Joint models for discrete longitudinal outcomes in ageing research

Ardo van den Hout
Department of Statistical Science, University College London
1-19 Torrington Place, London WC1E 7HB, UK
Phone: +44(0)20 3108 3243, E-mail: ardo.vandenhout@ucl.ac.uk

Graciela Muniz-Terrera
MRC Lifelong Health and Ageing Unit at UCL, London, UK

Abstract

Given the ageing population in the UK, statistical modelling of cognitive function in the older population is of interest. Joint models are formulated for survival and cognitive function in the older population. Because tests of cognitive function often result in discrete outcomes, binomial and beta-binomial mixed-effects regression models are applied to analyse longitudinal measurements. Dropout due to death is accounted for by parametric survival models, where the choice of a Gompertz baseline hazard and the specification of the random-effects structure are of specific interest. The measurement model and the survival model are combined in a shared-parameter joint model. Estimation is by marginal likelihood. The methods are used to analyse data from the Cambridge City over-75s Cohort Study and the English Longitudinal Study of Ageing.

Key words: Beta-binomial distribution, Cognitive function, Gompertz distribution, Survival analysis.

1 Introduction

The structure of the UK population is changing, with a fast increase in the older segments. In 1985, the percentage of the population aged 65 and over was 15%. According to a projection by the Office for National Statistics, in the year 2035, 23% of the population will be comprised by individuals aged 65 years and over (Office for National Statistics 2011). Most western societies face similar changes in their population structure, although the pace of increase is faster in developing countries than in developed countries (United Nations 2011). These structural changes in society impose serious challenges to governments, and health and social care services.

In ageing research, decline in cognitive function is of specific interest as it is typical of the ageing process and often results in need for care. Statistical modelling of cognitive function and its potential decline over time makes it possible to describe cognitive function in older age, to understand it, and to predict it.

This paper proposes joint models for survival and cognitive function in the older population. Dementia is at the extreme of the declining process and, in clinical settings, part of its diagnosis involves the crossing of threshold values in cognitive tests such as the Mini-Mental State Examination (MMSE, Folstein *et al.* 1975). Typically, cognitive function is measured using a questionnaire test with an integer scale. The MMSE is an example of this with a scale from 0 up to 30. The proposed models cover a broad definition of cognitive impairment. Instead of only modelling whether impairment is present or not, it is worthwhile to model cognitive function on an extended scale.

Dropout due to death cannot be ignored when elderly are followed up with respect to a process that is associated with ageing and the proximity of death. Hence a joint model is needed for the process of interest and survival.

A good overview of past and current research in joint models is presented in Chapters 13-16 of *Longitudinal Data Analysis* (Fitzmaurice *et al.* 2009). Typical applications of joint models are statistical analyses of longitudinal measurements in the presence of dropout during follow-up. If the dropout is related to the longitudinal process of interest, then ignoring the dropout may bias statistical results (Henderson *et al.* 2000). The joint models that will be presented in this paper build upon an established framework; see Rizopoulos (2012). However, model formulation is geared up to the specific features of modelling cognitive function data in an older population. First, cognitive function is often measured with a test resulting in an integer score. Non-linear mixed-effects models are therefore investigated as alternatives to linear mixed-effects models where the

conditional distribution of the outcome is continuous. Tests of cognitive function often result in skewed distributions due to ceiling effects, and this further undermines the assumptions of linear mixed models (Proust *et al.* 2011). Second, non- or semi-parametric modelling of the time-to-event is the default in many joint models. However, when used in a joint model, semi-parametric survival models have hardly any of the computational advantages that recommend their use in isolation. In the present context, the event is defined as dropout due to death and since prediction is of specific interest, parametric models will be formulated where the choice of the parametric shape is part of the model comparison. The Gompertz baseline hazard is especially useful in the context of ageing; see, for example, Hougaard (2000, Section 2.2). For cognitive function, Van den Hout and Matthews (2008) showed that a multi-state survival model for the onset of cognitive impairment fitted better with Gompertz baseline hazards than with Weibull baseline hazards. Third, delayed entry (left truncation) is an essential feature of longitudinal data for ageing research when age is the time scale: individuals are only in the data set if they survived up to the required baseline age of the study.

The general framework in the current paper is a shared-parameter model, where the models for the longitudinal response and the hazard model are conditionally independent given the random-effects. Shared-parameters models with linear mixed-effects regression have been applied in, for example, longitudinal HIV studies (De Gruttola and Tu 1994; Faucett and Thomas 1996). The proposed models will extend the shared-parameter model by combining models for discrete longitudinal response with parametric hazard models. The time scale will be age. For the discrete response binomial regression and beta-binomial regression models will be formulated. Marginal likelihood will be applied, where the random effects are integrated out of the likelihood.

Section 2 introduces the joint models and Section 3 discusses maximum likelihood estimation, derivation of fitted values, and prediction. In Section 4, a small simulation study is described. The applications are in Sections 5 and 6, where data are analysed from the Cambridge City over-75s Cohort (CC75C) and the English Longitudinal Study of Ageing (ELSA), respectively. Section 7 is the conclusion.

2 Models

A joint model can be defined by specifying two constituent submodels which share random effects. The following will define, firstly, a measurement model for the

longitudinal response and, secondly, a hazard model for the event time. The data for the first model consist of repeated measurements on individuals in the study, the data for the second model consist of age at the baseline of the study and age at death when observed during the follow-up.

We assume that the longitudinal response is discrete and ordinal, taking integer values in $\{0, 1, ..., m\}$. For example, in the applications in Sections 5 and 6, the response is the observed sum score on a cognitive test with test-specific m being the highest score attainable. For individual i, $i = 1, ..., N$, with longitudinal response $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})$ at times $(t_{i1}, ..., t_{in_i})$ the *measurement model* has a linear predictor given by

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\gamma} \\ \beta_{0i} &= \beta_0 + b_{0i} \\ \beta_{1i} &= \beta_1 + b_{1i}, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_{ij}$ is a vector with covariate values at t_{ij} and no intercept, and the distribution of the random effects is assumed to be bivariate normal, i.e., $\boldsymbol{b}_i = (b_{0i}, b_{1i}) \sim N(\boldsymbol{0}, \Sigma)$.

The inverse of the logit link is $\mu_{ij} = \exp(\eta_{ij})/\{1 + \exp(\eta_{ij})\}$ and the corresponding binomial distribution for Y_{ij} has probability of success μ_{ij}, with m the number of trials. We have $\mathbb{E}[Y_{ij}|\mu_{ij}] = m\mu_{ij}$ and $\mathrm{Var}[Y_{ij}|\mu_{ij}] = m\mu_{ij}(1-\mu_{ij})$. The beta-binomial distribution for Y_{ij} also has probability of success μ_{ij}, and m trials, but has an additional variance parameter θ. It follows that $\mathbb{E}[Y_{ij}|\mu_{ij}] = m\mu_{ij}$ and $\mathrm{Var}[Y_{ij}|\mu_{ij}] = m\mu_{ij}(1 - \mu_{ij})\{1 + (m - 1)\theta/(1 + \theta)\}$. Here it is assumed that θ is the same unknown constant for all individuals.

In a fixed-effects model, the beta-binomial distribution can be used when there is overdispersion with respect to a binomial distribution. If there is an observation-specific random effect in a binomial regression model, then switching to a beta-binomial model does not make sense as the overdispersion is dealt with by the random effect. However, in a model for longitudinal data with individual-specific random effects which are linked to more than one observation, using the beta-binomial distribution can lead to improved data analysis.

The binomial regression model is a generalised linear mixed model (GLMM). Since the beta-binomial distribution is not in the natural exponential family, the beta-binomial regression is not a GLMM; see also Agresti (2002).

Next, we define the hazard model for death as the event of interest. In addition to the covariate values, data for this model are t_{i1}, the time individual i enters the study, and t_i, the time at which either death is observed (denoted by $\delta_i = 1$),

or death is right-censored ($\delta_i = 0$).

The *hazard model* is a parametric regression model given by

$$h(t|\boldsymbol{\beta}_i) = h_0(t)\exp\{g(\alpha, \boldsymbol{\beta}_i, t) + \boldsymbol{x}_i^*\boldsymbol{\gamma}^*\}, \tag{2}$$

where $\boldsymbol{\beta}_i = (\beta_0 + b_{0i}, \beta_1 + b_{1i})$ and $\boldsymbol{x}_i^*$ is a vector with covariate values and without an intercept. Function g is used to include the random effects into the hazard model with additional parameter α.

Examples of parametric specifications of the baseline hazard function $h_0(t)$ are

$$\begin{aligned} \text{Exponential:}\quad h_0(t) &= \lambda \\ \text{Weibull:}\quad h_0(t) &= \lambda\tau t^{\tau-1} \\ \text{Gompertz:}\quad h_0(t) &= \lambda\exp(\xi t), \end{aligned}$$

where $\lambda > 0$ and $\tau > 0$. For numerical reasons, we will use the parametrisation $\lambda = \exp(\gamma_0^*)$ for $\gamma_0^* \in \mathbb{R}$. There is no formal restriction $\xi > 0$. However, if $\xi < 0$, then for t very large, the survivor function goes to $\exp(\lambda\xi^{-1}) > 0$, which implies that the event does not occur for a proportion of the population. In the applications in Section 6 where t is age in years, all estimated ξ are positive.

The hazard model defined by (2) allows for several specialisations. In joint models where the hazard model is a semi-parametric Cox model, $g(\alpha, \boldsymbol{\beta}_i, t)$ is often specified as $\alpha(\beta_{0i} + \beta_{1i}t)$; see, for example, Diggle *et al.* (2009). For the above parametric specifications of the baseline hazard, we make a few remarks—ignoring the covariate effects for ease of presentation.

(a) Using $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha(\beta_{0i} + \beta_{1i}t)$ and the Gompertz baseline hazard with $\lambda = \exp(\gamma_0^*)$ implies

$$\begin{aligned} h(t|\boldsymbol{\beta}_i) &= \exp\{\gamma_0^* + \xi t + \alpha(\beta_{0i} + \beta_{1i}t)\} &(3)\\ &= \exp\{\gamma_0^* + \alpha\beta_{0i} + (\xi + \alpha\beta_{1i})t\}. &(4) \end{aligned}$$

Equation (3) shows that the model defines the risk of death as a function of t, and allows this risk to change according to an individual-specific trajectory for the longitudinal response. Equation (4) shows that the model is still a Gompertz model. A possible model extension would be to use two α-parameters leading to

$$h(t|\boldsymbol{\beta}_i) = \exp\{\gamma_0^* + \alpha_0\beta_{0i} + (\xi + \alpha_1\beta_{1i})t\}, \tag{5}$$

which separates the effect of the initial level of the trajectory at time $t = 0$ from the effect of the slope of the trajectory.

(b) The exponential baseline hazard model with $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha(\beta_{0i} + \beta_{1i}t)$ defines a Gompertz hazard model $h(t|\boldsymbol{\beta}_i) = \exp\{\gamma_0^* + \alpha\beta_{0i} + \alpha\beta_{1i}t\}$.

(c) The Weibull baseline hazard model with $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha\beta_{0i}$ defines a Weibull model conditional on an individual-specific scale parameter:

$$h(t|\boldsymbol{\beta}_i) = \lambda_i \tau t^{\tau-1} \quad \text{for} \quad \lambda_i = \exp(\gamma_0^* + \alpha\beta_{0i}). \tag{6}$$

If $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha(\beta_{0i} + \beta_{1i}t)$ is included in the log-linear part of (2), then methods for time-varying covariates can be applied; see, e.g, Crowther *et al.* (2013). An effect of β_{0i} on the hazard can also be included to define an individual-specific shape parameter:

$$h(t|\boldsymbol{\beta}_i) = \lambda \tau_i t^{\tau_i - 1} \quad \text{for} \quad \tau_i = \tau \exp(\alpha_0 \beta_{0i}). \tag{7}$$

Conditional on β_{0i}, this is a Weibull model albeit not in the form of (2). An advantage of this is that the conditional survivor function is in closed form.

The interpretation of the α-parameter in specifications of (2) is similar to the interpretation of regression parameters in the Cox model. For example in (3), given individuals i and j at time t, if there is a one-unit difference $(\beta_{0i} + \beta_{1i}t) + 1 = (\beta_{0j} + \beta_{1j}t)$, then $\exp(\alpha)$ is the relative increase of the risk associated with that one-unit increase. Thus, conditional on a time t, model (3) is a proportional hazard model for the heterogeneity across individuals as measured by the random effects. In contrast, model (7) is a non-proportional hazard model for the heterogeneity, where the interpretation of α is via the change in the Weibull shape parameter.

The choice of the time scale is important in joint models. The applications in this paper investigate cognitive change in the older population. Given the association between ageing and cognitive change, the chosen time scale t is age. In other applications, for example, in studies where individuals are followed-up after a medical intervention, the chosen time scale will often be time since baseline.

In ageing research where older people are included in longitudinal studies conditional on having reached a specified age, it makes sense to define t as age minus the minimum age in the study. In the CC75C application, for example, we define t as age minus 70 years and the interpretation of a joint model which shares the random intercept by defining $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha\beta_{i0}$ is that information on cognitive function at age 70 is associated with survival. Note that in this case a model with $g(\alpha, \boldsymbol{\beta}_i, t) = \alpha\beta_{i0}$ would not make sense if t is untransformed age.

An alternative specification of the hazard in a joint model is to share the random effect via the baseline hazard. Including covariate effects, this leads to

$$h(t|\boldsymbol{\beta}_i) = h_0\{t|g_0(\alpha_0, \boldsymbol{\beta}_i, t)\} \exp(\boldsymbol{x}_i^* \boldsymbol{\gamma}^*),$$

which includes Weibull model (7) as a special case, and can also be used to define the Gompertz models.

The above framework allows for several extensions with respect to specifying the effects of the random effects in the hazard model; see Rizopoulos (2012, Chapter 5) for examples and further references.

3 Maximum likelihood inference

3.1 Marginal likelihood

Let $\boldsymbol{\omega}$ denote the vector with all the model parameters except the random effects. The likelihood contribution of individual i conditional on truncation time t_{i1} is

$$L_i(\boldsymbol{\omega}|\boldsymbol{y}_i, t_i, T \geq t_{i1}) = p(\boldsymbol{y}_i, t_i|T \geq t_{i1}, \boldsymbol{\omega}) = \frac{p(\boldsymbol{y}_i, t_i|\boldsymbol{\omega})}{p(T \geq t_{i1}|\boldsymbol{\omega})}. \tag{8}$$

The denominator in (8) is the survivor function evaluated at t_{i1} and can be derived by integrating out the random effects, i.e.,

$$P(T \geq t_{i1}|\boldsymbol{\omega}) = \int P(T \geq t_{i1}|\boldsymbol{b}_i, \boldsymbol{\omega}) p(\boldsymbol{b}_i|\boldsymbol{\omega}) d\boldsymbol{b}_i. \tag{9}$$

Assuming independence between the submodels conditional on random effects, the numerator in (8) can be written as

$$p(\boldsymbol{y}_i, t_i|\boldsymbol{\omega}) = \int p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\omega}) p(t_i|\boldsymbol{b}_i, \boldsymbol{\omega}) p(\boldsymbol{b}_i|\boldsymbol{\omega}) d\boldsymbol{b}_i. \tag{10}$$

Assuming independence of observations for the measurement model given individual-specific random effects implies $p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\omega}) = \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\omega})$. For the hazard model, it follows that

$$p(t_i|\boldsymbol{b}_i, \boldsymbol{\omega}) = h(t_i|\boldsymbol{b}_i, \boldsymbol{\omega})^{\delta_i} P(T \geq t_i|\boldsymbol{b}_i, \boldsymbol{\omega}).$$

For a similar handling of left-truncation in the context of frailty models; see Jensen *et al.* (2004) and Rondeau *et al.* (2006).

The integrands in (9) and (10) consist of closed-form expressions for the models in Section 2. For example, for the Gompertz model (3) we have

$$P(T \geq t_i | \boldsymbol{b}_i, \boldsymbol{\omega}) = \exp\left[-\lambda_i \xi_i^{-1} \{\exp(\xi_i t_i) - 1\}\right],$$

where $\lambda_i = \exp(\gamma_0^* + \alpha\beta_{0i})$ and $\xi_i = \xi + \alpha\beta_{1i}$.

For the maximum likelihood estimation, the integrals in the log-likelihood are approximated using Gauss-Hermite quadrature, where the two-dimensional integrals are approximated by one-dimensional integrals each with 13 nodes for the quadrature. To facilitate this approach, the bivariate normal distribution in the log-likelihood is formulated by using two univariate normal distributions. If $Y \sim N(\mu_Y, \sigma_Y^2)$, and $X \sim N(\mu_X, \sigma_X^2)$, then the density for (y, x) is equal to $\phi_{Y|X}(y|X = x)\phi_X(x)$, where $Y|X \sim N(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1-\rho^2))$ and ρ is the correlation of X and Y (Casella and Berger 2002, p. 177). For this reason, we do not estimate Σ directly, but work with σ_0, σ_1, and ρ instead.

The logarithm of the likelihood is maximised in the software environment R using the general-purpose optimiser optim, which uses the Nelder-Mead routine for finding an optimum of a multi-dimensional function (R Development Core Team 2011). The Nelder-Mead routine is chosen because it does not require derivatives and it is robust in the sense that it is good at dealing with irregular functions and rapidly changing curvature (Nelder and Mead 1965).

An alternative routine in optim is BFGS, which uses numerical gradients to build up a picture of the surface to be optimised. BFGS is less robust, but can lead to faster convergence. For the Nelder-Mead routine in optim we used the default setting, except for the maximum number of iterations, which was increased in some cases to obtain a report of successful convergence. The default setting in optim for convergence is that the optimisation algorithm stops if it is unable to increase the value of the likelihood by 1×10^{-8}. Given the complexity of the model, it is recommended to explore several sets of starting values to minimise the risk of ending up with a solution which is a local maximum.

In the maximisation, all model parameters with a restricted parameter space are transformed such that the resulting maximisation is over an unbounded parameter space. For example, $\sigma_0 > 0$ is estimated by maximising over $\log(\sigma_0)$, and $-1 < \rho < 1$ is estimated by maximising over $\log\{(1 + \rho)/(1 - \rho)\}$. The Hessian and the delta method are used after maximisation to compute standard errors.

There are alternatives for using marginal likelihood to estimate random-effects models. Methods such as restricted maximum likelihood and h-likelihood are possible; see, for example, Lee *et al.* (2006) for a comparison of these methods.

3.2 Fitted values

For the estimation of individual random effects, maximum *a posteriori* (MAP) estimation is undertaken. As before, let $\boldsymbol{\omega}$ denote the vector with all the model parameters. The density of $\boldsymbol{b}_i$ evaluated at the MLE of $\boldsymbol{\omega}$ is given proportionally by

$$p(\boldsymbol{b}_i|t_i, t_{i1}, \delta_i, \boldsymbol{y}_i; \boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) \;\propto\; p(t_i|\delta_i, \boldsymbol{b}_i; \boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) p(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}) p(\boldsymbol{b}_i|\boldsymbol{\omega}=\widehat{\boldsymbol{\omega}}). \qquad (11)$$

The proportionality follows from ignoring the normalising constant, which is not needed to estimate the random effects. For each i, random effects $\boldsymbol{b}_i$ can be estimated by maximising posterior (11) using a general-purpose optimiser such as optim in R.

An alternative method to estimate the random effects is to construct a Markov Chain Monte Carlo (MCMC) algorithm to draw values from (11). A basic Metropolis algorithm with a normal distribution as the symmetric jump distribution can be used. From a chain of sampled values, means and confidence intervals can be constructed for $\boldsymbol{b}_i$. Since a chain is needed for each i, this is a computationally intensive method. For the models in this paper, MAP as described above is faster.

With estimated random effects, fitted values for the longitudinal outcomes can be computed. It is recommended to plot observed trajectories versus fitted trajectories for a subsample of individuals. This will help to assess structural misfit if present.

Because observed data are not a random sample of the target population due to the non-random dropout, residuals derived after fitting the model may not follow their nominal distribution (Rizopoulos 2012). Nevertheless, we still recommend plotting residuals to check for problems with model fit such as the presence of outliers. Randomised quantile residuals can be used as proposed by Dunn and Smyth (1996) in the context of generalised linear models. In the absence of non-random dropout, the nominal distribution of quantile residuals is the standard normal.

3.3 Prediction

Consider prediction of longitudinal response based upon available individual data. Denote the response for individual i by $\tilde{\boldsymbol{y}}_i$. Let $\tilde{t}_{in_i}$ denote the age corresponding to the last element of $\tilde{\boldsymbol{y}}_i$, and let $\tilde{t}_{i1}$ denote the age corresponding to the first element of $\tilde{\boldsymbol{y}}_i$. If there is just one outcome, then $\tilde{t}_{i1} = \tilde{t}_{in_i}$. In order to be able

to predict an individual trajectory from $\tilde{t}_{in_i}$ onwards, values of random effects are needed. Given that data are available, it makes sense to derive the most likely value of the random effects given the model parameter estimates and the available data. We will use MAP estimation and estimate $\boldsymbol{b}_i$ by those values which maximise the conditional density

$$p(\boldsymbol{b}_i|, \tilde{t}_{i1}, \tilde{t}_{in_i}, \delta_i = 0, \tilde{\boldsymbol{y}}_i; \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) \propto \\ p(\tilde{t}_{in_i}|\delta_i = 0, \boldsymbol{b}_i; \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) p(\tilde{\boldsymbol{y}}_i|\boldsymbol{b}_i; \boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}) p(\boldsymbol{b}_i|\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}). \quad (12)$$

Given estimated random effects, both survival and longitudinal response can be predicted up to an assumed maximum age. This will be illustrated in the applications. Using MAP estimation implies that the uncertainty in the prediction of survival and the longitudinal response does not included the uncertainty induced by the estimated distribution for the random effects.

Parameter uncertainty can be included in the prediction by simulation using the maximum likelihood estimation. Consider the multivariate normal distribution with expectation equal to the maximum likelihood estimate of the parameter vector and the covariance matrix equal to the estimated covariance matrix at the optimum. From this distribution parameter vectors are sampled and for each of these vectors prediction is undertaken. The resulting set of predictions will reflect the uncertainty in the estimation of the model parameters.

4 Simulation study

A small simulation study was conducted to investigate parameter estimation for the joint model. For the simulation, we choose model (3) defined by the beta-binomial regression and the Gompertz hazard. Of specific interest is the inference for the α-parameter and the number of nodes needed for the Gauss-Hermite quadrature. Because the estimation by marginal likelihood is computationally intensive, the investigation is limited with respect to the chosen sample sizes.

The joint model used in the simulation is given by

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t \\ h_i(t|\boldsymbol{\beta}_i) &= \exp\{\gamma_0^* + \xi t + \alpha(\beta_{0i} + \beta_{1i} t)\}, \end{aligned}$$

where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}) = (\beta_0 + b_{0i}, \beta_1 + b_{1i})$, and the random-effect distribution for (b_{01}, b_{1i}) is specified by σ_0, σ_1, and ρ; see Section 3.1. The variance parameter for the beta-binomial distribution is θ.

For the parameters we choose values close to the estimates in the CC75C application in the next section. Specifically, we choose $\alpha = -0.40$; see Table 1 for the chosen values of the other parameters. In line with CC75C, we assume that the model describes cognitive function of individuals of 75 years and older.

In the simulation study, baseline age is in the range $[75, 90]$. The left truncation is taken into account by simulating longitudinal data for a large population and then analysing a random sample of those individuals who survived up to the truncation threshold.

To fix ideas, assume that the baseline of the study is the calendar year 2000. Using the model, we first simulate yearly life-long trajectories of individuals who are 75 years old in the calendar years 1985 up to 2000. In this simulation, the survivors in 2000 are in the age range $[75, 90]$. From these survivors, we take a random sample of size N and impose the follow-up according to the study design. The follow-up is fixed to $3, 6, 9, 12$, and 15 years after baseline 2000. Including the baseline data, this means that an individual who does not die before 2015 has six observations.

Given specified parameters and an age range, data for an individual i are simulated by drawing random effects β_{i0} and β_{i1} first. Using these effects the longitudinal trajectory is simulated using the beta-binomial distribution. Next, using the random effects again, the Gompertz parameters $\lambda_i = \lambda + \alpha\beta_{i0}$ and $\xi_i = \xi + \alpha\beta_{i1}$ are defined and survivor function $S(t)$ is computed for each grid point t in a discrete grid for age. Using the cumulative distribution function $F(t) = 1 - S(t)$, age at death (discretised) is drawn by the using the inversion method.

Reducing the number of nodes for the quadrature reduces the time needed for the maximum likelihood estimation, but using not enough nodes can bias results. After some explorative runs, we choose 13 nodes for the Gauss-Hermite quadrature. Table 1 shows the simulation results for the choices $N = 100, 200$, and 400.

As expected, there is a consistent reduction of the root mean square error when N is increased. Distributions of estimated values are slightly skewed and the median is chosen as the measure of location to asses the bias. Using the median in a simulation study is not common. Its use in the current setting is motivated by the limited number of repetitions (only 100) and the aim to minimise the influence of outliers in the estimation of the parameters values. For $N = 100$, the percentage bias shows that there is substantial bias for α. Also the correlation parameter ρ is hard to estimate with $N = 100$. There is clear improvement for these parameters when N is increased.

Table 1: Simulation study for $N = 100, 200$, and 400. Results based on 100 simulated data sets and estimation using 13 nodes in the Gauss-Hermite quadrature. Bias is computed with the median as the measure of location. Notation: *%* for percentage bias, *rMSE* for the root of the mean square error, and $\lceil x \rceil$ for absolute value less than x.

		Estimation $N = 100$			$N = 200$			$N = 400$		
	Value	*Bias*	*%*	*rMSE*	*Bias*	*%*	*rMSE*	*Bias*	*%*	*rMSE*
β_0	3.50	-0.083	2.4	0.308	-0.103	2.9	0.253	-0.048	1.4	0.211
β_1	-0.20	-0.002	1.0	0.026	0.006	3.0	0.021	0.006	2.8	0.018
σ_0	1.50	-0.025	1.7	0.276	-0.062	4.1	0.231	-0.053	3.5	0.177
σ_1	0.10	-0.005	5.0	0.029	-0.008	8.0	0.022	-0.005	4.5	0.018
ρ	-0.50	0.094	18.8	0.341	0.060	12.0	0.249	0.036	7.2	0.206
θ	0.02	-0.002	10.0	0.009	0.001	5.0	0.008	$\lceil 0.001 \rceil$	$\lceil 0.1 \rceil$	0.006
γ^*	-2.60	-0.093	3.7	0.460	-0.107	4.3	0.405	-0.125	5.0	0.362
ξ	0.10	0.003	3.5	0.028	0.005	5.5	0.026	0.007	7.5	0.022
α	-0.40	-0.057	14.2	0.157	-0.024	6.1	0.110	0.002	0.4	0.083

Comparing the percentage bias for $N = 100$ and $N = 400$, there is an unexpected increase for some parameters, including ξ which is the effect of age. However, with only 100 simulated data sets, some variation in the percentage bias is to be expected. The overall results for $N = 400$ are clearly an improvement compared to $N = 100$.

The simulation study shows that the estimation by marginal likelihood is able to reproduce the parameters that were used to generate the data. Choosing 13 nodes for the Gauss-Hermite quadrature seems to work well in this setting, but should not be used as a general guideline; see, for example, Lesaffre and Spiessens (2001). With only 100 simulated data sets, it is difficult to fully assess coverage rates of confidence intervals, or to compare empirical standard deviations (of estimated model parameters) with the average of the standard errors (as estimated from the model). However, looking tentatively at these statistics in the current simulation study (not reported), there seems to be a consistent underestimation of the standard errors. This aspect of the estimation needs more research. With respect to the interpretation of the estimation in the application in the following sections, reported 95% confidence intervals should be interpreted with care as there may be some underestimation of the uncertainty.

5 Cognitive function as a predictor in a survival model for CC75C

The Cambridge City over-75s Cohort Study (CC75C, `www.cc75c.group.cam.ac.uk`) is a UK population-based longitudinal study of ageing that started in 1985 with participants aged at least 75 years old in Cambridge city. Topics in the study are dementia, patterns of cognitive change, depression and depressive symptoms, socio-demographics and social contacts, falls and functional ability, and genetics. Here we focus on the measuring of cognitive function using the Mini Mental State Examination (MMSE). Because of the advanced age of CC75C participants at the baseline in 1985, there were fewer than ten survivors in 2010. The data of these survivors are not included in the current analysis. For all the other individuals, death times are available.

We analyse data from $N = 1932$ CC75C participants. A subset of this sample was analysed in Van den Hout *et al.* (2011) using years to death as the time scale in a mixed-effects model. In the present context we use age as the time scale in the joint model, which makes it possible to predict the process of interest.

With respect to the survey design, further interviews after baseline were conducted on average 2, 7, 9, 12, 17, and 21 years later, and the frequencies of the number of observations at baseline and the subsequent waves are 777, 485, 318, 196, 113, 37, and 6, respectively. Mean age at baseline is 81.4, median is 80.5. There are 698 men and 1234 women in the sample. The frequencies for dichotomised number of years of formal education are 1518 and 414 for fewer than ten years versus more than ten years.

The time scale t for the longitudinal MMSE outcome is age minus 70 years. Hence the interpretation of intercept β_0 is with respect to cognitive function at 70 years old. The time-to-event in the survival model is age of death minus 70 years, left truncated at age at baseline minus 70 years.

We are interested in how cognitive function changes over time, and how such a change affects survival. Two binary covariates are used in the model for the hazard: `sex` (0/1 for women/men) and `educ` (0/1 for fewer than ten years of education/ten or more years). Model A is specified using a binomial regression model and a Gompertz baseline hazard, where the sharing of the random effects is defined in (2) and (3). The regression equations are

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} \\ h(t|\boldsymbol{\beta}_i) &= h_0(t) \exp\{\alpha(\beta_{0i} + \beta_{1i} t) + \gamma_1^* \texttt{sex}_i + \gamma_2^* \texttt{educ}_i\} \end{aligned}$$

$$= \exp(\gamma_0^* + \alpha\beta_{0i} + (\xi + \alpha\beta_{1i})t + \gamma_1^*\texttt{sex}_i + \gamma_2^*\texttt{educ}_i\}. \qquad (13)$$

For Model B, the binomial regression is replaced by a beta-binomial regression. Model B is the better choice according to the Akaike information criterion: AIC = 33467 versus AIC = 33289, respectively.

Restricted versions of Model B were also investigated. In Model B_{R1}, the baseline hazard is defined by $h_0(t) = \exp\{\gamma_0^* + (\xi + \alpha\beta_{1i})t\}$, and in Model B_{R2}, the baseline hazard is $h_0(t) = \exp(\gamma_0^* + \alpha\beta_{0i} + \xi t)$. The log-linear model for the covariates is as in (13). The models B, B_{R1} and B_{R2} have the same number of parameters. AICs for B_{R1} and B_{R2} are 33462 and 33366, respectively. It is interesting to see that Model B_{R2} which shares the random intercept performs better than Model B_{R1} which shares the random slope. For CC75C, information on cognitive function at $t = 0$, i.e., at age 70, has a stronger association with survival than information on linear change after $t = 0$.

Models B_{R1} and B_{R2} have counterparts with Weibull specifications of the baseline hazard, namely $h_0(t) = \tau_i t^{\tau_i - 1}\exp(\gamma_0^*)$, for $\tau_i = \tau\exp(\alpha\beta_{1i})$, and $h_i(t) = \tau t^{\tau-1}\exp(\gamma_0^* + \alpha\beta_{0i})$, respectively; see remark (c) in Section 2. Pairwise, both these models have higher AICs (33641 and 33451, respectively) than Models B_{R1} and B_{R2}. This indicates better performance of the Gompertz specification compared to the Weibull.

Next Model C is defined by changing the hazard specification in Model B to

$$h(t|\boldsymbol{\beta}_i) \;=\; \exp(\gamma_0^* + \alpha_0\beta_{0i} + (\xi + \alpha_1\beta_{1i})t + \gamma_1^*\texttt{sex}_i + \gamma_2^*\texttt{educ}_i\}. \qquad (14)$$

Model C has AIC = 33273, and is with that the best model. According to Model C, the random intercept for cognitive level at $t = 0$ and the random slope for cognitive change are both important predictors for survival. The better performance of Model C compared to Model B shows that the effects of the random intercept and slope on survival are best described by estimating two distinct α-parameters.

For Model C, the fit to individual data is depicted in Figure 1 for a random subset of 16 individuals. Prediction is up to time of death, which is depicted by the vertical grey line. Each graph in Figure 1 shows the fitted mean trajectory, and ten sampled trajectories from the fitted beta-binomial distribution. Overall the model seems to capture the observed trajectories well. In general, if there is a large drop in observed MMSE scores, followed by a recovery, then there is some misfit. This is a direct consequence of our modelling which implies a monotonic mean trajectory for MMSE scores.

Randomised quantile residuals are presented in Figure 2. For the fitted values in the lower range of the MMSE there is a trend of more negative residuals. In general, the observation of low MMSE scores tends to be less reliable and hence harder to predict. A closer look at the residuals also shows that the model is not good at capturing sudden drops in MMSE scores. There are six residuals smaller than -6. These residuals are from fitting trajectories to data from five different individuals. Two of these individuals have very low MMSE scores, and for the other three low scores are observed together with high scores. An example is a woman with a response trajectory given by (27, 0, 20), which—as a trajectory—is an outlier. The zero in this trajectory may well be a data error. Overall, however, the residuals do not show signs of structural misfit.

Parameter estimates for Model C are presented in Table 2. The negative value for estimated α_0 implies that having better cognitive function at age 70 as measured by the MMSE is associated with a smaller hazard and thus with better survival. For estimated α_1, the negative value implies that a less negative slope for the change of MMSE over the years is associated with better survival. The estimates of γ_1^* and γ_2^* show that being a woman and having had more education are associated with better survival. Because of the α-parameters, these effects for gender and education are adjusted for cognitive trajectory.

Even though the point estimates of the α-parameters are close in value, they do not reflect similar strength in effect. Using the general Gompertz formulation in Section 2, parameter α_0 affects the λ-parameter, whereas α_1 affect the ξ-parameter. Consider individuals with the same hazard specification at age 70. If the first does not experience a change in cognition, i.e., $\beta_{i1} = 0$, and the other follows the fitted mean trend on the logit scale, i.e, $\beta_{i1} = \widehat{\beta}_1 = -0.135$, then the estimated relative increase in risk associated with the change is $\exp(-0.306 \times -0.135t) = \exp(0.041t)$ at $t > 0$. For example, at age 75, there is a relative increase of 1.23.

Prediction is depicted in Figure 3 for a woman aged 85 with fewer than ten years of education. Prediction in the present context is always with respect to both survival and the process of interest. Prediction in Figure 3 is conditional on current MMSE scores, in this case 28 and 20, and survival up to the time of the prediction. Note that the latter score is associated with poorer survival, and a bigger drop in mean MMSE scores over time than the former. The 95% confidence bands in Figure 3 are derived from the maximum likelihood estimation by simulation as explained in Section 3.3. The number of replications in the simulation is 1000, and 2.5% and 97.5% quantiles were derived to create the 95% confidence bands.

Table 2: CC75C data analysis with Gompertz baseline hazard in Model C. Point estimates (and 95%-confidence intervals) derived from the maximum likelihood estimation.

Measurement model			*Survival model*		
β_0	3.229	(3.089; 3.368)	γ_0^*	-2.596	(-2.821; -2.372)
β_1	-0.135	(-0.145; -0.126)	γ_1^*	0.459	(0.360; 0.559)
σ_1	1.300	(1.184; 1.428)	γ_2^*	0.242	(0.123; 0.362)
σ_2	0.083	(0.075; 0.092)	ξ	0.083	(0.073; 0.094)
ρ	-0.749	(-0.792; -0.699)	α_0	-0.462	(-0.531; -0.393)
θ	0.019	(0.016; 0.023)	α_1	-0.306	(-0.356; -0.256)

Figure 3 represents the prediction of a bivariate process. The prediction of the MMSE trajectory on the right-hand side cannot be interpreted on its own since that would imply immortality after the time of the prediction, which in this case is 85 years old. For this reason, a grey scale was added to the graph to depict predicted survival. The fading out of the gray illustrates that—given baseline age 85—predicting cognitive function up to a very old age is only relevant for a small part of the population.

Prediction can also be based on more than one observation. To illustrate the effect of cognitive decline on survival, consider again a woman with fewer than ten years of education. If observed MMSE scores at ages 85 and 89 are both 28, then predicted survival probability at age 95 is about 0.5. But if observed scores at the same ages are 28 and 20, then the predicted probability at age 95 is about 0.2. In the latter case, the cognitive decline has a strong effect on predicted survival.

The right-hand side panel of Figure 3 shows marginal trajectories conditional upon survival up to age 85. On an individual level, it is of interest to predict cognitive function conditional on survival up to a specified age in the future. For example, for someone who takes the test at age 85, it is of interest to predict MMSE performance at age 95 conditional on reaching that age. This prediction can be undertaken in the same way as above using MAP estimation. For the example, times $\tilde{t}_{i1}$ and $\tilde{t}_i$ in (12) correspond to ages 85 and 95, respectively, and $\tilde{\boldsymbol{y}}_i$ contains one entry only, namely the MMSE score at age 85. The predicted MMSE conditional on survival is higher than the marginal prediction. This is because the marginal prediction takes into account the association between the

increasing hazard of death and decreasing MMSE scores. For predicted MMSE at age 95 given MMSE scores 28 and 20 at age 85, the differences are very small (less than one MMSE unit). Differences increase with longer predictions. For example, given an MMSE score 28 at age 75, marginally predicted score at 95 is 14.6, but predicted score conditional upon survival at 95 is 17.9.

6 Cognitive function in ELSA

To illustrate data analysis where the measurement model is of primary interest we discuss briefly longitudinal data from the English Longitudinal Study of Ageing (ELSA, `www.ifs.org.uk/ELSA`). The ELSA baseline (1998-2001) is a representative sample of the English population aged 50 and older. ELSA contains information on health, economic position, and quality of life. Longitudinal data on cognitive function are available in the waves 1 - 5 (2002-2011). Data from ELSA can be obtained via the Economic and Social Data Service (`www.esds.ac.uk`). We use data from individuals who are interviewed in wave 1, and thus ignore the refreshments samples in wave 3 and 4.

There are a number of questions in ELSA that concern cognitive function. Here we focus on the number of words remembered in a delayed recall from a list of ten: "A little while ago, you were read a list of words and you repeated the ones you could remember. Please tell me any of the words that you can remember now." The test score is equal to the number of words remembered $\in \{0, 1, \ldots, 10\}$. We are interested in the effect of sex on cognitive change over time when controlling for education.

There are 11828 individuals who are interviewed in wave 1. For the analysis in this section, individuals who were interviewed only once with missing data on the number of words recalled are not included. Likewise the individuals without information on the year of birth are not included. In addition to the so-called *core sample members* in ELSA, cohabiting spouses or partners of core sample members are also included in ELSA. This inclusion is irrespective of the age of the spouse or partner and because of this there are individuals who were younger than 50 at baseline wave 1. Data from these younger individuals are ignored in the analysis to ensure a representative sample of the population aged 50 and older. Lastly, individuals with censored age at baseline are not included. The resulting sample size is $N = 10852$.

Of the 10852 individuals, 1884 die during follow-up. In the joint model, this dropout is modelled by using age of death as the time-to-event, left truncated at

age at baseline. A dropout rate around 17% is too much to ignore in the data analysis, especially in this case where the process of interest is associated with ageing.

At baseline, there are 5946 women and 4906 men. Highest educational qualification was dichotomised with value 1 for NVQ2/GCE O Level equivalent or higher, and 0 otherwise. At baseline there are 4699 individuals with the higher education level.

Our joint modelling starts with defining Model I, which includes binomial regression and a Gompertz hazard. This model is given by

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i}t_{ij} + \gamma_1 \texttt{sex}_i + \gamma_2 \texttt{educ}_i + \gamma_3(\texttt{sex}_i \times t_{ij}) + \gamma_4 \texttt{yob}_i \\ h(t|\boldsymbol{\beta}_i) &= \exp\{\gamma_0^* + \alpha\beta_{0i} + (\xi + \alpha\beta_{1i})t + \gamma_1^* \texttt{sex}_i + \gamma_2^* \texttt{bmi}_i\}, \end{aligned}$$

where `sex` = 1 for men, and `educ` = 1 for the higher education level. Covariate `bmi` is body mass index (BMI) grouped into < 20, $20-25$, $25-30$, $30-35$, $35-40$, 40+, with coded values -2, -1, 0, 1, 2, 3, respectively. Missing BMI is imputed by the value 0 for the $25-30$ group, which corresponds to median BMI group at baseline. Higher values of `bmi` are probably associated with poorer survival and `bmi` is therefore included in the hazard model. Covariate `yob` is year of birth minus 1900, which is added to take into account a potential cohort effect. Possible change of the gender effect over the age range is taken into account by including the interaction in the binomial regression. Covariate `sex` is included twice as we expect it to be explanatory for both cognitive function and survival. The time scale t is age minus 49 years given that the minimum of observed age at baseline is 50 years.

Model I has AIC = 162643.4. The model is extended by switching to beta-binomial regression, which defines Model II with AIC = 162643.2. Model II has one extra parameter but does not lead to a substantial AIC improvement. Variance parameter θ in Model II is estimated close to the boundary of the parameter space at 9.998×10^{-05}. Because of this boundary solution for θ and the minor difference in AICs, we select the more parsimonious Model I as the better model.

The sample size combined with the integral approximation is computationally intensive. For this reason, various models were investigated using a random subset of a 1000 individuals. The Weibull hazard was compared with the Gompertz hazard for a shared random-intercept model similar to the comparison in Section 5. According to the AICs, the Gompertz hazard was the better choice for the subset. By using a subset, this comparison of AICs is not based upon

Table 3: ELSA data analysis. Estimated parameters for Model I with the Gompertz baseline hazard. Point estimates (and 95%-confidence intervals) derived from the maximum likelihood estimation.

Measurement model			*Survival model*		
β_0	-1.618	(-1.771; -1.465)	γ_0^*	-6.630	(-6.829; -6.431)
β_1	-0.009	(-0.012; -0.006)	γ_1^*	0.486	(0.395; 0.577)
γ_1	-0.245	(-0.296; -0.193)	γ_2^*	0.022	(-0.031; 0.075)
γ_2	0.468	(0.440; 0.496)	ξ	0.113	(0.107; 0.119)
γ_3	0.001	(-0.002; 0.004)			
γ_4	0.033	(0.030; 0.036)	α	-0.112	(-0.176; -0.048)
σ_1	0.491	(0.446; 0.541)			
σ_2	0.025	(0.022; 0.029)			
ρ	-0.235	(-0.417; -0.035)			

all information available. The assumption underlying the comparison is that the size of the subset is adequate to decide upon the best parametric shape.

For Model I, fit to individual data is depicted in Figure 4 for a random subset of 16 individuals. Prediction is up to time of the last observation. Note that the scale on the horizontal axis varies. Observed trajectories are fitted well. The variability in the observed sum scores within individuals is captured by the variance of the fitted distribution.

Table 3 presents the inference for the parameters for Model I. According to the fitted model, the mean trend of cognitive function is downward ($\widehat{\beta}_1 = -0.009$). Both the fitted trajectories and the estimate $\widehat{\beta}_1$ show a difference with the inference for the CC75C data, where the downward trend was more pronounced. There may be various reasons for this. Given the difference in minimal age at baseline, it is to be expected that there is more cognitive decline among the individuals in CC75C. Furthermore, the two cognitive tests in the longitudinal studies are not the same. The MMSE used in CC75C is specifically developed as a multifaceted test for cognitive impairment, whereas the recall test is testing just one specific skill. Already from the limited data presented in Figures 1 and 4 it is clear that the recall test in ELSA is more variable than the MMSE in CC75C. Because of this, and the limited years of follow-up in ELSA, it is harder to detect change of cognitive function over time at an individual level.

Given that the estimated interaction ($\widehat{\gamma}_3$) is close to zero, we interpret the covariate main effects in the measurement model. There is a clear positive effect of more education ($\widehat{\gamma}_2 > 0$). Given the effect of education, there is an additional gender effect: women (`sex = 0`) tend to be better at remembering words than men ($\widehat{\gamma}_1 < 0$). There is also an effect of year of birth, i.e., being born later is associated with being better in remembering words ($\widehat{\gamma}_4 > 0$). In general, birth cohort effects on cognitive function remain an active research topic. For example, Jagger *et al.* (2007) did not found any evidence of an effect in the UK Cognitive Function and Ageing Study, but Christensen *et al.* (2013) did find an effect for two Danish cohorts, where being born later is associated with better MMSE performance.

In the survival model, the negative value for estimated α implies that better performance in the word recall is associated with a lower hazard and $\widehat{\xi} > 0$ implies that the hazard for death increases with age. Both these results are according to expectation. There is also an expected gender effect: men have a higher hazard compared to women of the same age ($\widehat{\gamma}_1^* > 0$). The estimated positive BMI effect ($\widehat{\gamma}_2^* > 0$) is unexpected at first sight as it would imply that for the current population, and controlling for gender, and for cognition via α, a higher BMI is associated with better survival. However, it might be that this protective effect of higher BMI is specific for an elderly population in the sense that it reflects changing from underweight towards a more healthy weight. Either way, the corresponding 95%-confidence interval implies that the estimated effect does not differ from zero significantly.

7 Conclusion

Joint models are presented for survival and discrete longitudinal outcomes in ageing research. The applications concern cognitive function as measured with discrete-valued tests. The statistical methods that are used build upon an established framework for shared-parameter models, but extend this framework to ageing research by specifying binomial and beta-binomial mixed-effects regression models and by exploring the Gompertz baseline hazard as a parametric choice for the survival model. In the model comparison for the CC75C data, the beta-binomial model is more able to capture overdispersion than the binomial model.

As stated in Section 2, the choice of the time scale is important in joint models for ageing research. Given the process of interest, it makes sense to use age as the basic time scale. Combining this choice with parametric models means that

estimated models can be used for prediction; see for instance Figure 3. In general, prediction beyond the age range in the data is not without danger. CC75C include data from centenarians, but age in the ELSA data as used in the current paper only goes up to ninety. Hence, prediction based upon the model for ELSA should be interpreted with care.

Closely related to the discussion in the current paper is the research based on the terminal decline hypothesis. This hypothesis states that individuals in the older population experience a change in the rate of decline of cognitive function before death (Riegel and Riegel, 1972). Investigating this behaviour with age as the time scale is problematic as it requires death as a reference point; see Van den Hout et al. (2013) for change-point modelling using years-to-death as the time scale.

The focus of the presentation in this paper is on statistical modelling. Regression models are formulated such that various random-effects structures can be specified. Model comparison is undertaken by applying the Akaike information criterion. The resulting models can be used to detect risk factors for the process of interest. Although the focus is on the modelling, we envisage that the framework can form a basis for applications in health economics, where long-term predictions are linked to cost functions. To enable prediction, parametric models are specified, but the same framework can be used to investigate semi-parametric models. As an example, the intercept-slope measurement model can be changed to a model with a smooth spline for the effect of age. Given a fixed number of knots for the spline, the general-purpose optimiser can be applied for maximum likelihood estimation.

Our statistical methods provide an alternative to the model presented by Proust-Lima *et al.* (2009), who discuss a joint model for cognitive function and survival using latent classes. Proust-Lima *et al.* propose to transform the scores on cognitive tests in order to deal with the skewness of the scores on the original scale and to be able to use the linear mixed model framework for the latent process underlying the manifest scores. This transformation hampers interpretation of the regression parameters. Our approach takes the scores at face value, using the original scale of the tests. Nevertheless, also in our model there is a transformation; interpretation of the parameters in the measurement model has to take into account the logit link. Another difference with Proust-Lima *et al.*, is that they do not consider the use of survival models with a Gompertz baseline hazard. Their parametric choice is the Weibull.

In this paper, the sum score is the longitudinal outcome variable. The binomial distribution assumption is not invalidated if there is variation in the success

probabilities for the individual questions (McCullagh and Nelder, 1989, p. 103). However, if there is dependency between the questions, then this violates the assumption of independent Bernoulli trials. A solution would be to replace the observed sum score by a latent variable which is linked to question-specific scores via an item response theory model; see Fox (2010) who discusses a linear mixed-effects model for longitudinal questionnaire data. A disadvantage of working with the latent variable is increased model complexity and the lack of a straightforward interpretation of the parameters for the regression model.

The joint models in this paper are estimated by maximising the marginal likelihood, which implies integrating out the random effects. For linear measurement models with time-independent random effects, Tsiatis and Davidian (2004) show that the assumption underlying marginal likelihood is that the censoring and timing of longitudinal measurement are uninformative. Their argument is also valid for a model with a non-linear link between mean and linear predictor.

Estimating model parameters using marginal likelihood may not always be the best approach. As an example, if the population consists of two classes, where—within the same age range—one class is characterised by downwards trajectories of longitudinal measurements and poor survival, and the other class by stable trajectories and good survival, then the marginally fitted trend is not a good representation of the trajectories in the data. However, we think that using marginal likelihood works well for the applications in this paper. Given that the observed individual trajectories show substantial variation across a wide age range in the older population, the marginally estimated association between trends of cognitive function and survival seems a good representation of the information in the data. To use a fitted model for prediction conditional on individual data, maximum *a posteriori* estimation can be used as shown at the end of Section 5.

We have specified the measurement model for the longitudinal outcome using time-independent random intercept and random slope. In the words of Tsiatis and Davidian (2004, p. 815), this specification implies that the "smooth trend followed by the subject's trajectory is an 'inherent' characteristic of the subject that is fixed throughout time". This allows us to investigate the association between the dominant trend and survival. For the applications at hand, this seems reasonable in the context of cognitive function as it allows us to ignore within-subject fluctuations (such as good day/bad day variation). The association is measured by the α-parameters which make it possible to investigate how the random effects for the longitudinal trend are linked to the hazard of death.

The formulation of the joint model in this paper is general and we think that it is applicable to a range of discrete-valued tests in ageing research. We advocate

binomial and beta-binomial regression models for discrete-valued longitudinal outcomes. Using mixed-effects linear models where the conditional outcome is assumed to be normally distributed is problematic when the observed response is limited to a finite set of integers, especially when the observed responses show a ceiling effect.

Acknowledgements

The authors would like to thank the CC75C group for providing the data; see `www.cc75c.group.cam.ac.uk` for a list of contributors and funding organisations. The ELSA data were made available through the UK Economic and Social Data Service. ELSA was developed by a team of researchers based at the National Centre for Social Research, University College London and the Institute for Fiscal Studies. The developers and funders of ELSA do not bear any responsibility for the analyses or interpretations presented here. In addition, the authors would like to thank the editors and two anonymous referees for substantial feedback that has helped to improve the manuscript.

References

Agresti, A. (2002) *Categorical Data Analysis. 2nd edition.* New York: Wiley.

Casella, G. and Berger, R.L. (2002) *Statistical Inference, 2nd edition.* Australia: Duxbury.

Christensen, K., Thinggaard, M., Oksuzyan, A., Steenstrup, T., Andersen-Ranberg, K., Jeune, B., McGue, M. and Vaupel J.W. (2013) Physical and cognitive functioning of people older than 90 years: a comparison of two Danish cohorts born 10 years apart. *The Lancet.* Published online. http://dx.doi.org/10.1016/ S0140-6736(13)60777.

Crowther, M.J., Abrams, K.R. and Lambert, P.C. (2013) Joint modelling of longitudinal and survival data. *Stata J.* **13**, 165–184.

De Gruttola, V. and Tu , X.M. (1994). Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. *Biometrics*, **50**, 1003–1014.

Diggle, P., Henderson, R. and Philipson, P. (2009) Random-effects models for joint analysis of repeated-measurement and time-to-event outcomes, in *Longitudinal Data Analysis*, Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds), pp. 349–366.

Dunn, P.K. and Smyth, G.K. (1996) Randomised quantile residuals. J. Comput. Graph. Statist., **5**, 236–244.

Faucett, C.L. and Thomas, D.C. (1996) Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statist. Med.*, **15**, 1663–1685.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Editors) (2009) *Longitudinal Data Analysis*. Boca Raton: Chapman & Hall/CRC.

Folstein, M.F., Folstein, S.E. and McHugh, P.R. (1975) Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J. of Psych. Research*, **12**, 189–198.

Fox, J.-P. (2010) *Bayesian Item Response Modeling*. New York: Springer.

Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **4**, 465–480.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. New York: Springer.

Jagger C., Matthews, R.J., Matthews, F.M., Spiers N.A., Judith Nickson, J., Eugene S Paykel, E.S., Huppert, F.A., Brayne C. and MRC-CFAS (2007) Cohort differences in disease and disability in the young-old: findings from the MRC Cognitive Function and Ageing Study (MRC-CFAS), *BMC Public Health*, **7**:156.

Jensen, H., Brookmeyer, R., Aaby, P. and Andersen, P.K. (2004) Shared frailty model for left-truncated multivariate survival data, Research Report 04/2, Department of Biostatistics, University of Copenhagen.

Lee, H., Nelder, J.A. and Pawitan Y. (2006) *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Boca Raton: Chapman & Hall/CRC.

Lesaffre, E. and Spiessens, B. (2001) On the effect of the number of quadrature points in a logistic random effects model: an example. *J. R. Statist. Soc. C*, **50**, 325-335.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Boca Raton: Chapman & Hall/CRC.

Nelder, J. A. and Mead, R. (1965) A simplex algorithm for function minimization. *Computer J.*, **7**, 308-313.

Office for National Statistics (2011) *Statistical Bulletin*, Office for National Statistics, London 29 September 2011. (Available from `http://www.ons.gov.uk/ons/publications/`)

Pinheiro, J.C. and Bates D. (2000) *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Proust-Lima, C., Joly, P., Dartigues, C.-F. and Jacqmin-Gadda, H. (2009) Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Comput. Statist. Data Anal.*, **53**, 1142–1154.

Proust-Lima, C., Dartigues, C.-F. and Jacqmin-Gadda, H. (2011) Misuse of the Linear Mixed Model When Evaluating Risk Factors of Cognitive Decline. *Am. J. Epidem.*, **174**, 1077–1088.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, `www.R-project.org`.

Riegel, K.F. and Riegel, R.M. (1972). Development, drop, and death. *Developm. Psych.*, **6**, 306–319.

Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data. With applications in R.* Boca Raton: CRC Press.

Rondeau, V., Filleul, L. and Joly P. (2006) Nested frailty models using maximum penalized likelihood estimation. *Statist. Med.*, **25**, 4036–4052.

Tsiatis, A.A. and Davidian, M. (2004) An overview of joint modeling of longitudinal and time-to-event data. *Statist. Sinica*, **14**, 793–818.

United Nations (2001). World Population Ageing: 1950-2050, Report ST/ESA/SER.A/207, New York: United Nations Publications.

Van den Hout, A. and Matthews, F.E. (2008). Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model, *Statist. Med.*, **27**, 5440–5455.

Van den Hout, A., Muniz-Terrera, G. and Matthews, F.E. (2011). Smooth random change point models. *Statist. Med.* **30**, 599–610.

Van den Hout, A., Muniz-Terrera, G. and Matthews, F.E. (2013). Change point models for cognitive tests using semi-parametric likelihood. *Comput. Statist. Data Anal.*, **57**, 684–698.

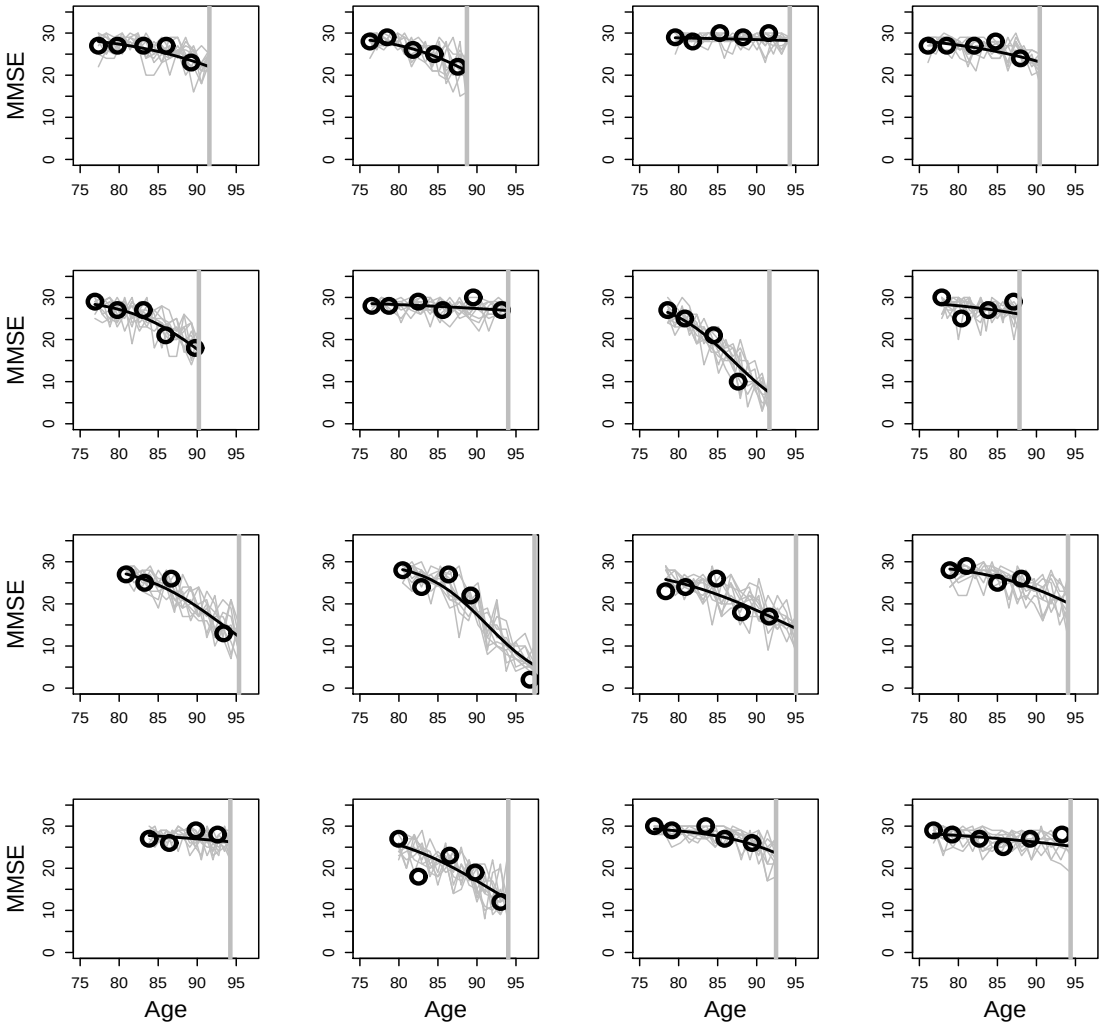

Figure 1: For CC75C data and Model C, observed MMSE (black circles) and fitted trajectories for 16 individuals randomly chosen from those with 4 or more observations. Black line for fitted mean trajectory, and grey lines for sampled trajectories from the fitted distribution. Vertical line for time of death.

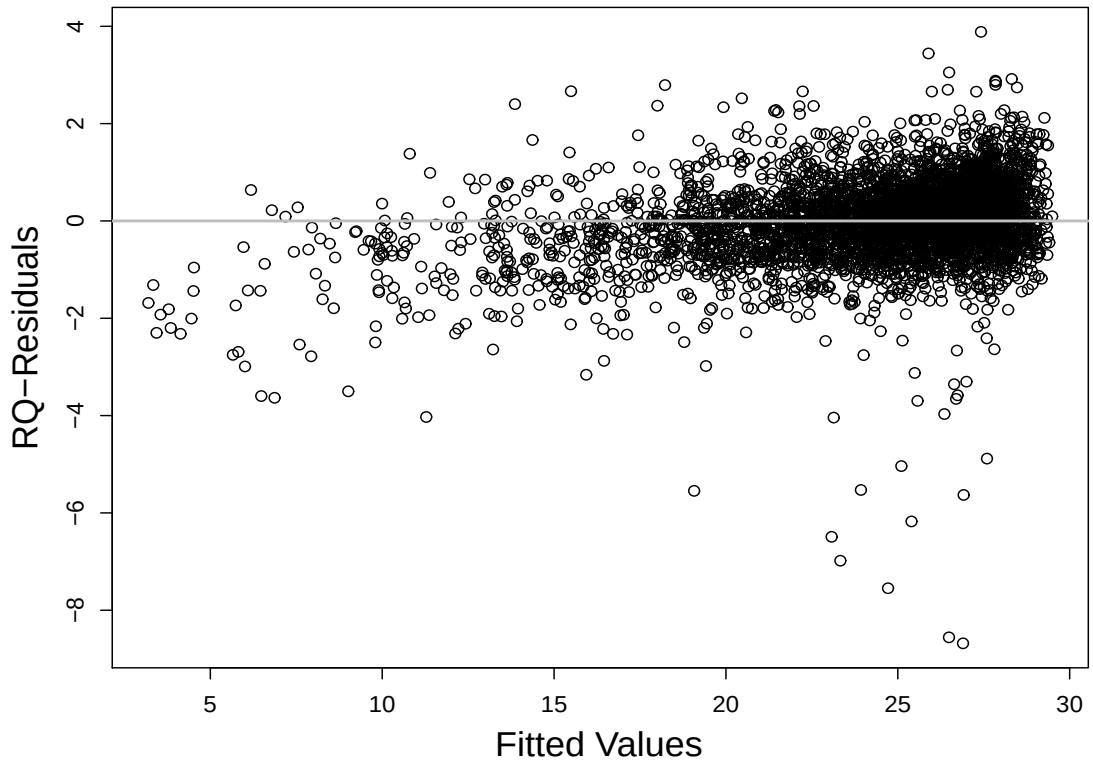

Figure 2: For CC75C data and Model C, randomised quantile residuals versus fitted values.

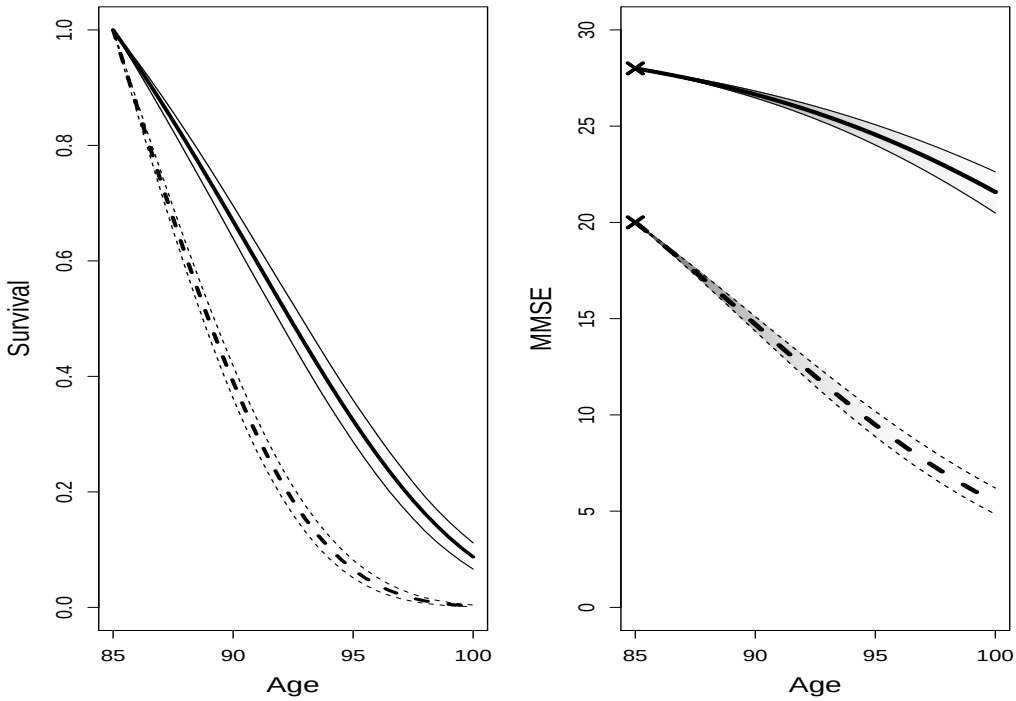

Figure 3: Prediction of survival and mean MMSE score for a woman aged 85 with fewer than ten years of education and current score of 28 (solid lines) or 20 (dashed lines). Derived from Model C for CC75C, with 95% confidence bands from MLE simulation with 1000 replications. Right-hand side: probability of survival depicted by grey scale from black (probability 1) to white (probability 0).

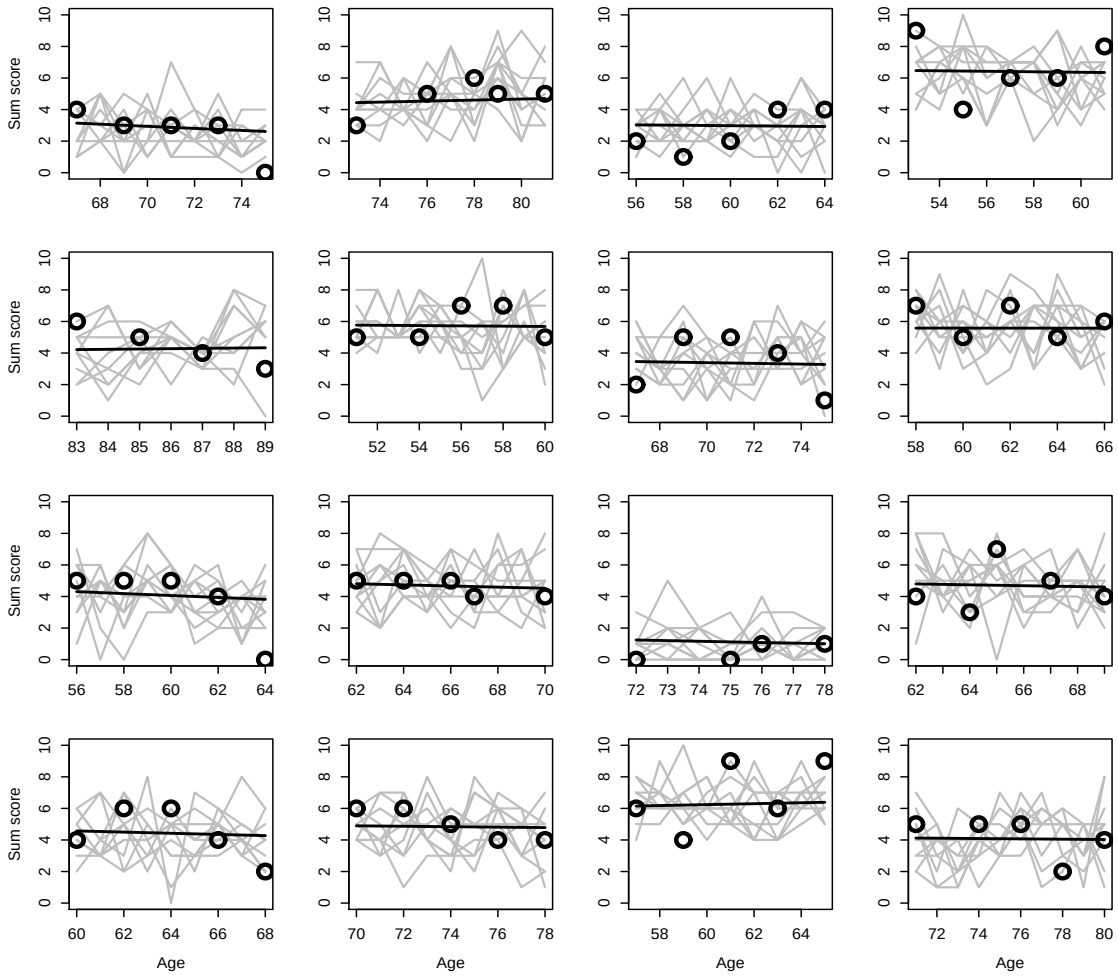

Figure 4: For ELSA data, observed sum score (black circles) and fitted trajectories for 16 individuals randomly chosen from those with four or more observations. Black lines for fitted mean trajectories, and grey lines for sampled trajectories from the fitted distribution using a one-year grid.