Bayesian inference for an illness-death model for stroke with cognition as a latent time-dependent risk factor

Ardo van den Hout

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K. E-mail: ardo.vandenhout@mrc-bsu.cam.ac.uk *tel. +44 (0)1223 330369*

Jean-Paul Fox

Department of Research Methodology, Measurement, and Data Analysis

Faculty of Behavioral Sciences, Twente University, The Netherlands

Rinke H. Klein Entink

TNO Quality and Safety, Zeist, The Netherlands

Abstract

Longitudinal data can be used to estimate transition intensities between healthy and unhealthy states prior to death. An illness-death model for history of stroke is presented where time-dependent transition intensities are regressed on a latent variable representing cognitive function. The change of this function over time is described by a linear growth model with random effects. Occasion-specific cognitive function is measured by an item response model for longitudinal scores on the Mini-Mental State Examination, a questionnaire used to screen for cognitive impairment. The illness-death model will be used to identify and to explore the relationship between occasion-specific cognitive function and stroke. Combining a multi-state model with the latent growth model defines a joint model which extends current statistical inference regarding disease progression and cognitive function. Markov chain Monte Carlo methods are used for Bayesian inference. Data stem from the Medical Research Council Cognitive Function and Ageing Study in the UK (1991-2005).

1 Introduction

The Medical Research Council Cognitive Function and Ageing Study (MRC CFAS, [1]) has longitudinal information on progression of cardiovascular diseases and information on cognitive function as measured by the Mini-Mental State Examination (MMSE, [2]). One of the interests is to evaluate whether cognitive function can be identified as a risk factor for cardiovascular diseases.

With regard to cardiovascular diseases, we use data on stroke. Occasion-specific cognitive function is modeled as a latent variable and its effect as a risk factor for stroke is investigated by combining a multi-state model for stroke and survival with a growth model for cognition. The relevance of this joint model will be illustrated by addressing survival after a stroke given various trends in cognitive decline, and by estimating the probability of having a stroke in a specified time interval conditional on an MMSE score at the start of the interval and survival up to the end of the interval. In both these cases, the change of cognitive function has an effect and thus illustrates the importance of modeling cognitive function jointly with the multi-state process.

The Bayesian framework is used for statistical inference. It allows individual-specific parameters for cognitive function to be estimated using information from both the multi-state data and the longitudinal MMSE data. Combining the growth model for latent cognitive function with a multi-state model has not been described before, and seems a promising way to handle questionnaire data and related latent variable information in an investigation of a multi-state process.

A continuous-time multi-state model can be used to describe disease progression over time. If one of the states is the death state, the model is called an illness-death model. In the analysis of the CFAS data, individuals are classified in state one if they never had a stroke, and in state two if they experience one or more strokes. State three is the death state. An intensity (hazard) of a transition from one state to another can be linked via a regression equation to risk factors for the transition such as age or sex. We will investigate the effect of cognitive function by modeling it as a risk factor for the transitions in the three-state model for stroke.

Frequentist continuous-time multi-state models can be found in Kalbfleisch and Lawless [3] and Jackson *et al.* [4]. Bayesian inference for parametric multi-state models is discussed in Sharples [5], Welton and Ades [6], Pan, Wu,Yen, and Chen

[7], and Van den Hout and Matthews [8]. Semi-parametric Bayesian methodology can be found in Kneib and Hennerfeind [9].

When risk factors are manifest and time-dependent, and a piecewise-constant approximation of the values seems reasonable, frequentist multi-state models can be fitted by using existing methodology. Jackson [10] provides an R package that can fit a broad range of multi-state models. Prediction in the presence of time-dependent risk factors is, however, not straightforward as the prediction of the multi-state process depends on the distribution of the risk factor.

Specific to the application, cognitive function is a latent time-dependent risk factor and we assume that changes in the function over time can be described by a random-effects linear growth model. Typically, the MMSE response data consist of dichotomous and polytomous item scores. Therefore, a generalized item response theory model will be used for the mixed-response types longitudinal MMSE data. The longitudinal item-based MMSE data are used to measure individual continuous-valued cognitive function scores.

An item response theory (IRT) model [11] assumes that certain observed discrete values are manifestations of an underlying latent construct. With regard to the MMSE, the discrete values are responses to a series of binary questions and one question with five ordered categories, and represent aspects of cognitive functioning. The time-dependent IRT model for longitudinal MMSE data relates the probability of the discrete values to the underlying occasion-specific cognitive function to explain MMSE performance.

Traditionally, the MMSE sum score is used as an estimate of cognitive function. However, using IRT has several advantages. Firstly, item response data contain more information than sum scores and this allows the IRT model to parameterize the items individually. Secondly, the IRT model is better equipped to handle missing data. Thirdly, IRT is more flexible with regard to incomplete designs and different number of items.

A specific problem with the MMSE sum score is that there is often a ceiling effect: many observed sum scores are close to the upper bound. Hence, the standard assumption that the conditional distribution of the observed response in the related growth model is normal is problematic. When cognitive function is assessed using IRT, the ceiling effect is less of a problem since cognitive function is modeled as a latent variable on a continuous scale.

Fox and Glas [12] defined a multilevel population model for a latent variable to account for the nesting of students in schools. This multilevel IRT measurement model is here extended to account for the nesting of time-dependent measurements within subjects and to account for mixed response types (dichotomous and polytomous items).

To summarize, a joint model is proposed for the multi-state data and the MMSE data, where cognitive function is the continuous latent variable that explains variation in the longitudinal MMSE scores and - potentially - variation in the transitions between the states.

For Bayesian inference, Markov chain Monte Carlo methods (MCMC) are used to sample values from the posterior density of the overall model that includes the multi-state model and the IRT growth model. The sampled values are used to compute posterior means, credible intervals, and other posterior quantities of interest.

The overall approach is very flexible and can therefore be used in other applications as well. Because MCMC is applied, random effects are estimated along with population parameters and dealing with missing MMSE item scores is relatively straightforward. In addition, in the estimation of the parameters it is possible to specify the information flow: in our joint model, the parameters for the covariate process are sampled using multi-state data. Both for the growth model and the multi-state model, the number of observations and the times of interview can vary within and between individuals.

The paper is organized as follows. Section 2 introduces the CFAS data and presents some basic descriptive statistics. Section 3 discusses the methods for data analysis: the multi-state model, the IRT linear growth model, model identification, and prior densities. In Section 4, the handling of missing MMSE scores is explained. Section 5 briefly discusses the MCMC that is used for Bayesian inference. The data analysis can be found in Section 6. Section 7 concludes the paper. The MCMC in Section 5 is detailed in the appendix.

2 Data

The Medical Research Council Cognitive Function and Ageing Study (CFAS) is a UK population based study in which individuals have been followed from baseline 1991-1992 ([1], `www.cfas.ac.uk`) up to the last interviews in 2004. All participants

Table 1: For men in CFAS data from Newcastle, frequencies of number of times each pair of states was observed at successive observation times.

	To state			
	1 = *Healthy*	2 = *History of Stroke*	*Death*	*Right-censored*
From state 1	836	49	549	239
2	0	75	116	21

are aged 65 years and above, and all deaths up to the end of 2005 have been included.

The three-state model for stroke is defined as follows. State 1 is the healthy state (no history of stroke), individuals in state 2 have had one or more strokes, and state 3 is the death state. Transitions from 1 to 2 are interval censored (exact times of strokes are not available), but death times are known. By definition, transitions from state 2 to state 1 are not possible.

Cognitive impairment was measured using the MMSE with sum scores in the range 0-30. There are 25 binary questions and one which has a scale from 0 up to 5. The latter is about counting backwards, where a score of 5 is given if the counting is flawless. This question is considered an important item in the MMSE. Note that when working with sum scores, the question can add 5 points to a scale with a total range of 0 up to 30. To simplify the model slightly, we take scores 0 and 1 together in category 1, resulting in ordered scores 1, 2, 3, 4, and 5. An alternative would be to dichotomize the scale but that would mean that the relative importance of the question is lost.

In this paper, we describe and analyze the data for men in Newcastle. The sample size is 925 and in total there are 2810 observations (total number of interviews, right-censored states, and observed deaths). In this data set, the median age at baseline is 73. Time between interviews varies between and within individuals. The median length of the time between two consecutive interviews is 26 months. The median number of interviews is 2.

The frequencies in Table 1 are the number of times each pair of states was observed at successive observation times. The table shows that for all individuals the state in the last record in the study is the death state or a right-censored state: $549 + 116 + 239 + 21 = 925$.

Originally, the MMSE was designed to screen for dementia. It contains questions

on memory, language and orientation. Most of the questions are relatively easy for individuals with average cognition. MMSE sum scores below 10 are indicative of dementia. Individuals with scores in the range 25-30 are said to have normal cognitive functioning. Currently, the MMSE is also widely used to measure overall cognitive function. When the MMSE is applied in a population based study such as CFAS, a large proportion of the observed MMSE sum scores will be in the range 25-30. In the data for men in Newcastle, the median of the MMSE sum score at baseline is 27.

MMSE scores are not always observed. There are 298 missing binary item scores in the records of 28 men. Nine men have a missing score for the five-category question.

3 Methods

In this section, the joint modeling framework is presented for latent growth trajectories and multi-state processes. Firstly, the multi-state model is discussed, followed by the latent growth model part. The derivation of the joint posterior distribution concludes this section.

3.1 The multi-state model

This section presents the likelihood of the continuous-time multi-state model. The basic ideas can be found in Kalbfleisch and Lawless [3] and Jackson *et al.* [4]. The formulation of the likelihood is different from the one in Van den Hout and Matthews [8] where an approximation with regard to exact death times was used. Transition probabilities in the likelihood are conditional on the current state *and* current values of risk factors. Commenges [13] uses the term partial-Markov to denote this kind of multi-state model since using the time-dependent risk factors implies that the process is not first-order Markov.

Let the interval-censored multi-state data be given by $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, where N is the number of individuals in the study. The trajectory of individual i is given by $\boldsymbol{x}_i = (x_{i1}, ..., x_{in_i})$, where n_i is the number of observed states, and state $x_{ij} \in \{1, ..., S\}$, where $j = 1, ..., n_i$ indexes the consecutive times of measurement. Times of observation - not necessarily equidistant - are given by $t_{i1}, ..., t_{in_i}$, where $t_{i1} = 0$,

for all i, denotes the start of the study. For individual i we have observed risk factor values $\boldsymbol{w}_i = (\boldsymbol{w}_{i1}, ..., \boldsymbol{w}_{in_i})$ at times $t_{i1}, ..., t_{in_i}$.

Let $(t, u]$ denote a generic time interval. For a continuous-time multi-state model, transition probabilities $p_{rs}(t, u) = P(x_u = s|x_t = r)$ are the entries of transition matrix $\boldsymbol{P}(t, u)$. Likelihood contributions are formulated using the transition matrices for the observed intervals, but the model itself is defined using intensity matrices which are matrices with transition intensities as entries. The transition matrix $\boldsymbol{P}(t, u)$ is derived from intensity matrix $\boldsymbol{Q}(t)$ by means of $\boldsymbol{P}(t, u) = \exp[(u-t)\boldsymbol{Q}(t)]$, where $\exp[\cdot]$ is the matrix exponential [14]. Off-diagonal entries of $\boldsymbol{Q}(t)$ not restricted to zero can be related to risk factors $\boldsymbol{w}$ by means of a log-linear model $\log[q_{rs}(t_{ij})] = \boldsymbol{\beta}_{rs}^{\top}\boldsymbol{w}_{ij}$. For example, a progressive three-state model where state 3 is the death state has vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{13}, \boldsymbol{\beta}_{23})$.

We assume a piecewise-constant multi-state model where individual trajectories through the states are conditionally independent. For individual i, the likelihood contribution is

$$p(\boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{w}_i) = P(x_{in_i}|x_{i,n_i-1}, \boldsymbol{\beta}, \boldsymbol{w}_{i,n_i-1}) \times ... \times P(x_{i2}|x_{i1}, \boldsymbol{\beta}, \boldsymbol{w}_{i1}).$$

This follows by conditioning on the first state, that is, by restricting $P(x_{i1}|\boldsymbol{\beta}, \boldsymbol{w}_i) = 1$. The likelihood is given by $p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{w}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{w}_i)$. See Appendix A for the construction of the likelihood of the three-state model that is used in the application and which takes into account exact death times and right-censoring and the end of the follow-up.

As implied by the above, we assume that given the current state and the current values of the risk factors, the distribution of the next state does not depend on the states visited before the current state. In addition, we assume that factor values are constant between consecutive observation times. Within each individually observed time interval $(t_{ij}, t_{i,j+1}]$, this defines a time-homogeneous process. By using age as a piecewise-constant time-dependent risk factor, possible dependence of transition intensities on changing age are taken into account [15].

If there are no other risk factors besides age, the model for the intensities is given by $\log[q_{rs}(t_{ij})] = \beta_{rs.1} + \beta_{rs.2}Age(t_{ij})$. This can also be formulated as $q_{rs}(t_{ij}) = \lambda_{rs}\exp[\gamma_{rs}Age(t_{ij})]$, for $\lambda_{rs} > 0$, which shows that the change of the intensities over time follows a Gompertz model with age as the time-scale.

3.2 Linear growth model for latent cognitive function

In our modeling, cognitive function is a latent time-dependent risk factor in the multi-state model. We assume that cognitive function is continuous and that the time-dependency can be described by a linear growth model. In the growth model, the function is represented by the variable θ.

For individual i with observation times $t_{i1}, ..., t_{in_i}$, let $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{in_i})$. The growth model is given by

$$\theta_{ij} = \eta_{1i} + \eta_{2i} t_{ij} + e_{ij} \qquad \begin{aligned} \boldsymbol{\eta}_i = (\eta_{1i}, \eta_{2i}) &\sim MVN(\boldsymbol{\nu}, \boldsymbol{\Sigma}) \\ e_{ij} &\sim N(0, \sigma^2). \end{aligned}$$

That is, random effects $\boldsymbol{\eta}_i$ are multivariate normally distributed with unknown mean $\boldsymbol{\nu} = (\nu_1, \nu_2)$ and 2×2 variance-covariance matrix $\boldsymbol{\Sigma}$. The conditional distribution of θ_{ij} is normal with unknown variance σ^2. Random intercept η_{1i} is the value of θ_{ij} at the start of the study at time $t_{ij} = 0$. Random slope η_{2i} reflects the change of θ_{ij} over time, where a negative value corresponds to a decline of ability over time.

Cognitive function is a latent variable as it cannot be observed directly but is measured by the MMSE. At every observation time, the MMSE consists of $K = 25$ binary items (questions) and one item with five ordered answer categories. Item Response Theory (IRT) models are used to link the observed discrete values to latent function $\boldsymbol{\theta}$.

For individual i, the data for the binary response IRT model are given by $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, ..., \boldsymbol{y}_{in_i})$ with $\boldsymbol{y}_{ij} = (y_{ij1}, ..., y_{ijK})$. The probability of individual i answering binary item k correctly at time t_{ij} given item parameters $\boldsymbol{a} = (a_1, ..., a_K)$ and $\boldsymbol{b} = (b_1, ..., b_K)$ is defined using the probit model

$$P(y_{ijk} = 1 | \theta_{ij}, a_k, b_k) = \Phi(a_k \theta_{ij} - b_k), \tag{1}$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal. The probit model is well established in the IRT literature for cross-sectional binary response data. The logit model is sometimes used as an alternative, but in practice results for both models are similar. We prefer the probit model because it has a more simple implementation in MCMC.

For $k = 1, ..., K$, parameter a_k is called a *discrimination parameter* and is the effect of a unit change in cognitive function θ on the success probability for item k.

Parameter b_k is a *difficulty parameter* and is the effect on the success probability when $\theta = 0$. Note that a large negative value of b_k corresponds to a relative easy question.

Time-specific response data are assumed to be independent given time-specific cognitive function. This makes it possible to factorize the likelihood and we obtain

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \prod_{k=1}^{K} P(y_{ijk} = 1|\theta_{ij}, a_k, b_k)^{y_{ijk}} \left(1 - P(y_{ijk} = 1|\theta_{ij}, a_k, b_k)\right)^{(1-y_{ijk})}.$$

For the item with the five ordered response categories we use the graded response model [16]. Let $\boldsymbol{u}_i = (u_{i1}, ..., u_{in_i})$ denote the polytomous data for individual i. Given response categories 1 up to 5 (with the latter denoting the best score), the model has four ordered thresholds parameters $d_1, .., d_4$. Together with the bounds $d_0 = -\infty$ and $d_5 = \infty$, and the ordering $d_0 < d_1 < d_2 < d_3 < d_4 < d_5$, these thresholds define five segments on the real line. The graded response model written in cumulative normal response probabilities has parameters c and $\boldsymbol{d} = (d_1, d_2, d_3, d_4)$, and is given by

$$P(u_{ij} = m|\theta_{ij}, c, \boldsymbol{d}) \quad = \quad \Phi(c\theta_{ij} - d_{m-1}) - \Phi(c\theta_{ij} - d_m), \tag{2}$$

for $m = 1, .., 5$ [17]. The model defines the probabilities of the five answer categories. Parameter c is the discrimination parameter, and $\boldsymbol{d}$ is the difficulty parameter. As an example, when d_1 is a large positive number, the first segment from $-\infty$ up to d_1 is large compared to the other segments. This implies that category 1 corresponds to a high probability and this reflects a difficult item. When d_4 is a large negative number, it is relative easy to obtain a score of 5. Notice that for an item with two categories, the thresholds would be $-\infty = d_0 < d_1 < d_2 = \infty$ and the graded response model reduces to the two-parameter (normal ogive) IRT model (1).

Fox [17] formulates this model for cross-sectional data, but - as above - given the conditioning on θ_{ij}, the same model can be used for longitudinal data. The likelihood is

$$p(\boldsymbol{u}|\boldsymbol{\theta}, c, \boldsymbol{d}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \sum_{m=1}^{5} P(u_{ij} = m|\theta_{ij}, c, \boldsymbol{d}) \delta(u_{ij} = m),$$

where $\delta(u = m) = 1$ if $u = m$ and 0 otherwise.

Analogous to the standard cross-sectional IRT model, we identify the growth model by fixing the scale of cognitive function $\boldsymbol{\theta}$. Note that for this variable only differences are important - values considered at face value are not informative. The mean and the variance of $\boldsymbol{\theta}$ are fixed to zero and one, respectively (cf. [17], sec. 4.4.2).

3.3 Posterior and prior densities

Bayesian inference is based on the posterior density of the model parameters. The posterior density is proportional to the likelihood of the data times the prior density of the model parameters. Ignoring manifest risk factors in the notation, the posterior of our model is given by

$$\begin{aligned} &p(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2 | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}) \\ &\quad \propto p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u} | \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2) p(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2), \end{aligned} \tag{3}$$

where $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u} | \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2)$ is the overall likelihood of the multi-state data $\boldsymbol{x}$, and MMSE data $\boldsymbol{y}$ and $\boldsymbol{u}$. Given the model specification in Section 3.2, it follows that

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u} | \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2) = p(\boldsymbol{x} | \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{y} | \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}) p(\boldsymbol{u} | c, \boldsymbol{d}, \boldsymbol{\theta})$$

The prior density for the parameters in (3) is given by

$$\begin{aligned} &p(\boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}, \sigma^2) \\ &\quad = p(\boldsymbol{\theta} | \boldsymbol{\eta}, \sigma^2) p(\boldsymbol{\eta} | \boldsymbol{\nu}, \boldsymbol{\Sigma}^{-1}) p(\boldsymbol{\beta}) p(\boldsymbol{a}) p(\boldsymbol{b}) p(c) p(\boldsymbol{d}) p(\boldsymbol{\nu}) p(\boldsymbol{\Sigma}^{-1}) p(\sigma^2), \end{aligned}$$

where the conditional distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are specified in Section 3.2.

For the parameter $\boldsymbol{\beta}$ of the three-state model, we use a non-informative (improper) prior density: $p(\boldsymbol{\beta}) \propto 1$. For the parameters of the growth model, the prior densities are given by

$$\begin{aligned} \boldsymbol{\nu} &\sim MVN(\boldsymbol{\nu}_0, \boldsymbol{C}) \\ \boldsymbol{\Sigma}^{-1} &\sim Wishart((\rho \boldsymbol{R})^{-1}, \rho) \\ \sigma^2 &\sim Inv\text{–}Gamma\,(\xi, \xi)\,, \end{aligned}$$

see Gelfland *et al.* [18]. These conjugate priors allow a straightforward implementation of the Gibb sampler that we use for the growth model. The choice of the hyper parameters is discussed in the application. For the IRT model, we use non-informative prior densities for the item parameters: $p(\boldsymbol{a}), p(\boldsymbol{b}), p(c), p(\boldsymbol{d}) \propto 1$.

4 Missing scores on test items

In CFAS, not all the MMSE questions are answered by all the individuals. Missing values are ubiquitous in statistical analysis, and we are not the first to point out the Bayesian framework is very suitable for dealing with certain forms of missingness.

We will assume that values are missing at random [19], i.e., the missingness does not depend on the missing value itself, but may depend on observed data. It will further be assumed that the parameters for the distribution of $\boldsymbol{\theta}$ and the parameters for the distribution of the missing-data mechanism are *a priori* independent. With these two assumptions, the missing-data mechanism is assumed to be ignorable for Bayesian inference ([20], def. 6.5). Given this assumption, Bayesian inference for the IRT model is relatively easy when item scores are missing. If, for example, for individual i at time t_{ij}, the value of y_{ijk} is missing, then the likelihood contribution for the items scores at t_{ij} can be formulated by using the model for the items $1, ..., k-1, k+1, ..., K$.

This flexible structure with respect to missing values is one of the reasons why we prefer to use an IRT model instead of using observed sum scores. The definition of a sum score is problematic when one or more item scores are missing.

Although we can estimate the model by ignoring the missing item scores, the Markov chain Monte Carlo (MCMC) method in the next section is easier to implement when we sample the scores along the way. In the MCMC algorithm, the missing scores are sampled first, after which the sampling of the model parameters proceeds as in the complete data case.

We illustrate the procedure for the binary response data. Given the probit model, latent cognitive function $\boldsymbol{\theta}$, and item parameters $\boldsymbol{a}$ and $\boldsymbol{b}$, sampling missing values is undertaken by using Bernoulli trials. If at time t_{ij}, the binary value of y_{ijk} is missing, then we use a trial with success probability $\Phi(a_k\theta_{ijk} - b_k)$. By sampling missing values in each iteration of the MCMC algorithm, the uncertainty with regard to the missing values is propagated into the sampling of the model parameters.

For a missing values of polytomous u_{ij}, values are sampled in a similar way using the multinomial distribution and parameters c and $\boldsymbol{d}$.

5 Bayesian inference

Markov chain Monte Carlo (MCMC) methods are used to sample from the posterior distribution over the unknown parameters. The algorithm we use is a Gibbs sampler [21] where each parameter is sampled conditional on the other parameters and the data. In case there is no closed form of the conditional probability distribution, Metropolis [22] or Metropolis-Hasting sampling [23] is undertaken. This scheme is sometimes known as *Metropolis-within-Gibbs* although some authors dislike this term, see the discussion in Carlin and Louis ([24], sec. 3.4.4).

To summarize, data of individual i at time t_{ij} consist of observed states x_{ij}, binary response y_{ijk} for item k, and polytomous response u_{ij}. Latent cognitive function is denoted θ_{ij}. The parameter vector for the three-state model is $\boldsymbol{\beta}$. Item parameters are $\boldsymbol{a} = (a_1, ..., a_K)$ and $\boldsymbol{b} = (b_1, ..., b_K)$, for the dichotomous item response model, and c and $\boldsymbol{d} = (d_1, ..., d_4)$ for the polytomous item response model. Parameters for the growth model are given by $\boldsymbol{\Omega} = (\boldsymbol{\nu}, \boldsymbol{\eta}, \boldsymbol{\Sigma}, \sigma)$. Conditioning on manifest risk factors $\boldsymbol{w}$ is ignored in the following notation.

Sampling the parameters of the IRT model for the dichotomous response is undertaken by using an auxiliary variable $\boldsymbol{z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)$. This is a continuous representation of binary data $\boldsymbol{y}$ which makes it possible to formulate a Gibbs sampler [25]. Corresponding to each y_{ijk} we define the latent variable z_{ijk} which is normally distributed with mean $a_k\theta_{ijk} - b_k$ and standard deviation 1. Value $y_{ijk} = 1$ is observed when $z_{ijk} > 0$, and $y_{ijk} = 0$ is observed, when $z_{ijk} \leq 0$.

An innovative step in our Gibbs sampler is the sampling of $\boldsymbol{\theta}$. This parameter vector is sampled using a Metropolis step where the sampling is informed by both the IRT data and the multi-state data. This illustrates the flexibility and the strength of MCMC.

Here, we enumerate the main steps of the Gibbs sampling, where conditioning on all other parameters is indicated by three dots, e.g., $p(\boldsymbol{a}|...)$. Details of each step and further references can be found in Appendix B.

1. Sample missing binary scores y_{ijk}^{mis} from $p(y_{ijk}^{\text{mis}}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})$.

2. Sample missing polytomous scores u_{ij}^{mis} from $p(u_{ij}^{\text{mis}}|\boldsymbol{\theta}, c, \boldsymbol{d})$.
3. Sample $\boldsymbol{z}$ from $p(\boldsymbol{z}|...) \propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y})$.
4. Metropolis sampling of $\boldsymbol{\theta}$.

    - A proposal distribution is specified by sampling from $p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}) \propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})p(\boldsymbol{\theta}|\boldsymbol{\Omega})$.
    - The vector $\boldsymbol{\theta}$ sampled from the proposal distribution is re-scaled such that the resulting values have mean 0 and variance 1.
    - Sampled and re-scaled $\boldsymbol{\theta}$ is the candidate for sampling from $p(\theta_{ij}|...) \propto p(y_{ij}|\theta_{ij}, \boldsymbol{a}, \boldsymbol{b})p(u_{ij}|\theta_{ij}, c, \boldsymbol{d})p(x_{i,j+1}|x_{ij}, \theta_{ij}, \boldsymbol{\beta})p(\theta_{ij}|\boldsymbol{\Omega})$.

5. Sample $\boldsymbol{a}$ from $p(\boldsymbol{a}|...) \propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})p(\boldsymbol{a})$.
6. Sample $\boldsymbol{b}$ from $p(\boldsymbol{b}|...) \propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})p(\boldsymbol{b})$.
7. Sample c from $p(c|...) \propto p(\boldsymbol{u}|\boldsymbol{\theta}, c, \boldsymbol{d})p(c)$.
8. Sample $\boldsymbol{d}$ from $p(\boldsymbol{d}|...) \propto p(\boldsymbol{u}|\boldsymbol{\theta}, c, \boldsymbol{d})p(\boldsymbol{d})$.
9. Sample $\boldsymbol{\Omega}$ using a standard scheme for a linear mixed model where $\boldsymbol{\theta}$ is the response variable.
10. Sample $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|...) \propto p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta})$.

Posterior inference with regard to means, credible intervals, and other derived quantities is based upon two chains, each with a burn-in of 5000 and an additional 15000 updates. Convergence of the chains for the item parameters and the parameters for the growth model are assessed by visual inspection of the chains and by diagnostics tools provided in the R-package coda [26] such as the convergence diagnostic by Geweke [27].

To compare models, we used the Deviance Information Criterion (DIC, [28]). The DIC comparison is based on a trade-off between the fit of the data to the model and the complexity of the model. Models with smaller DIC are better supported by the data. The deviance of interest is the deviance of the multi-state model given by

$$D(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}) = -2 \log p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{\theta}).$$

The DIC for the multi-state model is given by

$$DIC_{msm} = \widehat{D} + 2p_D,$$

where $\widehat{D} = D(\boldsymbol{x}, \boldsymbol{w}, E(\boldsymbol{\beta}), E(\boldsymbol{\theta}))$ and p_D denotes the effective number of parameters in the multi-state model. The latter can be estimated by $\overline{D} - \widehat{D}$, where $\overline{D} = M^{-1} \sum_{m=1}^{M} D(\boldsymbol{x}, \boldsymbol{\beta}^m, \boldsymbol{w}, \boldsymbol{\theta}^m)$ with m denoting the iterations in the MCMC algorithm. The DIC is therefore estimated by $DIC_{msm} = 2\overline{D} - \widehat{D}$, where $E(\boldsymbol{\beta})$ and $E(\boldsymbol{\theta})$ are estimated using the posterior means.

6 Application

The longitudinal MMSE data and multi-state data from the 925 men in CFAS in Newcastle will now be analyzed. As stated before, in the three-state model, state 1 is the healthy state (no history of stroke), individuals in state 2 have had one or more strokes, and state 3 is the death state. In the MMSE, there are 25 binary questions and one which is scored from 1 up to 5.

6.1 Estimation

Although the focus of the analysis is the three-state model, we briefly discuss inference for the growth model for the MMSE data.

The choice of the hyper parameters for the prior densities is $\boldsymbol{\nu}_0 = (0, 0)$, $\boldsymbol{C}^{-1} = \boldsymbol{0}$, $\xi_0 = 1/100$, $\rho = 2$, and $\boldsymbol{R} = 10\boldsymbol{I}_2$, where $\boldsymbol{I}_2$ is the 2×2 identity matrix. This choice defines vague priors.

Posterior means and credible intervals (CIs) for the parameters of the growth model are presented in Table 2. The negative posterior mean -0.036 for ν_2 which is the mean of the random slopes in the growth model concurs with our expectations. In the older population, if there is a change of cognitive function over a long time, then this will be a decline. The posterior mean 0.097 for $\boldsymbol{\Sigma}_{22}$ reflects the heterogeneity that is present in the data with regard to these slopes. Interesting is also the negative posterior mean of covariance $\boldsymbol{\Sigma}_{12}$, which means, for example, that a high intercept (high cognitive function) correlates with a small slope (less decline over time).

We do not aim to investigate the effect of the individual items in the MMSE. Nevertheless, it is interesting to see that there is indeed variation in the item-specific characteristics. For the parameters for the binary items see Figure 1. This illustrates why we are using an IRT model in the first place: assuming for instance that all

Table 2: Posterior inference for model parameters with 95% credible intervals in parentheses.

Three-state model					
Intercept	$\beta_{12.1}$	-3.740 (-4.079; -3.437)	Cognitive	$\beta_{12.3}$	-0.502 (-0.884; -0.120)
	$\beta_{13.1}$	-2.717 (-2.846; -2.590)	function	$\beta_{13.3}$	-0.524 (-0.663; -0.381)
	$\beta_{23.1}$	-1.766 (-2.007; -1.543)		$\beta_{23.3}$	-0.181 (-0.309; -0.056)
Age	$\beta_{12.2}$	0.062 (0.003; 0.115)			
	$\beta_{13.2}$	0.020 (-0.005; 0.044)			
	$\beta_{23.2}$	0.024 (-0.007; 0.054)			
Growth model					
	ν_1	0.098 (0.009; 0.188)		Σ_{11}	0.264 (0.209; 0.329)
	ν_2	-0.036 (-0.078; 0.006)		Σ_{12}	-0.025 (-0.047; -0.006)
	σ	1.422 (1.338; 1.511)		Σ_{22}	0.097 (0.083; 0.110)

questions are equally difficult is clearly incorrect (bottom part of Figure 1). Note that all difficulty parameters have a posterior mean smaller than zero. This reflects that for most people the MMSE items are easy. And this is as expected since the MMSE is originally constructed to screen for dementia and the questions are relatively easy for the majority of the individuals in CFAS. The variation in the discrimination parameters (top part of Figure 1) shows that some items are better at discriminating individual cognitive function than others.

For the graded response model, the sampling of the threshold parameters d_1, d_2, d_3, and d_4 is depicted in Figure 2. The best way to sample threshold parameters has been a topic in the literature ([17], sec. 4.3.4) and the references therein. We used truncated normal distributions to generate new candidates in the Metropolis-Hasting step for $\boldsymbol{d} = (d_1, d_2, d_3, d_4)$, see Appendix B. Figure 2 illustrates that this sampling scheme works well. Numerical diagnostics for convergence as provided in coda [26] all indicate good convergence.

We now turn to the three-state model for stroke. The intensities are linked to age and cognitive function via the log-linear regression model given by

$$\log[q_{rs}(t_{ij})] = \beta_{rs.1} + \beta_{rs.2} Age(t_{ij}) + \beta_{rs.3}\theta(t_{ij}), \tag{4}$$

where $Age(t_{ij})$ is the age midway through the interval $(t_{ij}, t_{i,j+1}]$ minus 75 years,

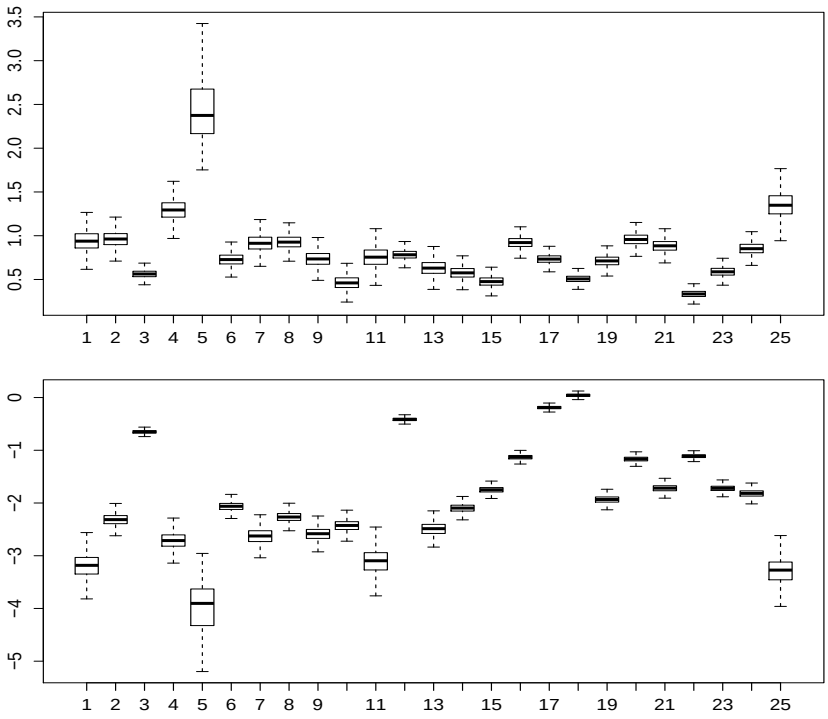

Figure 1: Posterior inference for item parameters using boxplots. Discrimination parameters $\boldsymbol{a}$ in top graph, difficulty parameters $\boldsymbol{b}$ in the bottom one.

and $\theta(t_{ij}) = \theta_{ij}$ denotes latent cognitive function at time t_{ij}.

We start by examining whether adding age and cognitive function as risk factors provides a better model than the intercept-only model. The latter has $DIC_{msm} = 4825$. The model with age but without cognitive function has $DIC_{msm} = 4777$. Clearly, we get a better model by adding age. The final model, i.e., (4) with no restrictions, has $DIC_{msm} = 4680$ which shows that taking cognitive function into account is worthwhile. Posterior inference for $\boldsymbol{\beta}$ in (4) is presented in Table 2.

The sign of the estimated effects of risk factors age and cognitive function are as expected: positive for age (getting older increases the risk of a transition) and negative for cognitive function (higher function is associated with a lower risk). Direct interpretation of the numerical results for the estimated effects is of limited use, see Section 6.4 for interpretation using estimated survival.

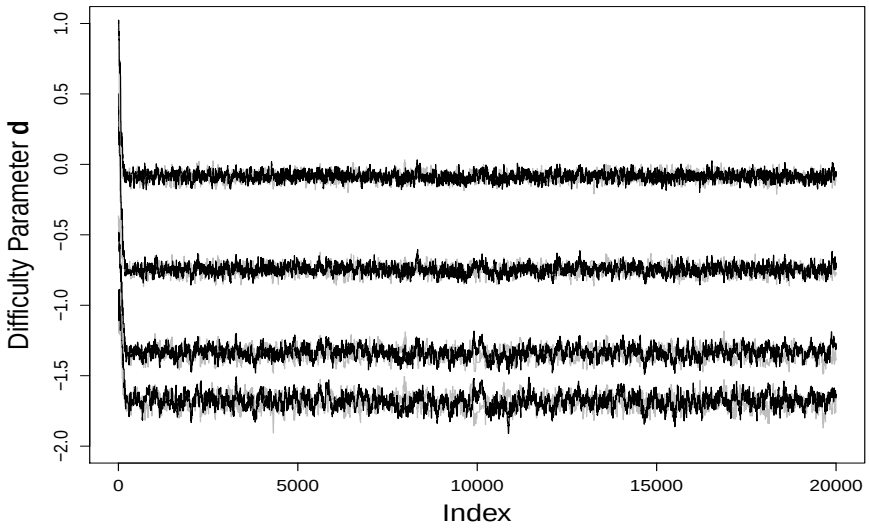

Figure 2: Monte Carlo Markov chains for the difficulty parameter vector $\boldsymbol{d}$ with thresholds d_1, d_2, d_3, and d_4. Burn-in included. Colors black and grey for the two set of starting values.

6.2 Goodness of fit

Model validation is undertaken by a *posterior predictive model check* [29]. Validation is hampered by the interval censoring of the transitions between the healthy state and the state defined by a history of stroke. Death times are, however, observed during the follow-up. We propose to validate the model by comparing deaths observed during follow-up with simulated deaths given the posterior distribution of the parameters. This does not capture all aspects of the three-state model, but nevertheless gives an idea of goodness of fit: if the simulated deaths differ significantly from observed deaths, then the model cannot be trusted.

We use a test statistic that depends both on observed deaths (say data $\boldsymbol{x}_d$) and on model parameters (denoted here by $\boldsymbol{\xi}$). For the time grid $0, 2, 4, 8, 10, 12, 14, 16$ in years since baseline, observed cumulative numbers of deaths at the grid points are given by $O = (0, 121, 250, 465, 552, 620, 664, 665)$. Notice that the last figure is the sum of the numbers of transitions into the death state in Table 1. Let E be the corresponding vector with the cumulative numbers of expected deaths given model parameters. We define the statistic $T(\boldsymbol{x}_d, \boldsymbol{\xi}) = \sum(O - E)^2/E$. The model check is

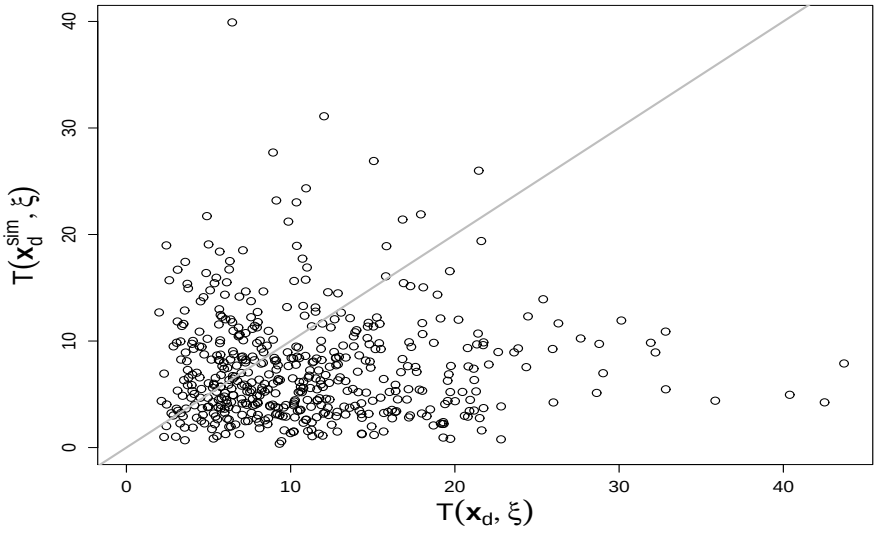

Figure 3: Posterior predictive model check. Comparing $T(\boldsymbol{x}_d^{sim}, \boldsymbol{\xi})$ and $T(\boldsymbol{x}_d, \boldsymbol{\xi})$ for 500 draws of $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ from its posterior distribution.

the comparison of $T(\boldsymbol{x}_d, \boldsymbol{\xi})$ with $T(\boldsymbol{x}_d^{sim}, \boldsymbol{\xi})$ where $\boldsymbol{\xi}$ varies according to its posterior distribution, and $\boldsymbol{x}_d^{sim}$ denotes simulated deaths given $\boldsymbol{\xi}$. The estimate of the p-value is the proportion of simulations where $T(\boldsymbol{x}_d^{sim}, \boldsymbol{\xi}) \geq T(\boldsymbol{x}_d, \boldsymbol{\xi})$. A p-value close to 0 or close to 1 means that the observed cumulative numbers of deaths are not very likely given the model. This would indicate a lack of model fit.

In the model check, given sampled $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\eta})$, deaths are simulated conditional on observed individual data (state and age) at baseline. At the grid points, age of individual i is known given age at baseline, and cognitive function $\boldsymbol{\theta}_i$ is derived given time and sampled random intercept η_{1i} and slope η_{2i}. Simulation of the three-state survival conditional on baseline state can then be undertaken and simulated death times are monitored. In this simulation, the intensities change piecewise-constantly from grid point to grid point. The algorithm is a Gillespie algorithm [30], and is also used and explained in Van den Hout and Matthews [15] where all risk factors are manifest.

We used 500 random samples from the MCMC for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, and obtained the p-value 0.30. Figure 3 depicts simulated $T(\boldsymbol{x}_d^{sim}, \boldsymbol{\xi})$ and $T(\boldsymbol{x}_d, \boldsymbol{\xi})$. With respect to observed deaths during follow-up, the model seems to fit the data well.

6.3 Prediction

Although posterior means for $\boldsymbol{\beta}$ are informative with regard to the direction of the effect of a risk factor, for a practical understanding of the effect it is more useful to investigate predicted survival. Examples will be presented for three individuals: A, B, and C, where the first two are hypothetical, and C is an individual in the study.

Consider the case of A who has had a stroke in the past. What is his survival curve (probabilities of not dying) for the next 15 years? According to our model this depends on current and future cognitive function. Assume that his current function is equal to the estimated population mean ($\eta_{A1} = \nu_1$). We consider baseline ages 65, 75, and 85. For each choice of baseline age, Figure 4 shows two survival curves conditional on assumptions with regard to the slope parameter in the growth model. For A we assume that the slope is equal to the mean of its population distribution *plus* one standard deviation of that distribution ($\eta_{A2} = \nu_2 + \Sigma_{22}^{1/2}$). The solid line is the estimated survival for A. Individual B is as A, except for his slope parameter which is equal to the mean of its population distribution *minus* one standard deviation ($\eta_{B2} = \nu_2 - \Sigma_{22}^{1/2}$). The dashed line is estimated survival for B. The uncertainty in the graph (the 95% CIs) is with regard to the posterior distribution of $\boldsymbol{\beta}$. Even though the CI-bands are quite wide, there is a clear and relevant difference in survival due to difference in future cognitive function.

When it comes to prediction in practice, we would like to predict survival conditional on observed MMSE scores at baseline. Individual C has baseline scores $\boldsymbol{y}_{C1}$ and $\boldsymbol{u}_{C1}$. The posterior of $\theta_{C1} = \eta_{C1}$ is given by

$$p(\eta_{C1}|\boldsymbol{y}_{C1}, \boldsymbol{u}_{C1}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\nu}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{y}_{C1}|\eta_{C1}, \boldsymbol{a}, \boldsymbol{b}) p(\boldsymbol{y}_{C1}|\eta_{C1}, c, \boldsymbol{d}) p(\eta_{C1}|\boldsymbol{\nu}, \boldsymbol{\Sigma}), \quad (5)$$

where $p(\boldsymbol{y}_{C1}|..)$ and $p(\boldsymbol{u}_{C1}|..)$ are likelihood contributions and $p(\eta_{C1}|..)$ is the density of the normal distribution with mean ν_1 and variance $\boldsymbol{\Sigma}_{11}$. Maximizing (5) yields the most likely value of η_{C1} conditional on the posterior means of the model parameters. This is called maximum a posterior (MAP) estimation.

C is an actual man in the data set. At baseline, he is 69 years old, has an MMSE sum score of 23, and has no history of stroke. The MAP estimate of baseline function is -0.670 which is in the lower part of the estimated population distribution with mean ν_1. Given baseline state 1 and assuming that the C's slope for the trend of cognitive function is the estimated mean ν_2 for the population, we can estimate survival. The bottom right graph in Figure 4 depicts this survival.

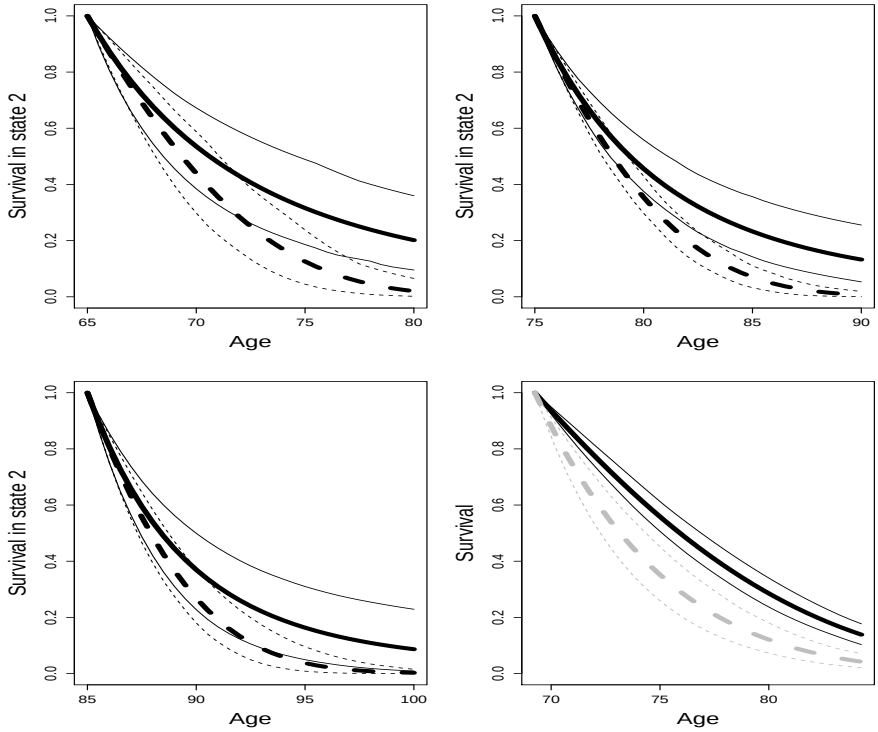

Figure 4: Prediction for men in state 2 at baseline, aged 65, 75, and 85 years old. Solid lines for survival if slope in growth model is equal to population mean plus one standard deviation, dashed lines for slope equal to population mean minus one standard deviation (thin lines for 95% CIs). Prediction of survival for selected individual who is in state 1 at baseline, aged 69 (grey lines if baseline state would have been 2).

We consider possible transition from state 1 to state 2. For C, the probability that he will be in state 2 after 15 years (estimated at 0.047) is less interesting than the probability of being in state 2 conditional on being still alive after 12 years. The latter is estimated at $0.047/(1-0.862) = 0.341$ with 95% CI (0.244; 0.521), where the uncertainty is with regard to the posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}$, and σ. Given the conditioning on baseline function η_{C1}, we used

$$\eta_{C2}|\eta_{C1}, \boldsymbol{\nu}, \boldsymbol{\Sigma} \;\; \sim \;\; N\left(\nu_2 + \frac{\sigma_{\nu_1}}{\sigma_{\nu_2}}\rho(\eta_{C1} - \nu_1), (1-\rho^2)\sigma^2_{\nu_2}\right)$$

where ρ is the correlation between intercept η_1 and slope η_2, derived from $\boldsymbol{\Sigma}$. This

conditional distribution follows from the distribution of $Z_2|Z_1 = z_1$ when both Z_1 and Z_2 are normally distributed ([31], sec. 3.5.2).

7 Conclusion

This paper presented an application where a three-state model for stroke and survival encompasses a latent growth model for time-dependent cognitive function using longitudinal MMSE data. The cognitive function was included in the joint analysis as a time-dependent risk factor for transitions in the three-state model.

Adding the MMSE sum score as a non-deterministic time-dependent risk factor is not a problem with respect to the estimation of a multi-state model when we assume that the piecewise-constant approximation is reasonable. However, for prediction we need a model for the time-dependent risk factor. A growth model with the MMSE sum score as response variable is problematic because the conditional distribution of the sum score is not normal, as the scale is discrete and there are ceiling effects. The binomial distribution is an alternative for the response distribution, but this distribution does not distinguish between the items (questions) that make up the sum score. It is only when IRT models are used that both the discrete nature of the MMSE and the item-specific characteristics are taken into account.

The presented growth model is an extension from the one introduced by Douglas [32]. Our model can deal with variation in time intervals between interviews and is more flexible due to the random-effects structure.

Both within the three-state model and the growth model we have used assumptions that are commonly made. In the multi-state process, the transition probabilities are conditional on the current state *and* current values of risk factors. Using the time-dependent risk factors implies that the process is not first-order Markov. The process is also not semi-Markov because time spent in the current state is not taken into account. Another important assumption is that the piecewise-constant approximation captures the essential part of time-dependent risk factors. The IRT for cognitive function in the growth model assumes local independence (given the item parameters, scores are independently distributed) and time-independent item parameters. A posterior model check was used to validate the model in the application.

In the three-state model for the history of stroke, each individually observed

interval (say $(t_{ij}, t_{i,j+1}]$ for individual i) is modeled in the likelihood as a homogenous process where values of risk factors at time t_{ij} are used to determine the distribution of the states at time $t_{i,j+1}$. It is because of this that we can say lower cognitive function is associated with a higher risk of stroke. Due to the piecewise-constant approximation, the model is not invalidated by the fact that a stroke often causes a drop in cognitive function. For example, if a stroke occurred within $(t_{ij}, t_{i,j+1}]$ and there is a drop in function, then the decreased function will only play a role in the modeling of the next interval $(t_{i,j+1}, t_{i,j+2}]$.

The use of MCMC methods ensures proper propagation of the uncertainty at the various levels of the model. By using a random-effects growth model, individual heterogeneity is taken into account. Given the general structure of the model, it can be extended easily, for example, with additional covariates in the growth model or in the multi-state model. Possible sub-models may also be of interest. For example, if there is no MMSE information available, the growth model can be dropped from the overall model, and θ_{ij} can take the role of a frailty which takes into account unobserved heterogeneity with regard to the risk of ill-health or death.

Acknowledgements

MRC CFAS is supported by major awards from the UK Medical Research Council and the Department of Health (grant MRC/G99001400). A. van den Hout is funded by the Medical Research Council WBS U.1052.00.013. The collaboration was supported by a grant from the British Council and Platform Beta Techniek (`www.britishcouncil.org/netherlands`).

References

1. Brayne C, McCracken C and Matthews FE. Cohort profile: the Medical Research Council Cognitive Function and Ageing Study (CFAS). *International Journal of Epidemiology* 2006; 35: 1140–1145.
2. Folstein MF, Folstein SE and McHugh PR. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; 12: 189–198.
3. Kalbfleisch J and Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 1985; 80: 863–871.

4. Jackson CH, Sharples LD, Thompson SG, Duffy SW and Couto E. Multi-state Markov models for disease progression with classification error. *Statistician* 2003; 52: 193–209.
5. Sharples LD. Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in Medicine*, 1993; 12: 1115–1169.
6. Welton NJ, and Ades AD. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making* 2005; 25: 633–645.
7. Pan SL, Wu HM, Yen AMF and Chen THH. A Markov regression random-effects model for remission of functional disability in patients following a first stroke: A Bayesian approach. *Statistics in Medicine* 2007; 26: 5335–5353.
8. Van den Hout A and Matthews FE (2009).Estimating dementia-free life expectancy for Parkinson's patients using Bayesian inference and micro-simulation. *Biostatistics* 2009; 10: 729–743.
9. Kneib T. and Hennerfeind A. Bayesian semi parametric multi-state models. *Statistical Modelling* 2008; 8: 169–198.
10. Jackson CH. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software* 2011; 38.
11. Van der Linden WJ and Hambelton RK. *Handbook of Modern Item Response Theory.* New York: Springer, 1997.
12. Fox J-P and Glas CAW. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 2001; 66: 271–288.
13. Commenges D. Multi-state models in epidemiology. *Lifetime Data Analysis* 1999; 5: 315-327.
14. Norris JR. *Markov Chains.* Cambridge: Cambridge University Press, 1997.
15. Van den Hout A and Matthews FE (2008). A piecewise-constant Markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Statistical Methods in Medical Research* 2008; 18: 145–162.
16. Samejima F. The graded response model. In: Van der Linden WJ and Hambleton RK (eds) *Handbook of modern item response theory.* New York: Springer, 1997, pp. 85–100.
17. Fox J-P. *Bayesian Item Response Modeling*, New York: Springer, 2010.
18. Gelfand AE, Hills SE, Racine-Poon A and Smith AFM. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 1990; 85: 972–985.
19. Rubin DB. Inference and missing data (with discussion). *Biometrika* 1976; 63: 581–592.
20. Little RJA and Rubin DB. *Statistical Analysis With Missing Data.* 2nd ed. Hoboken, New Jersey: Wiley, 2002.

21. Geman S and Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* 1984; 6: 721–741.
22. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E. Equation of state calculation by fast computing machines. *Journal of Chemical Physics* 1953; 21: 1087–1092.
23. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; 57: 97–109.
24. Carlin BP and Louis TA. *Bayesian Methods for Data Analysis.* 3rd ed. Boca Raton, Florida: Chapman and Hall/CRC, 2009.
25. Johnson VE and Albert JH. *Ordinal Data Modeling.* New York: Springer, 1999.
26. Plummer M, Best N, Cowles K and Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* 2006; 6: 7–11.
27. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (ed JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK. 1992: pp 169-193.
28. Spiegelhalter DJ, Best NG, Carlin BP and Van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 2002; 4: 583–640.
29. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* 1984; 12: 1151–72.
30. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 1977; 25: 2340–2361.
31. Rice JA. *Mathematical Statistics and Data Analysis. Second Edition.* Belmont: Duxbury Press, 1995.
32. Douglas JA. Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine* 1999; 18: 2917–31.
33. Gilks WR, Roberts GO and Sahu SK. Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association* 1998; 93: 1055–1067.

Appendix A: Likelihood Three-State Model

The following statements can be found in the literature referenced in Section 2. Presentation here is for convenience sake. The transition intensities $q_{rs}(t)$ are the entries of the transition intensity matrix $\boldsymbol{Q}(t)$, which for the three-state model in the paper is given by

$$\boldsymbol{Q}(t) = \begin{pmatrix} -q_{12}(t) - q_{13}(t) & q_{12}(t) & q_{13}(t) \\ 0 & -q_{23}(t) & q_{23}(t) \\ 0 & 0 & 0 \end{pmatrix}.$$

It is a general feature of intensity matrices that rows sum to zero. Transition probabilities for a time interval $(t, u]$ are given by the 3×3 matrix $\boldsymbol{P}(t, u) = \exp[(u - t)\boldsymbol{Q}(t)]$, with entries $p_{rs}(t, u) = P(x_u = s | x_t = r)$, for $r, s \in \{1, 2, 3\}$. Function $\exp[\cdot]$ is the matrix exponential. For the three-state model in the paper, $\boldsymbol{P}(t, u)$ is available in a closed-form. For $q_{rs} = q_{rs}(t)$ and $\Delta = u - t$, we have

$$\boldsymbol{P}(t,u) = \begin{pmatrix} e^{-(q_{12}+q_{13})\Delta} & p_{12}(t,u) & 1 - p_{11}(t,u) - p_{12}(t,u) \\ 0 & e^{-q_{23}\Delta} & 1 - p_{22}(t,u) \\ 0 & 0 & 1 \end{pmatrix}$$

where

$$p_{12}(t,u) = \frac{q_{12}(-1 + e^{(q_{12}+q_{13}-q_{23})\Delta})e^{-(q_{12}+q_{13})\Delta}}{q_{12} + q_{13} - q_{23}}.$$

Most of the more complex multi-state models require numerical approximations to derive $\boldsymbol{P}(t, u)$ from $\boldsymbol{Q}(t)$. This approximation is implemented in the R package msm [10].

Assume that an individual i has observations at times $t_{i1}, ..., t_{in_i}$, where the state at t_{ni} is either right-censored or death. Using the Markov assumption w.r.t. the states, the contribution of this individual to the likelihood is

$$\begin{aligned} p(\boldsymbol{x}_i | \boldsymbol{\beta}, \boldsymbol{w}) &= P(x_{in_i}, ..., x_{i2} | x_{i1}, \boldsymbol{\beta}, \boldsymbol{w}) P(x_{i1} | \boldsymbol{\beta}, \boldsymbol{w}) \\ &= P(x_{in_i} | x_{i,n_i-1}, \boldsymbol{\beta}, \boldsymbol{w}_{i,n_i-1}) P(x_{i,n_i-1} | x_{i,n_i-2}, \boldsymbol{\beta}, \boldsymbol{w}_{i,n_i-2}) \times ... \times P(x_{i2} | x_{i1}, \boldsymbol{\beta}, \boldsymbol{w}_{i1}). \end{aligned}$$

If the state observed at t_{in_i} is death, then, in shortened notation,

$$P(x_{in_i} | x_{i,n_i-1}) = P(x_{in_i} = 1 | x_{i,n_i-1}) q_{13}(t_{n_i}) + P(x_{in_i} = 2 | x_{i,n_i-1}) q_{23}(t_{n_i}).$$

So we assume an unknown state at time t_{in_i} and then an instant death. If the state is censored at t_{in_i}, then we assume that the individual is alive but with unknown state and we define $P(x_{in_i} | x_{i,n_i-1}) = P(x_{in_i} = 1 | x_{i,n_i-1}) + P(x_{in_i} = 2 | x_{i,n_i-1})$.

Appendix B: Gibbs Sampler

1. When missing, binary value y_{ijk} is sampled using a Bernoulli trial with success probability $\Phi(a_k \theta_{ijk} - b_k)$.

2. When missing, polytomous value u_{ij} is sampled using a multinomial distribution with probabilities given by (2).
3. Sample $\boldsymbol{z}$ from $p(\boldsymbol{z}|...) \propto p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{y})$. Value z_{ijk} is sampled from a truncated normal distribution with mean $a_k\theta_{ij} - b_k$ and variance 1, truncated from the left at zero if $y_{ijk} = 1$ and truncated from the right at zero if $y_{ijk} = 0$.
4. Metropolis sampling of $\boldsymbol{\theta}$.

   - A proposal distribution is specified by sampling from the conditional distribution of $\boldsymbol{\theta}$ with respect to the binary data (as represented by $\boldsymbol{z}$). It follows that

   $$p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}) \quad \propto \quad p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})p(\boldsymbol{\theta}|\boldsymbol{\Omega})$$

   Hence, for $\boldsymbol{X}_{ij}$ equal to 1×2 matrix $[1 \quad t_{ij}]$, we have

   $$p(\theta_{ij}|z_{ij}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}) \quad \propto \quad \exp\left[\frac{-1}{2}\sum_{k=1}^{K}(z_{ijk} + b_k - a_k\theta_{ij})^2\right] \exp\left[\frac{-1}{2\sigma^2}(\theta_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\eta}_i)^2\right].$$

   This is a normal regression model for $z_{ijk} + b_k$ on a_k, with coefficient θ_{ij}, variance known and equal to 1, and prior for θ_{ij} given by $N(\theta_{ij}|\boldsymbol{X}_{ij}\boldsymbol{\eta}_i, \sigma^2)$. It follows that $p(\theta_{ij}|z_{ij}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega})$ is a normal distribution with variance $V = (\sum_{k=1}^{K} a_k^2 + 1/\sigma^2)^{-1}$, and mean

   $$\frac{\sum_{k=1}^{K} a_k(z_{ijk} + b_k) + \boldsymbol{X}_{ij}\boldsymbol{\eta}_i/\sigma^2}{\sum_{k=1}^{K} a_k^2 + 1/\sigma^2}.$$

   - The vector $\boldsymbol{\theta}$ sampled from the proposal distribution is re-scaled such that the resulting values have mean 0 and variance 1.
   - Sampled and re-scaled $\boldsymbol{\theta}$ is the vector with the candidates for the Metropolis step which takes into account all data. The conditional distribution is given by

   $$\begin{aligned} &p(\theta_{ij}|\boldsymbol{y}_{ij}, u_{ij}, x_{ij}, x_{i,j+1}, \boldsymbol{a}, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{\beta}, \boldsymbol{\Omega}) \\ &\qquad \propto p(\boldsymbol{y}_{ij}|\theta_{ij}, \boldsymbol{a}, \boldsymbol{b})p(u_{ij}|\theta_{ij}, c, \boldsymbol{d})p(x_{i,j+1}|x_{ij}, \theta_{ij}, \boldsymbol{\beta})p(\theta_{ij}|\boldsymbol{\Omega}) \end{aligned}$$

   Because most of the information of θ_{ij} is contained in the binary data $\boldsymbol{y}$, the proposal is a good approximation of the posterior conditional distribution for θ_{ij}, and the acceptance rate is high.

5. Sample $\boldsymbol{a}$ from $p(\boldsymbol{a}|...) \propto p(\boldsymbol{z}|\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{b})p(\boldsymbol{a})$, where the prior is $p(\boldsymbol{a}) \propto 1$. Let the total number of records indexed over i and j be M. From $z_{ijk} = a_k\theta_{ijk} - b_k + e_{ijk}$ and $e_{ijk} \sim N(0, 1)$, it follows that $z_{ijk} + b_k = a_k\theta_{ijk} + e_{ijk}$. Treating a_k as a coefficient in an ordinary linear regression model, it follows that a_k can be sampled form a normal distribution with mean $\sum_{i,j} \theta_{ij}(z_{ijk} + b_k)/\sum_{i,j} \theta_{ij}^2$ and variance $1/\sum_{i,j} \theta_{ij}^2$.
6. Sample $\boldsymbol{b}$ from $p(\boldsymbol{b}|...) \propto p(\boldsymbol{z}|\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{b})p(\boldsymbol{b})$, where prior is $p(\boldsymbol{b}) \propto 1$. Let the total number of records indexed over i and j be M. From $z_{ijk} = a_k\theta_{ijk} - b_k + e_{ijk}$ and $e_{ijk} \sim N(0, 1)$, it follows that $b_k = a_k\theta_{ijk} - z_{ijk} + e_{ijk}$. Hence b_k can be sample from a normal distribution with mean $M^{-1}\sum_{i,j} a_k\theta_{ijk} - z_{ijk}$ and variance M^{-1}.
7. Sample c using a Metropolis step from $p(c|...) \propto p(\boldsymbol{u}|c, \boldsymbol{d}, \boldsymbol{\theta})p(c)$, where the prior is $p(c) \propto 1$ and the proposal is constructed using a normal distribution centered around the current value.
8. Sample $\boldsymbol{d}$ using a Metropolis-Hasting step from $p(\boldsymbol{d}|...) \propto p(\boldsymbol{u}|c, \boldsymbol{d}, \boldsymbol{\theta})p(\boldsymbol{d})$, where the prior is $p(\boldsymbol{d}) \propto 1$. The ordering in the parameter vector is maintained by generating an ordered candidate $\boldsymbol{d}^*$ conditional on current $\boldsymbol{d}$. This is established by sampling d_m^* sequentially from the truncated normal density

$$N(d_m, \tau^2)I(d_{m-1}^*, d_{m+1}) \qquad \text{for} \qquad m = 1, .., 4,$$

   where $d_0^* = -\infty$ and $d_5^* = \infty$. This density is not symmetric - hence the Hasting extension of the Metropolis algorithm.
9. Sample $\boldsymbol{\Omega} = (\boldsymbol{\nu}, \boldsymbol{\eta}, \boldsymbol{\Sigma}, \sigma)$ by following the scheme for a linear mixed model (where $\boldsymbol{\theta}$ is the response variable). These steps are Gibbs steps with conjugated priors. The parameters of the latter are ignored in the following notation.
   - Sample $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\Sigma}, \sigma)$.
   - Sample $\boldsymbol{\nu}$ from $p(\boldsymbol{\nu}|\boldsymbol{\eta}, \boldsymbol{\Sigma})$.
   - Sample $\boldsymbol{\Sigma}^{-1}$ from $p(\boldsymbol{\Sigma}^{-1}|\boldsymbol{\nu}, \boldsymbol{\eta})$.
   - Sample σ^2 from $p(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\eta})$.
10. Sample $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|...) \propto p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\theta})p(\boldsymbol{\beta})$ using three Metropolis steps, one for the intercepts, one for the slope for age, and one for the slope of $\boldsymbol{\theta}$. Candidates are sampled using multivariate normal distributions centered around the current values.

Steps 3, 5, and 6 are defined for the binary response IRT model and can be found for cross-sectional data in Johnson and Albert [25]. Fox [17] provides an overview of MCMC techniques for probit IRT models and logistic IRT models. The fact that we can formulate the steps with respect to longitudinal data is because of the conditioning on $\boldsymbol{\theta}$. The sampling scheme for the candidates in the first part of step 4 can be found in Fox and Glas [12] and Fox [17], but using this scheme to generate candidates for the Metropolis part has not been done before. Note that in the Metropolis, the sampling of $\boldsymbol{\theta}$ is informed by the multi-state data by including the transition probability $p(x_{i,j+1}|x_{ij}, \theta_{ij}, \boldsymbol{\beta})$. Step 8 can be found in Fox [17] for a cross-sectional model and is here used for a longitudinal model, and step 9 is an application of the scheme in Gelfand *et al.* [18]. In steps 7, 8, and 10, acceptance rates are monitored and adjusted during burn-in when necessary (*pilot adaption* [33]).