact timing of seroconversion was conveyed by the CD4 counts and VL measurements as modelled. Indeed, for the vast majority of patients the posterior mode for the date of seroconversion was between the 40 % and 60 % marks of the intervals between tests (6058/6082 (99.6 %)).

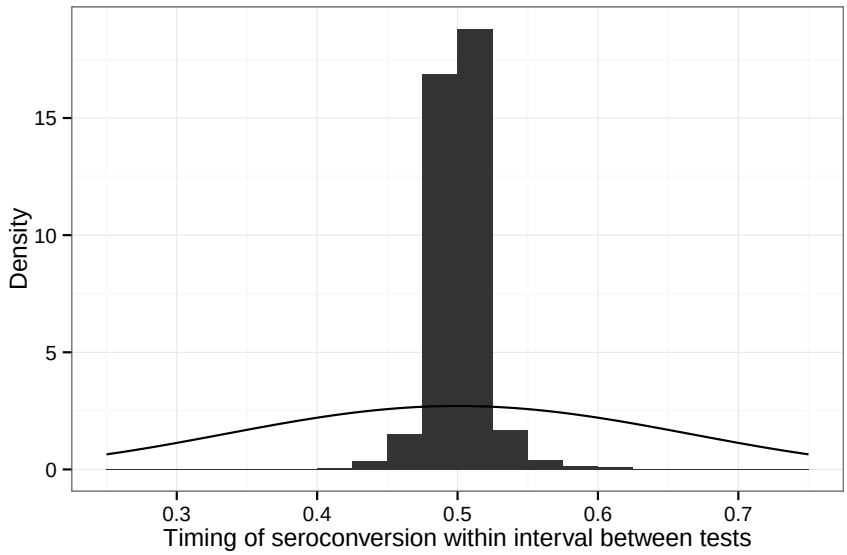

Figure 7.11. Histogram of posterior predictive modes relating to the timing of seroconversion for each patient in whom this is known to lie in an interval between last negative and first positive HIV tests, following from the fit of Mod_7' to the full dataset. In this plot the timing of seroconversion between negative and positive tests has not been scaled by the observed interval of potential dates for each patient, i.e. '0' indicates the time of last negative test, '1' is the time of first positive test and '0.5' is the mid-point between tests for all patients. The curved black line shows the probability density function for the specified prior beta distribution (with shape parameters both equal to 6).

Although the total is small, a greater number of patients had moderate evidence

that their true date of seroconversion was closer to their first positive test (n = 23 with posterior mode >60 % of the interval from negative to positive test) than had evidence that it was closer to their last negative test (n = 1 with posterior mode <40 % of the interval from negative to positive test). Plots of pre-treatment CD4 cell counts and VL measurements of the four patients with strongest evidence of a seroconversion date closer to their first positive test are shown in Figure 7.12. These plots demonstrate that, in principle, the model is functioning as intended in adjusting for any uncertainty in the exact date of seroconversion for these patients; the patients shown all have at least one high VL measurement (≥6 on the $\log_{10}$ scale) close to the date of their first positive HIV test followed by a drop in levels over the subsequent months, and the initial high values were likely because the patients was observed close to their true date of seroconversion as this provides the best match to the fitted model for pre-treatment VL measurements as shown in Figure 7.9.

The 95 % credible intervals of the approximate posterior predictive distribution of the true date of seroconversion for each of the patients shown in Figure 7.12 (and listed in the Figure caption) included the mid-point between negative and positive HIV tests, indicating that even for these patients the fitted model could not substantially narrow down the interval for their true date of seroconversion. This provides one explanation for the fact that the use of models accounting for uncertainty in the true date of seroconversion in this chapter did not lead to substantially different conclusions from the use of the fixed mid-point estimates of seroconversion date in Chapter 6. The use of a uniform prior distribution for the true date of seroconversion in each patient would have allowed greater deviations from the mid-point estimate, but was not successfully implemented within the structure developed.

7.8 Further sensitivity analyses

In order to check the finding that VL appears to be an important predictor of post-treatment CD4 recovery independent of the time elapsed from seroconversion, Mod_1' was refitted to only those patients with recorded date of seroconversion illness or lab evidence of seroconversion ($n_{patients}$ = 1707) and to a dataset also including patients with an interval between last negative and fist positive HIV tests of up to 6 months ($n_{patients}$ = 3479). The resulting parameter estimates are shown in Table 7.3 and these do not demonstrate an major discrepancies between the models fitted to the full and restricted datasets, notably higher VL was found to predict a more rapid response to treatment whether initiation was early (VB_{t1}) or late (VB_{t2}) for all of the model fits. This provides further evidence that the finding of a high predictive value of VL for post-treatment recovery in CD4 counts is not just a consequence of high VL being a marker for the acute period of infection.

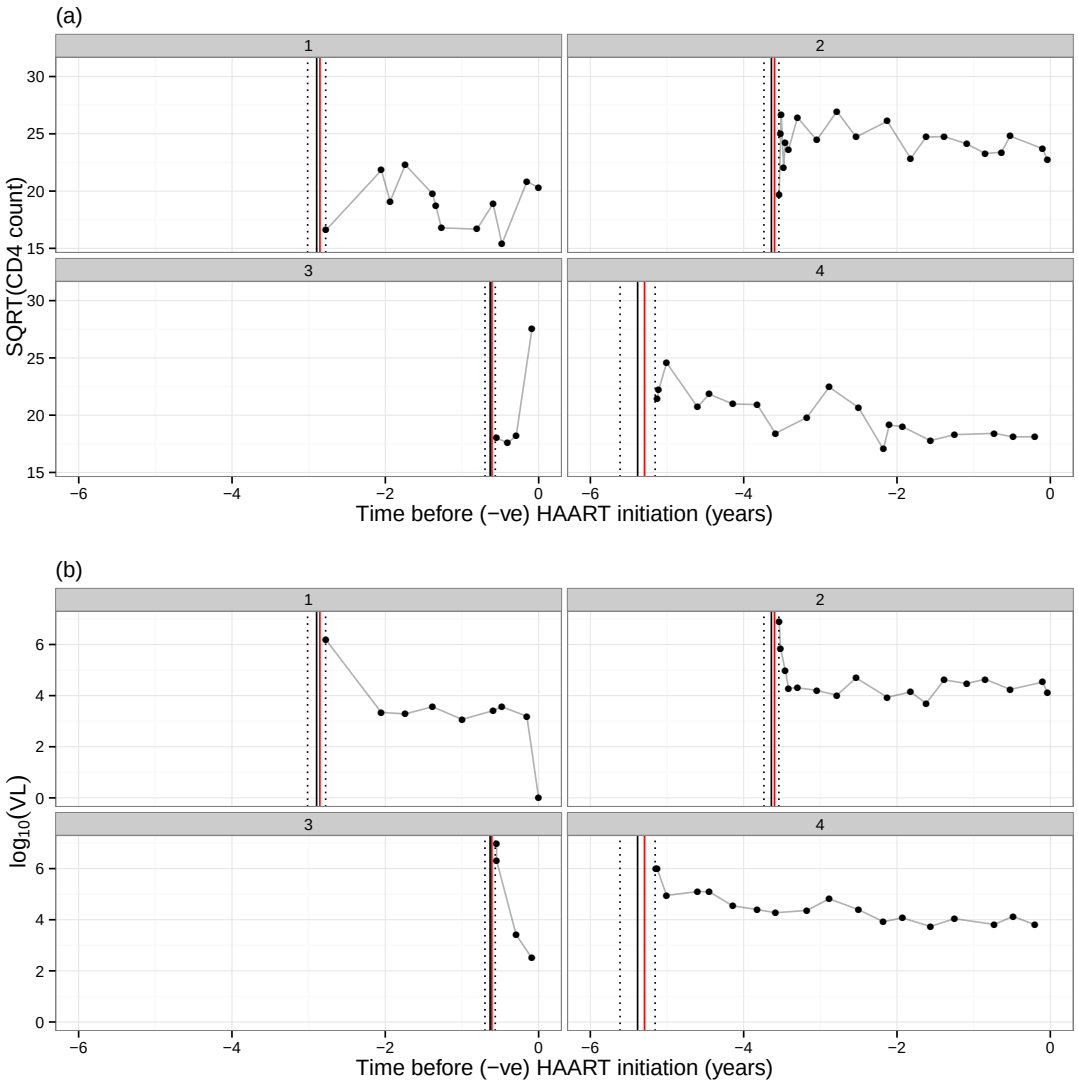

Figure 7.12. Pre-treatment (a) CD4 cell counts (on square-root scale) and (b) viral load measurements (on $\log_{10}$ scale) for the patients with strongest evidence of a seroconversion date closer to their first positive test, plotted against time prior to initiation of highly active antiretroviral therapy (HAART), following from the fit of Mod_7' to full dataset. Last negative and first positive HIV tests for each patient are shown as vertical dotted black lines, the midpoint between tests is shown as a vertical solid black line and the posterior predictive mode of time of seroconversion is shown as a vertical solid red line. The position of the predicted times of seroconversion (with approximate 95 % credible intervals) relative to last negative test (0) and first positive test (1) are: 1, 0.68 (0.38–0.90); 2, 0.71 (0.43–0.90); 3, 0.67 (0.39–0.88); 4, 0.70 (0.29–0.94).

Table 7.3. Parameter estimates for Mod_1' as applied to the full CASCADE dataset, and to versions restricted to those patients with recorded date of seroconversion (SC) illness or lab evidence of SC without or with patients with an interval between last negative and first positive HIV tests of ≤ 6 months.

Parameter	Full dataset	SC illness or lab evidence	SC illness or lab evidence, or test interval ≤ 6 months
$n_{patients}$	7789	1707	3479
β_0	22.236 (22.113 to 22.36)	22.254 (22.015 to 22.492)	22.469 (22.303 to 22.635)
β_1	-1.598 (-1.662 to -1.533)	-1.789 (-1.942 to -1.637)	-1.711 (-1.81 to -1.613)
U_{00}	20.963 (20.045 to 21.923)	21.265 (19.641 to 23.023)	19.537 (18.41 to 20.733)
ρ	-0.518	-0.390	-0.449
U_{11}	0.65 (0.532 to 0.793)	0.949 (0.676 to 1.332)	0.842 (0.662 to 1.072)
κ_{pre}	5.845 (5.532 to 6.175)	5.106 (4.463 to 5.842)	5.742 (5.281 to 6.243)
H_{pre}	0.294 (0.266 to 0.321)	0.283 (0.222 to 0.345)	0.25 (0.214 to 0.285)
σ	1.753 (1.718 to 1.79)	1.885 (1.811 to 1.961)	1.802 (1.748 to 1.858)
$At1_1$	14.981 (12.286 to 17.677)	16.937 (12.099 to 21.775)	14.415 (11.003 to 17.828)
$At1_2$	0.545 (0.432 to 0.657)	0.457 (0.265 to 0.649)	0.562 (0.416 to 0.708)
$At2_1$	21.637 (19.656 to 23.619)	24.222 (18.526 to 29.918)	22.806 (18.982 to 26.63)
$At2_2$	0.479 (0.398 to 0.561)	0.415 (0.201 to 0.629)	0.469 (0.31 to 0.628)
$Bt1_1$	-1.352 (-2.141 to -0.562)	-1.131 (-2.168 to -0.094)	-1.467 (-2.39 to -0.544)
$Bt1_2$	0.123 (0.085 to 0.16)	0.105 (0.062 to 0.148)	0.126 (0.082 to 0.17)
$Bt2_1$	-1.133 (-1.322 to -0.944)	-1.25 (-1.751 to -0.749)	-1.107 (-1.487 to -0.727)
$Bt2_2$	0.025 (0.015 to 0.036)	0.026 (0.002 to 0.049)	0.018 (-0.002 to 0.038)
Ω	9.596 (8.24 to 11.175)	12.446 (9.59 to 16.153)	9.456 (7.517 to 11.895)
κ_{post}	4.872 (4.59 to 5.172)	3.853 (3.259 to 4.555)	4.826 (4.396 to 5.298)
H_{post}	0.216 (0.2 to 0.232)	0.222 (0.179 to 0.266)	0.212 (0.188 to 0.236)
D	0.433 (0.417 to 0.449)	0.407 (0.375 to 0.442)	0.419 (0.397 to 0.443)
VA_{t1}	0.841 (-0.168 to 1.85)	0.61 (-1.369 to 2.589)	1.245 (0.011 to 2.479)
VA_{t2}	-1.768 (-2.556 to -0.981)	-3.639 (-5.856 to -1.421)	-2.215 (-3.579 to -0.85)
VB_{t1}	0.944 (0.576 to 1.313)	1.20 (0.684 to 1.716)	0.843 (0.455 to 1.23)
VB_{t2}	0.628 (0.552 to 0.704)	0.751 (0.587 to 0.915)	0.648 (0.526 to 0.769)
S	7.29 (5.954 to 8.926)	14.374 (6.678 to 30.938)	8.072 (6.418 to 10.153)
β_{0VL}	4.502 (4.48 to 4.524)	4.579 (4.534 to 4.624)	4.553 (4.52 to 4.587)
β_{1VL}	0.052 (0.046 to 0.058)	0.057 (0.044 to 0.069)	0.054 (0.045 to 0.063)
β_{2VL}	1.294 (1.227 to 1.362)	1.408 (1.328 to 1.488)	1.327 (1.257 to 1.398)
β_{3VL}	10.959 (9.783 to 12.277)	16.708 (14.913 to 18.72)	12.891 (11.617 to 14.304)
ϕ_{VL}	0.541 (0.519 to 0.564)	0.469 (0.428 to 0.515)	0.542 (0.509 to 0.578)
σ_{VL}	0.577 (0.573 to 0.582)	0.612 (0.602 to 0.623)	0.608 (0.601 to 0.615)
$\rho_{VL:CD4_0}$	-0.277	-0.376	-0.327
$\rho_{VL:CD4_1}$	-0.482	-0.411	-0.452

95% CIs are given in parentheses. Parameters are defined as in Table 7.2.

As in Chapter 6, we also attempted to fit models in which between-patient differences in variability were accounted for, using linked multivariate-t distributions for the pre- and post-treatment fractional Brownian motion components of the model as described in Section 5.7. This was carried out using the full dataset as described in Section 7.5. Convergence was not achieved when the extension was applied to the model including all patient and drug regimen characteristic, i.e. Mod_7'. However, maximum likelihood estimates were obtained when the extension was applied to the model with optimal BIC, i.e. Mod_1' including linear effects for baseline CD4 count and additional linear effects of baseline VL, varying by time since seroconversion, on the characteristics of response to treatment. A plot illustrating the associations between VL at treatment initiation, timing of treatment initiation and the characteristics of response from the resulting model is shown in Figure 7.13, showing a similar pattern to those obtained from Mod_7' (as shown in Figure 7.7).

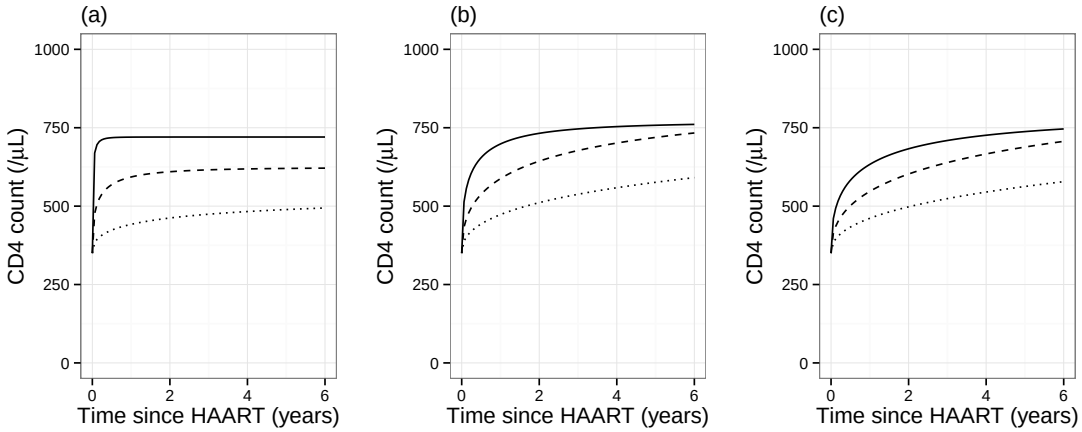

Figure 7.13. Plots of predicted median recovery in CD4 counts, based on Mod_1' with the addition of between-patient differences in variability, for patients with a 'true' baseline value of 350 cells/μL and a patient-specific viral load random intercept (on the $\log_{10}$ scale) corresponding to the 2.5$^{\text{th}}$ centile (·····, -1.44), 50$^{\text{th}}$ centile (- - -, 0) or the 97.5$^{\text{th}}$ centile (——, 1.44). Plots are shown of predictions for patients initiating treatment immediately at time of seroconversion (a), at 3 months (b) and at 1 year (c).

7.9 Discussion

In this chapter we have further developed the framework for the combined modelling of pre- and post-treatment data described in Chapters 5 and 6 in order to incorporate a second pre-treatment biomarker and also to allow for uncertainty in the pre-treatment 'zero' timepoint for each patient. The modelling strategy has been applied to the full CASCADE dataset as analysed in Chapter 6, with inclusion of a pre-treatment model for VL measurements and accounting for uncertainty in the true date of seroconversion for those patients in whom it is known to fall within

an interval between last negative and first positive tests for HIV. In this setting, the combined modelling framework allows maximum likelihood estimation to be carried out with integration over the range of possible seroconversion dates for each individual, which would be difficult to achieve using established statistical methods for analysing response to treatment. The modelling of pre-treatment biomarker with maximum likelihood estimates obtained by integration over possible dates of seroconversion was described by Sommen *et al.*[145] and Drylewicz *et al.*[146], but this technique has not previously been extended to the analysis of response to treatment.

The modelling strategy described in this chapter was developed with the aim of allowing the distinct factors that predict CD4 response to HAART to be better understood. By accounting for uncertainty in the true date of seroconversion for each patient, we hoped to be able to isolate the predictive value of pre-treatment VL level and to identify a time-interval within which early treatment initiation is associated with an improved CD4 cell recovery beyond that predicted by the baseline CD4 cell count alone, as indicated by the findings reported in Chapter 6. We have been partially successful in achieving these goals, but the high level of model complexity meant that not all desired features could be included within a single analysis of the data.

As found in the analysis reported in Chapter 6, the most important factor in predicting response to treatment was the CD4 count at initiation. Higher VL pre-treatment was found to predict more rapid recovery in CD4 counts with sustained higher values over the time-frame considered, whether or not HAART was initiated close to the date of seroconversion, in these analyses in which uncertainty in the exact timing of seroconversion was taken into account. This provides evidence that this association, which has been reported previously in the literature (e.g. Smith *et al.*[123], Florence *et al.*[132], Gras *et al.*[81]), is not an artefact resulting from the VL peak observed close to the time of seroconversion. In the present analyses, we found that initiation of HAART within around 4 months of seroconversion was associated with a more rapid initial improvement in CD4 counts, with the fitted models indicating an additional benefit (beyond that associated with higher baseline CD4 counts) over the first 2 years from the start of treatment. This is consistent with the findings reported by Le *et al.*[113]. However, the estimated benefit of early treatment (beyond that conveyed by higher baseline CD4 counts) was only moderate. The fitted models actually predict lower long-term CD4 counts beyond 2–3 years, for a given CD4 baseline value, for patients in whom HAART is initiated immediately at the time of seroconversion; this may be an artefact of the limitation in the modelling of recovery characteristics in terms of only linear functions of baseline CD4 and VL.

A previous study by Mussini *et al.*[136] analysing the CASCADE dataset found that, controlling for both baseline CD4 count and VL, steeper pre-treatment declines in CD4 count were associated with more rapid recovery once treatment was initiated.

Similarly Jarrin *et al.*[137], again using the CASCADE data, found that initial recovery at 1 month after treatment initiation was increased in 'rapid progressors' (patients with at least one CD4 count < 200 cells/μL within 12 months of seroconversion). It would have been interesting to have also tested for an independent effect of rate of pre-treatment CD4 decline in the present analysis, but this was not attempted following the failure to achieve convergence of parameter estimates for such models in the pilot investigation described in Chapter 5. The findings of Mussini *et al.*[136] and Jarrin *et al.*[137] are consistent with our findings that higher baseline VL and early treatment initiation are predictive of a more rapid recovery in the CD4 counts, as both factors are associated with a steeper pre-treatment decline. The correlation between these factors makes it difficult to separate out which is most important in predicting the characteristics of post-treatment recovery. It is also possible that 'regression to the mean' effects, as explored in Section 5.13.2, could have contributed to the findings of Mussini *et al.*[136] and Jarrin *et al.*[137].

A further limitation of the analyses that we have presented in this chapter is that we were not able to obtain maximum likelihood estimates for models that used a uniform prior distribution for the true date of seroconversion for those patients in whom this is not known, and instead implemented a beta distribution with a peak at the mid-point estimate. This contributed towards a clustering of posterior predictive modes for the true dates of seroconversion in these patients around the mid-point estimate, and thus it is not surprising that the results of the analyses in this chapter do not show any major discrepancies with those obtained using fixed mid-point estimates for the date of seroconversion as reported in Chapter 6. It was also necessary to limit the model for pre-treatment VL measurements to a relatively simple random intercept variance structure, and to limit the link between this part of the model and the model for pre-treatment CD4 counts to correlations between the random effect terms. It is possible that the use of a uniform prior distribution and/or a more fully developed model for pre-treatment VL could lead to different results to those obtained in the present analysis; achieving these goals might require a switch to a different technique for obtaining maximum likelihood estimates or to a fundamental change to a Bayesian modelling framework, which is further discussed in Chapter 8.

8 Discussion

In this thesis, we have developed several novel extensions to the statistical models available for longitudinal biomarker data. The work was been motivated by the analysis of CD4 cell counts before and after treatment in HIV patients but aspects of the modelling framework developed could be adapted for applications in other disease areas. When applied to CD4 cell count datasets, the use of models that combine stochastic process components with the multivariate-t distribution enabled novel observations regarding the patterns of between- and within-patient variability over time. The methodology developed for the combined analysis of pre- and post-treatment data produces inferences that are largely consistent with the existing literature, but provides a refined understanding of the predictive value of patient and drug regimen variables for the short- and long-term characteristics of response to treatment initiation. The combined model also provides a unified framework for simulation of patient cohorts. Furthermore the models that account for uncertainty in the exact timing of seroconversion allow estimation of the time period within which 'early treatment initiation' appears to have an additional benefit beyond that conferred by higher baseline CD4 counts, which has not previously been achieved.

The incorporation of fractional Brownian motion processes within the structure of the mixed effects model allows a mathematical description and quantification of the highly erratic nature of CD4 cell count trajectories over time within each patient. We have developed and published an R package, 'covBM'[35], that allows such models to be fitted within the familiar 'nlme'[6;34] framework, which we hope will facilitate its use in other settings. In Chapter 4, we showed that the use of overly simplistic covariance structures for longitudinal data can lead to biases in inferences when the attrition of patients from a dataset is dependent on observed values of the outcome variable for each individual; this is not a new finding in itself, but demonstrates the motivation for fitting statistical models that match the data under investigation as closely as possible.

When the multivariate-t distribution was also used for either the full pre-treatment model or for the fractional Brownian motion component of the model, substantial between-patient differences in variability were identified. Furthermore, for these models the estimated value of the H-index parameter was substantially lower, indicating a more substantial negative correlation between successive increments of the process and stronger reversion towards the underlying mean value for each patient. These are features that can be identified on visual inspection of the raw data, but which are not captured by standard mixed effects models. The use of the multivariate-t distribution was shown to provide a good description of the variability observed in the data through the use of novel diagnostic residual plots.

Interestingly, the use of the multivariate-t distribution for longitudinal data seemed to have a limited effect on inferences regarding the 'fixed effects' components of the models. In the simulations carried out in Chapter 4, with patient data generated using the full multivariate-t distribution model, estimation of the population mean slope showed no bias and appropriate confidence interval coverage when based on the equivalent model simplified to follow a marginal multivariate normal distribution. Similarly, in Chapter 5 when data were simulated from a combined pre- and post-treatment model using multivariate-t distributions for the stochastic process components, fitted models that did not incorporate the resulting between-patient differences in variability over time nonetheless provided appropriate estimation of the link between baseline values at treatment initiation and the characteristics of response to treatment. However, we note here again that the substantial between-patient differences in variability over time observed in CD4 cell counts of HIV patients represent an interesting finding in itself.

8.1 Alternative approaches and potential future research

8.1.1 Parameter estimation

For the combined analysis of pre- and post-treatment data developed in Chapters 5, 6 and 7, we found that it was not possible to obtain maximum likelihood estimates if too many features were added to a single model. For example, in Chapter 6 it was not possible to fit a model that included the full set of patient characteristics *and* the potential for between-patient differences in variability over time, whilst in Chapter 7 it was only possible to use linear functions to link the baseline CD4 count to treatment response characteristics for models in which uncertainty in the true date of seroconversion was taken into account. These problems result, at least partially, from the need to use the less accurate Laplace approximation to the marginal log-likelihood for the parameter estimation of models in which there are more than one latent variable term per patient, as described in Chapter 2. In Chapter 7 we were not able to fit models for which the prior distribution of the true date of seroconversion followed a uniform distribution, and it would be of interest to pursue this goal using different computational approaches in order to check whether this had a substantial impact on the inferences obtained from the fitted models.

An alternative approach to the maximum likelihood estimation employed throughout this thesis would be to fit equivalent models using a Bayesian framework. Such an approach can make complex models more computationally tractable, by avoiding the need to approximate the value of the marginal likelihood throughout the estimation procedure, and has the advantage that full posterior distributions can be obtained for latent variables of interest. However, the use of a Bayesian approach

requires specification of prior distributions for all model parameters, which is not straightforward for complex non-linear models, and also requires decisions to be made regarding the number of sample iterations and criteria to judge successful convergence to a posterior distribution[150]. The computational performance of Bayesian model fitting can be highly dependent on the way in which the statistical model is coded and incorporated within a sampling algorithm[151], and some of the expressions used for calculation of the marginal maximum likelihood in this thesis may also be of use within a Bayesian context.

8.1.2 Dynamic models

In this thesis, we have developed models based on the existing framework of parametric linear and non-linear mixed effects models. This approach has the advantages that the structure of the models can be clearly understood in terms of time elapsed since either infection or treatment initiation. In developing the combined models for pre- and post-treatment data we structured the models in such a way that the estimated parameters linked to response to treatment could be defined in terms of the important characteristics of speed of recovery and long-term maximum, although when the response model was generalised to a Janoshek–Sager curve interpretation was made more difficult due to the fact that maximal response was not achieved within the timeframe considered. However, an alternative approach is the use of dynamic models (sometimes also described as mechanistic models) based on systems of differential equations. When dynamic models include patient-specific random effect terms, they effectively form a subclass of non-linear mixed effects models[152].

The use of dynamic models for longitudinal data in a biomedical context was first developed for the analysis of PKPD data[153], for which this approach provided a natural framework for the estimation of drug absorption and clearance rates. Early application of such techniques in the context of HIV infection focused on estimation of viral replication rates and the lifespans of infected cells[154;155]. Subsequent research has involved more detailed investigation of both short- and long-term responses to antiretroviral treatments[14;156] using dynamic models. As well as providing estimates of parameters with direct biological interpretation, such as rates of viral replication, these models allow the influence of drug dosing, administration frequency and treatment interruptions[157] to be modelled directly; this has led some to suggest that dynamic models could be used to develop and implement personalised treatment regimes[14]. However, the dynamic approach to modelling longitudinal data does also have some disadvantages. The computational complexity for the estimation of dynamic models is greater than that for conventional non-linear mixed

effects models because, in most cases, the system of differential equations that defines the model cannot be solved exactly but rather requires numerical approximation in order to obtain the marginal likelihood (in addition to the integration over random effect terms required for all non-linear mixed effects models). There can also be problems regarding the identifiability of parameters even for relatively simple dynamic models[158]. The dynamic model defined in any situation, particularly for the complex virus–host interactions that are present in HIV infection, will always represent a substantial simplification of the biological system under investigation, and may not identify key features of the underlying biology.

We believe that the construction of 'empirical' models for longitudinal data (also sometimes termed 'descriptive' models), in which the structure is based on the patterns observed in the raw data rather than idealised mathematical relationships, can play a complementary role to the use of dynamic models. As we have done in this thesis, 'empirical' models can be structured to explore and answer key questions of interest, whilst retaining computational tractability for the dataset under investigation. There is the potential for the findings obtained using one approach to inspire further developments using the other; for example many functions derived from systems of differential equations can be used for non-linear regression modelling on the basis of empirical fit rather than an appeal to underlying mathematical relationships[130] whilst the implementation of joint models for longitudinal biomarker observations and time-to-event outcomes was first achieved for 'empirical' linear and non-linear mixed effects models before being extended to dynamic models[159].

In Chapter 7, we developed a model that included a bivariate random effects submodel for VL and CD4 cell counts measurements prior to treatment initiation based on that described by Pantazis *et al.*[127]. However, we were only able to achieve convergence of maximum likelihood estimation when the sub-model was restricted to only include a random intercept term for the VL measurements, and we did not attempt to fit a bivariate stochastic process component for CD4 cell counts and VL as has been described by Sy *et al.*[25]. One criticism of the use of 'empirical' models for bivariate (or multivariate) longitudinal data is that it is difficult to adequately capture the interactions between multiple biomarkers[14], and this limitation does indeed apply for the model presented in Chapter 7. However, with existing software and available computational resources, it may not have been possible to build a dynamic model that would allow investigation of the effects of a large number of patient characteristics on response to treatment and to fit this to the full CASCADE dataset.

8.1.3 Methods for time-dependent confounding

For the past three decades, there has been a great deal of methodological development for the estimation of causal effects in the presence of time-dependent confounding variables[160–162]. These methods allow estimation of causal effects of interventions with adjustment for variables that are both influenced by past treatment history and influence future treatment decisions. In the context of HIV research, the uses of such techniques have included estimation of the net effect of HAART on the risk of AIDS or death[163] and estimation of the effectiveness of different treatment initiation rules[164]. In the combined analyses of pre- and post-treatment data presented in this thesis, we avoided the problem of time-dependent confounding by considering only predictive factors at the time of treatment initiation for the recovery of CD4 cell counts of patients on continuous HAART. The framework that we used allowed for a rich description of the patterns of CD4 cell count trajectories over time and also allowed for uncertainty in the exact date of seroconversion to be incorporated into the model. However, the fact that techniques for dealing with time-dependent confounding were not employed meant that it was necessary to censor post-treatment observations at the occurrence of interruption of HAART, and it would not have been possible to assess the effects of changes to treatment regimes after the initiation of HAART.

Inverse probability of censoring weighting is a statistical technique related to the methods developed to adjust for time-dependent confounding. This technique was used by Lok *et al.*[110] to estimate median CD4 recovery in response to treatment stratified by baseline value. This methodology has the advantage that the median response can be estimated either on the condition that all patients were to receive continuous treatment, or for all patients averaged over the pattern of adherence observed in the studied population. It is also possible within this framework to conduct sensitivity analyses under different 'missing not at random'[100] (MNAR) scenarios, such as reduced access to treatment and increased levels of mortality in patients who are lost to follow-up[165]. In our analyses, response to HAART was estimated only for patients in care and receiving continuous treatment.

8.1.4 Joint modelling of time-to-event outcomes

A limitation of the analysis in this thesis of both pre- and post-treatment data is that the modelling of time-to-event outcomes was not considered. For the analysis of pre-treatment data alone in Chapter 4, the data for most patients were censored at treatment initiation and it is reasonable to make the MAR assumption, i.e. that the missingness of each potential observation is independent of its value conditional on the data that were observed, given that treatment initiation was largely conditional

on the observed CD4 counts. For the combined analyses of pre- and post-treatment data described in Chapters 5, 6 and 7 a similar assumption was made, but attrition from the dataset was due to treatment interruption, loss to follow-up, administrative censoring or death.

In effect the combined models presented represent response to HAART in a population of patients that remain in care and on continuous treatment and it seems reasonable to make the pragmatic assumption that in those patients who drop out of care or who have their treatment interrupted, the subsequent trajectory of CD4 counts that would have been observed had this event not occurred can be predicted by the observed data for that patient. The rate of death for patients on HAART in developed countries is low, and so although the MAR assumption is harder to interpret or justify for this outcome, the effect on the analysis is likely to be negligible. Sensitivity analyses were carried out in Chapters 6 and 7 in which patients were also censored at the observation of virological failure once on HAART. The motivation for this analysis was the assumption that most cases of virological failure result from imperfect adherence to treatment, and so the results obtained can be interpreted as representing the predicted recovery on HAART for patients with perfect adherence.

Although the MAR assumption used in relation to censoring can be defended in each case, it would be of interest to further investigate the links between longitudinal biomarker observations and the occurrence of clinical events within the framework developed. As noted in Section 8.1.3, other methods for MNAR sensitivity analyses have been developed, but the joint modelling of longitudinal biomarker data and time-to-event data would provide a more natural fit to the parametric probability models investigated in this thesis.

Much of the development of methods for the joint modelling of longitudinal biomarker data and time-to-event data has been conducted within the context of HIV research. We focus our discussion here on 'shared parameter' models[166], for which the sub-models for both the longitudinal and time-to-event outcomes are conditioned on a shared set of patient-specific latent variable terms. This class of models was introduced by De Gruttola and Tu[167] for the joint modelling of serial CD4 cell counts and death in HIV patients treated with zidovudine monotherapy (which has only limited efficacy), who found a strong association between the patient-specific slope of CD4 trajectory and survival time (those with a steeper decline had lower expected survival time). This statistical approach is now widely accessible for situations in which the longitudinal data sub-model can be expressed as a linear mixed model, with well developed packages available for both R[168] and STATA[169;170] that allow a range of options for the time-to-event sub-model.

'Shared parameter' joint models can be implemented for non-linear mixed effects models and for models that include a multivariate longitudinal outcome[127;148]

without a substantial increase in the computational complexity of maximum likelihood estimation. However, the incorporation of stochastic process components into the longitudinal data model does lead to computational challenges if realisations of the process are thought to be associated with the hazard of the event. Henderson *et al.*[171] considered a joint model in which a stationary stochastic process could influence both the longitudinal data and survival models, noting that for a fixed set of d time points for observations this would require integration over d latent variables (representing the sum of patient-specific random effects and stochastic process realisation at each time point) per patient. They addressed this problem using a combination of numerical integration through importance sampling and an EM algorithm. The authors note that the same approach could also be used for the inclusion of non-stationary stochastic processes. However, the requirement for integration of the maximum likelihood function over latent variable terms corresponding to every single observation in a dataset could lead to impossible computational requirements for a large dataset and would also likely be associated with poor computational stability in some settings. The use of a semi-parametric Cox regression model for the event outcome would also be highly problematic if the timings of events were recorded as continuous rather than only at predefined points, as the value of the stochastic process for every individual patient would have to be considered at all distinct event times in the dataset.

Wang and Taylor[172], using HIV cohort data that pre-dates the availability of effective treatment, described the implementation of a joint model for CD4 cell counts and the occurrence of AIDS incorporating both a subject-specific random intercept and either an IOU or Brownian motion process. Model fitting was conducted with a Bayesian approach using Markov chain Monte Carlo techniques to obtain an estimate of the posterior distribution for each parameter. However, although stochastic processes are defined in continuous time, in the survival model they were converted to a step process as this was necessary to obtain a tractable form for the likelihood of this section of the model. A similar approach was described by Struthers and McLeish[173].

In this thesis we have described the use of fractional Brownian motion, a non-stationary stochastic Gaussian process, as a novel component of linear or non-linear mixed effects models. It is interesting that when between-patient differences in the variance of this model component were included, the estimated H-index for the fractional Brownian motion process was substantially lower, indicating a stronger tendency of reversion towards an underlying mean value within each patient. This finding could be used as justification to fit joint models for time-to-event data that do not include realisations of the fractional Brownian motion process in the hazard function, which would greatly simplify model fitting; the argument being that the

stochastic process only represents short-to-medium term variation from the underlying long-term mean for any given patient and so might have less of an influence on the event outcome considered.

The latent variables used to enable patient-specific differences in variability over time could also themselves be included within the hazard function of a joint model for time-to-event data. For HIV patients on HAART, the rates of AIDS events and death are low, which makes it difficult to fit complex models for the risks of such events. However, the occurrence of a detectable VL for patients on HAART is a more common clinically relevant event, termed 'virological failure'. We propose that it would be interesting to investigate whether there is a link between CD4 cell count variability and the risk of virological failure for patients on HAART, as this could add to our scientific understanding of virological suppression and could also influence clinical monitoring strategies.

8.2 Publication plan

As noted earlier in the thesis, the contents of Chapter 4 form the basis for a publication in *Statistics in Medicine*[47] and the contents of Chapter 5 form the basis for a publication in *BMC Medical Research Methodology*[105], which are provided in Appendix C and Appendix D, respectively. We plan to submit a third article to an applied HIV research journal based on the analyses presented in Chapters 6 and 7; we are currently in the final stages of drafting this manuscript.

8.3 Conclusions

In the statistical analysis of longitudinal biomedical data there are competing pressures for conceptual and computational simplicity on the one hand and for adequate description of the structure and characteristics of the data under investigation on the other. Linear mixed effects models have gained widespread use because they provide a framework that reflects the dependency between observations in longitudinal datasets whilst being readily understood as an extension of linear regression. Furthermore, the marginal multivariate normal distribution provided by linear mixed effects models has formed the basis for the development of computationally efficient techniques for maximum likelihood estimation which have been implemented in all major statistical software packages. However, there are situations in which the data justify the implementation of more complex statistical models, and this has been facilitated in recent years by the development of very flexible statistical software tools such as ADMB for maximum likelihood estimation[50] and BUGS[151] and Stan[174] for Bayesian analyses.

In this thesis we have contributed to the development of statistical models that reflect the characteristics of CD4 cell count observations before and after treatment initiation in patients with HIV, allowing a richer description of within and between-patient patterns of variability over time and enabling a novel approach to the investigation of factors that predict the characteristics of response to HAART. In this setting the additional computational and conceptual complexity of the modelling framework proposed is justified by the features of the data under investigation, as it is particularly important for statistical models to reflect the structure of the data when observation schedules, timings of treatment initiation and attrition from the dataset are all highly variable between patients. We have also made some progress in addressing the problem of uncertainty in seroconversion dates for the analysis of response to treatment. Our findings relating to the recovery of CD4 cell counts in response to HAART are largely consistent with those that have been previously reported, but we provide evidence regarding the additional benefit of early initiation of HAART beyond that conveyed by the associated higher baseline CD4 count that constitutes a novel contribution to the applied literature.

Whilst the motivation for methodological developments and the applied analyses throughout the thesis have been focused on the HIV setting, we hope that the work will provide useful ideas for the development of models for biomarkers in other fields of research. Longitudinal biomedical data often has a highly complex structure, and in most cases it is not possible to incorporate all desirable features of interest within a single statistical model. However, the continuing development of statistical methodology for the analysis of longitudinal data, coupled with the availability of more flexible software, will allow researchers to more closely tailor their methods of analysis to the data and the research questions under investigation.

References

[1] Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[2] Verbeke G and Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer, 2000.

[3] Diggle PJ, Heagerty P, Liang K-Y, and Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, second edition, 2002.

[4] Lindstrom MJ and Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

[5] Pinheiro JC and Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.

[6] Pinheiro J and Bates D. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[7] Davidian M and Giltinan DM. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:387–419, 2003.

[8] Taylor JMG, Cumberland WG, and Sy JP. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, 89:727–736, 1994.

[9] Wang W-L and Fan T-H. Estimation in multivariate t linear mixed models for multiple longitudinal data. *Statistica Sinica*, 21:1857–1880, 2011.

[10] Liang KY and Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā: The Indian Journal of Statistics, Series B*, 62:134–148, 2000.

[11] Liu GF, Lu K, Mogg R, Mallick M, and Mehrotra DV. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine*, 28:2509–2530, 2009.

[12] Senn S. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25:4334–4344, 2006.

[13] Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[14] Prague M, Commenges D, and Thiébaut R. Dynamical models of biomarkers and clinical progression for personalized medicine: the HIV context. *Advanced Drug Delivery Reviews*, 65:954–965, 2013.

[15] Carroll RJ and Lin X. *Longitudinal Data Analysis*, chapter 9. Non-parametric and semi-parametric regression methods for longitudinal data, pages 199–221. Chapman and Hall/CRC, 2008. Edited by Verbeke G, Davidian M, Fitzmaurice G and Molenberghs G.

[16] CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe) Collaboration. *CASCADE: Participating cohorts*. Available at: http://www.cascade-collaboration.org [accessed 9 August 2016].

[17] Wolbers M, Babiker A, Sabin C, Young J, Dorrucci M, Chêne G, Mussini C, Porter K, Bucher HC, and CASCADE Collaboration Members. Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy–the CASCADE collaboration: a collaboration of 23 cohort studies. *PLoS Medicine*, 7:DOI: 10.1371/journal.pmed.1000239, 2010.

[18] Grimmett G and Stirzaker D. *Probability and Random Processes*, page 370. Oxford University Press, third edition, 2001.

[19] Mandelbrot B and van Ness JW. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[20] Coeurjolly JF. Simulation and identification of the fractional Brownian motion: a bibliographical and comparative study. *Journal of Statistical Software*, 7:1–53, 2000.

[21] Molz FJ, Liu HH, and Szulga J. Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. *Water Resources Research*, 33:2273–2286, 1997.

[22] Norros I. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13:953–962, 1995.

[23] Guasoni P. No arbitrage under transaction costs, with fractional Brownian motion and beyond. *Mathematical Finance*, 16:569–582, 2006.

[24] Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics*, 44:959–971, 1988.

[25] Sy JP, Taylor JMG, and Cumberland WG. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics*, 53:542–555, 1997.

[26] Taylor JM and Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine*, 17:2381–2394, 1998.

[27] Babiker AG, Emery S, Fätkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, and Vjechaj MJ; INSIGHT START Study Group. Considerations in the rationale, design and methods of the strategic timing of antiretroviral treatment (START) study. *Clinical Trials*, 10 (1 Suppl):S5–S36, 2013.

[28] Asar O and Diggle PJ. *lmenssp: Linear Mixed Effects Models with Non-Stationary Stochastic Processes*, 2015. R package version 1.1. https://cran.r-project.org/web/packages/lmenssp/index.html.

[29] Diggle PJ, Sousa I, and Asar Ö. Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, 16:522–536, 2015.

[30] Pinheiro JC and Bates DM. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.

[31] Gould W, Pitblado J, and Poi B. *Maximum Likelihood Estimation with Stata*. Stata Press, fourth edition, 2010.

[32] Dempster AP, Laird NM, and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[33] Liu C and Rubin DB. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648, 1994.

[34] Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. R package version 3.1-117. http://CRAN.R-project.org/package=nlme.

[35] Stirrup O. *covBM: Brownian Motion Processes for 'nlme'-Models*, 2015. R package version 0.1. https://cran.r-project.org/web/packages/covBM/index.html.

[36] Pinheiro JC. *Topics in Mixed Effects Models*. PhD thesis, University of Wisconsin – Madison, 1994.

[37] Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73:43–56, 1986.

[38] Goldstein H. Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76:622–623, 1989.

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

[40] Jacqmin-Gadda H, Sibillot S, Proust C, Molina J-M, and Thiébaut R. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis*, 51:5142–5154, 2007.

[41] Lange KL, Little RJA, and Taylor JMG. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84:881–896, 1989.

[42] Welsh AH and Richardson AM. Approaches to the robust estimation of mixed models. In Maddala GS and Rao CR, editors, *Handbook of Statistics (Vol. 15)*, pages 343–384. Amsterdam: Elsevier Science, 1997.

[43] Pinheiro JC, Liu C, and Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10:249–276, 2001.

[44] Kotz S and Nadarajah S. *Multivariate t-Distributions and Their Applications*. Cambridge University Press, 2004.

[45] Matos L, Prates M, and Lachos V. *tlmec: Linear Student-t Mixed-Effects Models with Censored Data*, 2012. R package version 0.0-2. https://cran.r-project.org/web/packages/tlmec/index.html.

[46] Matos LA, Prates MO, Chen M-H, and Lachos VH. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, 23:1323–1345, 2013.

[47] Stirrup OT, Babiker AG, Carpenter JR, and Copas AJ. Fractional Brownian motion and multivariate-t models for longitudinal biomedical data, with application to CD4 counts in HIV-patients. *Statistics in Medicine*, 35:1514–1532, 2016.

[48] Lin T-I. Longitudinal data analysis using t linear mixed models with autoregressive dependence structures. *Journal of Data Science*, 6:333–355, 2008.

[49] Griewank A and Walther A. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, second edition, 2008.

[50] Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson E, Maunder MN, Nielsen A, and Sibert J. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27:233–249, 2012.

[51] Sheiner LB and Beal SL. Evaluation of methods for estimating population pharmacokinetics parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8:553–571, 1980.

[52] Wang Y. Derivation of various NONMEM estimation methods. *Journal of Pharmacokinetics and Pharmacodynamics*, 34:575–593, 2007.

[53] Tierney L and Kadane JB. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

[54] Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80:791–795, 1993.

[55] Vonesh EF. A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83:447–452, 1996.

[56] MacKay DJC. *Information Theory, Inference, and Learning Algorithms*, chapter Laplace's Method. Cambridge University Press, 7.2 edition, 2005.

[57] Skaug HJ and Fournier DA. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51:699–709, 2006.

[58] Davidian M and Gallant AR. The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80:475–488, 1993.

[59] Kuhn E and Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49:1020–1038, 2005.

[60] Lavielle M and Mentré F. Estimation of Population Pharmacokinetic Parameters of Saquinavir in HIV Patients with the MONOLIX Software. *Journal of Pharmacokinetics and Pharmacodynamics*, 34:229–249, 2007.

[61] Wakefield J. The Bayesian analysis of population pharmacokinetic models. *Journal of the American Statistical Association*, 91:62–75, 1996.

[62] Song PXK, Zhang P, and Qu A. Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica*, 17:929–943, 2007.

[63] Song PXK, Fan Y, Kalbfleisch JD, Jiang J, Louis TA, Liao JG, Qaqish BF, and Ruppert D. Maximization by parts in likelihood inference. *Journal of the American Statistical Association*, 100:1145–1167, 2005.

[64] Skaug H and Fournier D. *Random Effects in AD Model Builder: ADMB-RE User Guide*, chapter Random Effects Modeling. ADMB Foundation: Honolulu, version 11.4 edition, 2015.

[65] Haslett J and Haslett SJ. The three basic types of residuals for a linear model. *International Statistical Review*, 75:1–24, 2007.

[66] DeGruttola V, Lange N, and Dafni U. Modeling the progression of HIV infection. *Journal of the American Statistical Association*, 86:569–577, 1991.

[67] Robinson GK. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6:15–32, 1991.

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

[69] Hilden-Minton JA. *Multilevel diagnostics for mixed and hierarchical linear models*. PhD thesis, University of California, Los Angeles, 1995.

[70] Nobre JS and Singer JM. Residual analysis for linear mixed models. *Biometrical Journal*, 49:863–875, 2007.

[71] Schützenmeister A and Piepho H-P. Residual analysis of linear mixed models using a simulation approach. *Computational Statistics and Data Analysis*, 56:1405–1416, 2012.

[72] Fitzmaurice G, Laird N, and Ware J. *Applied Longitudinal Analysis*, chapter Residual Analyses and Diagnostics. Hoboken, NJ: Wiley, 2004.

[73] Waternaux C, Laird NM, and Ware JH. Methods for analysis of longitudinal data: Blood-lead concentrations and cognitive development. *Journal of the American Statistical Association*, 84:33–41, 1989.

[74] Houseman EA, Ryan LM, and Coull BA. Cholesky residuals for assessing nor-

mal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, 99:383–394, 2004.

[75] Louis TA. General methods for analysing repeated measures. *Statistics in Medicine*, 7:29–45, 1988.

[76] Verbeke G, Lesaffre E, and Brant LJ. The detection of residual serial correlation in linear mixed models. *Statistics in Medicine*, 17:1391–1402, 1998.

[77] Gelman A. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13:755–779, 2004.

[78] Gelman A, Van Mechelen I, Verbeke G, Heitjan DF, and Meulders M. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61:74–85, 2005.

[79] Guihot A, Bourgarit A, Carcelain G, and Autran B. Immune reconstitution after a decade of combined antiretroviral therapies for human immunodeficiency virus. *Trends in Immunology*, 32:131–137, 2011.

[80] van der Helm JJ, Geskus R, Lodi S, Meyer L, Schuitemaker H, Gunsenheimer-Bartmeyer B, Monforte Ad, Olson A, Touloumi G, Sabin C, Porter K, and Prins M; CASCADE Collaboration in EuroCoord. Characterisation of long-term non-progression of HIV-1 infection after seroconversion: a cohort study. *Lancet HIV*, 1:e41–8, 2014.

[81] Gras L, Kesselring AM, Griffin JT, van Sighem AI, Fraser C, Ghani AC, Miedema F, Reiss P, Lange JM, and de Wolf F; ATHENA and Netherlands National Observational Cohort Study. CD4 cell counts of 800 cells/mm3 or greater after 7 years of highly active antiretroviral therapy are feasible in most patients starting with 350 cells/mm3 or greater. *Journal of Acquired Immune Deficiency Syndromes*, 45:183–192, 2007.

[82] Phillips A and Pezzotti P; CASCADE Collaboration. Short-term risk of AIDS according to current CD4 cell count and viral load in antiretroviral drug-naive individuals and those treated in the monotherapy era. *AIDS*, 18:51–58, 2004.

[83] Mellors JW, Margolick JB, Phair JP, Rinaldo CR, Detels R, Jacobson LP, and Muñoz A. Prognostic value of HIV-1 RNA, CD4 cell count, and CD4 Cell count slope for progression to AIDS and death in untreated HIV-1 infection. *JAMA*, 297:2349–2350, 2007.

[84] Egger M, May M, Chêne G, Phillips AN, Ledergerber B, Dabis F, Costagliola D, D'Arminio Monforte A, de Wolf F, Reiss P, Lundgren JD, Justice AC, Staszewski S, Leport C, Hogg RS, Sabin CA, Gill MJ, Salzberger B, and Sterne JA; ART Cohort Collaboration. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *Lancet*, 360:119–129, 2002.

[85] May M, Porter K, Sterne JA, Royston P, and Egger M. Prognostic model for HIV-1 disease progression in patients starting antiretroviral therapy was validated using independent data. *Journal of Clinical Epidemiology*, pages 1033–1041, 2005.

[86] Panel on Antiretroviral Guidelines for Adults, Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults, and adolescents. *Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents*. Department of Health and Human Services, 2016 [accessed 9 August 2016].

[87] Williams I, Churchill D, Anderson J, Boffito M, Bower M, Cairns G, Cwynarski K, Edwards S, Fidler S, Fisher M, Freedman A, Geretti AM, Gilleece Y, Horne R, Johnson M, Khoo S, Leen C, Marshall N, Nelson M, Orkin C, Paton N, Phillips A, Post F, Pozniak A, Sabin C, Trevelion R, Ustianowski A, Walsh J, Waters L, Wilkins E, Winston A, and Youle M. British HIV Association guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2012 (Updated November 2013). *HIV Medicine*, 15 Suppl 1:1–85, 2014.

[88] INSIGHT START Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. *New England Journal of Medicine*, 373:795–807, 2015.

[89] Churchill D, Waters L, Ahmed N, Angus B, Boffito M, Bower M, Dunn D, Edwards S, Emerson C, Fidler S, Fisher M, Horne R, Khoo S, Leen C, Mackie N, Marshall N, Monteiro F, Nelson M, Orkin C, Palfreeman A, Pett S, Phillips A, Post F, Pozniak A, Reeves I, Sabin C, Trevelion R, Walsh J, Wilkins E, Williams I, and Winston A. *BHIVA guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2015*. British HIV Association (BHIVA): London, 2015.

[90] Cohen MS, Shaw GM, McMichael AJ, and Haynes BF. Acute HIV-1 Infection. *New England Journal of Medicine*, 364:1943–1954, 2011.

[91] Busch MP, Lee LL, Satten GA, Henrard DR, Farzadegan H, Nelson KE, Read S, Dodd RY, and Petersen LR. Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors. *Transfusion*, 35:91–97, 1995.

[92] Tyrer F, Walker AS, Gillett J, and Porter K; UK Register of HIV Seroconverters. The relationship between HIV seroconversion illness, HIV test interval and time to AIDS in a seroconverter cohort. *Epidemiology & Infection*, 131:1117–1123, 2003.

[93] Wang W-L and Fan T-H. Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers. *Journal of Multivariate*

Analysis, 105:300–310, 2012.

[94] Bolker B, Skaug H, and Laake J. *R2admb: ADMB to R interface functions*, 2013. http://CRAN.R-project.org/package=R2admb.

[95] Cavanaugh JE and Neath AA. Generalizing the derivation of the schwarz information criterion. *Communications in Statistics - Theory and Methods*, 28:49–66, 1999.

[96] Abdool Karim SS. Overcoming impediments to global implementation of early antiretroviral therapy. *New England Journal of Medicine*, 373:875–876, 2015.

[97] Lodi S, Phillips A, Touloumi G, Geskus R, Meyer L, Thiébaut R, Pantazis N, Amo JD, Johnson AM, Babiker A, and Porter K; CASCADE Collaboration in EuroCoord. Time from human immunodeficiency virus seroconversion to reaching CD4+ cell count thresholds <200, <350, and <500 cells/mm^3: assessment of need following changes in treatment guidelines. *Clinical Infectious Diseases*, 53:817–825, 2011.

[98] World Health Organization. *Guideline on When to Start Antiretroviral Therapy and on Pre-Exposure Prophylaxis for HIV*, September 2015.

[99] Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[100] Rubin DB. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[101] Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, and Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, 58:621–630, 2002.

[102] Thiébaut R and Walker S. When it is better to estimate a slope with only one point. *QJM*, 101:821–824, 2008.

[103] Seaman S, Galat J, Jackson D, and Carlin J. What is meant by "missing at random"? *Statistical Science*, 28:257–268, 2013.

[104] UK Register of HIV Seroconverters Steering Committee. The AIDS incubation period in the UK estimated from a national register of HIV seroconverters. *AIDS*, 12:659–667, 1998.

[105] Stirrup OT, Babiker AG, and Copas AJ. Combined models for pre- and post-treatment longitudinal biomarker data: an application to CD4 counts in HIV-patients. *BMC Medical Research Methodology*, 16:121, 2016.

[106] Kenward MG, White IR, and Carpenter JR. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? by G. F. Liu, K. Lu, R. Mogg, M. Mallick and D. V. Mehrotra, Statistics in Medicine 2009; 28:2509-2530. *Statistics in Medicine*, 29:1455–1456, 2010.

[107] Tango T. On the repeated measures designs and sample sizes for randomized controlled trials. *Biostatistics*, 17:334–349, 2016.

[108] Kaufmann GR, Perrin L, Pantaleo G, Opravil M, Furrer H, Telenti A, Hirschel B, Ledergerber B, Vernazza P, Bernasconi E, Rickenbach M, Egger M, Battegay M, and Swiss HIV Cohort Study Group. CD4 T-lymphocyte recovery in individuals with advanced HIV-1 infection receiving potent antiretroviral therapy for 4 years: the Swiss HIV Cohort Study. *Archives of Internal Medicine*, 163:2187–2195, 2003.

[109] Moore RD and Keruly JC. CD4+ cell count 6 years after commencement of highly active antiretroviral therapy in persons with sustained virologic suppression. *Clinical Infectious Diseases*, 44:441–446, 2007.

[110] Lok JJ, Bosch RJ, Benson CA, Collier AC, Robbins GK, Shafer RW, Hughes MD, and ALLRT team. Long-term increase in CD4+ T-cell counts during combination antiretroviral therapy for HIV-1 infection. *AIDS*, 24:1867–1876, 2010.

[111] Hughes RA, Sterne JA, Walsh J, Bansi L, Gilson R, Orkin C, Hill T, Ainsworth J, Anderson J, Gompels M, Dunn D, Johnson MA, Phillips AN, Pillay D, Leen C, Easterbrook P, Gazzard B, Fisher M, and Sabin CA. Long-term trends in CD4 cell counts and impact of viral failure in individuals starting antiretroviral therapy: UK Collaborative HIV Cohort (CHIC) study. *HIV Medicine*, 12:583–593, 2011.

[112] Chambless LE and Davis V. Analysis of associations with change in a multivariate outcome variable when baseline is subject to measurement error. *Statistics in Medicine*, 22:1041–1067, 2003.

[113] Le T, Wright EJ, Smith DM, He W, Catano G, Okulicz JF, Young JA, Clark RA, Richman DD, Little SJ, and Ahuja SK. Enhanced CD4+ T-cell recovery with earlier HIV-1 antiretroviral therapy. *New England Journal of Medicine*, 368:218–230, 2013.

[114] Ding Y, Duan S, Wu Z, Ye R, Yang Y, Yao S, Wang J, Xiang L, Jiang Y, Lu L, Jia M, Detels R, and He N. Timing of antiretroviral therapy initiation after diagnosis of recent human immunodeficiency virus infection and CD4(+) T-cell recovery. *Clinical Microbiology and Infection*, 22:290.e5–8, 2016.

[115] Gazzola L, Tincati C, Bellistrì GM, Monforte AD, and Marchetti G. The absence of CD4+ T cell count recovery despite receipt of virologically suppressive highly active antiretroviral therapy: clinical risk, immunological gaps, and therapeutic options. *Clinical Infectious Diseases*, 48:328–337, 2009.

[116] Lewis J, Walker AS, Castro H, De Rossi A, Gibb DM, Giaquinto C, Klein N, and Callard R. Age and CD4 count at initiation of antiretroviral therapy in HIV-infected children: effects on long-term T-cell reconstitution. *The Journal of Infectious Diseases*, 205:548–556, 2012.

[117] Picat MQ, Lewis J, Musiime V, Prendergast A, Nathoo K, Kekitiinwa A, Nahirya Ntege P, Gibb DM, Thiebaut R, Walker AS, Klein N, Callard R, and

ARROW Trial Team. Predicting patterns of long-term CD4 reconstitution in HIV-infected children starting antiretroviral therapy in sub-Saharan Africa: a cohort-based modelling study. *PLoS Medicine*, 10:e1001542, 2013.

[118] Hastie T, Tibshirani R, and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter Basis Expansions and Regularization, pages 144–146. Springer, second edition, 2009.

[119] Balakrishnan N and Lai CD. *Continuous Bivariate Distributions*, chapter Bivariate Gamma and Related Distributions. Springer, second edition, 2009.

[120] Moran PAP. Statistical inference with bivariate gamma distributions. *Biometrika*, 56:627–634, 1969.

[121] Hernán MA and Taubman SL. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32:S8–S18, 2008.

[122] Wickham H. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[123] Smith CJ, Sabin CA, Youle MS, Kinloch de Loes S, Lampe FC, Madge S, Cropley I, Johnson MA, and Phillips AN. Factors influencing increases in CD4 cell counts of HIV-positive persons receiving long-term highly active antiretroviral therapy. *The Journal of Infectious Diseases*, 190:1860–1868, 2004.

[124] Geng EH, Neilands TB, Thiébaut R, Bwana MB, Nash D, Moore RD, Wood R, Zannou DM, Althoff KN, Lim PL, Nachega JB, Easterbrook PJ, Kambugu A, Little F, Nakigozi G, Nakanjako D, Kiggundu V, Ki Li PC, Bangsberg DR, Fox MP, Prozesky H, Hunt PW, Davies MA, Reynolds SJ, Egger M, Yiannoutsos CT, Vittinghoff EV, Deeks SG, and Martin JN. CD4+ T cell recovery during suppression of HIV replication: an international comparison of the immunological efficacy of antiretroviral therapy in North America, Asia and Africa. *International Journal of Epidemiology*, 44:251–263, 2015.

[125] Harrison L, Dunn DT, Green H, and Copas AJ. Modelling the association between patient characteristics and the change over time in a disease measure using observational cohort data. *Statistics in Medicine*, 28:3260–3275, 2009.

[126] Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

[127] Pantazis N, Touloumi G, Walker AS, and Babiker AG. Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:405–423, 2005.

[128] Janoschek A. Das reaktionshinetische grundgesetz und seire beziehungen zum wachstumsund ertragsgesetz. *Statistische Vierteljahresschrift*, 10:25–37, 1957.

[129] Sager G. Seasonally modified forms of the revised Janoschek growth function. *Gegenbaurs morphologisches Jahrbuch, Leipzig*, 130:659–669, 1984.

[130] Panik MJ. *Growth Curve Modeling: Theory and Applications*, chapter Parametric Growth Curve Modeling. Hoboken, NJ: Wiley, 2014.

[131] Hunt PW, Deeks SG, Rodriguez B, Valdez H, Shade SB, Abrams DI, Kitahata MM, Krone M, Neilands TB, Brand RJ, Lederman MM, and Martin JN. Continued CD4 cell count increases in HIV-infected adults experiencing 4 years of viral suppression on antiretroviral therapy. *AIDS*, 17:1907–1915, 2003.

[132] Florence E, Lundgren J, Dreezen C, Fisher M, Kirk O, Blaxhult A, Panos G, Katlama C, Vella S, and Phillips A; EuroSIDA Study Group. Factors associated with a reduced CD4 lymphocyte count response to HAART despite full viral suppression in the EuroSIDA study. *HIV Medicine*, 4:255–262, 2003.

[133] Moore DM, Harris R, Lima V, Hogg B, May M, Yip B, Justice A, Mocroft A, Reiss P, Lampe F, Chêne G, Costagliola D, Elzi L, Mugavero MJ, Monforte AD, Sabin C, Podzamczer D, Fätkenheuer G, Staszewski S, Gill J, and Sterne JA; Antiretroviral Therapy Cohort Collaboration. Effect of baseline CD4 cell counts on the clinical significance of short-term immunologic response to antiretroviral therapy in individuals with virologic suppression. *Journal of Acquired Immune Deficiency Syndromes*, 52:357–363, 2009.

[134] Bucy RP, Hockett RD, Derdeyn CA, Saag MS, Squires K, Sillers M, Mitsuyasu RT, and Kilby JM. Initial increase in blood CD4(+) lymphocytes after HIV antiretroviral therapy reflects redistribution from lymphoid tissues. *The Journal of Clinical Investigation*, 103:1391–1398, 1999.

[135] Diaz M, Douek DC, Valdez H, Hill BJ, Peterson D, Sanne I, Piliero PJ, Koup RA, Green SB, Schnittman S, and Lederman MM. T cells containing T cell receptor excision circles are inversely related to HIV replication and are selectively and rapidly released into circulation with antiretroviral treatment. *AIDS*, 17:1145–1149, 2003.

[136] Mussini C, Cossarizza A, Sabin C, Babiker A, De Luca A, Bucher HC, Fisher M, Rezza G, Porter K, and Dorrucci M; CASCADE Collaboration. Decline of CD4+ T-cell count before start of therapy and immunological response to treatment in antiretroviral-naive individuals. *AIDS*, 25:1041–1049, 2011.

[137] Jarrin I, Pantazis N, Dalmau J, Phillips AN, Olson A, Mussini C, Boufassa F, Costagliola D, Porter K, Blanco J, Del Amo J, and Martinez-Picado J; for CASCADE Collaboration in EuroCoord. Does rapid HIV disease progression prior to combination antiretroviral therapy hinder optimal CD4+ T-cell recovery once HIV-1 suppression is achieved? *AIDS*, 29:2323–2333, 2015.

[138] Greub G, Ledergerber B, Battegay M, Grob P, Perrin L, Furrer H, Burgisser P, Erb P, Boggian K, Piffaretti JC, Hirschel B, Janin P, Francioli P, Flepp M, and Telenti

A. Clinical progression, survival, and immune recovery during antiretroviral therapy in patients with HIV-1 and hepatitis C virus coinfection: the Swiss HIV Cohort Study. *Lancet*, 356:1800–1805, 2000.

[139] Tsiara CG, Nikolopoulos GK, Dimou NL, Bagos PG, Saroglou G, Velonakis E, and Hatzakis A. Effect of hepatitis C virus on immunological and virological responses in HIV-infected patients initiating highly active antiretroviral therapy: a meta-analysis. *Journal of Viral Hepatitis*, 20:715–724, 2013.

[140] Clotet B, Feinberg J, van Lunzen J, Khuong-Josses MA, Antinori A, Dumitru I, Pokrovskiy V, Fehr J, Ortiz R, Saag M, Harris J, Brennan C, Fujiwara T, and Min S; ING114915 Study Team. Once-daily dolutegravir versus darunavir plus ritonavir in antiretroviral-naive adults with HIV-1 infection (FLAMINGO): 48 week results from the randomised open-label phase 3b study. *Lancet*, 383:2222–2231, 2014.

[141] Raffi F, Babiker AG, Richert L, Molina JM, George EC, Antinori A, Arribas JR, Grarup J, Hudson F, Schwimmer C, Saillard J, Wallet C, Jansson PO, Allavena C, Van Leeuwen R, Delfraissy JF, Vella S, Chêne G, and Pozniak A; NEAT001/ANRS143 Study Group. Ritonavir-boosted darunavir combined with raltegravir or tenofovir-emtricitabine in antiretroviral-naive adults infected with HIV-1: 96 week results from the NEAT001/ANRS143 randomised non-inferiority trial. *Lancet*, 384:1942–1951, 2014.

[142] Muñoz A, Carey V, Taylor JM, Chmiel JS, Kingsley L, Van Raden M, and Hoover DR. Estimation of time since exposure for a prevalent cohort. *Statistics in Medicine*, 11:939–952, 1992.

[143] Geskus RB. On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion. *Statistics in Medicine*, 19:1753–1769, 2000.

[144] Taffé P, May M, and Swiss HIV Cohort Study. A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. *Statistics in Medicine*, 27:4835–4853, 2008.

[145] Sommen C, Commenges D, Vu SL, Meyer L, and Alioum A. Estimation of the distribution of infection times using longitudinal serological markers of HIV: implications for the estimation of HIV incidence. *Biometrics*, 67:467–475, 2011.

[146] Drylewicz J, Guedj J, Commenges D, and Thiébaut R. Modeling the dynamics of biomarkers during primary HIV infection taking into account the uncertainty of infection date. *The Annals of Applied Statistics*, 4:1847–1870, 2010.

[147] Thiébaut R and Jacqmin-Gadda H. Mixed models for longitudinal left-censored repeated measures. *Computer Methods and Programs in Biomedicine*, 74:255–260, 2004.

[148] Thiébaut R, Jacqmin-Gadda H, Babiker A, Commenges D, and CASCADE Collaboration. Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Statistics in Medicine*, 24:65–82, 2005.

[149] Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

[150] Bolker BM, Gardner B, Maunder M, Berg CW, Brooks M, Comita L, Crone E, Cubaynes S, Davies T, de Valpine P, Ford J, Gimenez O, Kéry M, Kim EJ, Lennert-Cody C, Magnusson A, Martell S, Nash J, Nielsen A, Regetz J, Skaug H, and Zipkin E. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, 4:501–512, 2013.

[151] Lunn D, Spiegelhalter D, Thomas A, and Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28:3049–3067, 2009.

[152] Davidian M. *Longitudinal Data Analysis*, chapter 5. Non-linear mixed-effects models, pages 107–141. Chapman and Hall/CRC, 2008. Edited by Verbeke G, Davidian M, Fitzmaurice G and Molenberghs G.

[153] Derendorf H and Meibohm B. Modeling of Pharmacokinetic/Pharmacodynamic (PK/PD) Relationships: Concepts and Perspectives. *Pharmaceutical Research*, 16:176–185, 1999.

[154] Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, and Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373:123–126, 1995.

[155] Perelson AS, Neumann AU, Markowitz M, Leonard JM, and Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271:1582–1586, 1996.

[156] Xiao Y, Miao H, Tang S, and Wu H. Modeling antiretroviral drug responses for HIV-1 infected patients using differential equation models. *Advanced Drug Delivery Reviews*, 65:940–953, 2013.

[157] Adams BM, Banks HT, Davidian M, and Rosenberg ES. Estimation and prediction with HIV-treatment interruption data. *Bulletin of Mathematical Biology*, 69:563–584, 2007.

[158] Guedj J, Thiébaut R, and Commenges D. Practical identifiability of HIV dynamics models. *Bulletin of Mathematical Biology*, 69:2493–2513, 2007.

[159] Guedj J, Thiébaut R, and Commenges D. Joint modeling of the clinical progression and of the biomarkers' dynamics using a mechanistic model. *Biometrics*, 67:59–66, 2011.

[160] Robins J. A new approach to causal inference in mortality studies with a sus-

tained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:9–12, 1986.

[161] Daniel RM, Cousens SN, De Stavola BL, Kenward MG, and Sterne JA. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32:1584–1618, 2013.

[162] Robins JM and Hernán MA. *Longitudinal Data Analysis*, chapter 23. Estimation of the causal effects of time-varying exposures, pages 553–599. Chapman and Hall/CRC, 2008. Edited by Verbeke G, Davidian M, Fitzmaurice G and Molenberghs G.

[163] Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel J, Detels R, Ervin C, Feldman J, Greenblatt R, Kingsley L, Lai S, Young M, Cohen M, and Muñoz A. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158:687–694, 2003.

[164] Cain LE, Logan R, Robins JM, Sterne JA, Sabin C, Bansi L, Justice A, Goulet J, van Sighem A, de Wolf F, Bucher HC, von Wyl V, Esteve A, Casabona J, del Amo J, Moreno S, Seng R, Meyer L, Perez-Hoyos S, Muga R, Lodi S, Lanoy E, Costagliola D, and Hernan MA; HIV-CAUSAL Collaboration. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Annals of Internal Medicine*, 154:509–515, 2011.

[165] Kiragga AN, Lok JJ, Musick BS, Bosch RJ, Mwangi A, Wools-Kaloustian KK, Yiannoutsos CT, and East Africa IeDEA Regional Consortium. CD4 trajectory adjusting for dropout among HIV-positive patients receiving combination antiretroviral therapy in an East African HIV care centre. *Journal of the International AIDS Society*, 17:18957, 2014.

[166] Verbeke G and Davidian M. *Longitudinal Data Analysis*, chapter 13. Joint Models for Longitudinal Data: Introduction and Overview, pages 319–326. Chapman and Hall/CRC, 2008. Edited by Verbeke G, Davidian M, Fitzmaurice G and Molenberghs G.

[167] De Gruttola V and Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

[168] Rizopoulos D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software*, 35, 2010.

[169] Pantazis N and Touloumi G. Analyzing longitudinal data in the presence of informative drop-out: The jmre1 command. *The Stata Journal*, 10:226–251, 2010.

[170] Crowther MJ, Abrams KR, and Lambert PC. Joint modeling of longitudinal and

survival data. *The Stata Journal*, 13:165–184, 2013.

[171] Henderson R, Diggle P, and Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000.

[172] Wang Y and Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96:895–905, 2001.

[173] Struthers CA and McLeish DL. A particular diffusion model for incomplete longitudinal data: application to the multicenter AIDS cohort study. *Biostatistics*, 12:493–505, 2011.

[174] Carpenter B, Lee D, Brubaker MA, Riddell A, Gelman A, Goodrich B, Guo J, Hoffman M, Betancourt M, and Li P. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, (in press), 2016.

Appendix A covBM R package vignette

covBM: incorporating Brownian motion components into 'nlme' models

Oliver Stirrup
MRC Clinical Trials Unit at UCL, University College London, London, UK

1 Introduction

Longitudinal data are now widely analysed using linear mixed models, with 'random slopes' models particularly common. These models can successfully account for the dependency that arises when repeated observations are made over time on each individual in a dataset, but make strong assumptions regarding the nature of this dependency. In the context of modelling CD4 cell counts over time in human immunodeficiency virus (HIV)-positive patients, it has been shown that the incorporation of non-stationary stochastic processes such as Brownian motion or integrated Ornstein–Uhlenbeck (IOU) processes into the modelling framework can lead to a very substantial improvement in model fit [1;2]. Recently, the use of a fractional Brownian motion component has been shown to provide a further improvement [3]. However, these extensions to the standard linear mixed model have not been widely used in practice, and are not readily implemented in current statistical software programs. The presence of such a component in a model for longitudinal data implies that the progress of the state of the underlying biological system for each individual does not follow a deterministic relationship with time, but rather follows an unpredictable stochastic path.

The `nlme` package [4] for R allows the user to fit a wide range of linear and non-linear mixed effects models, with in-depth documentation and a wealth of examples provided in the accompanying book by Pinheiro and Bates [5]. As well as incorporating within-subject dependence resulting from the inclusion of 'random effects' in a specified model, `nlme` also allows for a correlation structure to be imposed on the residual error terms (with estimation of any associated parameters) and for the residual error variance to be modelled as a function of variables in the data under consideration. It is even possible for the user to create their own correlation structures or variance functions for inclusion in the estimation of models in `nlme`.

It is possible to implement user-defined correlation structures in `nlme` to obtain point estimates of the parameters in linear and non-linear mixed effects models incorporating Brownian motion or IOU processes. However, some further additions to the original `nlme` code are required to obtain confidence intervals for the natural model parameters and to return a fitted model object that reports the natural parameters upon use of `print` or `summary`. The `covBM` package provides wrappers for the standard `nlme` functions in order to achieve these goals.

In Section 2, the characteristics of the statistical models under consideration are specified, and in Section 3, examples are provided to illustrate use of the functions provided in `covBM` to fit such models.

2 Model description

2.1 Scaled Brownian motion

Brownian motion (also known as a Wiener process) is a non-stationary stochastic process that constitutes a continuous-time generalisation of a simple random walk [6], in which successive increments are independent of the history of the process. When considered in terms of a given set of

observation points, a scaled Brownian motion process, denoted W_t at time t, is defined by the properties:

$$\begin{aligned} W_0 &= 0 \\ W_t - W_s &\sim N(0, \kappa(t-s)) \text{ for } 0 \leq s < t. \end{aligned}$$

The process starts at zero at time (t) zero, and increments of the process are stationary, independent (for disjoint periods of time) and normally distributed with mean zero and variance equal to the difference in time between observation points scaled by a constant factor κ. These conditions lead to the following characteristics:

$$\begin{aligned} \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \kappa t \\ \mathrm{Cov}[W_s, W_t] &= \kappa * \min(s,t). \end{aligned}$$

The distribution of a set of n observations relating to a given series of time points therefore follows a multivariate normal distribution with a mean vector of n zeros and covariance matrix defined by the formulae given above.

2.2 Scaled fractional Brownian motion

Fractional Brownian motion represents a generalisation of a Brownian motion process in which increments for disjoint time periods are not constrained to be independent, although they do remain stationary. The process was introduced by Mandelbrot and van Ness [7]. The characteristics of a fractional Brownian motion process are determined by an additional parameter, referred to as H or 'the Hurst index', that may take a value in the range (0,1). Standard Brownian motion represents a special case of fractional Brownian motion, corresponding to $H = \frac{1}{2}$. As for standard Brownian motion, the expectation of the value of the process is zero for all points in time.

When $H < \frac{1}{2}$, successive increments of the process are negatively correlated. This has the consequence, firstly, that the path of the trajectory appears 'jagged' and, secondly, that realisations of the process tend to revert towards the mean of zero. For $H > \frac{1}{2}$, successive increments of the process are positively correlated. This means that the path of the process has a relatively 'smooth' appearance, and also that realisations of the process tend to diverge away from zero.

As for Brownian motion, a scale parameter (κ) can be added to the standard definition of fractional Brownian motion, corresponding to the variance of the process at $t = 1$. We may then characterise the properties of the process as follows:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \kappa \left|t\right|^{2H} \\ \mathrm{Cov}[W_s, W_t] &= \frac{\kappa}{2}\left(\left|s\right|^{2H} + \left|t\right|^{2H} - \left|t-s\right|^{2H}\right). \end{aligned}$$

2.3 Integrated Ornstein–Uhlenbeck process

The IOU process is another non-stationary Gaussian stochastic process that has also been used to model CD4 counts in HIV-positive patients, a full description is provided by Taylor *et al.* [1]. The process has the following characteristics:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \frac{\kappa}{\alpha^3}\left(\alpha t + e^{-\alpha t} - 1\right) \\ \mathrm{Cov}[W_s, W_t] &= \frac{\kappa}{2\alpha^3}\left(2\alpha * \min(s,t) + e^{-\alpha t} + e^{-\alpha s} - 1 - e^{-\alpha|t-s|}\right). \end{aligned}$$

We have used the symbol κ to denote the variance scaling parameter (σ^2 was used by Taylor *et al.*[1]). The α parameter determines the extent to which the process reverts towards its mean value. For values of α approaching infinity, the process is equivalent to scaled Brownian motion, whereas for values of α approaching zero the process is equivalent to a random slopes model (without a random intercept)[1].

2.4 Marginal distribution

For models incorporating Gaussian processes such as Brownian motion, the fact that the marginal distribution of the full vector of observations of the outcome variable is multivariate normal (MVN) means that parameter estimation can be achieved through adjustment of the methods used for standard linear mixed models. The linear mixed model for longitudinal data can be expressed in the form[8]:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\, \boldsymbol{\Psi}) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\, \mathbf{R}_i). \end{aligned} \tag{1}$$

Here, $\mathbf{y}_i$ represents the vector of n_i observations for the i^{th} individual, $\mathbf{X}_i$ represents their design matrix for the 'fixed effects' parameters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_i$ is the vector of residual errors for each measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2 ... \mathbf{b}_N$ and residual errors $\mathbf{e}_1, \mathbf{e}_2 ... \mathbf{e}_N$ for each of the N individuals are independent of one another. It can be easily shown that this formulation leads to the following marginal distribution for $\mathbf{y}_i$:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{R}_i).$$

When linear mixed models are fitted to longitudinal data, it is common to assume that the residual errors for each observation within each individual, $\mathbf{e}_i$, are independent and with constant variance, σ^2, i.e. $\mathbf{R}_i$ as defined in (1) is equal to $\sigma^2\mathbf{I}_{n_i}$. However, other forms for $\mathbf{R}_i$ are widely used, particularly for the analysis of longitudinal or spatial data, for example the exponential correlation structure[5].

The remaining variability in the model, once the random effects have been accounted for, can also be subdivided into a component relating to a Gaussian process (independent of other model components) with expectation zero for all time points and an independent residual error for each observation. Defining $\boldsymbol{\Sigma_i}$ as the covariance matrix resulting from the chosen Gaussian process and set of time points $\mathbf{t}_i$ for the i^{th} individual, the linear mixed model can then be expressed as:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + W_i[\mathbf{t}_i] + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\, \boldsymbol{\Psi}) \\ W_i[\boldsymbol{t}_i] &\sim MVN(\mathbf{0},\, \boldsymbol{\Sigma}_i) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\, \sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{2}$$

with marginal distribution:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_i + \sigma^2\mathbf{I}_{n_i}).$$

Although here we have focused on the marginal distribution for linear mixed models that incorporate a stochastic process, similar adjustment of the multivariate normal residual error distribution (i.e. $\mathbf{R}_i$) can also be made for non-linear mixed effects models.

3 Examples

3.1 lmeBM function

The `lmeBM` function is a wrapper for the `lme.formula` function from the `nlme` package, i.e. the `lme` function as used with a formula argument to specify the desired model; and the various

arguments can be used in exactly the same way as the original `nlme` function. However, `lmeBM` allows Brownian motion, fractional Brownian motion or IOU process components to be added to a model.

Included in the `covBM` package is a dataset of serial CD4 counts obtained in HIV-positive children. This dataset is discussed in *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill[9], and the original is available online from the home page of this book. In the present package, rows with missing values of 'CD4CNT' (CD4 count on original scale), 'visage' (age of child in years at given visit) or 'baseage' (age of child in years at initial visit) have been removed.

```
> library(covBM)
> head(cd4)
```

```
  newpid   visage treatmnt CD4CNT  baseage   sqrtcd4         t
1      1 5.330833        1    626 3.910000 25.019992 1.4208333
2      1 5.848333        1    220 3.910000 14.832397 1.9383333
3      2 3.565000        2     30 3.565000  5.477226 0.0000000
4      2 3.778333        2      4 3.565000  2.000000 0.2133333
5      3 6.124167        1    714 6.124167 26.720778 0.0000000
6      3 6.354167        1    523 6.124167 22.869193 0.2300000
```

We will consider models for square root-transformed CD4 counts 'sqrtcd4', as this provides a better approximation to the normal distribution, in terms of the time elapsed in years since the initial visit 't'. The variable 'newpid' provides unique patient identifiers. The 'treatmnt' variable indicates whether that child was a control (==1) or given a zinc supplment (==2). However, this variable is not considered below.

First, we fit a standard 'random slopes' linear mixed model, using the `lme` function from the `nlme` package. We choose here to obtain the maximum likelihood parameter estimates throughout, although restricted maximum likelihood estimation could also be implemented using the argument `method=="REML"`.

```
> RS_model<-lme(sqrtcd4~t, data=cd4, random=~t|newpid, method="ML")
> RS_model
```

```
Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3424.766
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.664754   -5.556963

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.606187 (Intr)
t            5.792576 -0.375
Residual     5.354330

Number of Observations: 976
Number of Groups: 226
```

We then fit a 'random slopes' linear mixed model with additional inclusion of a scaled Brownian motion process. This requires the `covariance=covBM` argument using the `lmeBM` function, which exactly follows the `lme` syntax. The parameter estimates for the model do not converge when using the default optimiser in this dataset, but the model can be successfully fitted using the `control=list(opt="optim")` argument.

```
> BM_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                 covariance=covBM(form=~t|newpid), method="ML",
+                 control=list(opt="optim"))
> BM_model

Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3421.276
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.726746   -5.505073

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.675137 (Intr)
t            3.362038 -0.732
Residual     4.850621

Stochastic process component: covBM
 Formula: ~t | newpid
 Parameter estimate(s):
   Kappa
34.92393
Number of Observations: 976
Number of Groups: 226
```

A further generalisation of the model to incorporate a fractional Brownian motion process can also be considered:

```
> fBM_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                 covariance=covFracBM(form=~t|newpid), method="ML",
+                 control=list(opt="optim"))
> fBM_model

Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3420.997
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.763016   -5.479037

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.727100 (Intr)
t            3.272245 -0.83
Residual     4.551875

Stochastic process component: covFracBM
 Formula: ~t | newpid
 Parameter estimate(s):
      Kappa Hurst index
```

```
 40.8411823   0.3776367
Number of Observations: 976
Number of Groups: 226
```

The fitted model objects created using the `lmeBM` function are of class `"lme"`, and so all the usual `nlme` Methods can be used to extract and view useful information. For example, `anova.lme` can be used to compare a set of fitted models:

```
> anova(RS_model, BM_model, fBM_model)
```

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
RS_model      1  6 6861.531 6890.832 -3424.766
BM_model      2  7 6856.552 6890.736 -3421.276 1 vs 2 6.979464  0.0082
fBM_model     3  8 6857.993 6897.061 -3420.997 2 vs 3 0.558621  0.4548
```

Both the likelihood ratio tests and a comparison of Akaike's information criterion (AIC) values suggest that the model including a Brownian motion process should be chosen above a standard random slopes model, but that there is not evidience to support the generalisation to a fractional Brownian motion process. This conclusion is also supported by inspection of the approximate 95 % confidence intervals of parameter estimates for the fractional Brownian motion model, as the confidence interval for the H-index is inclusive of 0.5 (the value for a standard Brownian motion process).

```
> intervals(fBM_model)$corStruct
```

```
                 lower       est.     upper
Kappa       18.92012487 40.8411823 88.160210
Hurst index  0.06491599  0.3776367  0.841357
attr(,"label")
[1] "Correlation structure:"
```

The random slopes model incorporating an IOU process returns a high estimate of the α parameter, and does not show an improvement in fit relative to the scaled Brownian motion model.

```
> IOU_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                  covariance=covIOU(form=~t|newpid), method="ML",
+                  control=list(opt="optim"))
> IOU_model
```

```
Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3421.164
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.721825   -5.490878

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.655067 (Intr)
t            2.879292 -0.877
Residual     4.886538

Stochastic process component: covIOU
```

```
 Formula: ~t | newpid
 Parameter estimate(s):
      Kappa       Alpha
23758.19550    24.62635
Number of Observations: 976
Number of Groups: 226
```

```
> anova(BM_model, IOU_model)
```

```
          Model df      AIC      BIC    logLik   Test   L.Ratio p-value
BM_model      1  7 6856.552 6890.736 -3421.276
IOU_model     2  8 6858.327 6897.395 -3421.164 1 vs 2 0.2243718  0.6357
```

3.2 nlmeBM function

The `nlmeBM` function is a wrapper for the `nlme.formula` function from the `nlme` package. As for `lmeBM`, `nlmeBM` allows Brownian motion or fractional Brownian motion components to be added to a non-linear mixed effects model.

As an illustrative example, we consider the `Milk` dataset available in the `nlme` package. This dataset is discussed in Chapter 5 of Diggle *et al.*[10], and contains measurements of the protein concentration of the milk of a number of cows assessed weekly following calving. The cows are divided into groups according to diet, but we ignore this for the sake of simplicity. We fit an asymptotic regression function, using `SSasmyp` from `nlme`, with three fixed effects parameters: `Asym` representing the horizontal asymptote for large values of the time variable, `R0` representing the response at time zero and `lrc` representing the natural logarithm of the rate constant (see Pinheiro and Bates[5] for further details). We consider an initial model with independent errors of constant variance and a second model with correlated errors following a continuous autoregressive process, both fit using the `nlme` function. Thirdly, we consider a model including a fractional Brownian motion process within each cow in addition to independent residual errors, using the `covariance=covFracBM` argument for `nlmeBM`. A subject-specific 'random effect' is assigned to the asymptote parameter in each of the models.

```
> Model_1<-nlme(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                 fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                             start = c(Asym = 3.5, R0 = 4, lrc = -1))
> Model_2<-nlme(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                 fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                                 correlation=corCAR1(form=~Time|Cow),
+                             start = c(Asym = 3.5, R0 = 4, lrc = 0))
> Model_3<-nlmeBM(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                 fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                                 covariance=covFracBM(form=~Time|Cow),
+                             start = c(Asym = 3.5, R0 = 4, lrc = -1))
> AIC(Model_1)

301.4711

> AIC(Model_2)

-18.96245

> AIC(Model_3)

-23.20265

> Model_3
```

```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: protein ~ SSasymp(Time, Asym, R0, lrc)
  Data: Milk
  Log-likelihood: 18.60133
  Fixed: Asym + R0 + lrc ~ 1
     Asym        R0       lrc
3.3489469 4.7281304 0.0381144

Random effects:
 Formula: Asym ~ 1 | Cow
                Asym     Residual
StdDev: 2.281751e-07 0.0001115028

Stochastic process component: covFracBM
 Formula: ~Time | Cow
 Parameter estimate(s):
      Kappa Hurst index
 0.07054056  0.16214435
Number of Observations: 1337
Number of Groups: 79
```

On the basis of the AIC values, the model including the fractional Brownian motion component provides the best fit to the data of those considered here.

References

[1] Taylor JMG, Cumberland WG, and Sy JP. A stochastic model for analysis of longitudinal AIDS data. *J Am Stat Assoc*, 89, 727–736 1994.

[2] Babiker AG, Emery S, Fätkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, and Vjechaj MJ; INSIGHT START Study Group. Considerations in the rationale, design and methods of the strategic timing of antiretroviral treatment (START) study. *Clin Trials*, 10 (1 Suppl):S5–S36, 2013.

[3] Stirrup OT, Babiker AG, Carpenter JR, and Copas AJ. Fractional brownian motion and multivariate-t models for longitudinal biomedical data, with application to cd4 counts in hiv-patients. *Statistics in Medicine*, page (in press), 2015.

[4] Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. R package version 3.1-117.

[5] Pinheiro J and Bates D. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[6] Grimmett G and Stirzaker D. *Probability and Random Processes*, page 370. Oxford University Press, third edition, 2001.

[7] Mandelbrot B and van Ness JW. Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[8] Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[9] Gelman A and Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Home page: http://www.stat.columbia.edu/∼gelman/arm/. Cambridge University Press, 2006.

[10] Diggle PJ, Heagerty P, Liang K-Y, and Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, second edition, 2002.

Appendix B MLwiN macro for Brownian motion model

Once a random effects model has been defined in MLwiN, a scaled Brownian motion component can be added using a macro of the following form:

```
SUBS c1 -2 c2 c3 c10
ABSO c10 c10
SETD 2 C10
BATC 1
MAXI 50
STAR
FIXE
RAND
LIKE
```

Where 'c1' is a column containing the identification codes for each individual corresponding to each observation, 'c2' is a column containing the time points for each observation and 'c3' is a column containing the value zero for each observation. The matrices necessary to introduce the correct terms into the iterative generalised least squares procedure are stored in the column 'c10'.

The command SUBSymmetric creates a set of half-symmetric matrices by subtracting the time point for each observation from zero for each respective column. This creates a negative version of the required structure, and so the command ABSOlute values is used to make all values ≥ 0. The command SETDesign adds the covariance structure that has been defined into the model that is going to be estimated; the value '2' is used to denote that this relates to level 2 of the model (in the terminology of MLwiN), i.e. the first level of grouping above the residual observation-specific variance.

BATCh 1 sets batch mode on, and MAXIterations 50 sets the maximum number of iterations in the model estimation procedure to 50. STARt sets the estimation procedure running for the specified model. Once convergence has been achieved, the commands FIXEd, RANDom and LIKElihood print summaries of the fitted model.

Appendix C Statistics in Medicine paper

Statistics in Medicine

Research Article

Received 10 February 2015, Accepted 13 October 2015 Published online 10 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6788

Fractional Brownian motion and multivariate-t models for longitudinal biomedical data, with application to CD4 counts in HIV-positive patients

Oliver T. Stirrup,[a*†] Abdel G. Babiker,[a] James R. Carpenter[a,b] and Andrew J. Copas[a]

Longitudinal data are widely analysed using linear mixed models, with 'random slopes' models particularly common. However, when modelling, for example, longitudinal pre-treatment CD4 cell counts in HIV-positive patients, the incorporation of non-stationary stochastic processes such as Brownian motion has been shown to lead to a more biologically plausible model and a substantial improvement in model fit. In this article, we propose two further extensions. Firstly, we propose the addition of a fractional Brownian motion component, and secondly, we generalise the model to follow a multivariate-t distribution. These extensions are biologically plausible, and each demonstrated substantially improved fit on application to example data from the Concerted Action on SeroConversion to AIDS and Death in Europe study. We also propose novel procedures for residual diagnostic plots that allow such models to be assessed. Cohorts of patients were simulated from the previously reported and newly developed models in order to evaluate differences in predictions made for the timing of treatment initiation under different clinical management strategies. A further simulation study was performed to demonstrate the substantial biases in parameter estimates of the mean slope of CD4 decline with time that can occur when random slopes models are applied in the presence of censoring because of treatment initiation, with the degree of bias found to depend strongly on the treatment initiation rule applied. Our findings indicate that researchers should consider more complex and flexible models for the analysis of longitudinal biomarker data, particularly when there are substantial missing data, and that the parameter estimates from random slopes models must be interpreted with caution. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: CD4 counts; HIV; longitudinal data; missing data; random effects models; residuals

1. Introduction

Longitudinal data are commonly analysed using linear mixed models, as formalised by Laird and Ware [1], with 'random slopes' models (also including random intercepts) particularly common in the biomedical literature. However, the standard random slopes model makes a strong assumption about the relationship between the outcome variable and time, that is, that this follows a separate linear trajectory for each individual with independent normally distributed errors for each observation point. This underlying assumption is implausible in many biomedical scenarios, and the use of more realistically complex models to account for patterns of variability in the data may allow more information to be gained and lead to a reduction of variance and bias in the estimation of model parameters, particularly in the presence of missing data.

In this paper, we focus on modelling the progression of CD4 cell counts in human immunodeficiency virus (HIV)-positive patients prior to treatment. These are a type of white blood cell for which counts

[a]*MRC Clinical Trials Unit at UCL, University College London, London, U.K.*
[b]*Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, U.K.*
**Correspondence to: Oliver T. Stirrup, MRC Clinical Trials Unit at UCL, University College London, London, U.K.*
†*E-mail: oliver.stirrup.13@ucl.ac.uk*

are monitored over time in order to evaluate the progress of the disease and state of the immune system. Statistical analyses of CD4 cell count data are used to evaluate the natural history of HIV infection and to inform epidemiological simulations. Observational datasets of pre-treatment CD4 cell counts obtained in clinical practice are usually subject to a high degree of attrition with increasing time from diagnosis, as patients drop out of the cohort because of treatment initiation, loss to follow-up or death. Furthermore, the timing of observations can be very irregular between and within patients, meaning that flexible statistical structures are required in order to adequately describe patterns of variability in the data.

Taylor *et al.* [2] proposed the addition of a scaled Brownian motion component to a random slopes linear mixed model, finding that this led to a significant improvement in model fit in terms of Akaike's information criterion for a dataset of 722 measurements obtained from 87 seroconverters, patients who had been observed to transition from an HIV-negative to HIV-positive state. Babiker *et al.* [3] fitted such a model to a dataset of CD4 observations from over 15 000 seroconverters and used this to generate CD4 data for simulated cohorts of patients in order to carry out sample size and power calculations for a clinical trial randomising subjects to different treatment initiation rules. Taylor *et al.* [2, 4] also investigated the use of an integrated Ornstein–Uhlenbeck process, of which Brownian motion is a special case, as did Wolbers *et al.* [5]. Fractional Brownian motion is an alternative flexible generalisation of the standard Brownian motion process [6], but its use within the linear mixed model framework has not been investigated. Fractional Brownian motion may be useful for modelling CD4 or other biomarker data as, unlike the integrated Ornstein–Uhlenbeck process, it can allow more erratic variation over time than does simple Brownian motion.

A common finding when assessing the goodness of fit of a statistical model based on the normal distribution, including linear mixed models for the analysis of longitudinal data, is the observation of heavier tails than expected on diagnostic plots of residuals. A natural extension to the standard linear mixed model is to allow the set of observations for each individual as a whole to follow a multivariate-t distribution. The use of such a model for multivariate regression analysis was proposed by Lange *et al.* [7] and was further developed as an extension of the linear mixed model by Welsh and Richardson [8] and Pinheiro *et al.* [9]. None of these papers included the use of non-stationary stochastic process components for the modelling of longitudinal biomarker data.

The multivariate-t distribution was used by Wang and Fan [10] to model CD4 counts in a small sample of 30 HIV-positive patients taken from a historic trial of antiretroviral (ART) medication. Here, observations were recorded on a regular schedule, and Wang and Fan used a random slopes structure with an additional first-order autoregression parameter for the residual error. The same authors have also reported the fitting of a similar multivariate-t model for both CD4 and CD8 cell counts with a second-order autoregressive structure to a sample of 50 patients from the same historic dataset using a Bayesian approach [11]. Matos *et al.* [12] reported the use of a multivariate-t model for right-censored HIV RNA assays in untreated patients with acute infection using a nonlinear random effects model for the mean with independent error terms; their model was fitted to 830 observations in 320 individuals. We hypothesised that combining the use of a multivariate-t model with the addition of a non-stationary stochastic process component would lead to a further substantial improvement in model fit for pre-treatment CD4 data. The inclusion of a stochastic process component in the model is important to reflect the erratic trajectories of the CD4 counts of individual patients over time.

Verbeke and Lesaffre found that estimation of fixed effects parameters using linear mixed models is consistent in the presence of non-normal distributions for the random effects, although they presented a correction to the estimated covariance matrix for the parameter estimates when non-normality of random effects is suspected [13]. Jacqmin-Gadda *et al.* used simulations to show that inference for fixed effects is robust to misspecification of the error distribution when using linear mixed models in some situations [14]. However, these analyses did not take into account the potential for missing or unbalanced data where this is dependent on the observed values of the outcome variable (i.e. data that are 'missing at random' (MAR) in Rubin's terminology [15]). Gurka *et al.* showed that using overly simplistic covariance structures for linear mixed models can lead to inflation of the Type I error rate even for large samples in the absence of missing data [16]. There is therefore a motivation to further investigate biases that may arise from the application of overly simplistic models to realistic datasets that include censoring. This is an important issue for the analysis of observational pre-treatment CD4 counts in which the timing of censoring from the dataset due to treatment initiation is likely to be strongly linked to the preceding observed values for each individual and the statistical inferences drawn may be more dependent on model choice.

We aimed to further develop the available statistical models for longitudinal biomedical data, incorporating both fractional Brownian motion processes for flexible modelling of intra-individual variation and multivariate-t distributions to relax the assumption of multivariate normality. The motivating dataset

Statistics
in Medicine

of pre-ART CD4 counts used for analysis is introduced in Section 2. Theoretical characteristics of the models fitted and methods for maximum likelihood estimation are described in Section 3. Checking of model adequacy for the data under investigation is crucial, particularly in the presence of missing data. Residual diagnostics for models based on the multivariate-t distribution are discussed, and novel methods are proposed for the critical evaluation of such models in Section 4. Application of the models developed to the dataset of pre-ART CD4 counts is described in Section 5, informing simulation studies that are presented to demonstrate differences in predictions made by the more complex models regarding the timing of treatment initiation in population cohorts and to show that the application of simpler models can lead to substantial bias in parameter estimates when there is censoring dependent on observed values of the outcome variable. Practical and methodological implications of the work are discussed in Section 6.

2. Dataset

We demonstrate the use of the statistical methods developed through a reanalysis of the dataset of pre-ART CD4 counts described by Babiker *et al.* [3], comprising all available measurements prior to the occurrence of acquired immune deficiency syndrome (AIDS)-defining illness or initiation of ART up to December 2007 from 21 cohorts (originating from 12 countries) participating in the Concerted Action on SeroConversion to AIDS and Death in Europe (CASCADE) study [17]. Only patients with a well-estimated date of HIV seroconversion are included in the CASCADE study, providing a natural 'zero' time in each patient for statistical modelling. The total dataset includes 89 176 CD4 count observations in 15 274 individuals. However, only 3955 (4.4%) measurements from 789 (5.2%) patients were recorded at a time of more than 10 years, and so we chose to model only those CD4 measurements obtained up to 10 years from the time of seroconversion. This resulted in a dataset of 85 221 measurements in 15 164 individuals. A further 365 observations were excluded for which an identical CD4 measurement was recorded only 1 day after the previous count for that patient, as these were found to cause problems with model estimation and were assumed to result from data-entry errors, resulting in a dataset of 84 856 measurements for analysis.

The CD4 cell counts are measured as cells per microlitre, and we followed established practice in modelling the counts on a square-root scale [3]. As an illustrative example, the CD4 measurements were modelled only in terms of time from seroconversion, expressed as continuous in years, although it would be possible to include other predictive variables. The median number of CD4 observations per individual in the analysed dataset was 4, with a range of 1–57 and an interquartile range of 2–8. There was no rigid pattern to the timing of observations in each patient, with a median interval between measurements of 112 days (interquartile range, 70–182). The highly unbalanced nature of the dataset and the irregular observation schedule necessitate the use of flexible modelling strategies that can accommodate such features. Visual inspection of the CD4 data suggests that the trajectories over time for each individual do not follow predictable paths and that there may be between-patient differences in variability over time, motivating the combination of stochastic process components and the multivariate-t distribution, respectively, as presented in this paper. A total of 9831 (64.8%) patients were censored from the dataset at initiation of ART, 1111 (7.3%) because of a recorded AIDS event and 318 (2.1%) at death. Two thousand four hundred and forty-four (16.1%) patients can be considered lost to follow-up (with no clinic visit recorded for 12 months and no censoring event), and the remaining 1460 (9.6%) were in follow-up at the time that the data were gathered.

We hope that the models developed will form the basis for improved epidemiological simulations, as required for the planning of clinical trials and population health analyses, and provide more accurate estimates of the mean CD4 count over time were there to be no censoring of data. Furthermore, the characterisation and quantification of within-patient and between-patient variability in CD4 count trajectories may help develop understanding of the natural history of untreated HIV.

3. Stochastic process and multivariate-t models

3.1. Characteristics of Brownian motion and related processes

3.1.1. Scaled Brownian motion. In a mathematical sense, Brownian motion (also known as a Wiener process) is a non-stationary stochastic process that constitutes a continuous-time generalisation of a simple random walk [18], in which successive increments are independent of the history of the process.

When considered in terms of a given set of observation points (these may be irregularly spaced in time), a scaled Brownian motion process, denoted W_t at time t, is defined by the following properties:

$$W_0 = 0$$
$$W_t - W_s \sim N(0, \kappa(t-s)) \text{ for } 0 \leqslant s < t, 0 < \kappa.$$

The process starts at zero at time zero ($t = 0$), and increments of the process are stationary, independent (for disjoint periods of time) and normally distributed with mean zero and variance equal to the difference in time between observation points scaled by a positive constant factor κ. The following characteristics arise from these conditions:

$$\mathrm{E}[W_t] = 0$$
$$\mathrm{Var}[W_t] = \kappa t$$
$$\mathrm{Cov}[W_s, W_t] = \kappa \times \min(s, t).$$

The distribution of a set of n observations relating to a given series of time points therefore follows a multivariate normal (MVN) distribution with a mean vector of n zeros and covariance matrix defined by the formulae given. As such, Brownian motion is an example of a Gaussian process and can be readily incorporated into the theoretical framework of linear mixed models, as will be discussed in Section 3.2.

3.1.2. Scaled fractional Brownian motion. Fractional Brownian motion represents a generalisation of a Brownian motion process in which increments for disjoint time periods are not constrained to be independent, although they do remain stationary. The process was introduced by Mandelbrot and van Ness[6]. The characteristics of a fractional Brownian motion process are determined by an additional parameter, referred to as H or 'the Hurst index', that may take a value in the range (0,1). Standard Brownian motion represents a special case of fractional Brownian motion, corresponding to $H = \frac{1}{2}$. As for standard Brownian motion, the expectation of the value of the process is zero for all points in time.

When $H < \frac{1}{2}$, successive increments of the process are negatively correlated. This has the consequence, firstly, that the path of the trajectory appears 'jagged' and, secondly, that realisations of the process tend to revert towards the mean of zero. For $H > \frac{1}{2}$, successive increments of the process are positively correlated. This means that the path of the process has a relatively 'smooth' appearance, and also that individual realisations of the process tend to diverge away from the mean of zero. Illustrative simulated realisations of fractional Brownian motion processes generated with varying values of H are shown in Figure 1.

As for Brownian motion, a positive scale parameter (κ) can be added to the standard definition of fractional Brownian motion, corresponding to the variance of the process at $t = 1$. We may then characterise the properties of the process as follows:

$$W_0 = 0$$
$$\mathrm{E}[W_t] = 0$$
$$\mathrm{Var}[W_t] = \kappa |t|^{2H}$$
$$\mathrm{Cov}[W_s, W_t] = \frac{\kappa}{2}\left(|s|^{2H} + |t|^{2H} - |t-s|^{2H}\right).$$

Fractional Brownian motion is defined as a continuous-time stochastic process. However, as we are concerned with modelling biomedical measurements obtained at specific time points, we have focused here on the properties of the process relating to a finite set of observations. As for simple scaled Brownian motion, scaled fractional Brownian motion is a Gaussian process that follows a MVN distribution for any given set of observation points, with expectation zero and covariance matrix as defined.

3.2. Marginal distribution for stochastic process models

For models incorporating Gaussian processes such as Brownian motion, the fact that the marginal distribution of the full vector of observations of the outcome variable is MVN means that parameter estimation can be achieved through adjustment of the methods used for standard linear mixed models. The linear mixed model for longitudinal data can be expressed in the following form [1]:

Statistics
in Medicine

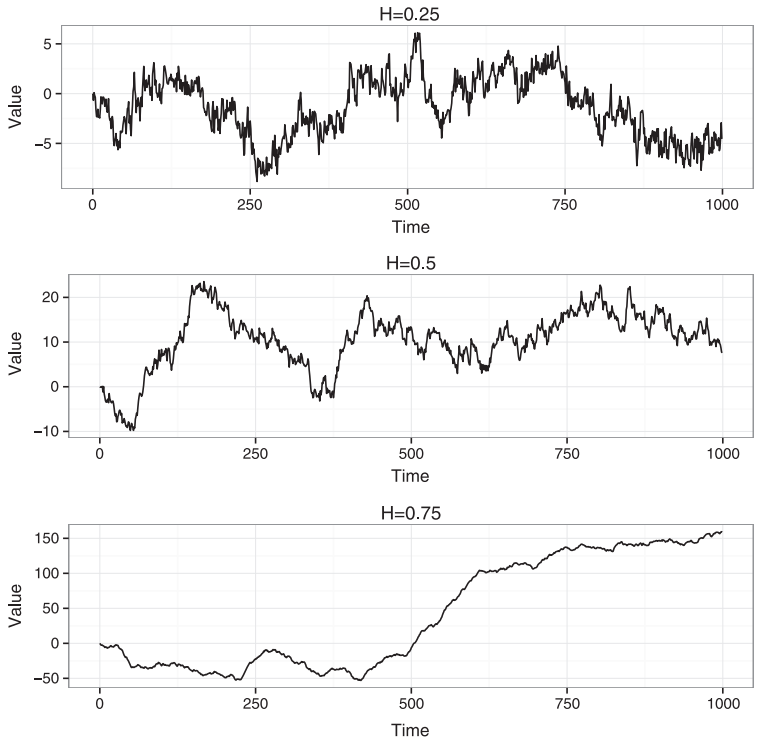

Figure 1. Realisations of fractional Brownian motion processes with varying values of H and scale parameter fixed at 1. A finite set of 1000 observations was generated in each case.

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\, \boldsymbol{\Psi}) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\, \mathbf{R}_i). \end{aligned} \tag{1}$$

Here, $\mathbf{y}_i$ represents the vector of n_i observations for the ith individual, $\mathbf{X}_i$ represents their design matrix for the 'fixed effects' parameters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_i$ is the vector of residual errors for each measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_N$ and residual errors $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N$ for each of the N individuals are independent of one another. It can be easily shown that this formulation leads to the following marginal distribution for $\mathbf{y}_i$:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{R}_i).$$

When linear mixed models are fitted to longitudinal data, it is common to assume that the residual errors for each observation within each individual, $\mathbf{e}_i$, are independent and with constant variance, σ^2, that is, $\mathbf{R}_i$ as defined in (1) is equal to $\sigma^2\mathbf{I}_{n_i}$. However, other forms for $\mathbf{R}_i$ are widely used, particularly for the analysis of longitudinal or spatial data. An example is provided by the exponential decay correlation structure [19], for which the elements (r_{jk}) of $\mathbf{R}_i$ are calculated as a function of the 'distance' s between each pair of observations (in the context of longitudinal data this would be the time difference) and a 'range' parameter γ as follows:

$$r_{jk} = \sigma^2\, exp\left(-\frac{s_{jk}}{\gamma}\right).$$

Alternatively, the remaining variability in the model, once the random effects have been accounted for, can be subdivided into a component relating to a Gaussian process (independent of other model components) with expectation zero for all time points and an independent residual error for each observation

(here assumed to have constant variance); this effectively just creates a class of parameterisations for $\mathbf{R}_i$. Defining $\boldsymbol{\Sigma_i}$ as the covariance matrix resulting from the chosen Gaussian process and set of time points $\mathbf{t}_i$ for the ith individual, the linear mixed model can then be expressed as follows:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + W_i[\mathbf{t}_i] + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\ \boldsymbol{\Psi}) \\ W_i[\mathbf{t}_i] &\sim MVN(\mathbf{0},\ \boldsymbol{\Sigma}_i) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\ \sigma^2\mathbf{I}_{n_i}) \end{aligned} \tag{2}$$

with marginal distribution

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_i + \sigma^2\mathbf{I}_{n_i}).$$

3.3. Multivariate-t distribution for longitudinal data

There are a number of multivariate generalisations of the univariate-t distribution, and a thorough review of this topic is provided by Kotz and Nadarajah [20]. However, we shall refer to the *multivariate-t distribution* as that with the probability density function as follows:

$$f\left(\mathbf{y}_i; \boldsymbol{\mu}_i, \mathbf{V}_i, v\right) = \frac{\boldsymbol{\Gamma}\left(\left(v + n_i\right)/2\right)}{\boldsymbol{\Gamma}\left(v/2\right) v^{n_i/2}\pi^{n_i/2}\left|\mathbf{V}_i\right|^{1/2}\left(1 + \frac{1}{v}\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right)^{\mathrm{T}}\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right)\right)^{(v+n_i)/2}}.$$

Where n_i represents the length of the random vector $\mathbf{y}_i$ ($\in \mathbb{R}^{n_i}$), $\mathbf{V}_i$ is a $n_i \times n_i$ positive-definite scale matrix, $\boldsymbol{\mu}_i$ is a $n_i \times 1$ location vector and v is a degrees of freedom parameter. The mean of the distribution is $\boldsymbol{\mu}_i$ if $v > 1$ and otherwise undefined, and the variance of the distribution is $\frac{v}{v-2}\mathbf{V}_i$ if $v > 2$ and otherwise undefined. This is the most commonly used definition of the multivariate-t distribution.

In the present context, the mean vector $\boldsymbol{\mu}_i$ will be represented as $\mathbf{X}_i\boldsymbol{\beta}$, that is, a function of a design matrix $\mathbf{X}_i$ and vector of parameters $\boldsymbol{\beta}$. As for linear mixed models based on the normal distribution, the scale matrix $\mathbf{V}_i$ can be divided into components relating to a random effects structure and a residual error structure, $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}}$ and $\mathbf{R}_i$, respectively. Pinheiro *et al.* consider the situation in which the degrees of freedom parameter may vary between subgroups of individuals, but we shall assume that this is a single constant [9].

If a vector of observations $\mathbf{y}_i$ follows a multivariate-t distribution

$$\mathbf{y}_i \sim t_{n_i}\left(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{V}_i,\ v\right),$$

then this can alternatively be represented as a hierarchical model in which $\mathbf{y}_i$ follows a MVN distribution conditional on a gamma-distributed variable τ_i (with parameters given for 'shape' and 'rate', respectively) as follows [9]:

$$\begin{aligned} \mathbf{y}_i|\tau_i &\sim MVN\left(\mathbf{X}_i\boldsymbol{\beta},\ \frac{1}{\tau_i}\mathbf{V}_i\right) \\ \tau_i &\sim \text{gamma}\left(\frac{v}{2},\ \frac{v}{2}\right). \end{aligned} \tag{3}$$

In the context of the models proposed, combining variance components related to random effects, stochastic processes and measurement error (i.e. $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_i + \sigma^2\mathbf{I}_{n_i}$), this is equivalent to

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + W_i[\mathbf{t}_i] + \mathbf{e}_i \\ \mathbf{b}_i|\tau_i &\sim MVN\left(\mathbf{0},\ \frac{1}{\tau_i}\boldsymbol{\Psi}\right) \\ W_i[\mathbf{t}_i]|\tau_i &\sim MVN\left(\mathbf{0},\ \frac{1}{\tau_i}\boldsymbol{\Sigma}_i\right) \\ \mathbf{e}_i|\tau_i &\sim MVN\left(\mathbf{0},\ \frac{1}{\tau_i}\sigma^2\mathbf{I}_{n_i}\right) \\ \tau_i &\sim \text{gamma}\left(\frac{v}{2},\ \frac{v}{2}\right). \end{aligned}$$

Statistics
in Medicine

As noted by Pinheiro *et al.* [9], it directly follows from the hierarchical form of the model that

$$\tau_i|\mathbf{y}_i \sim \text{gamma}\left(\frac{v+n_i}{2}, \frac{v+\delta_i^2(\boldsymbol{\theta})}{2}\right), \quad \text{where} \quad \delta_i^2(\boldsymbol{\theta}) = \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right). \tag{4}$$

Here, $\boldsymbol{\theta}$ represents the parameter vector that includes $\boldsymbol{\beta}$ and determines the construction of $\mathbf{V}_i$. From the standard properties of a gamma distribution, it can be seen that

$$\mathrm{E}\left(\tau_i|\mathbf{y}_i\right) = \frac{v+n_i}{v+\delta_i^2(\boldsymbol{\theta})}.$$

3.4. Maximum likelihood estimation and software

As the likelihood function for the multivariate-normal or multivariate-t linear mixed-effects model has a closed form, whatever the structure of $\mathbf{V}_i$, it is possible to directly apply Newton–Raphson-type optimisation procedures. Although finite differencing can be employed, the use of analytically derived exact gradients (with respect to the model parameters) in Newton–Raphson-type procedures typically greatly improves stability and speed of convergence. However, in some situations, such as incorporating stochastic process components into the multivariate-t linear mixed effects model, the analytic derivation of the gradients is not trivial. In addition, once an analytic form for each of the gradient terms has been derived, it is required that this be programmed into the computational procedure for the optimisation in an efficient manner.

An alternative method is provided by automatic differentiation, whereby a computer program is structured in such a way that it can automatically calculate the derivatives of a mathematical function to the same degree of accuracy as analytical derivatives (to machine precision) [21]. In essence, this is achieved through application of the chain rule to each of the elementary operations that comprise the calculation of the objective function (i.e. the log-likelihood function). The open-source Automatic Differentiation Model Builder (ADMB) software (ADMB Foundation, Honolulu, HI, USA) allows optimisation for any statistical model that has a closed form differentiable log-likelihood function [22] (the software also includes functionality for models without a closed form for the likelihood that is not employed in this paper). For any given model, the user is required to write a 'template' file defining a program to calculate the log-likelihood in terms of the data and the set of unknown parameters to be estimated based on the C++ language; additional statistical and mathematical functions (including matrix and vector functions and operations) are provided by the software to facilitate this. A zip file containing several example template files and a simulated dataset is provided online (Supplementary Data File S1).

For all models presented in Section 5, maximum likelihood estimates of the parameters were obtained using the ADMB software (Version 10.1). The 'R2admb' package [23] was used to run analyses and manage results through the R statistical computing environment. Starting values are required for all parameters when using ADMB. These were obtained by using approximate values from a model fit for the initial 'random slopes' linear mixed model (including random intercepts) from the *nlme* package for R, and subsequent models were fitted using parameter estimates from the previous simpler model as the initial value. When fitting models with a Brownian motion component, an initial value of 1 was used for the scale parameter, and for models with fractional Brownian motion, an initial value of 0.5 was used for the H index. For models based on the multivariate-t distribution, an initial value of 10 was used for the degrees of freedom parameter. An R package (covBM) that will allow the implementation of all MVN models described in this paper is under development by the authors.

The 'fixed effects' for each model are the intercept (β_0) and a slope (β_1) parameter. For the 'random effects' covariance/scale matrix ($\boldsymbol{\Psi}$) for each model, U_{00} and U_{11} represent the variance of the random intercepts and random slopes, respectively, for each individual, with ρ representing the correlation between them. For the multivariate-t models, this interpretation holds conditional on scaling by the vector of unobserved latent variables $\boldsymbol{\tau}$. Models were parameterised using log-transformations of U_{00} and U_{11} and a generalised logistic transformation of ρ. For all models, the residual error term was parameterised using $\log(\sigma)$ (i.e. the log of the residual standard deviation). The exponential decay correlation structure was parameterised using the log of the range parameter (γ), and Brownian motion models (including fractional) used the log of the scale parameter (κ). Fractional Brownian motion was parameterised using the logistic transformation of H. A log transformation was used for the degrees of freedom parameter in multivariate-t models. For all model parameters, confidence intervals are reported derived from the estimated asymptotic MVN distribution based on the observed information on the transformed scales.

4. Residual diagnostics for multivariate-t models

The evaluation of diagnostic plots of the residuals resulting from fitted statistical models forms an important part of model criticism and development. Such plots can be used to check the adequacy of fitted models to describe the data under investigation and, when problems are observed, to suggest how further improvements might be made. This is particularly important in the present context in which there is interest in understanding patterns of variability within and between individuals as well as ensuring correct inference for fixed effects parameters.

4.1. Subject-level residuals

Much of the focus regarding the use of multivariate-t linear mixed effects models has been on providing robust inference for the fixed effects; this follows from the fact that individuals with observations that are further from the mean are down-weighted in the estimation of the fixed effects parameters. Lange *et al.* were concerned with achieving robust multivariate regression and suggested the use of diagnostic residual plots that indicated whether the fitted model adequately reflected the presence of outlying sets of measurements (i.e. corresponding to the various measurements conducted on a single individual) [7]. They point out that for a normal linear mixed model, the statistic

$$\hat{\delta}_i^2(\boldsymbol{\theta}) = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)$$

for each individual would asymptotically follow a χ^2 distribution with n_i degrees of freedom. However, under a multivariate-t model, the statistic $\frac{\hat{\delta}_i^2(\boldsymbol{\theta})}{n_i}$ would asymptotically follow an F-distribution with n_i and $\hat{\nu}$ degrees of freedom. Lange *et al.* transform these statistics to standard normal deviates and then use quantile–quantile (Q–Q) plots to assess model fit. A similar approach was used by Wang and Fan [10]. Such plots can demonstrate the inadequacy of the normal linear mixed-effects model to describe the observed data. However, the plots do not directly show whether the multivariate-t model correctly describes variability between individual measurements.

4.2. Measurement-level residuals

We propose that the gamma–normal formulation of the multivariate-t model, as given in (3), can be also used to assess whether the multivariate-t distribution fully describes the patterns of variability observed for all individual measurements in a dataset. As the observations for the ith individual are assumed to follow a MVN distribution conditional on τ_i, one option is to use empirical Bayes estimates (i.e. the mean of the predicted posterior distribution) of the τ_i as follows:

$$\hat{\tau}_i = \frac{\hat{\nu} + n_i}{\hat{\nu} + \hat{\delta}_i^2(\boldsymbol{\theta})}$$

to estimate the normal covariance matrix ($\hat{\mathbf{V}}_i'$) for each individual

$$\hat{\mathbf{V}}_i' = \frac{1}{\hat{\tau}_i}\hat{\mathbf{V}}_i.$$

This could then be used to transform the marginal residuals for the ith individual (i.e. $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$) as for a normal linear mixed model using the inverse of a Cholesky decomposition of the covariance matrix (as suggested by Fitzmaurice, Laird and Ware [24]), with the transformed residuals for all individuals displayed in a Q–Q plot. However, assuming the empirical Bayes estimates of the τ_i to be correct for all individuals might result in misleading conclusions in a similar manner to that which can be observed when evaluating the empirical Bayes estimates of random effects in a normal linear mixed model (e.g. as reported by Verbeke and Lesaffre [25]). An alternative would be to draw a number of repeated samples from the predicted posterior distribution of the full vector of $\boldsymbol{\tau}$, using each sample to generate a full set of $\hat{\mathbf{V}}_i'$ matrices and corresponding Cholesky-transformed marginal residuals. The sets of transformed marginal residuals could then be used individually to generate multiple Q–Q plots or used together to derive a single Q–Q plot showing the distribution of 'observed quantiles' over multiple realisations of the $\boldsymbol{\tau}$.

Statistics
in Medicine

The gamma–normal formulation of the multivariate-t model provides another route to model checking through the separate evaluation of each individual in the dataset. Assuming that the model parameters are known, then the transformed marginal residuals using the inverse of the Cholesky decomposition of the scale matrix for each individual, $\mathbf{V}_i$, are normally and independently distributed with mean $\mathbf{0}$ and variance $\frac{1}{\tau_i}$ (conditioned on τ_i) as follows:

$$\mathbf{V}_i = \mathbf{L}_i \mathbf{L}_i^{\mathrm{T}}$$
$$\mathbf{L}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} \right) | \tau_i \sim MVN \left(\mathbf{0}, \frac{1}{\tau_i} \mathbf{I}_{n_i} \right).$$

Hence, for a model that correctly describes the data, separate Q–Q plots (with respect to the standard normal distribution) of these transformed residuals for each individual should each indicate a normal distribution (with differing variance). For small datasets, it may be possible to create multipanel graphics that simultaneously display the Q–Q plots for all individuals, but for larger datasets, it would be necessary to select a random sample of individuals for inspection. This approach will be more effective when there are a greater number of observations per individual, as it is difficult to assess the assumption of normality for very small samples. This reflects the fact that the presence of a greater number of observations per individual in a dataset will provide more information as to whether there truly is a difference in underlying variability between individuals, as represented by the values of τ_i. This technique could also be used for fitted MVN models, using a Cholesky decomposition of the marginal covariance matrix for each individual, in order to assess whether the multivariate-t distribution might be appropriate for the data.

The assessment of measurement-level residuals is particularly important when the motivation for an analysis is to be able to make predictions regarding future individual measurements or to simulate datasets in which the exact pattern of values within each individual is critical. The use of subject-level residuals may be sufficient for multivariate regression analysis (for example, a defined set of different patient characteristics at a single time point in each individual), but for the analysis of longitudinal data we believe that measurement-level residuals should also be investigated. Examples of the residual plots proposed are presented in Section 5. For these plots, calculations were carried out in R, and graphics were generated using the ggplot2 package for R (Version 0.9.3.1) [26].

5. Application and implications of modelling strategy

5.1. Set of models fitted

The initial model fitted was a standard linear mixed-effects model including correlated random intercept and slope terms and independent measurement error terms of constant variance. An exponential delay correlation structure was considered for the error terms of this model, and the initial model was then extended to also include either a scaled Brownian motion process or a scaled fractional Brownian motion process. The equivalent set of four models was then fitted using a marginal multivariate-t distribution, that is, with the scale matrix $\mathbf{V}_i$ structured in the same manner but assuming an unobserved scaling variable for each individual as described in Section 3.3.

5.2. Results and diagnostic checks

Table I shows the results of linear mixed models (including stochastic process extensions), with marginal MVN distribution, fitted to the pre-ART CASCADE data. Nested models are compared using the likelihood ratio test; as only a single parameter is being added to the model in each of the comparisons presented, the critical value for change in 2×log-likelihood ($2\Delta\ell$) at the 5 % significance level is only 3.84. Non-nested models are compared using the Bayesian information criterion (BIC) statistic, using the total number of observations in the dataset for the calculation of the penalty term; this is supported by the derivation of Cavanaugh and Neath [27].

The addition to the initial random slopes model of an exponential decay correlation structure for the residual variance resulted in a significant improvement in model fit ($2\Delta\ell$ 460 for 1 degree of freedom (df), $P<0.001$). However, the addition of a Brownian motion component to the random slopes model led to a greater increase in log-likelihood ($2\Delta\ell$ 4940 for 1 df, $P<0.001$), with a subsequently lower value of BIC for this model. A further improvement in model fit was observed when the Brownian motion component was generalised to a fractional Brownian motion process ($2\Delta\ell$ 160 for 1 df, $P<0.001$). As

Table I. Summaries of extended linear mixed models (with marginal multivariate-normal distribution) fitted to square-root transformed pre-antiretroviral therapy CD4 measurements from the Concerted Action on SeroConversion to AIDS and Death in Europe dataset.

	Random slopes + measurement error	Random slopes + exp. cor. + measurement error	Random slopes + Brownian motion+ measurement error	Random slopes + fBM + measurement error
β_0	24.13 (24.02 to 24.24)	24.12 (24.01 to 24.23)	23.81 (23.7 to 23.92)	23.82 (23.71 to 23.92)
β_1	−1.36 (−1.4 to −1.33)	−1.35 (−1.38 to −1.31)	−1.15 (−1.18 to −1.11)	−1.15 (−1.19 to −1.12)
U_{00}	33.68 (32.65 to 34.73)	33.22 (32.2 to 34.27)	28.69 (27.72 to 29.7)	27.46 (26.46 to 28.51)
ρ	−0.39 (−0.41 to −0.36)	−0.38 (−0.41 to −0.35)	−1 (−1 to 1)	−0.59 (−0.63 to −0.54)
U_{11}	1.62 (1.54 to 1.71)	1.54 (1.46 to 1.62)	0.20 (0.16 to 0.24)	0.58 (0.49 to 0.68)
σ	2.76 (2.74 to 2.77)	2.79 (2.77 to 2.81)	2.28 (2.26 to 2.29)	2.01 (1.94 to 2.07)
γ	—	0.03 (0.03 to 0.03)	—	—
κ	—	—	7.00 (6.78 to 7.22)	9.32 (8.78 to 9.91)
H	—	—	—	0.30 (0.27 to 0.33)
ℓ	−232 579	−232 349	−230 109	−230 029
BIC	465 226	464 777	460 297	460 149

Parameter estimates are given with 95 % confidence intervals in parentheses. BIC, Bayesian information criterion; exp. cor., exponential decay correlation structure for residual error term; fBM, fractional Brownian motion; ℓ, log-likelihood.

such, the fractional Brownian motion model was found to have the lowest BIC of the fitted linear mixed models. A 'random slopes + integrated Ornstein–Uhlenbeck process + measurement error' model was also considered but was found to return the special case of a Brownian motion process (i.e. with a very large estimate for the α parameter [2]).

It is of particular interest that the estimate of the H parameter for the model incorporating a fractional Brownian motion process is below 0.5, indicating that successive increments of the process are negatively correlated and hence that the process will tend to revert towards its mean. The mean in this case would include the subject-specific random effects for the intercept and slope. The correlation between the random intercept and random slope for each individual for the model incorporating a scaled standard Brownian motion process is estimated to be −1.00, which seems rather unnatural. However, when the process is generalised to a fractional Brownian motion, an estimate of -0.59 (95 % CI, -0.63 to -0.54) is obtained for this correlation. The Cholesky-transformed residuals of the commonly used random slopes model and of the best-fitting linear mixed model, incorporating a fractional Brownian motion component, were analysed to assess the goodness of fit. For both of these models, the Q–Q plot of the Cholesky residuals indicates that their distribution is markedly heavy-tailed in comparison to the expected standard normal under a correctly specified model (Figure 2).

Summaries of the multivariate-t distribution models fitted to the pre-ART CASCADE data are provided in Table II. As for the MVN models, the fractional Brownian motion model was found to have the lowest BIC of the fitted multivariate-t distribution models. Furthermore, all of the multivariate-t models were found to have lower BIC values than all of the normal linear mixed models. The difference in 2ℓ between the normal and the multivariate-t 'random slopes + fractional Brownian motion + measurement error' models is 8298, indicating a significant and substantial improvement in model fit (1 df, $P<0.001$). Note that these models can be considered nested as the multivariate-t model is equivalent to the MVN model as the degrees of freedom parameter tends to (positive) infinity.

The estimated degrees of freedom parameter was between 5 and 6 for all of the fitted multivariate-t models, as expected given the heavy tails observed in the Q–Q plots for the normal linear mixed models. However, the heavy tails could be due to distributional structures other than the multivariate-t distribution employed, for example the random effects and any Gaussian processes included could follow MVN distributions whilst the residual error terms followed independent t-distributions. As such, there is a need for further investigation to assess the goodness of fit of the chosen multivariate-t model with respect to the data. As described in Section 4.2, for the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model, 1000 simulations of the vector of latent variables $\boldsymbol{\tau}$ were generated, based on the predicted posterior distribution in each individual and used to calculate sets of Cholesky-transformed residuals for the model. The Q–Q plot of the Cholesky residuals derived using the empirical Bayes

Statistics
in Medicine

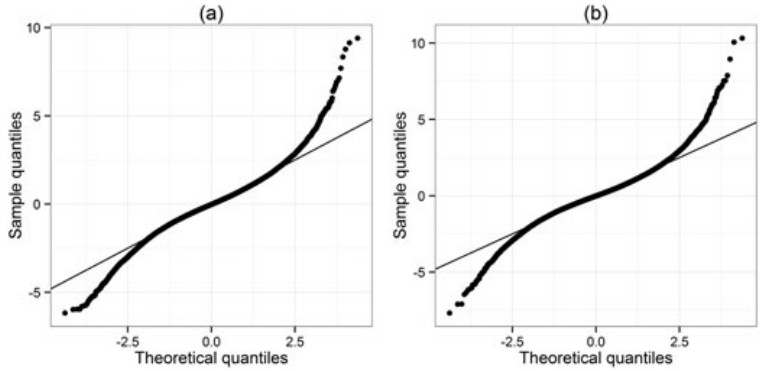

Figure 2. Quantile–quantile plots of Cholesky-transformed residuals from (a) the 'random slopes + measurement error' and (b) the 'random slopes + fractional Brownian motion + measurement error' linear mixed model fitted to the pre-antiretroviral therapy CD4 counts from the Concerted Action on SeroConversion to AIDS and Death in Europe dataset. Plots are generated with respect to a standard normal distribution, and the line of equality is shown.

Table II. Summaries of multivariate-t distribution models fitted to square-root transformed pre-antiretroviral therapy CD4 measurements from the Concerted Action on SeroConversion to AIDS and Death in Europe dataset.

	Random slopes + measurement error	Random slopes + exp. cor. + measurement error	Random slopes + Brownian motion+ measurement error	Random slopes + fBM + measurement error
β_0	23.77 (23.67 to 23.87)	23.76 (23.66 to 23.86)	23.57 (23.47 to 23.67)	23.59 (23.49 to 23.69)
β_1	−1.27 (−1.31 to −1.24)	−1.23 (−1.27 to −1.2)	−1.10 (−1.13 to −1.07)	−1.11 (−1.14 to −1.07)
U_{00}	23.82 (22.99 to 24.69)	22.83 (22 to 23.68)	20.3 (19.5 to 21.14)	18.82 (17.98 to 19.7)
ρ	−0.37 (−0.4 to −0.34)	−0.36 (−0.39 to −0.33)	−1 (−1 to 1)	−0.51 (−0.55 to −0.47)
U_{11}	1.17 (1.1 to 1.23)	1.01 (0.95 to 1.08)	0.12 (0.1 to 0.15)	0.49 (0.43 to 0.55)
σ	2.25 (2.23 to 2.27)	2.32 (2.3 to 2.35)	1.88 (1.86 to 1.9)	1.45 (1.35 to 1.55)
γ	—	0.07 (0.06 to 0.07)	—	—
κ	—	—	5.17 (4.98 to 5.36)	8.02 (7.44 to 8.64)
H	—	—	—	0.23 (0.21 to 0.26)
df	5.64 (5.4 to 5.88)	5.34 (5.12 to 5.57)	5.83 (5.58 to 6.09)	5.76 (5.52 to 6.02)
ℓ	−228 221	−227 705	−226 015	−225 880
BIC	456 521	455 501	452 121	451 862

Parameter estimates are given with 95 % confidence intervals in parentheses. BIC, Bayesian information criterion; df, degrees of freedom parameter; exp. cor., exponential decay correlation structure for residual error term; fBM, fractional Brownian motion; ℓ, log-likelihood.

estimate ($\hat{\tau}_i$) for each individual shows a near perfect fit to the standard normal distribution (Figure 3). However, taking quantiles over multiple simulations of $\boldsymbol{\tau}$ indicates the presence of slightly heavier tails than expected.

The goodness of fit of the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model was further investigated by inspection of Q–Q plots of residuals for individual patients transformed by the inverse of the Cholesky decomposition of their estimated scale matrix ($\hat{\mathbf{V}}_i$) without any correction for τ_i. As little would be gained by evaluating patients with very few observations, only those with greater than 15 measurements in the dataset were considered; one thousand and forty-four (6.9%) individuals in the dataset met this criterion. Q–Q plots for 25 randomly selected individuals are shown in Figure 4. Under a correctly specified model, each of the plots should approximately show a straight line of points, with differing slopes between individuals; for the ith individual, the expected slope is a function of their unobserved scale variable: $\tau_i^{-1/2}$, where $\tau_i \sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, with ν being the degrees of freedom parameter in the multivariate-t model. These plots suggest that there are indeed differences in overall variability between individuals as implied by the multivariate-t model; for example, Plot 9 shows a clearly steeper slope than Plot 11. To further illustrate this, the raw data from

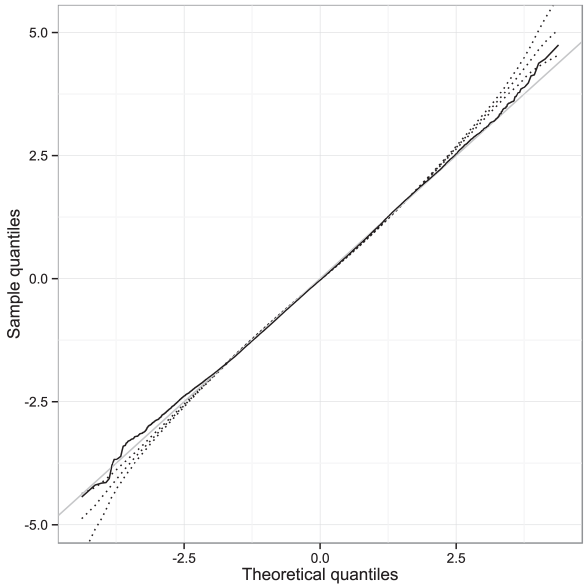

Figure 3. Composite quantile–quantile plot of the distribution of Cholesky-transformed residuals (for all measurements) from the 'random slopes + fractional Brownian motion + measurement error' multivariate-t distribution model fitted to the pre-antiretroviral therapy CD4 counts from the Concerted Action on SeroConversion to AIDS and Death in Europe dataset, based on 1000 simulations of the vector of latent variables $\boldsymbol{\tau}$. The dotted lines show the 2.5th, 50th and 97.5th percentiles of the sample quantiles for each theoretical quantile corresponding to the total number of observations; the solid black line shows the sample quantiles derived using the empirical Bayes estimate ($\hat{\tau}_i$) for each individual, with the line of equality also displayed in grey.

the 25 sampled patients are shown in Figure 5, with the observations for the patients corresponding to Plots 9 and 11 in Figure 4 made prominent. The 'Plot 9' patient has the lowest predicted latent scaling variable ($\hat{\tau} = 0.29$) amongst this subset, corresponding to high variability over time, whilst the 'Plot 11' patient has the highest predicted latent scaling variable ($\hat{\tau} = 2.33$) in this group, corresponding to low variability over time.

5.3. Simulation study: impact of model choice on treatment initiation predictions

The initiation of ART in HIV-positive patients is commonly based on the observations of a CD4 count below a given threshold, with the most appropriate cut-off (or whether treatment should be given immediately upon diagnosis) for any given setting still under debate. As such, there is interest in determining the proportion of patients that will cross any given threshold and initiate ART as a function of time from seroconversion, as this will impact on clinical practice and on the cost of different healthcare strategies. Lodi *et al.* [28] used random slopes linear mixed models fitted to over 175 000 CD4 measurements from the CASCADE cohort (including the data analysed in the present study) to predict the proportion of untreated patients reaching thresholds of <500, <350 and <200 cells/μL with respect to time from seroconversion, reflecting the cut-offs used in various versions of official guidelines. In this analysis, the distribution of subject-specific slopes was used to estimate the proportion of patients with 'true' CD4 count below each threshold value.

Using their fitted linear mixed model including a Brownian motion component, Babiker *et al.* [3] investigated the proportion of patients reaching a threshold of <350 cells/μL through simulation of sets of longitudinal measurements for tens of thousands of individuals. This approach has the advantage of allowing realistic assessment of the characteristics of a cohort in practice, and several regimes for the scheduling of measurements and initiation of ART were considered in their simulations. However, the predictions made from the simulations were not directly compared to those that would have been obtained using a normal random slopes model. We have therefore performed a similar analysis based on several of the fitted models in order to investigate this.

Simulated cohorts of individuals were generated based on three MVN models as follows: the random slopes model, the Brownian motion model and the fractional Brownian motion model (with the latter two

Statistics
in Medicine

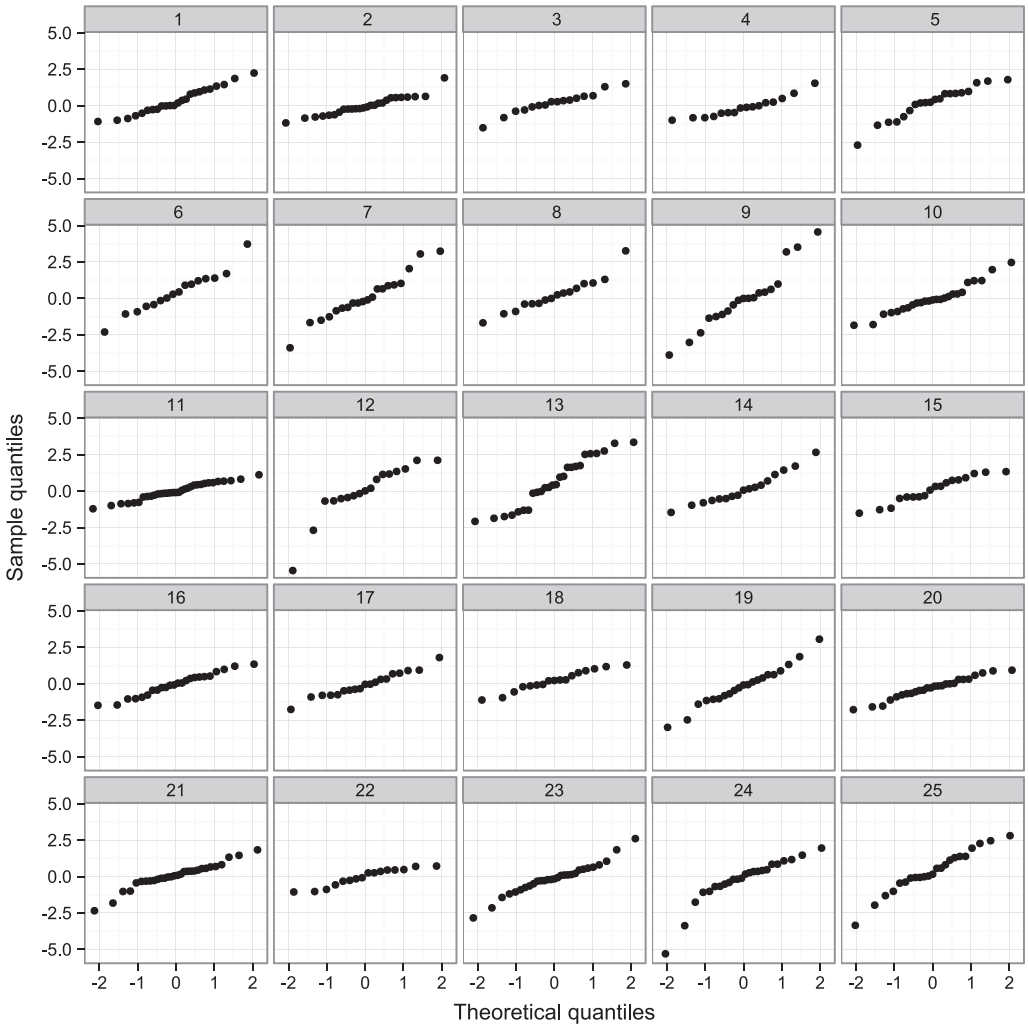

Figure 4. Quantile–quantile plots for the residuals under the 'random slopes + fractional Brownian motion + measurement error' multivariate-t model of 25 randomly selected individuals with greater than 15 observations. The residuals for individual patients have been transformed by the inverse of the Cholesky decomposition of their estimated scale matrix ($\hat{\mathbf{V}}_i$) without any correction for the unobserved scale variable τ_i. Theoretical quantiles in each case are those from the standard normal distribution.

also including a random slopes structure and all including measurement error). In addition, a cohort was generated using the fitted multivariate-t fractional Brownian motion model (again, including a random slopes structure and measurement error). For each of these models, data for five million individual patients were simulated based on scheduled measurements being obtained every 4 months for up to 10 years. Data were also generated for measurements 1 month after the scheduled observation in each case for use in the analysis, corresponding to a confirmatory test. CD4 thresholds of <500, <350 and <200 cells/μL for ART initiation were investigated. If a scheduled measurement was observed below a given threshold, then the value 1 month later was assessed to mimic the conduct of an additional confirmatory test as commonly performed in clinical practice. The patient was considered to initiate ART if this second value was also below the threshold.

The results of the analysis of the simulated cohorts are presented in Figure 6. The differences in predictions made by each of the fitted models are large enough to have practical implications particularly within a public health or health economics context; for example, using the <500 cells/μL threshold, the proportion of patients on ART 2 years after seroconversion is predicted to be 57% by the normal random

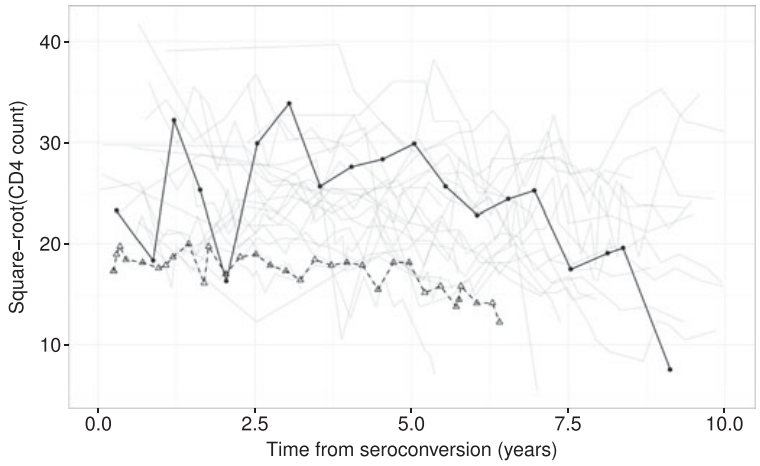

Figure 5. Line plot of the square-root transformed CD4 counts observed in the random sample of 25 patients with greater than 15 observations (as in Figure 4). The observations for the patients corresponding to Plot 9 (solid black line, filled circles for individual data points) and Plot 11 (dashed black line, open triangles for individual data points) in Figure 4 are made prominent.

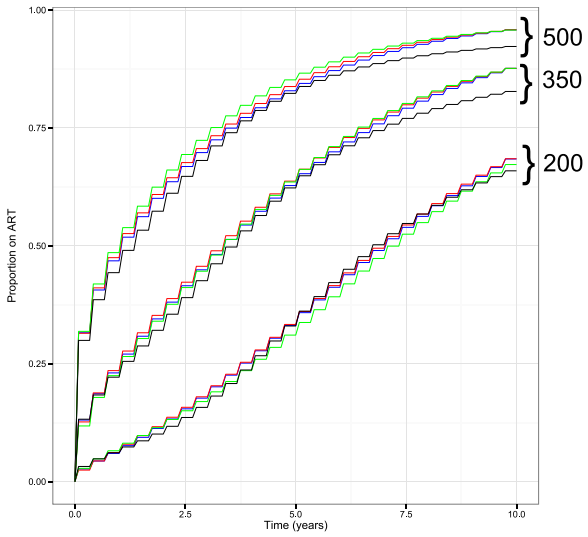

Figure 6. The proportion of HIV-positive patients predicted to have initiated antiretroviral therapy (ART) as a function of time since seroconversion, based on simulation from the fitted normal random slopes model (black line), Brownian motion model (blue line) and fractional Brownian motion model (red line) and the multivariate-t fractional Brownian motion model (green line). Results are presented using CD4 thresholds for ART initiation of <500, <350 and <200 cells/μL, as indicated at top right of the graph. Simulations are based on CD4 measurements being obtained every 4 months, with initiation of ART conditional on an additional observation below the cut-off concerned 1 month after the 'scheduled' measurement.

slopes model and to be 62% by the multivariate-t model with fractional Brownian motion. The planning of the Strategic Timing of AntiRetroviral Treatment trial described by Babiker *et al.* [3] made use of predictions of the proportion of patients initiating ART at the 350 cells/μL threshold for which we found only small differences between each of the models that included a stochastic process component (i.e. excluding the standard random slopes model). It is interesting to note that for the 500 and 350 cells/μL cut-offs, the predictions for the models incorporating stochastic process components converge as time increases towards 10 years, separate to the lower predictions made by the standard random slopes model.

Statistics
in Medicine

5.4. Simulation study: parameter bias in slope estimates

One interesting feature of the various models fitted to the CASCADE pre-ART CD4 data is that the mean slope (β_1) of CD4 decline is substantially less negative for the linear mixed models that include standard or fractional Brownian motion components (both –1.15) than for the random slopes model (–1.36). The estimated slopes for the equivalent multivariate-t models were also less steep in each case (Tables I and II). We performed a simulation study to assess the impact of model choice and missing data patterns on this difference, which may indicate apparent bias from the use of simpler models.

It follows from Liang and Zeger [29] that a linear mixed model analysis of longitudinal data will give consistent estimates of the fixed effects given that either there is no missing data or that data is 'missing completely at random' (MCAR) (following the terminology of Rubin [15]). This also requires the structure of the fixed effects to be correctly specified in the model, but not the exact distribution of observations or covariance between them. Hence, it seems that the substantial differences in slope estimates between different models fitted to pre-ART CD4 data are due to the presence of missing data for which the missingness is not MCAR, although the framework for missing data terminology is less clear for highly unbalanced datasets without a consistent observation schedule.

It is often postulated that the missingness of observations in pre-ART datasets can be treated as MAR, that is, that it is independent of the unobserved outcome variable conditional on the observed values of the outcome variable and other covariates included in the model, and that as such the missingness can be ignored under maximum likelihood estimation such as the use of linear mixed models. The MAR assumption is plausible if patients are thought to mainly drop out of the dataset upon initiation of ART, and if this is entirely dependent on their observed CD4 counts. However, the beneficial properties of maximum likelihood-based inference (i.e. consistency and asymptotic normality and efficiency of estimates) with respect to MAR data are dependent on a correctly specified model for the likelihood. The fact that adding stochastic process components and/or generalising to a multivariate-t distribution leads to a very substantial improvement in BIC indicates that the standard random slopes model does not correctly describe the covariance structure or probability model for pre-ART CD4 data.

To further investigate bias in parameter estimates resulting from overly simplistic models, the best-fitting model (i.e. multivariate-t with fractional Brownian motion) was assumed to be 'correct' and cohorts of patient data simulated from it. CD4 cell count observations were generated from 0 to 5 years, for groups of either 100 or 200 patients and with an annual observation frequency of 1 or 3; five hundred cohorts were generated for each combination. For each simulated cohort, models were first fitted to the complete uncensored data (although this would include impossible negative values), and subsequently to the data following censoring corresponding to ART initiation at CD4 cut-off values of 200, 350 and 500 cells/μL. The 'correct' multivariate-t model and three normal linear mixed models (the random slopes model, the Brownian motion model and the fractional Brownian motion model) were applied to each simulated cohort under each condition. For the analyses involving censoring, additional confirmatory measurements were generated 1 month after the 'scheduled' observations; these were only considered to be observed when the scheduled measurement was below the cut-off value, and the patient was only censored when the confirmatory value was also below the cut-off. The censored datasets could therefore be considered to correspond to observations being MAR but not MCAR. As the MAR condition holds for any possible realisation, this scenario meets the 'everywhere MAR' definition provided by Seaman *et al.* [30], allowing valid frequentist likelihood inference. Model fitting was considered to have failed when parameter estimates were not returned or when the covariance matrix of parameter estimates was not positive-definite.

Limited bias was observed in the estimation of the intercept term when using simplified models and so the results of this analysis are only presented for estimation of the slope parameter β_1. Bias in the estimation of β_1 and the coverage of 95% confidence intervals for this parameter are presented in Table III. As expected, a lack of bias (or only very minimal bias) and appropriate coverage intervals were observed when the correctly specified model was fitted, even in the presence of censoring. Interestingly, no or only minimal bias was observed when the equivalent normal linear mixed model (including a fractional Brownian motion component) was used. Linear mixed models including a Brownian motion component showed some downward bias in the presence of censoring, with this most marked when censoring was applied using the CD4 cut-off of 500 cells/μL. Substantial downward biases and poor coverage of confidence intervals were observed when a standard random slopes linear mixed model was applied in the presence of censoring, with the degree of bias clearly linked to the extent of censoring.

Statistics in Medicine

Table III. Summary of the results of simulation analyses to assess bias in the estimate of mean slope (β_1) when models that are simpler than the data-generating process are applied in the presence of 'missing at random' censoring.

			RS+ME		RS+BM+ME		RS+fBM+ME		MVT: RS+fBM+ME	
	Prop. cens. (median (IQR))	*n* obs. (median (IQR))	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)	Failed (%)	β_1 bias (coverage)
N=100, freq=1										
Uncensored	0 (0–0)	600 (600–600)	0	0.013 (93.8)	0.0	0.012 (94.6)	2.6	0.011 (94.0)	1.2	0.012 (96.0)
ART200	21 (18–24)	573 (567–579)	0	−0.100 (88.2)	0.4	−0.021 (96.8)	1.4	0.000 (95.3)	0.2	0.008 (95.4)
ART350	51 (48–55)	505 (493–515)	0	−0.244 (73.0)	1.2	−0.071 (95.1)	2.6	−0.001 (96.5)	1.4	0.000 (95.9)
ART500	77 (75–80)	400 (388–413)	0	−0.384 (72.0)	2.4	−0.146 (94.9)	5.2	−0.022 (94.3)	1.0	−0.019 (93.5)
N=100, freq=3										
Uncensored	0 (0–0)	1600 (1600–1600)	0	0.000 (95.4)	0.0	−0.001 (97.0)	4.0	−0.002 (96.0)	2.0	−0.001 (96.7)
ART200	30 (27–34)	1414 (1386–1441)	0	−0.161 (84.4)	0.8	−0.054 (95.8)	4.8	−0.009 (96.0)	3.4	−0.004 (96.1)
ART350	63 (60–66)	1095 (1060–1131)	0	−0.289 (71.8)	0.4	−0.125 (95.0)	7.0	−0.002 (97.8)	2.6	−0.002 (97.9)
ART500	85 (82–87)	724 (687–761)	0	−0.322 (84.6)	2.2	−0.270 (92.0)	9.8	−0.025 (94.5)	3.0	−0.027 (94.2)
N=200, freq=1										
Uncensored	0 (0–0)	1200 (1200–1200)	0	0.004 (94.6)	0.0	0.005 (94.4)	1.0	0.004 (93.7)	0.0	0.001 (94.4)
ART200	22 (20–24)	1144 (1136–1152)	0	−0.106 (82.4)	0.2	−0.019 (95.4)	1.2	0.000 (94.1)	0.6	0.000 (95.0)
ART350	52 (49–54)	1008 (993–1023)	0	−0.234 (56.0)	0.4	−0.057 (95.8)	2.4	0.007 (95.3)	0.8	0.005 (95.2)
ART500	78 (76–80)	799 (780–816)	0	−0.360 (56.2)	0.4	−0.118 (92.6)	6.8	0.007 (94.6)	1.6	−0.001 (94.5)
N=200, freq=3										
Uncensored	0 (0–0)	3200 (3200–3200)	0	−0.002 (94.6)	0.0	−0.001 (95.6)	2.4	−0.001 (94.7)	0.4	0.000 (92.2)
ART200	30 (28–32)	2840 (2804–2874)	0	−0.161 (66.6)	0.2	−0.054 (93.0)	2.2	−0.012 (94.5)	0.8	−0.005 (92.1)
ART350	62 (60–65)	2197 (2142–2244)	0	−0.300 (44.6)	0.4	−0.127 (88.0)	4.4	−0.014 (96.2)	2.4	−0.007 (96.1)
ART500	84 (83–86)	1454 (1404–1508)	0	−0.337 (70.2)	1.0	−0.243 (86.3)	12.4	−0.004 (93.8)	5.8	−0.017 (93.8)

Bias is calculated as the mean estimate of β_1 minus the true value, and is presented with coverage of nominal 95% confidence intervals in parentheses. For each combination of number of simulated patients (N) and annual frequency of observation (freq), 500 cohorts were generated and analysed under different censoring regimes, corresponding to treatment initiation at CD4 cut-offs of 200 (ART200), 350 (ART350) or 500 (ART500). All cohorts were simulated with a follow-up of 5 years, including an observation at time zero for each patient. Data were generated according to a multivariate-t distribution (MVT) incorporating a fractional Brownian motion (fBM) process and measurement error (ME) and, alongside a model of the correct form, normal linear mixed models were fit with a random slopes (RS) structure alone and with RS in combination with Brownian motion (BM) and fBM processes. Model fitting was considered to have failed when parameter estimates were not returned or when the covariance matrix of parameter estimates was not positive-definite. IQR, interquartile range; *n* obs., total observations included in analysis per simulated cohort; Prop. cens., proportion of patients in simulated cohort subject to censoring before 5 years.

Statistics
in Medicine

A summary of the standard deviations of point estimates for the mean slope and the average estimated standard error for this parameter in the simulations is also provided as supplementary material (Table S1). There were not large discrepancies between these two measures of the standard error. The mean slope estimates from the correctly defined model showed slightly lower variance than the estimates from the incorrectly defined models in any given situation, but the scale of these differences seems relatively small compared with the large biases observed.

The differences in slope estimates observed between models under the censoring conditions in this simulation study correspond to the differences observed between the models when applied to the real dataset. This provides supporting evidence that special attention should be given to the probability model used, and in particular the covariance structure, when analysing a dataset for which there are substantial missing data that are not MCAR. These simulations imply that an analysis using a wrongly specified model might incorrectly indicate differences between two groups in their average rate of decline if they have been subject to different censoring mechanisms. We carried out an additional investigation in which two groups of either 100 or 200 patients each were simulated with three observations per year, with the first group subject to censoring at the '200 cut-off' whilst the '500 cut-off' was applied for the second group. Other details of the simulation and model fitting were as previously described, but two additional 'fixed effects' parameters were added to the models to allow the mean intercept (δ_0) and slope (δ_1) of the second group to differ from the first group (with the true value of these parameters being zero). These simulations confirmed that bias could occur in the estimation of between-group differences in slope within a single model (estimated bias for random slopes model with 200 patients per group: –0.163, Table S2).

6. Discussion

In this study, we have further developed the statistical modelling of longitudinal biomarker data, through application to pre-treatment CD4 counts in patients with HIV, in which we have shown that the combination of a fractional Brownian motion component and generalisation of the normal linear mixed model to a multivariate-t distribution leads to substantial improvements in model fit. This novel combination of model features provides additional information regarding the between-patient and within-patient variability in observations over time. Evidence is provided for the appropriateness of using a multivariate-t distribution in the studied dataset through evaluation of novel diagnostic plots. Furthermore, simulation studies are presented to demonstrate the impact of model choice on cohort-level predictions and on bias in mean slope estimates when data are MAR.

The presence of non-stationary stochastic process components in models for longitudinal data implies that the progress of the state of the underlying biological system for each individual does not follow a deterministic relationship with time, but rather follows an unpredictable path. This finding seems intuitive in the context of the extremely complex interactions between viral replication and immune system response that influence the CD4 count series that are observed in HIV-positive patients. When using a fractional Brownian motion component, the H values obtained were less than 0.5, indicating that the process is erratic but displays some reversion towards an underlying mean. The estimates of the degrees of freedom parameter for the multivariate-t models of between five and six indicate substantial between-patient differences in variability over time.

Through simulations based on generating data from the more complex fitted model, it is demonstrated that the use of a normal random slopes model is associated with substantial bias in the estimation of the mean slope parameter in the presence of censoring, with the degree of bias strongly dependent on the choice of censoring regime. This is important, as estimates of this parameter are often used as a proxy for rate of decline in health and compared between groups. As initiation of ART is usually dependent on observed CD4 values, the MAR condition is often invoked to argue that likelihood-based model estimation will lead to valid inferences, but this only holds conditional on the correct specification of the likelihood model. It can therefore be argued that in this context, greater effort should be made to make use of statistical models that adequately describe the distributional and covariance patterns present in the data.

Diagnostic Q–Q plots of Cholesky-transformed marginal residuals from MVN models fitted to square-root CD4 counts show very heavy tails, indicating clear violation of the modelling assumptions. We have demonstrated that the use of a multivariate-t distribution in combination with a non-stationary stochastic process component leads to a very substantial improvement in BIC with diagnostic Q–Q plots that only indicate relatively mild violation of the model's assumptions. Such models can be fit efficiently and to large datasets using the open-source ADMB software [22], with this task made easier by the fact that the

log-likelihood of the multivariate-t distribution is available in closed form. It would be of interest to investigate whether models comprised of different combinations of multivariate-t and normal distributions could provide a better fit to the data; such models have been previously discussed by Song *et al.* [31]. For example, it may be considered more biologically plausible to fit a statistical model in which the variability of the stochastic process component differs between individuals (i.e. follows a multivariate-t distribution) but the random effects and measurement error terms do not (i.e. they follow normal distributions). For such models, the likelihood function is not available in closed form, making the computations required for parameter estimation substantially more complex. The implementation and evaluation of such models will be the topic of further research.

Normal linear mixed models including simple or fractional Brownian motion processes cannot be fitted using standard routines in existing statistical software packages, and this is probably responsible for the fact that they have not been widely adopted in practice (at least in the setting of HIV-research). However, an R package (covBM) that will allow the implementation of such models is under development by the authors. Most software does not offer any standard function for fitting mixed models based on the multivariate-t distribution, although an R package 'tlmec' does exist for fitting models generalised from a normal model with independent error terms of constant variance [12].

Our research has been focused on CD4 cell counts in HIV-positive patients, but the modelling framework developed may be of use for the analysis of longitudinal data in other biomedical applications. For example, Diggle *et al.* recently described the use of an extended linear mixed model including another non-stationary stochastic process, integrated Brownian motion, for the analysis of estimated glomerular filtration rates in patients at risk for renal failure [32]. The authors provide plots of 'Cholesky-standardised' residuals produced from the application of the model, which show very heavy tails. The multivariate t-distribution implies differences in the volatility of observations between patients, which may by useful in planning and interpreting the monitoring of biomarkers in HIV and other disease areas.

Whilst it is arguably impossible to claim that any statistical model exactly represents the data-generating mechanism under investigation, it seems that both the addition of stochastic process components to the standard linear mixed model and the use of a multivariate-t distribution can be used to gain a greater understanding of longitudinal biomedical data. Such models provide greater flexibility, but require only a small number of additional parameters and follow a model specification that can be interpreted in terms of the underlying biological process; as such, the potential gains in inference and understanding through their use are likely to greatly outweigh any drawbacks of increased model complexity. There is therefore a motivation to develop more efficient methods of fitting such models and to make these more widely available.

Acknowledgements

We would like to thank the CASCADE investigators. The CASCADE research project has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under EuroCoord grant agreement no. 260694. O. T. S. is supported by a Medical Research Council PhD Studentship.

References

1. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
2. Taylor JMG, Cumberland WG, Sy J P. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* 1994; **89**:727–736.
3. Babiker AG, Emery S, Fätkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, Vjechaj MJ. INSIGHT START Study Group. Considerations in the rationale, design and methods of the Strategic Timing of AntiRetroviral Treatment (START) study. *Clinical Trials* 2013; **10**(1 Suppl):S5–S36.
4. Taylor JMG, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* 1998; **17**:2381–2394.
5. Wolbers M, Babiker A, Sabin C, Young J, Dorrucci M, Chêne G, Mussini C, Porter K, Bucher HC. CASCADE Collaboration Members. Pre-treatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy—the CASCADE collaboration: a collaboration of 23 cohort studies. *PLoS Medicine* 2010; **7**:e1000239.
6. Mandelbrot B, van Ness JW. Fractional Brownian motions, fractional noises and applications. *SIAM Review* 1968; **10**: 422–437.
7. Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association* 1989; **84**:881–896.
8. Welsh AH, Richardson AM. Approaches to the robust estimation of mixed models. In *Handbook of Statistics (Vol. 15)*, Maddala GS, Rao CR (eds). Elsevier Science: Amsterdam, 1997; 343–384.

Statistics
in Medicine

9. Pinheiro JC, Liu C, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate-t distribution. *Journal of Computational and Graphical Statistics* 2001; **10**:249–276.
10. Wang W-L, Fan T-H. Estimation in multivariate-t linear mixed models for multiple longitudinal data. *Statistica Sinica* 2011; **21**:1857–1880.
11. Wang W-L, Fan T-H. Bayesian analysis of multivariate-t linear mixed models using a combination of IBF and Gibbs samplers. *Journal of Multivariate Analysis* 2012; **105**:300–310.
12. Matos LA, Prates MO, Chen M-H, Lachos V H. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica* 2013; **23**:1323–1345.
13. Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis* 1997; **23**:541–556.
14. Jacqmin-Gadda H, Sibillot S, Proust C, Molina J-M, Thiébaut R. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis* 2007; **51**:5142–5154.
15. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
16. Gurka MJ, Edwards LJ, Muller KE. Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine* 2011; **30**:2696–707.
17. *CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe) Collaboration: Participating cohorts*. Available at: http://www.cascade-collaboration.org [accessed 10 July 2015].
18. Grimmett G, Stirzaker D. *Probability and Random Processes* 3rd ed. Oxford University Press: Oxford, 2001; 370.
19. Pinheiro J, Bates D. *Mixed-effects Models in S and S-PLUS*. Springer: New York, 2000.
20. Kotz S, Nadarajah S. *Multivariate t-distributions and Their Applications*. Cambridge University Press: Cambridge, 2004.
21. Griewank A, Walther A. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation 2nd edn*. Society for Industrial and Applied Mathematics: Philadelphia, 2008.
22. Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson E, Maunder M N, Nielsen A, Sibert J. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* 2012; **27**:233–249.
23. Bolker B, Skaug H, Laake J. R package *R2admb: ADMB to R interface functions* (version 0.7.11). http://CRAN.R-project.org/package=R2admb [accessed 10 July 2015].
24. Fitzmaurice G, Laird N, Ware J. Residual analyses and diagnostics. In *Applied Longitudinal Analysis*. Wiley: Hoboken; 2004.
25. Verbeke G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 1996; **91**:217–221.
26. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York, 2009.
27. Cavanaugh JE, Neath AA. Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics—Theory and Methods* 1999; **28**:49–66.
28. Lodi S, Phillips A, Touloumi G, Geskus R, Meyer L, Thiébaut R, Pantazis N, Amo JD, Johnson AM, Babiker A, Porter K. CASCADE Collaboration in EuroCoord. Time from human immunodeficiency virus seroconversion to reaching CD4+ cell count thresholds <200, <350, and <500 cells/mm^3: assessment of need following changes in treatment guidelines. *Clinical Infectious Diseases* 2011; **53**:817–825.
29. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
30. Seaman S, Galat J, Jackson D, Carlin J. What is meant by missing at random?. *Statistical Science* 2013; **28**:257–268.
31. Song PX-K, Zhang P, Qu A. Maximum likelihood inference in robust linear mixed-effects models using multivariate-t distributions. *Statistica Sinica* 2007; **17**:929–943.
32. Diggle PJ, Sousa I, Asar Ö. Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics* 2015; **16**:522–536.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Appendix D BMC Med Res Meth paper

Stirrup *et al. BMC Medical Research Methodology* (2016) 16:121
DOI 10.1186/s12874-016-0187-2

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Combined models for pre- and post-treatment longitudinal biomarker data: an application to CD4 counts in HIV-patients

Oliver T. Stirrup*, Abdel G. Babiker and Andrew J. Copas

Abstract

Background: There has been some debate in the literature as to whether baseline values of a measurement of interest at treatment initiation should be treated as an outcome variable as part of a model for longitudinal change or instead used as a predictive variable with respect to the response to treatment. We develop a new approach that involves a combined statistical model for all pre- and post-treatment observations of the biomarker of interest, in which the characteristics of response to treatment are treated as a function of the 'true' value of the biomarker at treatment initiation.

Methods: The modelling strategy developed is applied to a dataset of CD4 counts from patients in the UK Register of HIV Seroconverters (UKR) cohort who initiated highly active antiretroviral therapy (HAART). The post-HAART recovery in CD4 counts for each individual is modelled as following an asymptotic curve in which the speed of response to treatment and long-term maximum are functions of the 'true' underlying CD4 count at initiation of HAART and the time elapsed since seroconversion. Following previous research in this field, the models developed incorporate non-stationary stochastic process components, and the possibility of between-patient differences in variability over time was also considered.

Results: A variety of novel models were successfully fitted to the UKR dataset. These provide reinforcing evidence for findings that have previously been reported in the literature, in particular that there is a strong positive relationship between CD4 count at initiation of HAART and the long-term maximum in each patient, but also reveal potentially important features of the data that would not have been easily identified by other methods of analysis.

Conclusion: Our proposed methodology provides a unified framework for the analysis of pre- and post-treatment longitudinal biomarker data that will be useful for epidemiological investigations and simulations in this context. The approach developed allows use of all relevant data from observational cohorts in which many patients are missing pre-treatment measurements and in which the timing and number of observations vary widely between patients.

Keywords: CD4, HAART, HIV, Longitudinal data, Mixed effects models, Statistical methodology

Background

In medical research, there is often interest in evaluating response to treatment conditional on the baseline value at initiation of the biomarker under investigation. In the setting of randomised controlled trials (RCTs), designed primarily to assess the difference between treatment conditions, some authors have argued that optimal efficiency is gained by treating the baseline measurement as an outcome variable within a parametric model [1, 2], whilst Senn has argued that conditioning estimation of treatment effect on the baseline observation through the use of ANCOVA is preferable in most trial situations [3] and Kenward et al. demonstrated that with correct adjustments for sample size the two approaches have nearly identical properties [4]. However, both of these approaches can be problematic when applied to the estimation of response to treatment using longitudinal observational datasets, in which the timing and choice of

*Correspondence: oliver.stirrup.13@ucl.ac.uk
MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, 125 Kingsway, WC2B 6NH London, UK

treatment have not been randomised and in which baseline observations immediately prior to treatment may not be available for all patients. Furthermore, there is often substantial interest in the influence of the baseline value of the biomarker in itself in determining the level of response to treatment, rather than just using this to provide a better estimate of the differences between treatment choices. In this article we describe the development of flexible parametric models for this situation, providing a combined analysis of pre- and post-treatment data in which the response of the biomarker to treatment is dependent on a 'true' baseline value that is not directly observed; this combines elements of both previous approaches in that the pre-treatment data are modelled as 'response variables', but the trajectory of the biomarker after treatment initiation can also be modelled using flexible functions of the baseline value. The models developed are applied to CD4 cell counts in human immunodeficiency virus (HIV)-positive patients who initiate highly active antiretroviral therapy (HAART).

CD4 cells are a type of white blood cell for which counts are monitored over time both before and after treatment initiation in HIV patients in order to evaluate the progress of the disease and state of the immune system. Although the CD4 counts within an individual can vary erratically over time, on average the counts decline steadily from normal levels following HIV infection and then in most cases recover towards normal levels following initiation of HAART. Over the last 20 years, effective regimens of HAART have been developed for the treatment of HIV, allowing long-term management of the condition and greatly improving the life expectancy and quality of life of affected individuals, at least for those with the condition diagnosed in a resource-rich country. Until recently, clinical guidelines regarding the initiation of treatment varied between countries. In the USA, the Health and Human Services Panel on Antiretroviral Guidelines for Adults and Adolescents have for a number of years recommended immediate initiation of HAART for most patients newly diagnosed with HIV [5], whereas in Europe guidelines recommended monitoring of CD4 in most patients, with treatment initiated once this dropped below 350 [6]. However, a recent RCT has provided definitive evidence of the benefit of immediate initiation of HAART on diagnosis of HIV [7], leading to a shift in clinical guidelines towards early treatment initiation in all well-resourced countries, including the UK [8].

In observational datasets, the timing of recorded CD4 measurements can be highly variable between patients. In much of the existing literature about the long-term response of CD4 counts to HAART, the investigators have avoided any associated complications in their analyses by converting the available data into a set of discrete time points, typically corresponding to annual or 6-monthly observations. This has been done by linear interpolation (Kaufmann et al.) [9], selecting only the observation closest to the chosen time point (Moore and Keruly) [10] or taking the mean measurement within intervals (Lok et al.) [11]. Each of these studies included an analysis stratified by intervals of baseline CD4 count and, although the statistical methodology varied between studies, each found that higher baseline CD4 counts were associated with higher values after several years of HAART. A study by Le et al. suggested that the long-term response to HAART in HIV-positive patients is improved if it is initiated within the first few months after infection, with this effect independent of the CD4 count at baseline [12]. This analysis also relied on stratification of patients into groups.

We now also know that early treatment of HIV leads to a substantial reduction in the occurrence of both acquired immune deficiency syndrome (AIDS)-defining conditions and serious non-AIDS events [7], but there nonetheless remains clinical interest in understanding the factors that are predictive of the recovery in CD4 counts upon HAART initiation as for many patients there is a substantial delay between infection and diagnosis and suboptimal CD4 recovery remains a concern for patients and clinicians [13]. The principal aim of this research is the development of a flexible parametric framework for the combined modelling of pre- and post-treatment CD4 data in HIV positive individuals. This is motivated by the clinical interest in investigating the factors that determine the characteristics of long-term response to HAART, in particular the influences of baseline CD4 count and the time elapsed from infection to treatment initiation. However, the modelling strategy developed could also be used in other settings in which a biomarker is monitored prior to some treatment initiation or clinical intervention.

The modelling strategy described in this article represents a flexible extension of established non-linear mixed effects models, fitted through maximum likelihood estimation based on all observed data using time as a continuous variable. As well as allowing inclusion of all available data in its original format (other than global transformations for normalisation) and the combined assessment of multiple predictive factors, the approach will have the advantage that the characteristics of CD4 trajectories of individual patients over time will be quantified, creating a complete framework for epidemiological simulations or patient-specific predictions, whereas previously this has been done using separate models for pre- and post-treatment data [14]. The models developed are applied to CD4 data from the UK Register of HIV Seroconverters cohort [15]. Following previous work on the modelling of pre-treatment CD4 counts [16], we also incorporate stochastic process components and between-patient differences in variability over time into the models

developed. This is done with the aim of defining models that are as realistic as possible in representing the structure of the biological measurements under investigation, which is particularly important when considering analyses for datasets in which missing data and irregular follow-up times are a substantial concern.

Methods

Dataset

The UK Register of HIV Seroconverters is an observational cohort study of patients whose date of infection can be reliably estimated [15]. The UK Register of HIV Seroconverters has research ethics approval (MRC MREC: 04/Q2707/155). Recruitment to the cohort began in 1994, but, as we are interested in modelling the response to modern HAART regimens, we restrict our analysis to patients with an estimated date of HIV-1 seroconversion during or after 2003. Patients who started a suboptimal regimen of antiretroviral drugs prior to HAART were excluded, as were patients without at least one post-treatment CD4 count recorded. Patients without any pre-treatment CD4 counts were, however, included in the analysis. HAART is defined by a regimen of at least three antiretroviral drugs from at least two different classes (unless abacavir or tenofovir is used in a regimen with three nucleoside analog reverse-transcriptase inhibitors (NRTIs)).

Application of these conditions resulted in a study population of 852 patients, with a total of 5805 pre-HAART and 7302 post-HAART CD4 observations recorded. The median (interquartile range (IQR)) number of pre-HAART CD4 counts was 5 (3–10), whilst that for post-HAART observations was 6 (3–12). There were a total of 39 patients without any pre-HAART CD4 counts recorded. The median (IQR) time from estimated date of seroconversion to initiation of HAART was 1.3 (0.6–2.8) years, with 192 patients starting HAART within 6 months and 149 starting between 6 months and 1 year from seroconversion.

CD4 cell counts are measured as cells per microlitre, and we followed established practice in modelling the counts on a square-root scale [14, 16]. For the pre-treatment part of the model, time is measured in years from date of HIV seroconversion, whilst for the post-treatment part of the model it is measured in years from HAART initiation. We have censored patients at recorded interruption of HAART (including switch to suboptimal treatment) for more than 1 week, but have not censored according to viral load status or change to HAART regimen. Treatment interruption was recorded in 124 (14.6 %) patients, and there were a total of seven deaths recorded (three of which occurred after censoring due to interruption of HAART). Data from a random subset of 100 of the patients analysed are shown in Fig. 1.

Baseline state as a latent variable

It can be shown that in situations in which the initiation of treatment is conditional on a biomarker that is monitored over time, and which is measured with error, the observed value of the biomarker at the start of treatment provides a biased estimate of the 'true' underlying value [14]. This presents a problem when attempting to model treatment

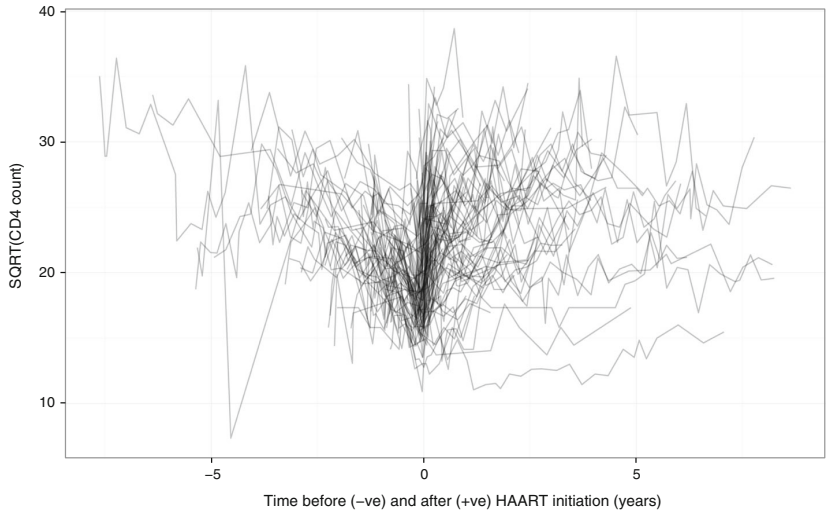

Fig. 1 'Spaghetti plot' of the square root of CD4 counts from a random sample of 100 patients. Patients are from the UK Register of HIV Seroconverters dataset. Lines are semi-translucent to aid visualisation. Time has been centred at the time of highly active antiretroviral therapy (HAART) initiation for each patient

response conditional on the baseline value. We propose that one option in this situation is to build a combined model for both the pre- and post-treatment data, allowing the response to treatment to be conditional on all available pre-treatment data rather than on just a single baseline value. Such an approach would also have the advantage that patients could be included for whom no measurement close to the start of treatment had been obtained. Additionally, fewer assumptions regarding the marginal distribution of 'true' baseline values of any given population would be required. For example, such an approach could appropriately deal with a set of distinct treatment initiation guidelines applied across different periods of time or sub-populations, which might lead to a multimodal distribution of baseline values in the total study population, whereas a standard mixed model approach would generally assume the observed baseline values to follow a normal distribution for the population as a whole.

Any linear mixed effects model implies a marginal multivariate normal distribution [17] (*MVN*), for which the log-likelihood function can be expressed in closed form. However, this is not true (except for some special cases) for non-linear mixed effects models [18]. For such models, numerical integration or analytical approximation of the log-likelihood is required at each iteration of any optimisation algorithm [19]. Among the available options, adaptive Gauss–Hermite quadrature is particularly attractive as an increasing number of quadrature points can be used for each random effect to ensure that the log-likelihood is evaluated to an adequate degree of accuracy. However, if more than one random effect is included in the model for each independent individual in the analysis then the number of points that need to be evaluated in the adaptive Gauss–Hermite quadrature algorithm increases exponentially with the number of random effects terms per individual. As such, adaptive Gauss–Hermite quadrature is not generally used when there are more than two or three random effects terms defined in a model, and the computational requirements to attain high accuracy in calculation of the log-likelihood function are lowest when there is only one random effect term per individual.

Because of the computational issues described, to undertake the combined modelling of pre- and post-treatment CD4 data we focus on the use of non-linear latent variable models that require numerical integration only over the unobserved 'true' CD4 count at treatment initiation (which we will term u). The rationale of this approach is that it will allow adequate flexibility in model structure without increasing the computational requirements to a level that will prevent application to the dataset available. In order to achieve this, we will specify linear mixed models for the pre-treatment data ($\boldsymbol{y}_{pre}$) and non-linear models for the post-treatment data ($\boldsymbol{y}_{post}$), conditioned on the 'true' baseline CD4 count, that *are* linear in any other random effects terms (allowing a closed form expression for each of these two parts of the model). Under such a scheme, the likelihood function for the combined pre- and post-treatment data for each individual can therefore be expressed as:

$$\begin{aligned} f\left(\boldsymbol{y}_{pre}, \boldsymbol{y}_{post}\right) &= \int_{-\infty}^{\infty} f_{pre,post,u}\left(\boldsymbol{y}_{pre}, \boldsymbol{y}_{post}, u\right) du \\ &= \int_{-\infty}^{\infty} f_{pre}\left(\boldsymbol{y}_{pre}\right) f_{post,u}\left(\boldsymbol{y}_{post}, u|\boldsymbol{y}_{pre}\right) du \\ &= \int_{-\infty}^{\infty} f_{pre}\left(\boldsymbol{y}_{pre}\right) f_{post}\left(\boldsymbol{y}_{post}|\boldsymbol{y}_{pre}, u\right) f_{u}\left(u|\boldsymbol{y}_{pre}\right) du. \end{aligned}$$

For simplicity above, we suppress notation to indicate that each element of the likelihood function is dependent on model parameters. However, we now consider calculation of the likelihood function dependent on the values of a parameter vector relating to the pre-treatment part of the model '$\boldsymbol{\theta}_{pre}$', a parameter vector relating to the post-treatment part of the model '$\boldsymbol{\theta}_{post}$' and a shared measurement error variance parameter 'σ^2'. If we assume that the post-treatment response depends on the pre-treatment data only though the true baseline value at treatment initiation, i.e. that $\boldsymbol{y}_{post}$ is independent of $\boldsymbol{y}_{pre}$ given u, then we may write:

$$\begin{aligned} f\left(\boldsymbol{y}_{pre}, \boldsymbol{y}_{post}\right) = \int_{-\infty}^{\infty} & f_{pre}\left(\boldsymbol{y}_{pre}|\boldsymbol{\theta}_{pre}, \sigma^2\right) \\ & f_{post}\left(\boldsymbol{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right) f_{u}\left(u|\boldsymbol{y}_{pre}, \boldsymbol{\theta}_{pre}, \sigma^2\right) du. \end{aligned}$$

This follows a similar form to the likelihood expression for standard random effects models but here the distribution of the latent variable u, which is integrated out to obtain the marginal likelihood, is conditioned on the pre-treatment data for each individual rather than following a pre-specified distribution across the population. For those patients in whom no pre-treatment observations were obtained, the likelihood contribution can be calculated solely for the post-treatment observations:

$$f\left(\boldsymbol{y}_{post}\right) = \int_{-\infty}^{\infty} f_{post}\left(\boldsymbol{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right) f_{u}\left(u|\boldsymbol{\theta}_{pre}, \sigma^2\right) du.$$

It should be pointed out here that, in practice, optimisation algorithms to obtain maximum likelihood estimates operate on the log-likelihood scale. In Subsection "Differences in variability between patients", we describe the addition of two further latent variables to the model for each individual in order to allow for between-patient differences in variability over time.

Pre-treatment model structure

At present we consider only linear mixed model formulations for the likelihood of $\boldsymbol{y}_{pre:i}$, representing the observed

vector of $n_{pre:i}$ pre-treatment observations for the i^{th} individual. However, this is inclusive of stochastic Gaussian process components, such as Brownian motion [20, 21] or fractional Brownian motion [16], as these do not prevent the use of a (multivariate normal) closed form for the pre-treatment likelihood function f_{pre}. Denoting the vector of values of the stochastic process $\mathbf{W}_{pre:i}$ at times $\mathbf{t}_{pre:i}$, and defining $\mathbf{\Sigma}_{pre:i}$ as the covariance matrix resulting from the chosen Gaussian process for the i^{th} individual, the linear mixed model can then be expressed as:

$$\begin{aligned}\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i}\\ \mathbf{b}_i &\sim MVN(\mathbf{0},\ \mathbf{\Psi})\\ \mathbf{W}_{pre:i} &\sim MVN(\mathbf{0},\ \mathbf{\Sigma}_{pre:i})\\ \mathbf{e}_{pre:i} &\sim MVN(\mathbf{0},\ \sigma^2\mathbf{I}_{n_{pre:i}}).\end{aligned}$$

Here, $\mathbf{X}_i$ represents the pre-treatment design matrix for the 'fixed effects' parameters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the pre-treatment 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_{pre:i}$ is the vector of residual errors for each pre-treatment measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2 \cdots \mathbf{b}_N$, residual errors $\mathbf{e}_{pre:1}, \mathbf{e}_{pre:2} \cdots \mathbf{e}_{pre:N}$ and stochastic process realisations $\mathbf{W}_{pre:1}, \mathbf{W}_{pre:2} \cdots \mathbf{W}_{pre:N}$ for each of the N individuals are independent of one another. It can be easily shown that this formulation leads to the following marginal distribution for $\mathbf{y}_{pre:i}$:

$$\mathbf{y}_{pre:i} \sim MVN\left(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{Z}_i\mathbf{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{\Sigma}_{pre:i} + \sigma^2\mathbf{I}_{n_{pre:i}}\right).$$

We shall use $\mathbf{V}_{pre:i}$ to denote the marginal covariance matrix for $\mathbf{y}_{pre:i}$.

In this analysis, we shall consider only a 'random intercepts and slopes' structure for the fixed and random effects parts of the pre-treatment model. We shall also include fractional Brownian motion as a Gaussian process component, along with an independent residual error term [16]. A Brownian motion process represents an unpredictable 'random walk', and it has been found that adding this as a further component to linear mixed models for pre-treatment CD4 counts in HIV patients leads to an improvement in model fit [20, 21]. Fractional Brownian motion is a generalisation of the standard Brownian motion process [22]. The characteristics of a fractional Brownian motion process are determined by an additional parameter, termed H or 'the Hurst index', that can take a value in the range (0,1). Standard Brownian motion represents a special case of fractional Brownian motion, corresponding to $H = \frac{1}{2}$. When $H < \frac{1}{2}$, successive increments of the process are negatively correlated. This leads to the path of the trajectory appearing 'jagged' and realisations of the process tend to revert towards the mean of zero.

As for standard Brownian motion, the expectation of a fractional Brownian motion process is zero for all points in time (0, s, t …). A positive scale parameter (κ) can be added to the standard definition of fractional Brownian motion, corresponding to the variance of the process at $t = 1$. Fractional Brownian motion is a Gaussian process, with the following properties (which determine the structure of $\mathbf{\Sigma}_{pre:i}$ and $\mathbf{\Sigma}_{post:i}$):

$$\begin{aligned}W_0 &= 0\\ \mathrm{E}[\,W_t] &= 0\\ \mathrm{Var}[\,W_t] &= \kappa\,|t|^{2H}\\ \mathrm{Cov}[\,W_s, W_t] &= \frac{\kappa}{2}\left(|s|^{2H} + |t|^{2H} - |t-s|^{2H}\right).\end{aligned}$$

Conditional distribution of 'true' baseline

The use of a pre-treatment model with marginal multivariate normal distribution means that the conditional distribution of the 'true' baseline value (u_i) at treatment initiation for each individual given their observed pre-treatment data can be readily obtained. We denote the time of treatment initiation from the start of observation (HIV seroconversion in this case) as $t_{trt:i}$. We shall assume that u_i is formed by the sum of the fixed effects parameter vector ($\boldsymbol{\beta}$) multiplied by a row vector ($\mathbf{X}_{trt:i}$) corresponding to an extension of the design matrix ($\mathbf{X}_i$) for that individual relating to variable values (e.g. time) at $t_{trt:i}$, the equivalent term for the subject-specific random effects (i.e. $\mathbf{Z}_{trt:i}\mathbf{b}_i$) and the realisation of the subject's stochastic process at $t_{trt:i}$:

$$u_i = \mathbf{X}_{trt:i}\boldsymbol{\beta} + \mathbf{Z}_{trt:i}\mathbf{b}_i + W_{trt:i}.$$

As such, the joint distribution $\mathbf{y}_{pre:i}$ and u_i is multivariate normal:

$$\begin{pmatrix}\mathbf{y}_{pre:i}\\ u_i\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mathbf{X}_i\boldsymbol{\beta}\\ \mathbf{X}_{trt:i}\boldsymbol{\beta}\end{pmatrix},\right.$$
$$\left.\begin{pmatrix}\mathbf{V}_{pre:i} & \mathbf{Z}_i\mathbf{\Psi}\mathbf{Z}_{trt:i}^{\mathrm{T}} + \mathrm{Cov}\left[\mathbf{W}_{pre:i}, W_{trt:i}\right]\\ \mathbf{Z}_{trt:i}\mathbf{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathrm{Cov}\left[W_{trt:i}, \mathbf{W}_{pre:i}\right] & \mathbf{Z}_{trt:i}\mathbf{\Psi}\mathbf{Z}_{trt:i}^{\mathrm{T}} + \mathrm{Var}\left[W_{trt:i}\right]\end{pmatrix}\right).$$

The variance and covariance terms for the stochastic component of the model can be calculated for any given Gaussian process based on $\mathbf{t}_{pre:i}$, $t_{trt:i}$ and any pre-treatment model parameters relating to the process. The conditional probability density function of u_i given $\mathbf{y}_{pre:i}$, $f_u\left(u_i|\mathbf{y}_{pre:i}, \boldsymbol{\theta}_{pre}, \sigma^2\right)$, can therefore be obtained using the standard result for a partitioned multivariate normal distribution. Using a simplified notation:

$$\begin{pmatrix}\mathbf{y}_{pre:i}\\ u_i\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mathbf{X}_i\boldsymbol{\beta}\\ \mathbf{X}_{trt:i}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{pre:i} & \mathbf{v}_{12:i}\\ \mathbf{v}_{21:i} & v_{22:i}\end{pmatrix}\right),$$

it is known that:

$$\begin{aligned}u_i|\mathbf{y}_{pre:i} &\sim N\left(\mu\prime, \nu\prime\right),\\ \text{where } \mu\prime &= \mathbf{X}_{trt:i}\boldsymbol{\beta} + \mathbf{v}_{21:i}\mathbf{V}_{pre:i}^{-1}\left(\mathbf{y}_{pre:i} - \mathbf{X}_i\boldsymbol{\beta}\right)\\ \text{and } \nu\prime &= v_{22:i} - \mathbf{v}_{21:i}\mathbf{V}_{pre:i}^{-1}\mathbf{v}_{12:i}.\end{aligned}$$

If a patient has no pre-treatment observations, then the probability density function for the baseline value is simply that for a normal distribution with mean $\mathbf{X}_{trt:i}\boldsymbol{\beta}$ and variance $v_{22:i}$.

The conditional distribution of each u_i is normal and so will include potential negative realisations, even if the probability of this is vanishingly small for most individuals. As such, we use the notation u_i^+ to indicate a latent variable for which all probability mass for values $u_i < 0$ is assigned instead to $u_i = 0$, i.e. $u_i^+ = Max(0, u_i)$. The coding used to achieve this is given in Additional file 1.

Post-treatment model structure

Mean response to treatment

Although a range of models could be considered for the post-treatment observations, we focus on the use of an asymptotic regression model for the underlying mean structure. Such models have been used to describe CD4 recovery over several years from treatment initiation in children [23, 24]. In our definition of this model, the mean value for the i^{th} individual at time after initiation of treatment t_{post}, conditional on the 'true' baseline value u_i^+, is given by the function:

$$g\left(t_{post}, u_i^+\right) = \phi_{1:i} + \left(u_i^+ - \phi_{1:i}\right) \exp\left(-\exp\left(\phi_{2:i}\right) t_{post}\right). \quad (1)$$

This function takes the value u_i^+ when $t_{post} = 0$ (i.e. at the exact time of treatment initiation), and it has a horizontal asymptote at $\phi_{1:i}$ as $t_{post} \to \infty$. The value of $\phi_{2:i}$ determines the speed of transition from u_i^+ to $\phi_{1:i}$, i.e. from the value of the response variable at baseline to its long-term mean, as t_{post} increases. The shape of the function is illustrated in Fig. 2. It is useful to note that, as this function involves a change from a baseline value to a long-term maximum that follows an 'exponential decay'-type curve, the 'half life' of this transition can be calculated as $\frac{\log(2)}{\exp(\phi_{2:i})}$; this facilitates interpretation of the estimated values of parameters that define $\phi_{2:i}$.

In models of this type, the place of u_i^+ in this function is usually taken by a single parameter (or a linear function of a set of parameters) to be estimated, potentially with an associated subject-specific random effect term. However, we instead make use of the fact that a subject-specific distribution for u_i^+ can be included in the model conditioned on the observed pre-treatment data for that individual. Similarly, we will consider $\phi_{1:i}$ and $\phi_{2:i}$ as potentially being determined as a function of u_i^+, alongside other variables, i.e. we will investigate whether the long-term average value of the response variable and the speed at which this is attained are predicted by the 'true' value of the variable at treatment initiation.

Long-term maximum

The simplest potential model for the long-term maximum response to treatment in each individual, i.e. the horizontal asymptote $\phi_{1:i}$, is to assume that this is equal to a single constant for the entire population:

$$\phi_{1:i} = A_1, \text{ for all } i.$$

The implication of this model is that the long-term response to treatment does not depend on the value of the variable in any given patient at treatment initiation, or on any other factors. This formulation also assumes

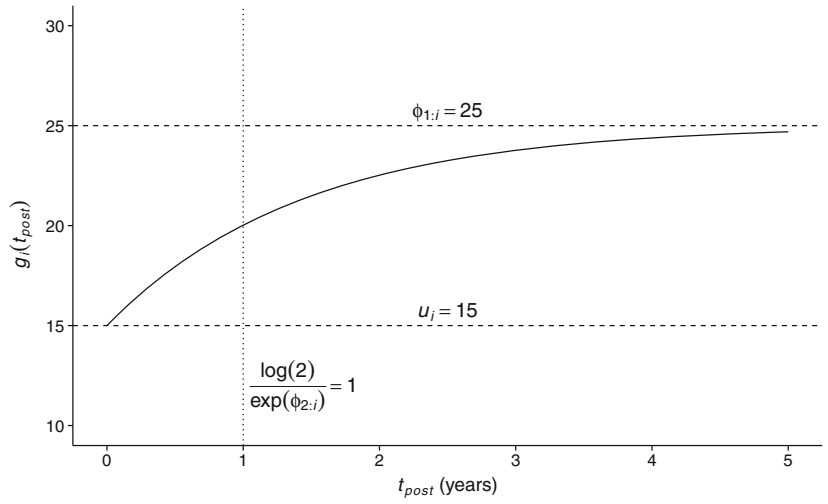

Fig. 2 Illustrative plot of an asymptotic regression curve. Here the baseline (u_i) is set to 15, the asymptotic maximum ($\phi_{1:i}$) is set to 25 and the rate of recovery parameter ($\phi_{2:i}$) is set to log(log(2)), leading to a 'half-life' of 1

that there is no random variation in the long-term maximum response between patients, but we will include a subject-specific random-effect term 'τ_i', alongside any deterministic function ($\phi_1(\ldots)$), throughout:

$$\phi_{1:i} = \phi_1(\ldots) + \tau_i, \text{where } \tau_i \sim N(0, P),$$

with the variance parameter P to be estimated. Although the post-treatment model defined in Eq. (1) is non-linear in terms of the parameters, using this formulation it *is* linear in terms of the subject-specific random effect. As such $f_{post}\left(\boldsymbol{y}_{post}|u, \boldsymbol{\theta}_{post}, \sigma^2\right)$ can be expressed in closed form as a multivariate normal distribution (assuming no further random effect terms are added to the model), even though it does not constitute a linear mixed effects model conditioned on the unobserved baseline variable. Further details are given in Additional file 1.

The next model considered is that the expected long-term maximum (working on the square-root scale for CD4 counts) for any given patient follows a linear dependence on their 'true' value at treatment initiation:

$$\phi_1\left(u_i^+\right) = A_1 + A_2 \times u_i^+.$$

Where A_1 and A_2 are parameters to be estimated.

We then wish to investigate whether ϕ_1 is a more complex, non-linear, function of u_i^+. One option would be to specify that ϕ_1 is some specific non-linear function of u_i^+. However, the fact that the relationship between $\phi_{1:i}$ and u_i^+ cannot be directly visualised using the raw data means that there is no obvious way to go about selecting the functional form. Another option is the use of cubic splines defined in terms of u_i^+, this approach has the advantage of allowing consideration of a wide variety of possible relationships between the predictive and outcome variable. In order to restrict the total number of model parameters and improve stability of optimisation, we make use of natural cubic splines derived from a truncated power series basis as described by Hastie, Tibshirani and Friedman [25]. We use knots at 15.5, 17.5, 19.5 and 22 in terms of square-root CD4, corresponding to approximately the 20^{th}, 40^{th}, 60^{th} and 80^{th} centiles of the last observed CD4 count before treatment initiation, when available, in the UK Register of HIV Seroconverters dataset.

We also consider models in which the relationship between the long-term maximum response and the baseline value (u_i^+) can vary according to the time elapsed between seroconversion and treatment initiation for each patient ($t_{trt:i}$). Although ideally this would be done using a smooth function of u_i^+ and $t_{trt:i}$, for computational stability we fit separate functions of u_i^+ stratified by $t_{trt:i}$ (in years) as follows: $0 \leq t_{trt:i} \leq 0.5$, $0.5 < t_{trt:i} \leq 1.0$ and $1.0 < t_{trt:i}$. These grouping were chosen based on a combination of findings reported previously in the literature, the level of uncertainty in terms of estimated dates of seroconversion in our study population and the need to ensure that an adequate number of patients were included in each group to allow parameter estimates to be obtained for the model.

Were patient characteristics (i.e. age, gender *etc.*) to be included in the model for $\phi_{1:i}$, and assuming a linear function in terms of u_i^+ for simplicity of exposition, we would have an extended function for ϕ_1 of the form:

$$\phi_1\left(u_i^+, \mathbf{x}_i\right) = A_1 + A_2 \times u_i^+ + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{\phi_1},$$

where $\mathbf{x}_i$ is the patient-specific vector of data specifying relevant characteristics and $\boldsymbol{\beta}_{\phi_1}$ is the associated vector of parameters that determines their effects.

Speed of response to treatment

As for the function for the long-term maximum value, we consider first a constant value for $\phi_{2:i}$ across the population ($\phi_{2:i} = B_1$) and secondly a linear dependence on u_i^+:

$$\phi_{2:i} = B_1 + B_2 \times u_i^+$$

where B_1 and B_2 are parameters to be estimated. We then consider a natural cubic spline function of u_i^+, including an analysis with stratification according to groups defined by the time elapsed from seroconversion to treatment. The addition of a subject-specific random effect to this function was also considered, this required integration of the log-likelihood function over an additional latent variable for each patient and so the Laplace approximation was used.

Residual variance structure

We propose the following model for the vector of post-treatment observations ($\mathbf{y}_{post:i}$) for the i^{th} individual, conditioned on their 'true' baseline value at treatment initiation (u_i^+):

$$\begin{aligned}
\mathbf{y}_{post:i}|_{U_i^+ = u_i^+} &= \mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\
\tau_i &\sim N(0, P) \\
\mathbf{W}_{post:i} &\sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{post:i}\right) \\
\mathbf{e}_{post:i} &\sim MVN\left(\mathbf{0}, \sigma^2 \mathbf{I}_{n_{post:i}}\right).
\end{aligned}$$

The vector of observation times $\mathbf{t}_{post:i}$ relates to time since treatment initiation, with $n_{post:i}$ post-treatment observations for the i^{th} subject. The function $\mathbf{g}$ here represents a vectorised version of g in Eq. (1), i.e.:

$$\mathbf{g}\left(\mathbf{t}_{post:i}, u_i^+, \tau_i\right) = \begin{pmatrix} g\left(t_{post:i1}, u_i^+, \tau_i\right) \\ g\left(t_{post:i2}, u_i^+, \tau_i\right) \\ \vdots \\ g\left(t_{post:in_{post:i}}, u_i^+, \tau_i\right) \end{pmatrix}.$$

For the stochastic process component $\mathbf{W}_{post:i}$, we include a 'new' fractional Brownian motion process with

value zero at time of treatment initiation and separate parameters to the pre-treatment process. The vector $\mathbf{e}_{post:i}$ represents independent residual measurement errors (or very short-term physiological variation), with a variance parameter (σ^2) that is shared with the pre-treatment model.

Differences in variability between patients

Previous work on pre-treatment CD4 counts in HIV patients has found that the generalisation of the model structure as described in "Pre-treatment model structure" to a multivariate-t distribution leads to a substantial improvement in model fit in terms of the log-likelihood and residual diagnostic plots [16]. However, the application of a marginal multivariate-t distribution is not possible in the current setting, in which a combined model is defined for pre- and post-treatment data. We instead consider models in which the stochastic process components before and after treatment each follow a marginal multivariate-t distribution, with correlated scaling variables.

There are a number of multivariate generalisations of the univariate t-distribution, and a thorough review of this topic is provided by Kotz and Nadarajah [26]. However, we refer to the *multivariate-t distribution* as that with the probability density function:

$$f(\mathbf{y}_i;\boldsymbol{\mu}_i,\mathbf{V}_i,\nu)=\frac{\boldsymbol{\Gamma}\left((\nu+n_i)/2\right)}{\boldsymbol{\Gamma}(\nu/2)\nu^{n_i/2}\pi^{n_i/2}|\mathbf{V}_i|^{1/2}\left(1+\frac{1}{\nu}(\mathbf{y}_i-\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{V}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right)^{(\nu+n_i)/2}},$$

where n_i represents the length of the random vector $\mathbf{y}_i$ ($\in \mathbb{R}^{n_i}$), $\mathbf{V}_i$ is a $n_i \times n_i$ positive-definite scale matrix, $\boldsymbol{\mu}_i$ is a $n_i \times 1$ location vector and ν is a degrees of freedom parameter. The mean of the distribution is $\boldsymbol{\mu}_i$ if $\nu > 1$ and otherwise undefined, and the variance of the distribution is $\frac{\nu}{\nu-2}\mathbf{V}_i$ if $\nu > 2$ and otherwise undefined.

If a vector of observations $\mathbf{y}_i$ follows a multivariate-t distribution:

$$\mathbf{y}_i \sim t_{n_i}\left(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{V}_i,\ \nu\right),$$

then this can alternatively be represented as a hierarchical model in which $\mathbf{y}_i$ follows a multivariate normal distribution conditional on a gamma-distributed variable w_i (with parameters given for 'shape' and 'rate', respectively) [27]:

$$\begin{aligned}\mathbf{y}_i|_{W_i=w_i} &\sim MVN\left(\mathbf{X}_i\beta,\ \frac{1}{w_i}\mathbf{V}_i\right) \\ W_i &\sim \text{gamma}\left(\frac{\nu}{2},\ \frac{\nu}{2}\right).\end{aligned} \tag{2}$$

The desired model structure for a combined analysis of pre- and post-treatment data requires the use of a bivariate gamma distribution, of which a number are available (as reviewed by Balakrishna and Lai [28]). Such models will include three latent variables per patient, and as such a Laplace approximation to the log-likelihood [19, 29, 30] rather than adaptive Gauss–Hermite quadrature will be used. Because of this, Moran's bivariate gamma distribution [28, 31] makes a natural choice. This distribution is defined by first transforming random variables (A and B) from the standard normal bivariate distribution with correlation ρ_{Moran} into a copula $C(\Phi(a),\Phi(b))$, where Φ is the standard normal cumulative distribution function, and secondly using the inverse cumulative distribution functions of univariate gamma distributions ($W_1 = F^{-1}(\Phi(A))$, $W_2 = G^{-1}(\Phi(B))$) to find the joint distribution function of W_1 and W_2 (each of which has a marginal univariate gamma distribution). F is here defined as the cumulative distribution function for gamma distribution with 'shape' and 'rate' parameters both equal to $\frac{\nu_1}{2}$, whilst G is that for the gamma distribution with parameters both equal to $\frac{\nu_2}{2}$.

Analogous to our previous work [16], the model for pre-treatment CD4 counts is then defined as:

$$\begin{aligned}\mathbf{y}_{pre:i} &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{W}_{pre:i} + \mathbf{e}_{pre:i} \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\ \boldsymbol{\Psi}) \\ \mathbf{W}_{pre:i}|_{W_{1:i}=w_{1:i}} &\sim MVN\left(\mathbf{0},\ \frac{1}{w_{1:i}}\boldsymbol{\Sigma}_{pre:i}\right) \\ \mathbf{e}_{pre:i} &\sim MVN\left(\mathbf{0},\ \sigma^2\mathbf{I}_{n_{pre:i}}\right),\end{aligned}$$

whilst, the model for post-treatment data is:

$$\begin{aligned}\mathbf{y}_{post:i}|_{U_i^+=u_i^+} &= \mathbf{g}\left(\mathbf{t}_{post:i},u_i^+,\tau_i\right) + \mathbf{W}_{post:i} + \mathbf{e}_{post:i} \\ \tau_i &\sim N\left(0,\ P\right) \\ \mathbf{W}_{post:i}|_{W_{2:i}=w_{2:i}} &\sim MVN\left(\mathbf{0},\ \frac{1}{w_{2:i}}\boldsymbol{\Sigma}_{post:i}\right) \\ \mathbf{e}_{post:i} &\sim MVN\left(\mathbf{0},\ \sigma^2\mathbf{I}_{n_{post:i}}\right),\end{aligned}$$

with the scaling factors jointly following Moran's bivariate gamma distribution:

$$\begin{pmatrix}W_{1:i}\\W_{2:i}\end{pmatrix} \sim Moran\left(\rho_{Moran};\frac{\nu_1}{2},\frac{\nu_1}{2};\frac{\nu_2}{2},\frac{\nu_2}{2}\right).$$

This specific bivariate gamma distribution is a natural choice because the marginal log-likelihood function for the model can be found by integrating out the latent variables on the standard normal scale, for which the Laplace approximation is optimally accurate [32], as follows (omitting indexing for each individual and dependence on model parameters):

$$\begin{aligned}&f\left(\mathbf{y}_{pre},\mathbf{y}_{post}\right) = \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{pre}\left(\mathbf{y}_{pre}|w_1=F^{-1}(\Phi(a))\right) f_{post}\left(\mathbf{y}_{post}|u,w_2=G^{-1}(\Phi(b))\right) \\ &f_u\left(u|\mathbf{y}_{pre},w_1=F^{-1}(\Phi(a))\right) f_{ab}(a,b)\,du\,da\,db,\end{aligned}$$

where f_{ab} is the probability density function for a standard bivariate normal distribution with correlation ρ_{Moran}.

The ρ_{Moran} parameter can be estimated from the data through maximum likelihood estimation as for other model parameters.

Overall model structure and interpretation

A directed acyclic graph depicting the proposed model structure is shown in Fig. 3. For simplicity, we omit here the extension to the basic model in which further latent variables are added to the model to allow between-patient differences in variability over time as described in Subsection "Differences in variability between patients". This diagram illustrates the fact that in the model, response to treatment is linked to pre-treatment data only through the 'true' baseline value u and the time from seroconversion to treatment initiation. These links are mediated through variables representing the long-term maximum response to treatment (ϕ_1) and the speed at which this is attained (ϕ_2) in each patient. When fitted to the dataset under investigation, this structure should allow estimates of individual parameters of the model to be interpreted in a meaningful way. Although in this article we do not consider further potential predictive variables, it would be relatively straightforward to extend the model to assess whether patient characteristics such as age and gender or drug regimen choice are independently predictive of response to treatment.

The primary interpretation of our models as presented is the prediction of the response to HAART in terms of prior CD4 counts and time from seroconversion. It has been argued that causal effects can only be estimated from observational studies with respect to clearly defined interventions [33]. Whilst interventions with regard to the monitoring of CD4 counts and guidelines for treatment initiation can be defined within the present context, it is not possible to begin treatment conditional on the 'true' value of a patient's CD4 count, as this cannot be observed directly. Furthermore it is not possible to define a treatment policy in terms of a specific simultaneous combination of 'time from seroconversion' and 'true CD4 count', when in a certain period a patient may only experience a limited range of CD4 counts.

As we have censored patients at recorded interruption of HAART but not according to viral load status, the fitted models can be taken to represent treatment response for all patients were they all to remain on HAART (regardless of success or failure of virological suppression). All included patients had at least one post-HAART CD4 observation, but beyond this the number and timing of CD4 cell counts recorded for each individual were highly variable. We have assumed that the missingness of observations can be treated as 'missing at random' (following the terminology of Rubin [34]), i.e. that the 'missingness' of any observation is independent of the unobserved data conditional on the observed values of the outcome variable and any other covariates included in the model. Similarly we assume that the timing of observations is dependent only on previously observed outcomes, under which condition maximum likelihood estimation of a model for the outcome variable alone is consistent, without the need for specification of a model for the distribution of follow-up times [35].

Maximum likelihood estimation

All models presented have been fitted by direct maximum likelihood estimation using the open source AD Model Builder software (Version 11.2; ADMB Foundation) [30]. This requires the user to write out the log-likelihood function for the model in terms of the data and unknown parameters to be estimated in the C++ language, with additional statistical and mathematical

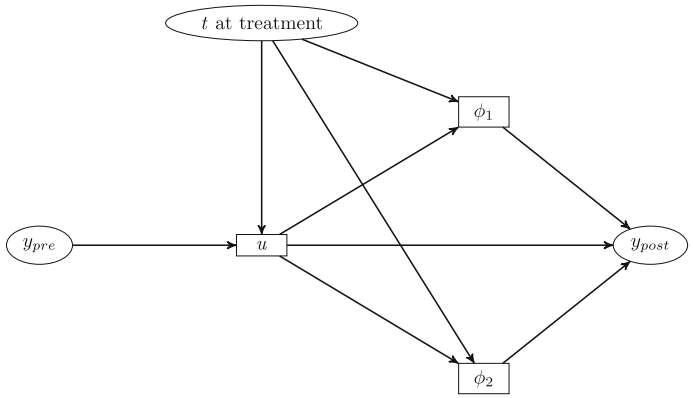

Fig. 3 Directed acyclic graph depicting the proposed model structure for each patient. Observed variables are shown within *ellipses*, whilst unobserved latent variables are shown within *rectangles*

functions (including matrix and vector functions and operations) provided by the software to facilitate this. The 'random effects' mode was used for ADMB, allowing optimisation of a log-likelihood function with automated integration over latent variables [29]. The log-likelihood function for each individual (for their complete pre- and post-treatment data) was defined using the 'separable function' utility, allowing computational efficiency to be gained from the modelled independence of each individual. 15-point adaptive Gauss–Hermite quadrature was used to obtain the maximum likelihood estimates for all models described in this report for which only one latent variable was included per individual (i.e. the 'true' baseline). However, for the models including additional latent variables associated with between-patient differences in variability over time, Gauss–Hermite quadrature was not feasible and the Laplace approximation was used.

Models were parameterised using logarithmic, logistic and generalised logistic transformations where appropriate such that parameter estimates could be obtained using unrestricted optimisation (e.g. maximum likelihood estimation was carried out using log-transformed variance parameters, with a parameter space of $(-\infty, +\infty)$ rather than $[0, +\infty]$). For all model parameters, confidence intervals are reported derived from the estimated asymptotic multivariate normal sampling distribution based on the observed information on the transformed scales. The 'R2admb' package [36] was used to output data files in the necessary format through the R statistical computing environment (R Foundation, Vienna, Austria). The ggplot2 package for R [37] was used for statistical graphics. All maximum likelihood estimates reported in this document were obtained using a computer cluster running with Linux operating systems. The authors acknowledge the use of the University College London (UCL) Legion High Performance Computing Facility (Legion@UCL), and associated support services, in the completion of this work. Fitting each of the models presented to the UK Register of HIV Seroconverters dataset took between 1 and $2\frac{1}{2}$ hours (using a core with 4GB RAM), whereas fitting one of the models using a mid–low specification personal laptop (4GB RAM, Celeron Dual-Core CPU T3500 @ 2.1 GHz) required around 10 h.

When considering only a single latent variable per patient, nested models are compared using the generalised likelihood ratio test, comparing the change in 2×log-likelihood ($\Delta 2\ell$) to a χ^2 distribution. Non-nested models are compared using the Bayesian information criterion (BIC) statistic, using the total number of observations in the dataset for the calculation of the penalty term. It is worth noting that these methods are only valid because adaptive Gauss–Hermite quadrature can be used to calculate the log-likelihood of the fitted models to a high degree of accuracy; this is not the case for less computationally intensive approximations of the log-likelihood.

Results

Model fitting

Summaries of the set of models fitted to the UK Register of HIV Seroconverters dataset are presented in Table 1, and to facilitate their interpretation Table 2 provides a description of each model parameter. The most basic model considered included constant parameters for the mean long-term maximum CD4 count (on square-root scale) and the rate of recovery from baseline at treatment initiation, without division of patients according to time from seroconversion to initiation of HAART (Model_1 in Table 1). Modelling the long-term maximum (ϕ_1) and speed of response to treatment (ϕ_2) as linear functions of the baseline value in each individual (u_i^+) led to a significant improvement in model fit (Model_2 *vs* Model_1, $\Delta 2\ell$ 460.4 for 2 parameters; $P < 0.0001$). A model equivalent to Model_2 but without pre- and post-treatment stochastic process components was also fitted for comparison and was found to have a much higher BIC value (64398); correspondingly the model including stochastic processes showed a significant improvement in fit ($\Delta 2\ell$ 844.8 for 4 parameters; $P < 0.0001$). The extension of *Model*$_2$ to allow natural cubic spline functions to define the relationships between u_i^+ and ϕ_1 and ϕ_2 led to a further significant improvement in model fit (Model_3 *vs* Model_2, $\Delta 2\ell$ 31.4 for 4 parameters; $P < 0.0001$).

Fitting a model with separate linear relationships between u_i^+ and ϕ_1 and ϕ_2 according to timing of HAART subgroup (Model_4) led to a reduction in BIC relative to the single-group natural cubic spines model. It was not possible to obtain a model fit for natural cubic spline functions defined separately for each subgroup (due to lack of convergence), but allowing linear functions in the early start subgroups in combination with natural cubic spline functions for the remaining patients led to a further improvement in model fit (Model_5 *vs* Model_4, $\Delta 2\ell$ 16.0 for 4 parameters; $P = 0.003$). However, Model_4, with linear link functions for all subgroups, retained the lowest BIC value and so we have focused on interpretation of this model.

It is harder to make a direct comparison for Model_6, which matches Model_4 with the addition of jointly distributed latent scaling variables for the pre- and post-treatment fractional Brownian motion processes. Because of the need to integrate the log-likelihood function over multiple latent variables, parameter estimates for Model_6 were obtained using the Laplace approximation, meaning that generalised likelihood ratio tests or comparisons of the BIC statistic are not appropriate. However, the low values obtained for the estimates of the pre- and

Table 1 Summary of the results of combined models for pre- and post- highly active antiretroviral therapy (HAART) CD4 cell count data, after square root transformation, for patients from the UK Register of HIV Seroconverters dataset

	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$
β_0	22.44 (22.13 to 22.74)	22.45 (22.16 to 22.74)	22.44 (22.15 to 22.73)	22.26 (21.96 to 22.56)	22.26 (21.96 to 22.56)	22.23 (21.94 to 22.53)
β_1	−1.36 (−1.52 to −1.2)	−1.39 (−1.55 to −1.23)	−1.39 (−1.55 to −1.23)	−1.3 (−1.46 to −1.14)	−1.32 (−1.47 to −1.16)	−1.36 (−1.5 to −1.21)
U_{00}	12.37 (10.64 to 14.37)	13.39 (11.77 to 15.23)	13.42 (11.79 to 15.28)	14.43 (12.68 to 16.43)	14.53 (12.77 to 16.54)	12.92 (11.29 to 14.8)
ρ	−0.65 (−0.79 to −0.44)	−0.86 (−0.99 to 0.18)	−0.84 (−0.98 to −0.1)	−0.95 (−1 to 1)	−0.92 (−1 to 0.91)	−0.63 (−0.76 to −0.44)
U_{11}	0.55 (0.33 to 0.93)	0.25 (0.08 to 0.75)	0.28 (0.1 to 0.75)	0.2 (0.05 to 0.74)	0.21 (0.06 to 0.74)	0.49 (0.31 to 0.77)
κ_{pre}	9.68 (8.77 to 10.68)	5.91 (5.23 to 6.67)	5.9 (5.22 to 6.68)	5.99 (5.29 to 6.8)	5.92 (5.21 to 6.72)	5.37 (4.37 to 6.6)
H_{pre}	0.11 (0.09 to 0.14)	0.3 (0.25 to 0.37)	0.3 (0.24 to 0.36)	0.31 (0.25 to 0.37)	0.31 (0.25 to 0.38)	0.16 (0.13 to 0.19)
σ	1.25 (1.09 to 1.42)	1.95 (1.89 to 2.01)	1.94 (1.87 to 2)	1.92 (1.85 to 1.99)	1.92 (1.86 to 1.99)	1.32 (1.19 to 1.46)
ϕ_1 model:						
long-term maximum	Constant for all patients	Linear for all patients	NCS for all patients	Linear for all patients stratified by ART_t	Linear for early treatment groups or NCS for late treatment group	Linear for all patients stratified by ART_t
$At1_1$	—	—	—	7.04 (4.75 to 9.33)	7.06 (4.77 to 9.35)	8.44 (6.05 to 10.83)
$At1_2$	—	—	—	0.9 (0.79 to 1.01)	0.9 (0.79 to 1)	0.84 (0.72 to 0.95)
$At2_1$	—	—	—	10.73 (7.93 to 13.53)	10.68 (7.85 to 13.51)	12.32 (9.28 to 15.35)
$At2_2$	—	—	—	0.67 (0.54 to 0.81)	0.67 (0.53 to 0.81)	0.64 (0.47 to 0.8)
A_1	25.93 (25.49 to 26.36)	11.42 (9.74 to 13.09)	5.1 (0.3 to 9.9)	14.58 (12.3 to 16.86)	3.76 (−1.99 to 9.51)	14.35 (12.32 to 16.38)
A_2	—	0.69 (0.62 to 0.77)	1.14 (0.84 to 1.44)	0.55 (0.44 to 0.66)	1.23 (0.86 to 1.6)	0.57 (0.46 to 0.67)
A_3	—	—	-0.43 (−0.64 to −0.22)	—	−0.32 (−0.63 to −0.01)	—
A_4	—	—	0.82 (0.43 to 1.2)	—	0.52 (−0.07 to 1.11)	—
ϕ_2 model:						
recovery speed	Constant for all patients	Linear for all patients	NCS for all patients	Linear for all patients stratified by ART_t	Linear for early treatment groups or NCS for late treatment group	Linear for all patients stratified by ART_t
$Bt1_1$	—	—	—	2.66 (0.52 to 4.79)	2.8 (0.76 to 4.84)	5.68 (2.94 to 8.43)
$Bt1_2$	—	—	—	0.02 (−0.08 to 0.11)	0.01 (−0.08 to 0.1)	−0.14 (−0.29 to −1.98e-03)
$Bt2_1$	—	—	—	−0.99 (−3 to 1.02)	−0.92 (−2.97 to 1.13)	0.23 (−1.39 to 1.86)
$Bt2_2$	—	—	—	0.15 (0.05 to 0.26)	0.15 (0.04 to 0.26)	0.01 (−0.1 to 0.12)

Table 1 Summary of the results of combined models for pre- and post- highly active antiretroviral therapy (HAART) CD4 cell count data, after square root transformation, for patients from the UK Register of HIV Seroconverters dataset (*Continuation*)

B_1	-0.16 (−0.3 to −0.02)	−3.34 (−4.19 to −2.48)	1.82 (−0.23 to 3.87)	−3.64 (−4.7 to −2.59)	2.42 (0.26 to 4.58)	−2.25 (−3.3 to −1.21)	
B_2	—	0.24 (0.2 to 0.28)	−0.11 (−0.24 to 0.02)	0.23 (0.17 to 0.29)	−0.15 (−0.29 to −0.02)	0.13 (0.07 to 0.19)	
B_3	—	—	0.28 (0.19 to 0.38)	—	0.19 (0.04 to 0.33)	—	
B_4	—	—	−0.52 (−0.71 to −0.33)	—	−0.28 (−0.58 to 0.02)	—	
P	11.09 (8.76 to 14.03)	2.97 (2.09 to 4.23)	3.05 (2.13 to 4.38)	3.07 (2.19 to 4.31)	3.31 (2.39 to 4.59)	2.72 (1.71 to 4.31)	
κ_{post}	7.59 (6.79 to 8.49)	3.09 (2.46 to 3.89)	3.17 (2.53 to 3.98)	3.36 (2.7 to 4.18)	3.3 (2.66 to 4.11)	4.33 (3.5 to 5.36)	
H_{post}	0.08 (0.07 to 0.1)	0.42 (0.32 to 0.52)	0.4 (0.3 to 0.5)	0.38 (0.29 to 0.48)	0.39 (0.3 to 0.5)	0.13 (0.11 to 0.16)	
Differences in variability between patients	No	No	No	No	No	Yes	
df_{pre}	—	—	—	—	—	3.84 (3.06 to 4.82)	
df_{post}	—	—	—	—	—	4.28 (3.4 to 5.38)	
ρ_{Moran}	—	—	—	—	—	0.37 (0.19 to 0.52)	
n_{pars}	13	15	19	23	27	26	
ℓ	−31954.8	−31724.6	−31708.9	−31664.5	−31656.5	−31299.7[a]	
BIC	64032.85	63591.41	63597.94	63547.06	63568.98	62845.9[a]	

Parameter estimates are given with 95 % confidence intervals in parentheses. [a]Not comparable to other values in Table, as calculated using Laplace approximation. ART_i, time from seroconversion to treatment initiation; *BIC*, Bayesian information criterion; ℓ, log-likelihood; *NCS*, natural cubic spline; n_{pars}, number of parameters estimated in model
The interpretation of each model parameter is listed in Table 2

Table 2 Description of parameters for combined models of pre- and post-treatment data

Model parameter	*Description*
β_0	Pre-treatment mean intercept
β_1	Pre-treatment mean slope
U_{00}	Pre-treatment intercept subject-specific random effect variance
ρ	Correlation between pre-treatment intercept and slope subject-specific random effects
U_{11}	Pre-treatment slope subject-specific random effect variance
σ	Standard deviation of residual error term for each measurement, shared by pre- and post-treatment parts of model
κ_{pre}	Scale parameter for pre-treatment fBM process
H_{pre}	Hurst index for pre-treatment fBM process
ϕ_1 model	These parameters relate to the long-term maximum value of the response variable after treatment initiation
$At1_1, At1_2$	Intercept and slope terms in relationship with u_i^+ for patients treated within 6 months of seroconversion
$At2_1, At2_2$	Intercept and slope terms in relationship with u_i^+ for patients treated beyond 6 months but within 1 year of seroconversion
A_1, A_2	Intercept and slope terms in relationship with u_i^+ for linear or NCS models[a]
A_3, A_4	Third and fourth coefficients for NCS models[a]
ϕ_2 model	These parameters relate to the rate of recovery of the response variable after treatment initiation
$Bt1_1, Bt1_2$	Intercept and slope terms in relationship with u_i^+ for patients treated within 6 months of seroconversion
$Bt2_1, Bt2_2$	Intercept and slope terms in relationship with u_i^+ for patients treated beyond 6 months but within 1 year of seroconversion
B_1, B_2	Intercept and slope terms in relationship with u_i^+ for linear or NCS models[a]
B_3, B_4	Third and fourth coefficients for NCS models[a]
P	Residual variance for long-term maximum (ϕ_{1i}) not explained by u_i^+
κ_{post}	Scale parameter for post-treatment fBM process
H_{post}	Hurst index for post-treatment fBM process
df_{pre}	Degrees of freedom parameter for pre-treatment stochastic process
df_{post}	Degrees of freedom parameter for post-treatment stochastic process
ρ_{Moran}	Correlation parameter for latent scaling variables of pre- and post-treatment stochastic processes

[a]Only applicable to patients with treatment initiation more than 1 year after seroconversion when separate terms are included for earlier groups. *fBM*, fractional Brownian motion; *NCS*, natural cubic spline
Some of the parameters relate to the link functions between the 'true' value of the response variable at treatment initiation, u_i^+, and the post-treatment model

post-treatment degrees of freedom parameters (which are effectively fixed at $+ve\infty$ for the other models considered) indicate that this model may better reflect the structure of the observed data. Convergence of parameter estimates was not achieved when the same extension was made to Model_5.

Convergence of parameter estimates also failed when a subject-specific random effect was added to the speed of response to treatment function (ϕ_2) for Model_4, Model_5 or Model_6. We also attempted to extend each of these models to allow an independent linear effect of the patient-specific slope of pre-HAART decline (requiring an additional two latent variable per patient for their random intercept and slope terms), but convergence of parameter estimates was not achieved in each case. Using Model_4, we checked the assumption that the pre- and post-HAART measurement error variance can be treated as constant, and no significant improvement in model fit was observed when separate parameters were fitted for the two periods ($\Delta 2\ell$ 0.6 for 1 parameter; $P = 0.44$).

Plots of residuals derived from Model_6 are provided in Additional file 1 (based on Fitzmaurice et al. [38] and Stirrup et al. [16]), and these do not indicate substantial problems with the fitted model. As a further check of the model structure developed, the fitted Model_6 was used to simulate pre- and post-treatment CD4 counts for a cohort of 100 patients. The plot of these simulated data is visually consistent with the equivalent plot of 100 randomly selected patients from the real dataset. This comparison could be described as a posterior predictive check [39]. Additionally, a small simulation study was carried out to demonstrate that the use of a natural cubic spline basis for baseline CD4 count would be able to provide approximations to non-linear functions for the long-term maximum and speed of recovery following initiation of HAART, even if specification of the probability model as a whole is not completely correct; this is presented in Additional file 1.

An R script and ADMB template files are also provided in Additional file 2 to simulate data based on the structure and point estimates of Model_6, and to then refit Model_4 and Model_6 to these data.

Model interpretation

All models fitted (other than Model_1 by definition) showed a positive association between baseline CD4 count at HAART and the long-term maximum; this finding was consistent across subgroups of patients defined by timing of treatment initiation with only relatively small differences in the fitted functions for each group in models 4–6 (Figs. 4, 5 and 6). When modelled as a linear function across all patients (i.e. Model_2), the speed of response to treatment also showed a positive association with baseline CD4 count at HAART. However, when the link function was defined by HAART-timing subgroup, the speed of response to treatment was found to be substantially higher at moderate and lower baseline CD4 counts (below around 25 on the square-root scale) in those patients who started treatment within 6 months of seroconversion, with an intermediate difference observed for the subgroup who started treatment after 6 months but within 1 year. This overall pattern of findings was consistent across models 4–6, although the exact shape of the link functions showed some differences.

As the full vector of pre- and post-treatment data and u_i for each individual do not jointly follow a multivariate normal distribution, it is not possible to derive a closed form for the posterior predictive distribution of the u_i conditioned on the observed data in the way that would be done for the realizations of the random effects in a linear mixed model. However, the values of u_i for each individual that maximise $f\left(\boldsymbol{y}_{pre:i}, \boldsymbol{y}_{post:i}, u_i\right)$, $\hat{u}_i$, conditional on the current values of the model parameters, are calculated at each iteration of the adaptive Gauss–Hermite quadrature algorithm. The values of $\hat{u}_i$ corresponding to the final parameter estimates for each model are returned by ADMB, and these correspond to the posterior mode of $f_{u|\boldsymbol{Y}_{pre}=\boldsymbol{y}_{pre},\boldsymbol{Y}_{post}=\boldsymbol{y}_{post}}(u)$ for each individual. Kernal density plots for the u_i values for each subgroup in Model_4 are presented in Fig. 7, approximating the distribution for $f_{u|\boldsymbol{Y}_{pre}=\boldsymbol{y}_{pre},\boldsymbol{Y}_{post}=\boldsymbol{y}_{post}}(u)$ as normal and making use of subject-specific standard deviation estimates also resulting from the adaptive Gauss–Hermite quadrature algorithm. Equivalent plots for Model_5 and Model_6

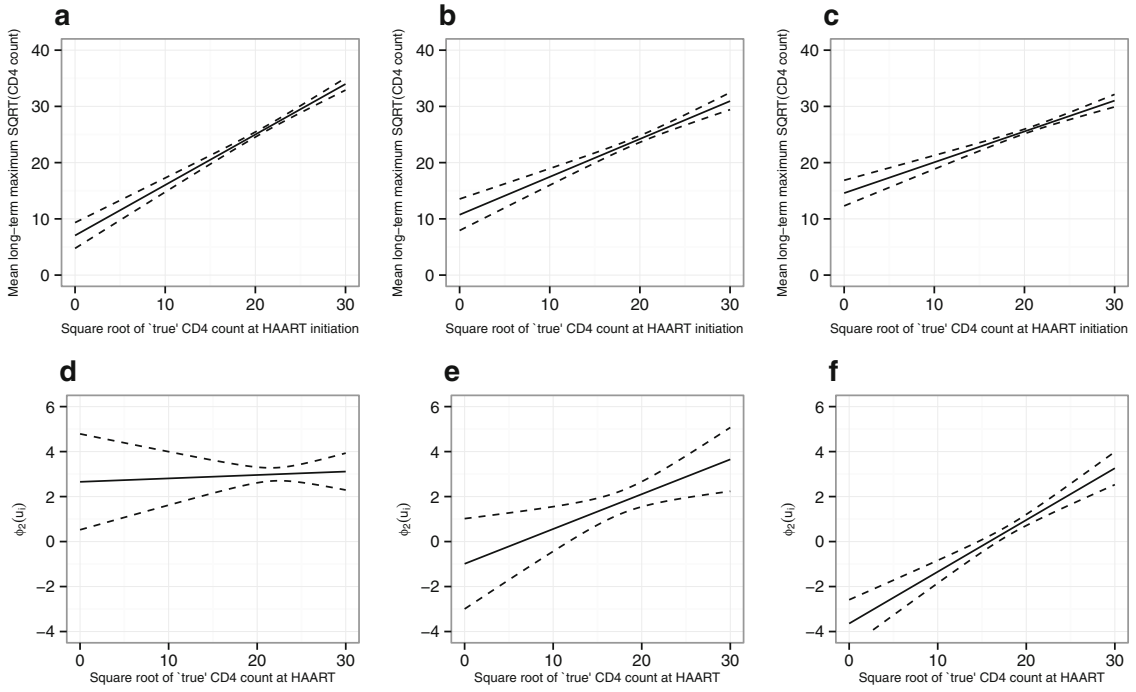

Fig. 4 Plots of $\phi_1\left(u_i^+\right)$ (**a**–**c**, relating to long-term maximum) and $\phi_2\left(u_i^+\right)$ (**d**–**f**, relating to speed of response) for Model_4. Graphs on the left of each row **a**, **d** show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre **b**, **e** show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right **c**, **f** show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown as dashed lines

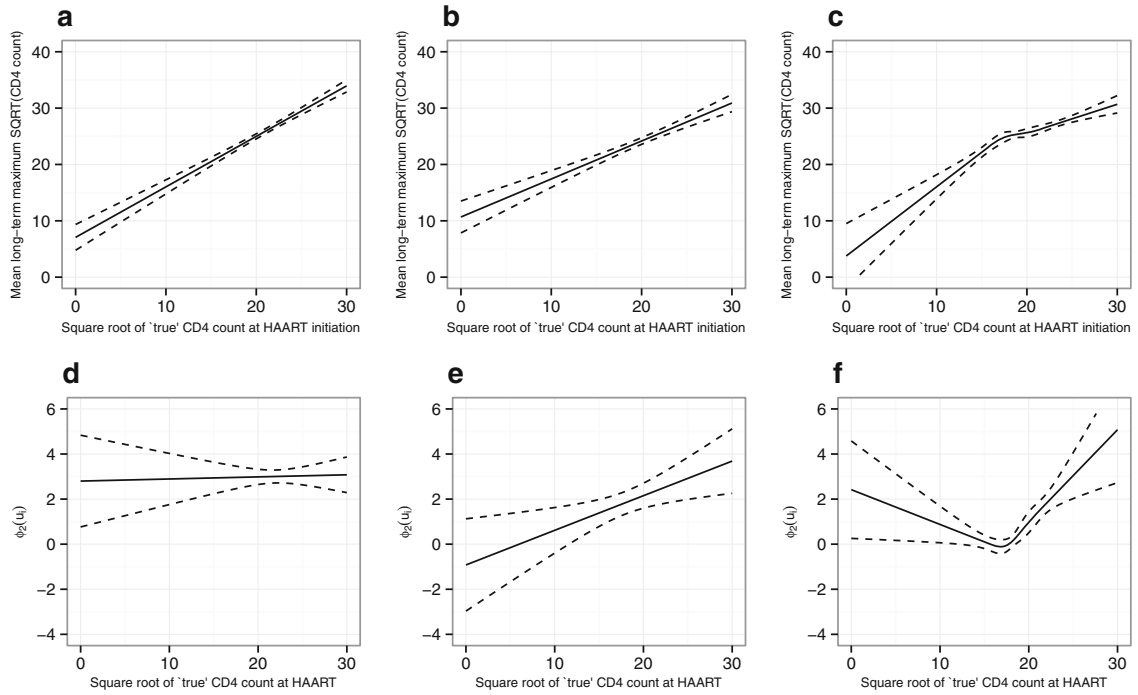

Fig. 5 Plots of $\phi_1\left(u_i^+\right)$ (**a**–**c**, relating to long-term maximum) and $\phi_2\left(u_i^+\right)$ (**d**–**f**, relating to speed of response) for Model_5. Graphs on the left of each row **a**, **d** show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre **b**, **e** show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right **c**, **f** show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown as dashed lines

did not show substantial differences. Histograms of the last observed square-root CD4 count before treatment for those individual in whom this was recorded within 6 months of treatment initiation are also presented in Fig. 7 for comparison, showing a similar shaped distribution in each subgroup. As expected given the results of previous simulations regarding treatment initiation based on observed CD4 cell counts [14], for more than half of patients (63 %) the mode of the posterior predictive distribution ($\hat{u}_i$) was greater than the last observed CD4 count (where available within 6 months); the median difference for $CD4_{last\_obs} - \hat{u}_i$ was -18 cells/μL when transformed back to the original measurement scale.

Predicted ranges for CD4 cell counts based on Model_4 are shown in Fig. 8 for patients with a 'true' CD4 counts at initiation of HAART of 200, 350 and 500 cells/μL. These charts further illustrate the model predictions that, in general, patients with a higher CD4 cell count at treatment initiation will go on to show a higher long-term maximum and will attain higher values more quickly after the start of treatment, but that response to treatment is rapid if it is initiated within 6 months of seroconversion regardless of baseline CD4. These charts also illustrate that the model predicts considerable variability in response to treatment between patients at any given baseline CD4 value. However, in the models presented we have not included variables such as patient age, gender and mode of infection that may also be predictive of response to treatment, and so it is possible that more fully developed models would include less unexplained variance in the long-term response to treatment. The inclusion of such potential confounding variables may also affect estimates of the influence of baseline value of CD4 at treatment initiation on each patient's response to treatment. Equivalent plots for Model_5 and Model_6 showed similar overall patterns of predictions.

For Model_6, estimates of the pre- and post-treatment degrees of freedom parameters (3.84 (95 % CI, 3.06–4.82) and 4.28 (3.4–5.38), respectively) indicate that there are considerable between-patient differences in the variability of observations over time. It is interesting to note that the correlation parameter between the pre- and post-treatment latent scaling variables was positive, but only of moderate magnitude ($\hat{\rho}_{Moran}0.37$ (0.19–0.52)), i.e. the degree of variability over time before and after treatment for each patient shows a moderate positive correlation.

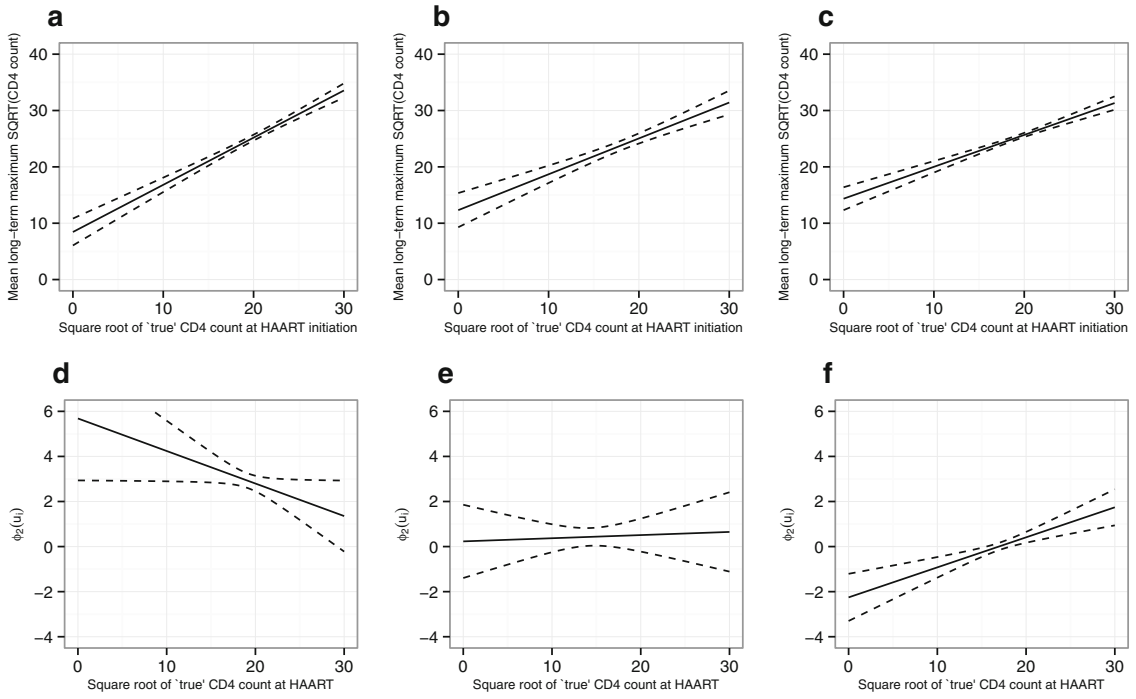

Fig. 6 Plots of $\phi_1\left(u_i^+\right)$ (**a**–**c**, relating to long-term maximum) and $\phi_2\left(u_i^+\right)$ (**d**–**f**, relating to speed of response) for Model_6. Graphs on the left of each row **a**, **d** show the fitted functions for patients initiating treatment within 6 months of seroconversion, those in the centre **b**, **e** show the functions for patients initiating treatment beyond 6 months but within 1 year and those on the right **c**, **f** show the functions for patients who started treatment beyond 1 year. Pointwise 95 % confidence intervals for the functions are shown as dashed lines

It is also of interest that the estimated H-index for the post-treatment fractional Brownian motion process in this model was much lower than that for the equivalent model without the latent scaling variables (0.13 (0.11–0.16) *vs* 0.38 (0.29–0.48)), indicating that although some patients show high variability in CD4 observations over time, successive increments of the stochastic process are strongly negatively correlated and there is an associated reversion of the process towards the underlying mean in each patient. It is possible to use the modes of the posterior predictive distributions of the latent scaling variables for each patient to identify those individuals with particularly smooth or erratic patterns of CD4 counts over time; observations for the two patients with the most extreme values obtained for the post-treatment latent scaling variable are plotted in Fig. 9.

Discussion

The statistical methodology developed in this article provides a novel framework for the combined analysis of pre- and post-treatment longitudinal biomarker data. The approach proposed has the advantage of making use of all available data, does not require an a priori assumption regarding the distribution of baseline values at treatment across the studied population as a whole and allows a flexible choice of functions to link the pre- and post-treatment trajectories of the biomarker under investigation for each patient. When applied to CD4 data from the UK Register of Seroconverters cohort, the resulting fitted models provide evidence of a positive association between baseline CD4 count at initiation of HAART and the long-term maximum achieved by each patient, which is consistent with previous published literature on this topic [9–11]. In addition the fitted models suggest that initiation of HAART closer to the date of HIV seroconversion is associated with a more rapid response to treatment, regardless of the baseline CD4 value. This finding warrants further investigation in larger datasets, with inclusion of additional factors that are thought to be associated with response to treatment into the modelling framework; this extension would be straightforward using the methodology developed.

The standard non-linear mixed effects model approach in this situation, ignoring observations before the start of treatment, would require rigid assumptions regarding the distribution of the biomarker variable at treatment

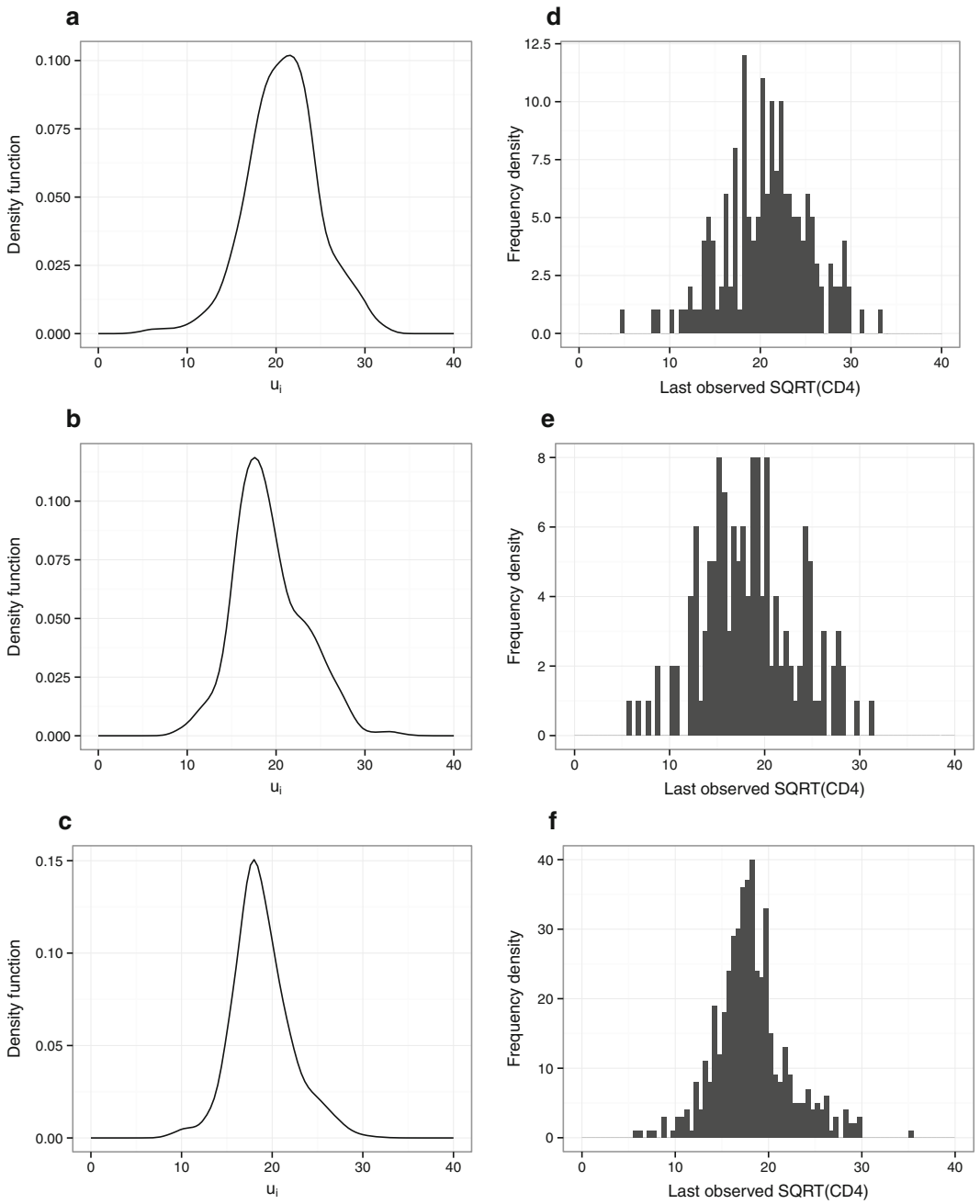

Fig. 7 Kernel density plots (**a**–**c**) for the 'true' baseline square root CD4 counts and (**d**–**f**) histograms of the last observed square-root CD4 count before treatment. **a**–**c** Kernel density plots for the 'true' baseline square root CD4 counts for each individual (u_i), approximating the posterior distribution of each as normal (with subject-specific standard deviation as estimated during model fitting), and **d**–**f** histograms of the last observed square-root CD4 count before treatment for those individuals in whom this was recorded within 6 months of treatment initiation ($n = 170$, $n = 141$ and $n = 486$, respectively). Graphs in the top row **a**, **d** relate to patients initiating treatment within 6 months of seroconversion, those in the centre row **b**, **e** relate to patients initiating treatment beyond 6 months but within 1 year and those on the lower row **c**, **f** are for patients who started treatment beyond 1 year

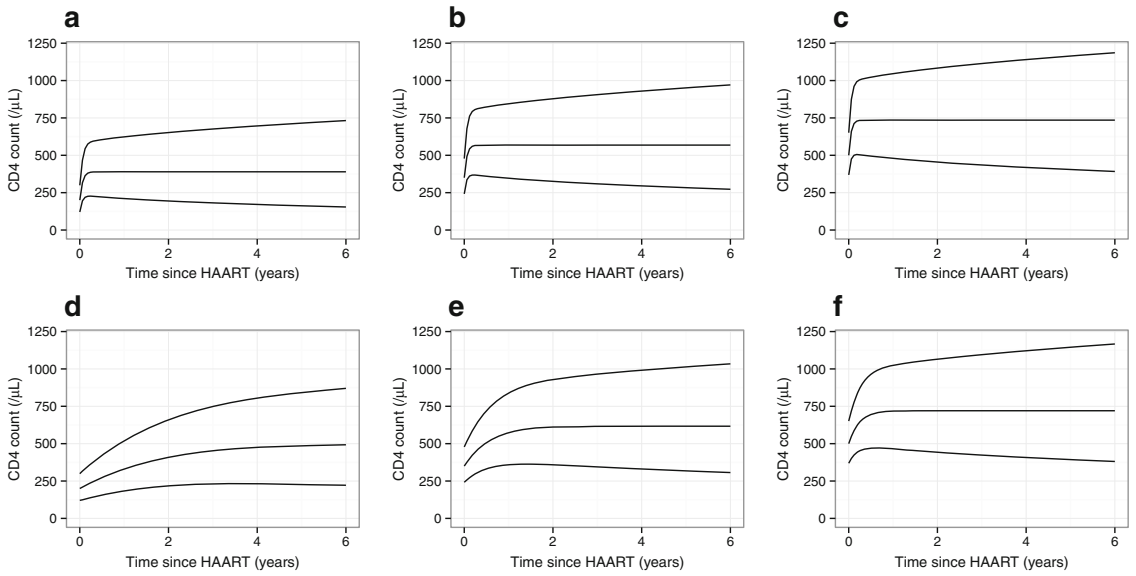

Fig. 8 Predictions for hypothetical patients made from fitted model. Plots of the 90 % range of CD4 counts predicted by Model_4 for a population of patients initiating highly active antiretroviral therapy (HAART) either within 6 months (**a**–**c**) or more than 1 year (**d**–**f**) from seroconversion, with 'true' CD4 counts at treatment initiation of: **a**, **d** 200, **b**, **e** 350 and **c**, **f** 500 cells/μL. The predicted ranges include measurement error (alongside the stochastic process component and variance in the subject-specific long-term maximum), explaining the variance present at time zero. The ranges shown have been back-transformed from the model predictions generated on the square-root scale

initiation and its relationship to subsequent post-treatment observations, i.e. typically that baseline values and the long-term maximum value for each patient follow a bivariate normal distribution. The modelling strategy that we have developed allows greater flexibility in the link between baseline and post-treatment maximum values of the biomarker, and does not restrict the shape of the overall marginal distribution of baseline values in the studied population. Alternatively, the standard use of baseline observations as a predictive variable would also discard any information from measurements obtained prior to this point in time and would require a separate

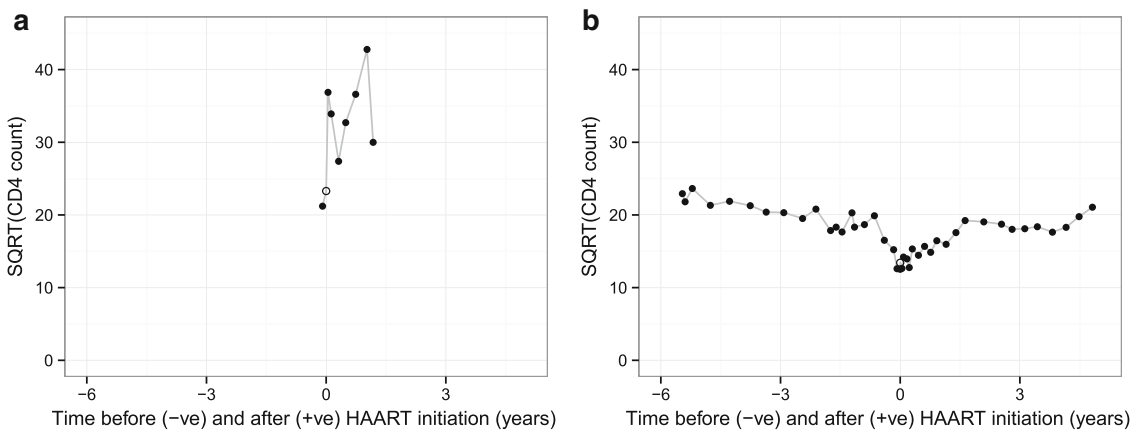

Fig. 9 Plots of CD4 counts (•) observed in the two patients with the most (**a**) and least (**b**) erratic response to highly active antiretroviral therapy (HAART). Variability of response was assessed as indicated by the modes of the posterior predictive distributions of the post-treatment latent scaling variables ($\widehat{w_{2i}}$) obtained from Model_6. The mode of the posterior predictive distribution for the 'true' baseline value ($\widehat{u_i}$) is also shown in each case (○)

imputation model for missing values of the baseline measurement, which would not be straightforward to define for observational data with highly irregular number and timing of measurements for each patient. Furthermore, it is not obvious how the primary model for multiple post-treatment observations should be structured in this context, as it would be overly restrictive to assume a constant fixed effect coefficient for the baseline observation for all time points after the initiation of treatment.

The proposed model for the analysis of pre- and post-treatment CD4 data has been structured so that the estimated parameters of the different components of the model each have a clear practical interpretation, i.e. it is of direct interest to clinicians to know how baseline CD4 and time from seroconversion at initiation of HAART are associated with the speed and maximal level of treatment response that can be expected. If further patient variables were added to the functions that determine the characteristics of response to treatment then the modelled effects would be independent of the influence of the true baseline value of the biomarker, making interpretation of estimated coefficients relatively simple. If a mixed effects model is fitted to only baseline and post-treatment measurements, then assessment of the influence of a covariable on treatment response conditional on a baseline observation requires an additional stage of statistical adjustment [40].

The cost of using a combined model for pre- and post-treatment data is that we are required to assume that the proposed model structure provides an adequate description of the data under analysis. The requirement for strong assumptions regarding the correctness of model structure has been used as an argument against the use of integrated models for baseline and treatment response data [3]. In the present study, the motivations for the inclusion of pre- and post-treatment stochastic process components in the models and for the use of natural cubic spline functions to link baseline CD4 and characteristics of the treatment response trajectory were to maximise model flexibility and therefore provide an optimal fit to the data. However, we plan to investigate further extensions of the model structure using larger datasets, which would be able to support a greater number of parameters in model-fitting. As such, the scientific results from the present study can only be taken as preliminary findings.

An advantage of the extension of the non-linear mixed effects modelling approach as developed in this paper is that the nature of the variability in biomarker observations over time within each patient can be investigated, whereas this is often lost when using approaches that only consider population mean values or the marginal distribution of observations across the population at each point in time. A focus on realistic modelling of the patterns of variation in the data is also required in order to provide valid inference under the 'missing at random' assumption for missing data and when the timing of observations is dependent on previous outcomes [35]. A limitation of the present analysis is that we have not considered the possibility of censoring being related to underlying latent variable terms rather than just the observed CD4 counts. Such joint modelling of longitudinal and event time data [41, 42] would provide useful information regarding the patterns of drop-out from the cohort, but would add further to the computational complexity of estimation.

The fitted models in the present analysis show that there is considerable unexplained variance in the long-term asymptotic maximal response to treatment for each patient, even after accounting for baseline CD4 and time from seroconversion to initiation of HAART, although this might be reduced by the inclusion of additional patient and drug regimen variables into the model. There is also considerable erratic post-treatment variability over time, represented by the fractional Brownian motion process as previously introduced for the analysis of pre-treatment CD4 data [16]. The parameter estimates for the model in which the stochastic process components were generalised to follow marginal multivariate t-distributions indicate substantial between-patient differences in their variability over time, with a moderate positive association between the degree of pre- and post-treatment variability within each patient, which are novel findings in this context. The fact that the models fitted follow a structure that can accommodate any combination of number and timing of observations in each patient means that they can be readily used for simulation studies of patient cohorts.

Conclusions

We have developed a framework for the combined analysis of pre- and post-treatment longitudinal biomarker data and have successfully applied the novel methodology to CD4 data from a cohort of HIV-positive patients with well estimated date of seroconversion. The methodology developed could also be applied to other medical settings in which an intervention is triggered following monitoring of a biomarker of interest, and in which the response to treatment may be conditional on the state of the patient (as indicated by the value of the biomarker) at the time of treatment initiation. Seroconverter cohorts have a special status in HIV research, and in other disease settings the 'zero time' for pre-treatment observations might be time of diagnosis or another clinically significant event. The framework proposed could be applied with different choice of pre- and post-treatment model components, but those demonstrated may be a natural choice in many settings.

Additional files

Additional file 1: Appendices containing (1) details of marginal distribution for post-treatment model and coding for positive-only latent variable, (2) residual plots for $Model_6$ and (3) results of a simulation study demonstrating approximation of non-linear link functions using natural cubic splines. (PDF 1290 kb)

Additional file 2: Tar file containing an R script and ADMB template files to simulate data based on the structure and point estimates of $Model_6$, as described in Results section , and to then refit $Model_4$ and $Model_6$ to these data. (TAR 205 kb)

Abbreviations

ADMB, AD model builder; AIDS, acquired immune deficiency syndrome; BIC, Bayesian information criterion; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus; IQR, interquartile range; MVN, multivariate normal; RCT, randomised controlled trial; UCL, University College London; UK, United Kingdom

Acknowledgements

We would like to thank all the UK HIV Seroconverters Cohort participants for allowing their routine clinical data to be included. We gratefully acknowledge the work of the members of the Steering Committee and colleagues at the clinical centres. Members of the UK Register of HIV Seroconverters Steering Committee are: Andrew Phillips (Chair), University College London (UCL), London; Abdel Babiker, MRC CTU at UCL, London; Valerie Delpech, Public Health England, London; Sarah Fidler, St. Mary's Hospital, London; Amanda Clarke, Brighton & Sussex University Hospitals NHS Trust, Brighton; Julie Fox, Guys and St Thomas' NHS Trust/Kings College, London; Richard Gilson, UCL, London; David Goldberg, Health Protection Scotland, Glasgow; David Hawkins, Chelsea & Westminster NHS Trust, London; Anne Johnson, UCL, London; Margaret Johnson, UCL and Royal Free NHS Trust, London; Ken McLean, West London Centre for Sexual Health, London; Eleni Nastouli, UCL, London; Frank Post, King's College, London. Ronald Geskus provided comments on an early draft of this work and suggested the use of natural cubic spline link functions.

Funding

OTS is supported by a Medical Research Council PhD Studentship and the UK Register of HIV Seroconverters cohort study is funded by the Medical Research Council.

Availability of data and materials

Further details for the UK Register of HIV Seroconverters cohort can be found at: 'http://www.ctu.mrc.ac.uk/our_research/research_areas/hiv/studies/ukr/'. The dataset is not publicly available, as access is only granted for relevant academic research, but proposals outlining the aims and methodology of a research project with a request for access can be sent to the Principal Investigator, Kholoud Porter, via 'enquiries@ctu.mrc.ac.uk'. An R script and ADMB template files are provided to simulate data based on the structure and point estimates of a fitted model, and to then refit correct and simplified models to these data.

Authors' contributions

All authors collaborated in developing the modelling strategy reported. The programming and running of the analysis was carried out by OTS. OTS wrote the first draft of the manuscript, with revisions provided by AGB and AJC. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The UK Register of HIV Seroconverters study has research ethics approval (Medical Research Council Multicentre Research Ethics Committee (MRC MREC), Health Research Authority Research Ethics Service Committee West Midlands - South Birmingham: 04/Q2707/155) and patients provide written informed consent at enrolment.

Received: 5 January 2016 Accepted: 8 July 2016
Published online: 15 September 2016

References

1. Liang KY, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhyā: Indian J Stat Series B. 2000;62:134–48.
2. Liu GF, Lu K, Mogg R, Mallick M, Mehrotra DV. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Stat Med. 2009;28:2509–530.
3. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. 2006;25:4334–44.
4. Kenward MG, White IR, Carpenter JR. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? by G. F. Liu, K. Lu, R. Mogg, M. Mallick and D. V. Mehrotra, Stat Med 2009; 28:2509-2530. Stat Med. 2010;29:1455–6.
5. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in HIV-1-infected Adults and Adolescents. Bethesda: Department of Health and Human Services; 2014. accessed 27 Oct 2014.
6. Williams I, Churchill D, Anderson J, Boffito M, Bower M, Cairns G, Cwynarski K, Edwards S, Fidler S, Fisher M, Freedman A, Geretti AM, Gilleece Y, Horne R, Johnson M, Khoo S, Leen C, Marshall N, Nelson M, Orkin C, Paton N, Phillips A, Post F, Pozniak A, Sabin C, Trevelion R, Ustianowski A, Walsh J, Waters L, Wilkins E, Winston A, Youle M. British HIV Association guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2012 (Updated November 2013). HIV Med. 2014;15 Suppl 1:1–85.
7. INSIGHT START Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. N Engl J Med. 2015;373:795–807.
8. Churchill D, Waters L, Ahmed N, Angus B, Boffito M, Bower M, Dunn D, Edwards S, Emerson C, Fidler S, Fisher M, Horne R, Khoo S, Leen C, Mackie N, Marshall N, Monteiro F, Nelson M, Orkin C, Palfreeman A, Pett S, Phillips A, Post F, Pozniak A, Reeves I, Sabin C, Trevelion R, Walsh J, Wilkins E, Williams I, Winston A. BHIVA Guidelines for the Treatment of HIV-1-positive Adults with Antiretroviral Therapy 2015. London: British HIV Association (BHIVA); 2015.
9. Kaufmann GR, Perrin L, Pantaleo G, Opravil M, Furrer H, Telenti A, Hirschel B, Ledergerber B, Vernazza P, Bernasconi E, Rickenbach M, Egger M, Battegay M, Swiss HIV Cohort Study Group. CD4 T-lymphocyte recovery in individuals with advanced HIV-1 infection receiving potent antiretroviral therapy for 4 years: the Swiss HIV Cohort Study. Arch Intern Med. 2003;163:2187–95.
10. Moore RD, Keruly JC. CD4+ cell count 6 years after commencement of highly active antiretroviral therapy in persons with sustained virologic suppression. Clin Infect Dis. 2007;44:441–6.
11. Lok JJ, Bosch RJ, Benson CA, Collier AC, Robbins GK, Shafer RW, Hughes MD, ALLRT team. Long-term increase in CD4+ T-cell counts during combination antiretroviral therapy for HIV-1 infection. AIDS. 2010;24:1867–76.
12. Le T, Wright EJ, Smith DM, He W, Catano G, Okulicz JF, Young JA, Clark RA, Richman DD, Little SJ, Ahuja SK. Enhanced CD4+ T-cell recovery with earlier HIV-1 antiretroviral therapy. N Engl J Med. 2013;368:218–30.
13. Gazzola L, Tincati C, Bellistrì GM, Monforte AD, Marchetti G. The absence of CD4+ T cell count recovery despite receipt of virologically suppressive highly active antiretroviral therapy: clinical risk, immunological gaps, and therapeutic options. Clin Infect Dis. 2009;48:328–37.
14. Babiker AG, Emery S, Fätkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, Vjechaj MJ, INSIGHT START Study Group. Considerations in the rationale, design and methods of the strategic timing of antiretroviral treatment (START) study. Clin Trials. 2013;10 (1 Suppl):5–36.
15. UK Register of HIV Seroconverters Steering Committee. The AIDS incubation period in the UK estimated from a national register of HIV seroconverters. AIDS. 1998;12:659–67.
16. Stirrup OT, Babiker AG, Carpenter JR, Copas AJ. Fractional Brownian motion and multivariate-t models for longitudinal biomedical data, with application to CD4 counts in HIV-patients. Stat Med. 2016;35:1514–32.
17. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38:963–74.

18. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics. 1990;46:673–87.
19. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat. 1995;4:12–35.
20. Taylor JMG, Cumberland WG, Sy JP. A stochastic model for analysis of longitudinal AIDS data. J Am Stat Assoc. 1994;89:727–36.
21. Wolbers M, Babiker A, Sabin C, Young J, Dorrucci M, Chêne G, Mussini C, Porter K, Bucher HC, CASCADE Collaboration Members. Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy–the CASCADE collaboration: a collaboration of 23 cohort studies. PLoS Med. 2010;7:e1000239.
22. Mandelbrot B, van Ness JW. Fractional brownian motions, fractional noises and applications. SIAM Rev. 1968;10:422–37.
23. Lewis J, Walker AS, Castro H, De Rossi A, Gibb DM, Giaquinto C, Klein N, Callard R. Age and CD4 count at initiation of antiretroviral therapy in HIV-infected children: effects on long-term T-cell reconstitution. J Infect Dis. 2012;205:548–56.
24. Picat MQ, Lewis J, Musiime V, Prendergast A, Nathoo K, Kekitiinwa A, Nahirya Ntege P, Gibb DM, Thiebaut R, Walker AS, Klein N, Callard R, ARROW Trial Team. Predicting patterns of long-term CD4 reconstitution in HIV-infected children starting antiretroviral therapy in sub-Saharan Africa: a cohort-based modelling study. PLoS Med. 2013;10:1001542.
25. Hastie T, Tibshirani R, Friedman J. Basis expansions and regularization. In: The elements of statistical learning: data mining, inference, and prediction. 2nd edn. New York: Springer; 2009. p. 144–6.
26. Kotz S, Nadarajah S. Multivariate t-Distributions and Their Applications. Cambridge: Cambridge University Press; 2004.
27. Pinheiro JC, Liu C, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. J Comput Graph Stat. 2001;10:249–76.
28. Balakrishnan N, Lai CD. Bivariate gamma and related distributions. In: Continuous Bivariate Distributions. 2nd edn. New York: Springer; 2009.
29. Skaug HJ, Fournier DA. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. Comput Stat Data Anal. 2006;51:699–709.
30. Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson E, Maunder MN, Nielsen A, Sibert J. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods Softw. 2012;27:233–49.
31. Moran PAP. Statistical inference with bivariate gamma distributions. Biometrika. 1969;56:627–34.
32. Skaug H, Fournier D. Random effects modeling. In: Random Effects in AD Model Builder: ADMB-RE User Guide. Version 11.4 edn. Honolulu: ADMB Foundation; 2015.
33. Hernán MA, Taubman SL. Does obesity shorten life? the importance of well-defined interventions to answer causal questions. Int J Obes. 2008;32:8–18.
34. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
35. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. Biometrics. 2002;58:621–30.
36. Bolker B, Skaug H, Laake J. R2admb: ADMB to R Interface Functions. 2013. http://CRAN.R-project.org/package=R2admb. Accessed 5 Jan 2016.
37. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York: Springer; 2009.
38. Fitzmaurice G, Laird N, Ware J. Residual analyses and diagnostics. In: Applied Longitudinal Analysis. Hoboken, NJ: Wiley; 2004.
39. Gelman A. Exploratory data analysis for complex models. J Comput Graph Stat. 2004;13:755–79.
40. Harrison L, Dunn DT, Green H, Copas AJ. Modelling the association between patient characteristics and the change over time in a disease measure using observational cohort data. Stat Med. 2009;28:3260–75.
41. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53:330–9.
42. Pantazis N, Touloumi G, Walker AS, Babiker AG. Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. J R Stat Soc Series C (Appl Stat). 2005;54:405–23.