DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 30, ARTICLE 42, PAGES 1219-1244
PUBLISHED 15 APRIL 2014
http://www.demographic-research.org/Volumes/Vol30/42/
DOI: 10.4054/DemRes.2014.30.42

Research Article

Investigating healthy life expectancy using a multi-state model in the presence of missing data and misclassification

Ardo van den Hout **Ekaterina Ogurtsova**

Jutta Gampe **Fiona E. Matthews**

This publication is part of the Special Collection on "Multistate Event History Analysis," organized by Guest Editors Frans Willekens and Hein Putter.

Table of Contents

Demographic Research: Volume 30, Article 42
Research Article

Investigating healthy life expectancy using a multi-state model in the presence of missing data and misclassification

Ardo van den Hout [1]

Ekaterina Ogurtsova [2]

Jutta Gampe [3]

Fiona E. Matthews [4]

Abstract

BACKGROUND
A continuous-time three-state model can be used to describe change in cognitive function in the older population. State 1 corresponds to normal cognitive function, state 2 to cognitive impairment, and state 3 to dead. For statistical inference, longitudinal data are available from the UK Medical Research Council Cognitive Function and Ageing Study.

OBJECTIVE
The aim is statistical analysis of longitudinal multi-state data taking into account missing data and potential misclassification of state. In addition, methods for long-term prediction of the transition process are of interest, specifically when applied to the study of healthy life expectancy.

METHODS
Cognitive function in the older population is assumed to be stable or declining. For this reason, observed improvement of cognitive function is assumed to be caused by misclassification of either state 1 or 2. Regression models for the transition intensities are formulated to incorporate covariate information. Maximum likelihood is used for statistical inference.

RESULTS
It is shown that missing values for the state at a pre-scheduled time can easily be taken into account. Long-term prediction is explained and illustrated by the estimation of state-

[1] Department of Statistical Science, University College London, UK. E-mail: ardo.vandenhout@ucl.ac.uk.
[2] Max Planck Institute for Demographic Research, Rostock, Germany.
[3] Max Planck Institute for Demographic Research, Rostock, Germany.
[4] MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

specific life expectancies. In addition, it is shown how microsimulation can be used to further explore predictions based on a fitted multi-state model.

CONCLUSIONS
Statistical analysis of longitudinal multi-state data can take into account missing data and potential misclassification of state. With respect to long-term prediction, microsimulation is a useful tool for summarising and displaying characteristics of cognitive decline and survival.

1. Introduction

Statistical models for multi-state processes can be found in many applications. In demography, the models can be used to study transition processes, such as change in marital status, relocating region of residence, or change of employment status. In biostatistics, multi-state models are used to study health-related processes. An illness-death model describes a process with two or more living states, which represent different stages of a condition or disease, and an absorbing state that corresponds to dead.

Practical applications of multi-state models face several complications, most of them relating to missing information. When individuals are followed up, they are usually not monitored continuously, but examined or interviewed at pre-scheduled times. If the state of a multi-state process is only observed intermittently, then the data are called panel data and times of transitions between states are interval-censored. When information on the state at the pre-scheduled times is missing for some individuals, then this should be taken into account in the data analysis.

This paper presents a three-state illness-death model for cognitive ability in older ages. The three states correspond to normal cognition (state 1), cognitive impairment (state 2), and dead (the absorbing state 3), see Figure 1. Cognitive decline in the older population is assumed to be a progressive process. Once an individual becomes cognitively impaired, no return to the state 1 is possible. Cognitive ability is usually assessed by standardised instruments, such as the Mini-Mental-State Examination (MMSE, see Folstein, Folstein, and McHugh 1975), in which higher scores represent better performance. Although cognitive ability is assumed to be stable or declining over time, test scores can show higher values at later interviews for some individuals. People can vary in their performance from day to day, and the scores on tests for cognition may fluctuate accordingly. Correspondingly, test scores can be seen as assessments of the (latent) cognitive status, and may be affected by measurement error. These measurement errors lead to misclassification of the

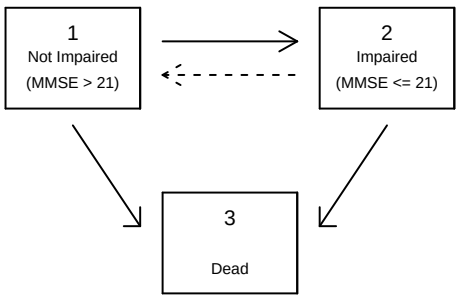
actual state. The model that will be presented accounts for this kind of misclassification. Additionally, the model allows for missing data on state at pre-scheduled observation times by adopting a relatively simple *missing at random* (Little and Rubin 2002) model, in which it is assumed that missingness may depend on observed states, but not on unobserved states.

Panel data are available for the study that will be presented. Times of transitions from state 1 to state 2 are interval-censored, but death dates (transitions to state 3) are known exactly. The presence of an absorbing state with known entry time is not essential to the presented methodology, but the fact that the process is observed via panel data is an intrinsic part of the model in this paper.

The risk of transitions between the three states in the model for cognitive decline and death depends on the age of the individual. We consider the process to be time continuous, and age is the time-scale in the model. Continuous-time multi-state models are formulated by specifying transition hazards. For human mortality, the Gompertz distribution, whose hazard is exponentially increasing, is known to be an appropriate model. We therefore specify the hazards for transitions into state 3 by a Gompertz model. Also, the transition from the cognitively intact to the impaired state (from state 1 to state 2) is modelled via a Gompertz hazard.

In addition to the age-specific hazards for the different transitions, effects of risk factors are of importance. This leads to regression models for the transition intensities. In this paper we will incorporate sex, education and birth cohort as additional covariates. From the estimated models, remaining life expectancy at different ages (overall and in the two states of cognitive function) can be derived.

Figure 1: Three-state model for cognitive function in the older population. Dashed arrow for the transition that is observed only because of misclassification of state



The estimation of the multi-state model parameters is based on existing methodology, see Kalbfleisch and Lawless (1985), Satten and Longini (1996), and Jackson (2011). Statistical inference is extended beyond the work in Van den Hout, Jagger, and Matthews (2009), and Van den Hout and Matthews (2010). Model fit is explored in more detail, and a comparison to alternative models is undertaken. The fitted model is used to compute life expectancies, which are investigated graphically to better illustrate the effect of risk factors.

Specific to the investigation of cognition, the model allows for misclassification of the living states: an observed state may not correspond to the underlying true state. For this reason, the initial state distribution is estimated using logistic regression.

Once the transition rates are estimated, additional results can be produced by microsimulation. Fur this purpose, individual trajectories from the multi-state model are created by Monte Carlo simulation. The resulting virtual population can be analysed with respect to many additional features. This paper uses the microsimulation tool from the MicMac software (Zinn et al. 2009). This software was implemented to simulate individual life courses from continuous-time multi-state models, with special emphasis on population projections (Willekens 2005). It provides various tools to summarise and display characteristics of the virtual population, some of which are shown for the cognitive decline and death process.

The rest of the paper is organised as follows. Section 2 describes the statistical modelling, which includes the time-dependent hazards, the accommodation of missing state information, and the modelling of misclassification of states. The calculation of residual life expectancies is discussed in Section 3. Section 4 describes the microsimulation tool that is employed for producing additional results. The actual application is presented in Section 5, where data are analysed from the MRC Cognitive Function and Ageing Study (CFAS). A discussion in Section 6 complements the paper.

2. Statistical modelling

The continuous-time multi-state model for panel data is discussed by Kalbfleisch and Lawless (1985). Methods for taking into account the availability of exact times of death can be found in Kay (1986). Satten and Longini (1996) and Jackson (2011) present models for dealing with misclassification of states. In case all transition times for the continuous-time process are observed (no interval-censoring), other methodology is available, see, e.g., Putter, Fiocco, and Geskus (2007). This case will not be discussed.

This section will explain three specific aspects of the three-state model in the application: the piecewise-constant approximation of the time-dependent hazards, the presence of missing states, and the modelling of misclassification. See Figure 1 for the three-state

model where interval-censored transitions from state 2 to state 1 are assumed to be the result of misclassification.

The models presented in this section can be fitted using the free R-package msm (Jackson 2011). In msm, the parameters in the multi-state models are estimated by maximum likelihood, where the optimisation is undertaken by a call to the R-routine optim. All the models in the application in Section 5 are fitted using msm, where the optimisation by optim is specified by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) routine.

2.1 Time-dependent hazards

Continuous-time multi-state models can be specified by linking covariates to transition intensities. A transition intensity $q_{rs}(t)$ is the instantaneous risk (hazard) of moving from state r to state s at time t. In the three-state model that we consider in the application, there are three transition-specific hazards: $q_{12}(t)$, $q_{13}(t)$, $q_{23}(t)$, and age is the time scale t. Since a hazard is non-negative, a log-link is used to relate the hazard to covariates. The loglinear model is given by

$$\log[q_{rs}(\textit{age})] = \beta_{rs.0} + \beta_{rs.1}\textit{age} + \beta_{rs.2}\textit{ybirth} + \beta_{rs.3}\textit{sex} + \beta_{rs.4}\textit{educ}, \tag{1}$$

where *age* is age minus 75, *ybirth* is year of birth minus 1900, *sex* is gender (0 = women, 1 = men), and *educ* is education (0 = less than ten year of education, 1 = ten or more years of education).

The loglinear model (1) can also be written as

$$q_{rs}(t) = \lambda_{rs}\exp[\gamma_{rs}t]\exp[\beta_{rs.2}\textit{ybirth} + \beta_{rs.3}\textit{sex} + \beta_{rs.4}\textit{educ}], \tag{2}$$

where t is *age*. This formulation better illustrates that the multi-state model can be seen as an extension of the standard survival model (state 1 = alive, state 2 = dead). The baseline hazard in (2) is Gompertz with parameters $\lambda_{rs} > 0$ and γ_{rs}. Other choices are possible for the baseline hazard (such as the Weibull hazard or a piecewise-constant hazard), see, e.g., Van den Hout and Matthews (2008) who investigate the Weibull/Gompertz choice for a three-state survival model for cognitive function in the older population.

To calculate the individual contributions to the likelihood function, the Gompertz baseline hazards will be approximated using piecewise-constant hazards. The approximation will vary across individuals and will be determined by the data. We explain this by an example.

If the observation times for individual i are given by $t_{i1}, t_{i2}, t_{i3}, t_{i4}$, then the transition-specific hazards (2) are evaluated at the starting time of the intervals and are held constant throughout the intervals. That is, the hazards are held constant during the interval $(t_{ij}, t_{ij+1}]$ at the level $q_{rs}(t_{ij})$ for $j = 1, 2, 3$. As long as the intervals are not too long (with respect to the process under investigation), this will provide a good approximation.

In our experience, defining the piecewise-constant hazards using the time midway zo the interval, i.e., using $q_{rs}\left((t_{ij}+t_{ij+1})/2\right)$, does not induce a difference that is relevant for practice. This is of course data dependent. An example in which using the time at the start of the interval may introduce bias is a setting where all individuals are of the same age at baseline, and for all individuals the first interval between observations is considerably longer than the later intervals. This may introduce bias in the approximation of the specified parametric shape.

Alternatively, a grid for the piecewise-constant approximation can be defined independently from the data, in which case, the observed times are embedded within this grid and hazards change from grid point to grid point (Van den Hout and Matthews 2008). This is not investigated in the current application.

In the piecewise-constant approximation explained above, hazards change piecewise-constantly according to the parametric shape of the Gompertz hazard. This approximation of a parametric shape should be distinguished from a non-parametric piecewise-constant hazard model, where for each time interval in a grid, a separate parameter is estimated. The advantage of the latter is that it can approximate any shape. The disadvantage of the non-parametric approach is that it may require many parameters, and that prediction beyond the time of follow up is not straightforward.

2.2 Missing data on state

It is well know that ignoring missing data can lead to biased results. In longitudinal surveys, missing data are ubiquitous. Here we define a relative simple model to deal with missing states at pre-scheduled observation times. We explain it by an example. If observation times for individual i are given by t_{i1}, t_{i2}, t_{i3}, then the likelihood contribution for times t_{i1}, t_{i2}, t_{i3} with missing state at t_{i2} is given by

$$\mathbb{P}(X_{t_{i1}}, X_{t_{i3}}) = \sum_{x=1,2} \mathbb{P}(X_{t_{i1}}, X_{t_{i2}}=x, X_{t_{i3}}), \tag{3}$$

where the sum is over all possible states at time t_{i2}. This model assumes that the missingness is independent from data that are not observed, but may depend on observed data (*missing at random*, Little and Rubin 2002). This is a strong assumption. In case it is assumed that the missingness is dependent on missing data, additional modelling has to be undertaken (Van den Hout and Matthews 2010).

The effect of (3) on the estimation of the transition hazard is hard to quantify in general. It is of course very much related to the extent of the missingness, and whether the missingness is dependent on unobserved data. Using a simulation study for a three-state illness-death model, Van den Hout and Matthews (2010) show that ignoring the missing

states by undertaking a complete-data analysis underestimates the risk of moving from state 1 to state 2 and overestimates the risk of moving from state 1 to state 3.

A direct result of taking into account missing states via (3) is that it improves the piecewise-constant approximation of the hazards. By adding t_{i2} in the model, the grid in the piecewise-constant approximation for the data from individual i consists of two intervals $(t_{i1}, t_{i2}]$ and $(t_{i2}, t_{i3}]$, instead of one interval $(t_{i1}, t_{i3}]$. Note that such an approach can also be used to improve the piecewise-constant approximation even when there are no missing states in the data.

2.3 Misclassification

We assume that cognitive function in the older population is stable or decreasing. However, when cognitive function is measured over time, individual performance may vary from day to day and test scores may fluctuate accordingly. The three-state model assumes that a transition from the impaired state back to the non-impaired state is not possible and a misclassification model is used to deal with backward transitions. We see the misclassification as a way to smooth the data: the measurement of cognition over time may result in a backward transition, but the underlying latent process is assumed to be stable or decreasing. Whether this assumption is realistic for cognitive impairment in the older population is a topic of an ongoing debate, see, e.g., Le Couteur et al. (2013).

Misclassification is taken into account by estimating

$$\theta_{rs} = \mathbb{P}(X^* = s | X = r) = \mathbb{P}(\text{Observed state} = s | \text{Latent state} = r), \tag{4}$$

for $(r, s) \in \{(1, 2), (2, 1)\}$, which is the probability of observing state s given a latent true state r. Death is not misclassified. We assume that the misclassification is independent between individuals, and also independent across times of observation. As a result the likelihood contribution of individual i for times $t_{i1}, ..., t_{in_i}$ is given by

$$\begin{aligned} \mathbb{P}(X^*_{t_{i1}}, .., X^*_{t_{in_i}}) &= \sum \mathbb{P}(X^*_{t_{i1}}, .., X^*_{t_{in_i}} | X_{t_{i1}}, .., X_{t_{in_i}}) \mathbb{P}(X_{t_{i1}}, .., X_{t_{in_i}}) \\ &= \sum \mathbb{P}(X^*_{t_{in_i}} | X_{t_{in_i}}) \times .. \times \mathbb{P}(X^*_{t_{i1}} | X_{t_{i1}}) \mathbb{P}(X_{t_{i1}}, .., X_{t_{in_i}}), \end{aligned} \tag{5}$$

where the summation is over all possible paths of latent states $X_{t_{i1}}, .., X_{t_{in_i}}$.

Using the first-order Markov assumption, $\mathbb{P}(X_{t_1}, .., X_{t_{n_i}})$ in (5) becomes $\mathbb{P}(X_{t_{n_i}} | X_{t_{n_i - 1}}) \times .. \times \mathbb{P}(X_{t_2} | X_{t_1}) \mathbb{P}(X_{t_1})$. Note that $\mathbb{P}(X_{t_1})$ is unknown and has to be estimated. In the three-state model for cognition, there are only two living states and a

standard logistic regression model can be used to estimate $\mathbb{P}(X_{t_1})$. We define

$$\begin{aligned} \mathbb{P}(X_{age}=2) &= \frac{\exp[\mu]}{1+\exp[\mu]} \\ \mu &= \alpha_0+\alpha_1 age+\alpha_2 sex\,. \end{aligned} \tag{6}$$

The linear predictor can of course be extended by adding more covariates. Here we restrict the model by using only *age* and *sex*. This choice is partly driven by the aim for parsimony, but also because age and sex are important factors for the onset of cognitive impairment (see the analysis in Section 5.2). If there are more than 2 living states, multinomial logistic regression models can be applied in a similar manner.

3. Life expectancies

We consider residual life expectancy (LE) at a given age t_0. Given a fitted multi-state model, the estimation of LEs is established methodology, see, e.g., Izmirlian et al. (2000) and Van den Hout, Jagger, and Matthews (2009). This section provides the formulas and some extra explanation.

Total LE is expected stay in the living states, but also of interest is expected stay in a specific living state. In our model, LE in state 1 is *impairment-free life expectancy* and LE in state 2 is *impaired life expectancy*. We will show how LEs can be derived from the parameters of the multi-state model. This will be explained by a comparison with mean survival in a standard survival analysis.

In standard survival (state 1 = alive, state 2 = dead) LE is the expectation of the remaining years spent alive (U) given by

$$\begin{aligned} \mathbb{E}(U|t_0,\mathcal{Z}) &= \int_0^\infty u f(u|t_0,\mathcal{Z})du = \int_0^\infty S(u|t_0,\mathcal{Z})du \\ &= \int_0^\infty \mathbb{P}(X_{t_0+u}=1|t_0,\mathcal{Z})du, \end{aligned}$$

where $\mathcal{Z}$ is the covariate history which is assumed to be deterministic. This is called mean survival. In a multi-state survival model, LE is defined in a similar way. LE in state s given state r at t_0 is given by

$$e_{rs}(t_0)=\int_0^\infty \mathbb{P}(X_{t+t_0}=s|\ X_{t_0}=r,\mathcal{Z})dt. \tag{7}$$

LE conditional on a state at t_0 is an integral where the integrand is a transition probability. The latter can be estimated from the fitted multi-state model, see also Izmirlian et al.

(2000). Approximation of the integral can thus be undertaken by numerical methods using the estimated parameters from the multi-state model.

Expected stay in a state is sometimes called *occupancy time* in the literature on stochastic processes, see for example Kulkarni (2011). Occupancy time is defined up to a given time T. LE in a specified state in an illness-death model can therefore be seen as occupancy time in that state for $T = \infty$.

It is useful to define marginal LE. For our three-state model, this is defined by $e_{\bullet s}(t_0) = \sum_{r=1}^{2} \mathbb{P}(X_{t_0} = r|\mathcal{Z}) e_{rs}(t_0)$. This is LE in state s irrespective of the state at t_0. Note that the distribution of the living states (6) at age t_0 is needed to compute this quantity. Total LE at age t_0 is now given by $e(t_0) = \sum_{s=1}^{2} e_{\bullet s}(t_0)$.

To extrapolate the uncertainty in the estimation of the model parameters to the estimation of LEs, we consider the multivariate normal distribution with expectation equal to the maximum likelihood estimate of the model parameter vector, and the covariance matrix equal to the estimated covariance matrix at the optimum. By drawing parameter values from this distribution and computing the LEs for each of the drawn values, the uncertainty in the estimation of the model parameters will be propagated (cf. Aalen et al. 1997).

4. Microsimulation from multi-state models

While many characteristics of multi-state models can be derived from analytic expressions, still more detailed results can be obtained by Monte Carlo simulation from the estimated model. For a starting population with specified characteristics (that is, the age distribution of the individuals as well as the states they occupy) individual trajectories are simulated according to the stochastic process that the model describes. Commonly, such an approach is called microsimulation.

What is needed for this purpose, besides the estimated model, is a software tool that allows us to flexibly specify the model, to efficiently simulate a potentially large number of individuals, and some means to conveniently summarise the characteristics of interest from the simulated population (which is also called the virtual population). The MicMac software, which was developed during the MicMac-project (see Willekens 2005, and www.micmac-projections.org), is based on a generic continuous-time multi-state model with transition intensities that can depend both on the age of the individuals and on calendar time.

The software consists of three components. The so-called pre-processor, which allows the researcher to estimate and prepare input data for the simulation, the MicMac-Core, which performs the simulation, and finally, a post-processor, which provides versatile tools to summarise the simulation output in tables, summary statistics and figures. The pre- and post-processor are written in R (R Development Core Team 2012). The

MicMac-Core (Zinn et al. 2009) is implemented as a plug-in in JAMES II (JAva based Multipurpose Environment for Simulation), which is a multi-agent system and simulation platform, see Himmelspach and Uhrmacher (2007). In Section 5.5 we will use this microsimulation tool to illustrate and supplement analytical results of the model that was presented in Section 2. Using a fitted multi-state model and specified covariate values, transition intensities are derived for the age range, and fed into MicMac. On the bases of this information, MicMac simulates individual trajectories.

5. Application

5.1 CFAS data

The MRC Cognitive Function and Ageing Study (CFAS, www.cfas.ac.uk) is a population-based longitudinal study of cognition and health conducted between 1991 and 2004 in the older population of England and Wales (Brayne, McCracken, and Matthes 2006). Data were collected in six centres, five in England and one in Wales. Here we analyse the CFAS subset defined by the data from rural Cambridgeshire with respect to cognitive function and survival. Using CFAS, Matthews et al. (2006) showed that there are regional differences in the UK in aspects of health. We would like to stress that results presented in this section pertain to the population of rural Cambridgeshire only. The states of the model are defined in Figure 1.

The sample size for rural Cambridgeshire is 2600, with 1493 women and 1107 men, who at the baseline of the study were aged 65 years or older. The frequencies of the five age categories $(\leq 70, (70,75], (75,80], (80,85], > 85)$ at baseline are (861, 533, 569, 373, 264). In the sample there are 845 individuals who were observed once in the alive states, 611 were observed twice, 455 three times, and 612 four times. There are 77 individuals in the data who where not observed in state 1 or 2. Most of these individuals have a missing state at baseline followed by death before the next scheduled interview. The total number of records (including the ones with missing states, deaths, and right-censored states) is 10,525.

The living states 1 and 2 are defined using 21 as cut point in the Mini-Mental State Examination (MMSE) scale, see Figure 1, in which cognitive impairment is defined by an MMSE of 21 or lower. The state table is given in Table 1, which presents the number of times a pair of states is observed at successive observation times. There are 68 interval-censored transitions from state 2 back to state 1. According to our model, these transitions are caused by misclassification.

Table 1: State table for the CFAS data from rural Cambridgeshire. Number of times a pair of states is observed at successive observation times and at the end of the follow-up

	To				
From	**1**	**2**	**Dead**	**Missing**	**Right-censored**
1	2862	229	737	857	444
2	68	169	280	188	46
Missing	28	16	680	908	413

5.2 Models

This section will investigate three formulations of a three-state model for the data at hand. Models denoted with A are progressive models with misclassification. Model B is a model with misclassification *and* a backward transition from state 2 to state 1. Model C is a model with a backward transition from state 2 to state 1, but without misclassification.

We start with the progressive model defined by (1), (4), and (6), and investigate restrictions on the regression parameters. Model A_{01} is the model with intercepts and only *age* in (1), i.e., with restrictions $\beta_{rs.2} = \beta_{rs.3} = \beta_{rs.4} = 0$, and with intercept and *age* in (6), i.e., with restriction $\alpha_2 = 0$. The value -2 $\times$ maximised loglikelihood ($-2LL$) for Model A_{01} is 14822. The model without any restrictions (Model A) has $-2LL = 14643$. The difference in $-2LL$ is significant according to the likelihood ratio test. The Akaike Information Criterions (AIC) for the models are respectively 14842 and 14683 - further proof of modelling improvement. Table 2 presents $-2LLs$ and AICs, also for intermediate Models A_{02} and A_{03}. As was to be expected, adding *sex* to the model leads to substantial improvement. The parameters of the three-state model with recovery *and* misclassification may be hard to identify. Allowing backward transitions and misclassification makes it difficult to estimate the model, as it is not clear which process underlies individually observed interval-censored transitions.

Table 2: Information criteria for the maximum likelihood estimation: -2 $\times$ maximised loglikelihood (-$2LL$) and the Akaike Information Criterion (AIC). The number of parameters is #p

Model	Covariates		-$2LL$	#p	AIC
	Multi-state	*Baseline*			
A_{01}	*age*	*age*	14822	10	14842
A_{02}	*age*, *ybirth*	*age*	14815	13	14841
A_{03}	*age*, *ybirth*, *sex*	*age*, *sex*	14665	17	14699
A	*age*, *ybirth*, *sex*, *educ*	*age*, *sex*	14643	20	14683

Table 3: Parameter estimates for Model A. Estimated standard errors in parentheses

Multi-state model						
Intercept	$\beta_{12.0}$	-3.720 (0.684)	$\beta_{13.0}$	-3.050 (0.273)	$\beta_{23.0}$	-2.873 (0.282)
age	$\beta_{12.1}$	0.164 (0.035)	$\beta_{13.1}$	0.050 (0.013)	$\beta_{23.1}$	0.082 (0.014)
ybirth	$\beta_{12.2}$	0.003 (0.033)	$\beta_{13.2}$	-0.031 (0.013)	$\beta_{23.2}$	0.029 (0.013)
sex	$\beta_{12.3}$	-0.693 (0.217)	$\beta_{13.3}$	0.844 (0.098)	$\beta_{23.3}$	0.455 (0.107)
educ	$\beta_{12.4}$	-0.657 (0.195)	$\beta_{13.4}$	-0.092 (0.090)	$\beta_{23.4}$	0.204 (0.113)

Logistic regression mode			**Misclassification model**	
Intercept	α_0	-2.032 (0.130)	θ_{12}	0.016 (0.003)
age	α_1	0.161 (0.012)	θ_{21}	0.172 (0.030)
sex	α_2	-0.349 (0.155)		

For the CFAS data, we were not able to identify the parameters in the model defined by (1), (4), and (6), where (1) is also defined for $q_{21}(age)$. The maximisation of the likelihood in msm provides a maximum, but does not provide a Hessian matrix from which uncertainty can be derived. By restricting the misclassification, the identifiability problem can be solved for the data at hand: in Model B we allow only misclassification of state 2 as state 1 (θ_{21}), but not vice versa ($\theta_{12} = 0$ restriction). The restriction is motivated by the fit of Model A, where $\widehat{\theta}_{12} = 0.016$. For Model B, the probability of misclassifying state 2 is estimated at $\widehat{\theta}_{21} = 0.195$. Model B has $-2LL = 14656$ and AIC = $14656 + 2 \times 24 = 14704$, so the model does not perform better than Model A.

Note that we cannot use the likelihood ratio test to compare Model B with Model A. Because of the restriction on the misclassification, the latter model is not a restricted version of the former.

We define Model C as Model B without misclassification. The likelihood ratio test cannot be used to compare Model C to a three-state model with misclassification. Models with misclassification include logistic regression models for the states at baseline, whereas Model C does not. The likelihood of a model with misclassification is a combined likelihood of the three-state model and the logistic regression model. These two submodels are not independent; they share the parameters for the misclassification. In addition, the modelling of a missing baseline state for an individual is based on the whole observed trajectory for that individual, and is therefore not independent from the three-state model. Model C has 20 parameters (4 transitions with 5 parameters per transition). Estimated LEs using the models A, B, and C will be compared in the next section.

Note that a model without recovery and without misclassification is not possible for the CFAS data, since such a model cannot deal with observed interval-censored transitions from state 2 to state 1.

We briefly discuss the parameter estimates for Model A as presented in Table 3. All estimated coefficients for age are positive, which reflects the fact that with increasing age a transition to a next state becomes more likely - as was to be expected. The estimated effect of education implies that more education delays the onset of cognitive impairment but shortens the stay in the impaired state. This agrees with the results presented in Reuser, Willekens, and Bonneux (2011). Men are less likely to move to the impaired state than women but have a higher mortality rate. The estimated effect of year of birth for the transition from the impaired state to the dead state implies that the younger generation spends less time in the impaired state.

The estimated coefficients for the logistic regression model imply that the risk of impairment increases with age, and that women have a higher risk than men.

The estimated misclassification mainly indicates failure to detect an impaired state: there is a probability of 17% that an underlying impaired state is classified as a healthy state.

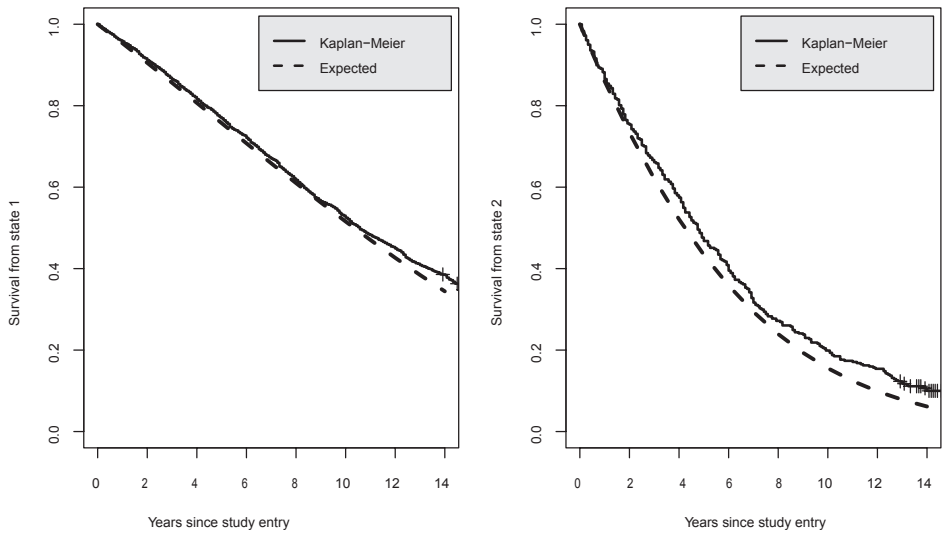
5.3 Model validation

Formal model validation for multi-state models with misclassification has not yet been developed. Aguirre-Hernandez and Farewell (2002) and Titman and Sharples (2010) provide tests for a set of multi-state models, but this set does not include our model. However, Titman and Sharples also review less formal ways of model validation, and we use their basic ideas in the following validation of Model A.

Figure 2 depicts Kaplan-Meier estimators and predicted survival conditional on ob-

served baseline state. For the predicted survival, individual trajectories are predicted conditional on observed baseline state and baseline covariate values, where not-observed baseline states are imputed (once) using the fitted logistic regression model. The potential misclassification for all baseline states is taken into account by imputing (once) the latent state conditional on the observed state and the fitted misclassification model. In the prediction, age-dependent intensities are modelled as piecewise-constant on a one-year grid for all individuals. The piecewise-constant approach allows a direct computation of the transition probabilities from the time-dependent intensities.

Figure 2: Survival conditional on baseline state, where expected is derived from Model A

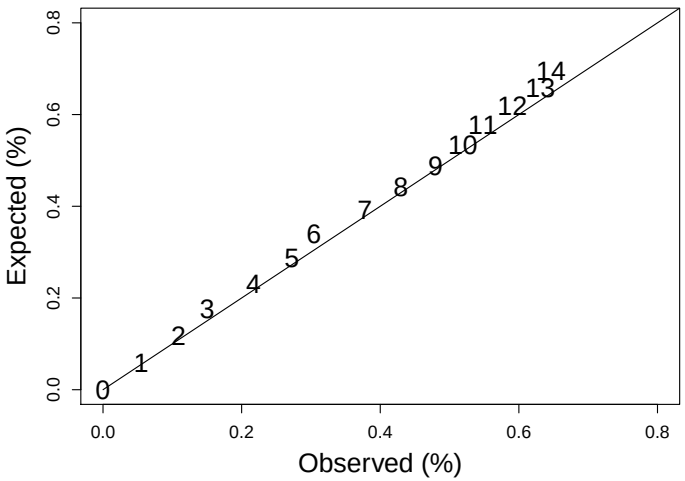


Because single imputation is used, graphs differ slightly when the process is rerun, but differences are minor. The lack of large differences after rerunning the process was the motivation to impute only once. The comparison to predicted survival only addresses part of the model and should not be seen as a definite assessment of goodness of fit. Nevertheless, from Figure 2 we may conclude that Model A captures well the survival as observed in the data, although there is some difference between the curves near the end of the follow-up.

As an alternative, we could investigate observed and expected prevalence in state 3 for pre-specified time grid. Figure 3 shows observed versus expected prevalence in per-

centages for the years since the start of the study (0 up to 14 years). At the start of the study, no deaths are observed and no deaths are expected. This is depicted by the 0 at location $(0, 0)$. Then, as the years accumulate, we see agreement between observed and expected deaths: the numbers for the subsequent years are close to the diagonal. Figure 3 is a coarser version of the information presented in Figure 2.

Figure 3: **Observed versus expected prevalence with regard to death according to Model A. The numbers correspond to the year-to-year grid starting at the time of entry to the study. The location of the numbers depicts expected versus observed**



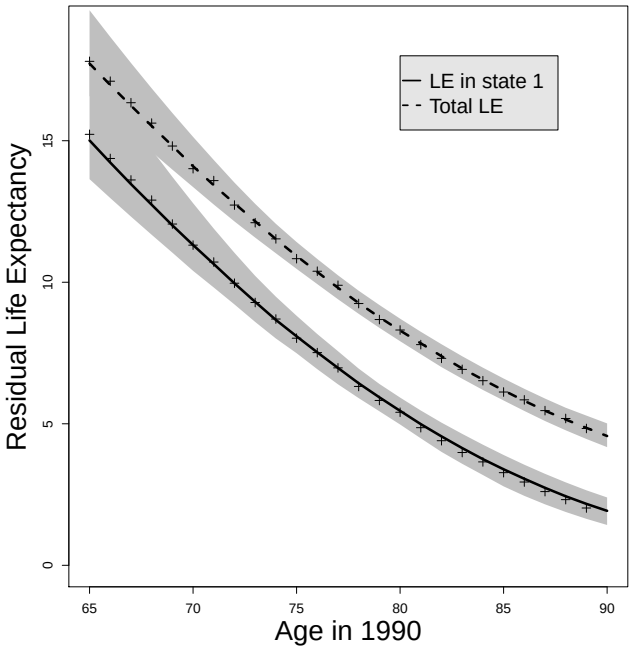
The model being validated is not strictly first-order Markovian. More than current state is used to predict future states - current age and current covariate values are also taken into account. Better modelling may still be possible by including state-specific length of stay, but this is not investigated here, as Figure 2 shows good agreement between predicted survival and the Kaplan-Meier estimates.

5.4 Life expectancies

We will not provide complete inference for estimated life expectancies (LEs), but as an illustration of the methodology, we will provide LEs estimates for women. Estimates from Model A are discussed first. At the end, a comparison with estimates from Models B and C is made.

Figure 4 depicts estimated non-impaired LE and total LE for women (with 95% confidence intervals estimated using 100 iterations in the simulation). The LEs are conditional on being in the non-impaired state at specified age in 1990, and pertain to women with less than ten years of education. The integral in (7) is numerically approximated. The grid for this approximation also defines the piecewise-constant approximation of the intensities.

Figure 4: **For women in the non-impaired state with less than ten years of education: estimated non-impaired LE and total LE (with estimated 95% confidence intervals) conditional on age in 1990 (Model A).** MicMac **results depicted by "+"**

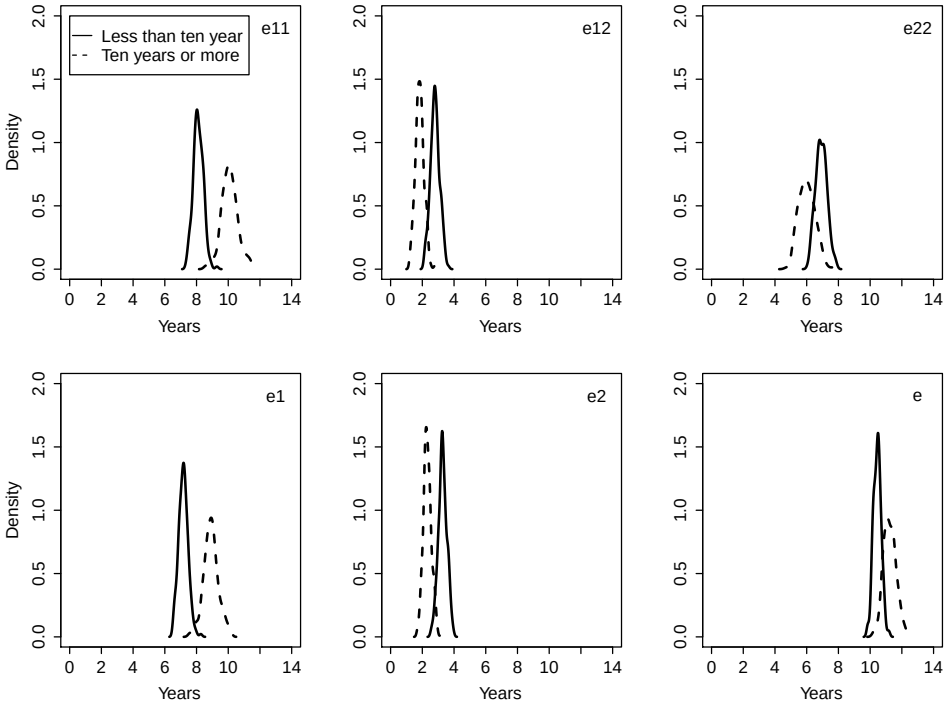


As can be deduced from the graph, time spent in the impaired state is fairly constant with age: the distance between the solid and the dashed curves does not change much along the age axis.

The graph also shows that estimation for the younger ages is more uncertain. This is understandable, as this prediction involves extrapolation over a long time - even beyond the follow-up time of the study.

The effect of education can be investigated by plotting the distributions of estimated LEs. Figure 5 shows the distributions of estimated LEs for women aged 75 in 1990. LEs are functions of model parameters, and Figure 5 depicts the uncertainty in the estimation of the LEs as derived from the uncertainty associated with the maximum likelihood estimator of the model parameters.

Figure 5: Uncertainty distribution of life expectancy estimates resulting from Model A, for women aged 75 in 1990 (by years of education)



The effect of education is most pronounced for e_{11}: the overlap of the (support of the) distributions is small. As there are no transitions from state 2 back to state 1, this effect of education is also present in $e_{\bullet 1}$. In state 2, the effect of education is not strong: more education is associated with less time in state 2, but there is some overlap between the distributions. As can be deduced from the distributions of estimated total LEs, there is a positive effect from more years of education, but it does not seem significant. For women aged 65 or 85 in 1990 (distributions not reported), the effect of education is similar.

With a larger sample size, there would be less uncertainty in the estimation of model parameters, and this would lead to less uncertainty in the LEs estimation. Assessing the effects of risk factors by Figure 5 is, of course, conditional on the sample size of the study.

Next we compare LEs estimated from Models A, B and C, see Table 4. Starting with Models A and B, we see that the total LEs are similar (taking into account the 95% confidence intervals), but that there are differences for the marginal LEs in the living states. The latter is understandable, since the assumptions for the possible transitions between the living states are different in the two models. Model B leads to longer estimated stay in the impaired state. This is the result of a higher estimated probability for misclassifying a latent state 2 as an observed state 1. Fitted Model B implies a higher unobserved prevalence of cognitive impairment than fitted Model A, and this leads to longer estimated stay in the impaired state.

Given the estimated misclassification in Models A and B, it is understandable that Model C (without misclassification) estimates a shorter stay in the impaired state than Models A and B. Total LEs estimated using Model C are similar to the estimates using Model A and B when taking the 95% confidence intervals into account, although the point estimates using Model C are consistently slightly higher. The latter is probably caused by the estimated misclassification for Models A and B; assuming that there are more individuals in state 2 than observed leads to a higher (latent) prevalence in state 2, which implies poorer overall survival.

External validation of the model is possible by comparing estimated total LEs with estimates provided by statistical agencies. The web site of the UK Office for National Statistics (ONS, www.statistics.gov.uk) provides "Historic Interim Life Tables" for England. For the ages 65, 75 and 85 with year of birth 1915, the estimates are 17.04, 10.98, and 6.15 respectively. To compare these figures with estimates from our models, we estimate total LEs using the level of education equal to the mean of *educ* at baseline, which is 0.3. The results for Model A are 17.22, 10.61, and 5.00, respectively. For Model C these results are 17.65, 10.92, and 5.71. This comparison should be made with care: e.g., ONS produces an estimate for the whole of England and we use data from rural Cambridgeshire. Only large differences may imply model misfit. The estimates for the 85 years old using Model A cause some concern in this respect.

Table 4: Estimated life expectancies for women born in 1915 and with less than ten years of education. Comparison between models. Point estimates and 95% confidence intervals

	To Model A		Model B		Model C	
Age 65						
e_1	13.80	(12.80, 14.79)	13.37	(12.28, 14.56)	14.43	(13.41, 15.22)
e_2	3.11	(2.46, 3.75)	3.55	(2.89, 4.36)	2.86	(2.49, 3.46)
e	16.91	(15.86, 18.03)	16.92	(16.04, 17.92)	17.28	(16.32, 18.03)
Age 75						
e_1	7.14	(6.54, 7.83)	6.85	(6.32, 7.49)	7.74	(7.22, 8.31)
e_2	3.27	(2.70, 3.77)	3.60	(3.06, 4.21)	2.88	(2.49, 3.28)
e	10.41	(9.98, 10.86)	10.44	(9.98, 10.91)	10.63	(10.16, 11.05)
Age 85						
e_1	2.12	(1.52, 2.78)	2.28	(1.66, 2.99)	3.32	(2.58, 4.20)
e_2	2.85	(2.19, 3.60)	2.92	(2.35, 3.72)	2.25	(1.71, 2.83)
e	4.97	(4.33, 5.50)	5.20	(4.62, 5.80)	5.57	(4.94, 6.19)

5.5 Inference using MicMac microsimulation

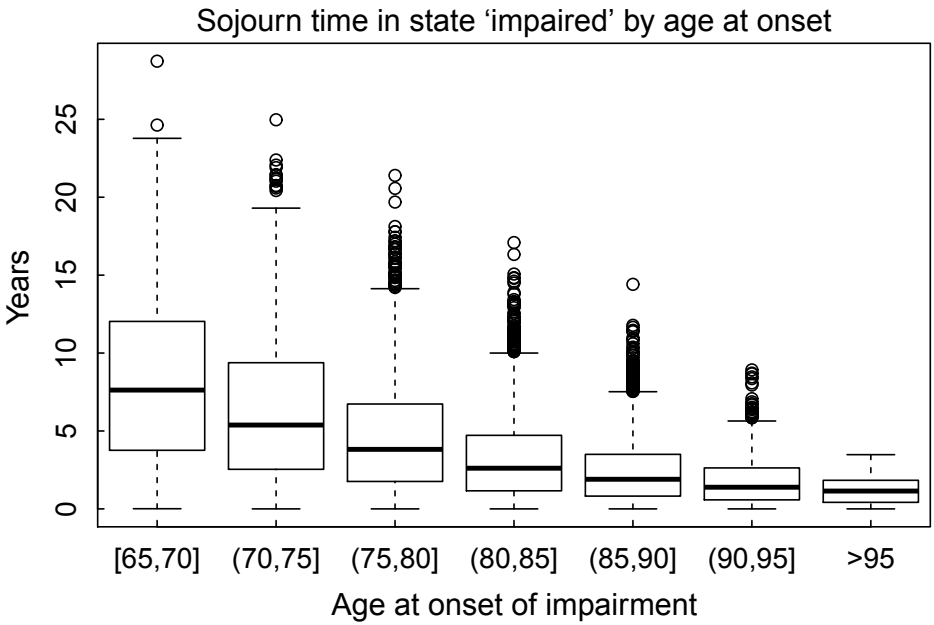
In this section we illustrate how the fitted multi-state model can be exploited further by feeding the estimated transition rates into a microsimulation. In this particular application we used the MicMac microsimulation software (see Section 4).

As a first step, we replicated the calculation of the residual impairment-free LE and total LEs that was presented in Section 5.4. For this purpose we simulated 25 cohorts of individuals who were between 65 and 90 years old in 1990. Each of these cohorts consisted of 20,000 women and 20,000 men, and the individuals were submitted to the hazards estimated in Section 5.2 and were followed for 50 years. From the simulated life courses summary statistics, such as the average length of life and the average duration of stay in state 1 (non-impaired) can be calculated. With large enough sample sizes, deviations of these averages from the analytically derived LEs should be minor. Figure 4 includes the results obtained from the microsimulation for impairment-free LE and total LE, and clearly demonstrates the agreement of the two approaches.

While in Figure 4 residual LE in the non-impaired and, implicitly, the cognitively impaired state is considered as a function of age (in 1990), we used the microsimulation approach to study the length of stay in the impaired state by the age at onset of the impairment. Again, we follow a cohort of 20,000 women, all of whom are aged 65 in 1990 and not impaired at this time. We focus on women with less than ten years of education. From the simulated life courses, we calculated the proportion of women who moved into

state 2 (impairment) by age, and we determined the duration of how long the individuals remained in this state before dying. The results are summarised in Table 5 and Figure 6.

Figure 6: Duration of cognitive impairment by age at onset for women aged 65 in 1990 with less than ten years of education



About 35% of the simulated individuals never entered the state of cognitive impairment. The incidence is low in the two youngest age-groups – about 5% and 9%, respectively – but the mean length of stay in state 2 is comparatively high – more than 8 and 6 years, respectively. Incidence of cognitive impairment peaks in the age-group 85 to 90, where about 40% of the survivors to age 85 move into state 2 during the following five years.

Table 5: Age at onset and duration of cognitive impairment (CI) for the cohort of women aged 65 in 1990 with less than ten years of education

Age at onset	To [65,70]	[70,75]	[75,80]	[80,85]	[85,90]	[75,80]	>95	never
Proportion who develop CI	0.048	0.080	0.134	0.183	0.157	0.050	0.002	0.346
Proportion who develop CI among those surviving to age range	0.048	0.088	0.170	0.293	0.401	0.346	0.121	
Duration								
Mean	8.36	6.29	4.62	3.30	2.44	1.81	1.24	
Standard deviation	5.57	4.59	3.59	2.72	2.10	1.57	0.90	

Model comparison in Section 5.3 indicated that the year of birth should be included in the final model. Later birth cohorts were estimated to have a similar risk of moving into the impaired state, but to suffer from a higher death rate out of this state (whereas the hazard of death for the non-impaired was estimated to decline with later birth years). In their joint effect, these different trends can have interesting consequences for the future prevalence of the cognitively impaired and the structure of the sub-population in state 2. These effects could be studied well by microsimulation, too, but this would need realistic assumptions about the cohort sizes arriving at age 65. This investigation is beyond the scope of the current paper and will be studied elsewhere.

6. Discussion

A continuous-time three-state model was used to describe change in cognitive function in the older population. Observed improvement of cognitive function was assumed to be caused by misclassification. Estimation of state-specific life expectancies (LEs) was illustrated and microsimulation was utilised to further explore the implications of the fitted model.

Regarding the application to the CFAS data, the presented methods can be used to explore extended transition-specific regression models. The analysis in Section 5.2 is limited in this aspect, as only age and sex are used as explanatory variables. Another extension of the analysis would be to define regression models for the misclassification probabilities, see, e.g., Jackson (2011).

When misclassification is present and modelled, LEs can be estimated conditional on observed (manifest) state or conditional on true (latent) state. We have chosen to do the latter. When it comes to practice with regard to the whole population, for example the planning of future health care, estimating LEs conditional on latent state makes more

sense as need for care will be induced by the true latent state. However, when state-specific LEs for a given individual are the primary quantities of interest, estimation will need to take misclassification into account.

The comparison between the models A, B, C, i.e., between models with and without misclassification, was undertaken for the data at hand. The conclusions from that comparison are tentative. For more insight, the differences should be investigated in a simulation study where the models are assessed in various scenarios.

Model validation is very important, especially when a multi-state model is used for prediction as in the case when life expectancies are computed. The estimation of life expectancies implies extrapolation of the model beyond the age range in the data. For any model, be it parametric or non-parametric, this kind of extrapolation is not without danger and model validation is of specific interest. However, model validation is still a subject of research (Titman and Sharples 2010) and has not yet caught up with the current flexibility to create extended models. Nevertheless, some heuristic methods to assess model fit where presented in this paper.

The link between the software msm for the fitting of continuous-time multi-state models and the MicMac software for microsimulation has great potential. Both tools are available in the free programming environment R (R Development Core Team 2012). The model formulation in msm is very flexible. Using microsimulation based on a fitted model makes it relatively easy to compute additional quantities, such as duration in a state by age at entering that state.

7. Acknowledgements

The authors would like to thank CFAS for the permission to use the data. CFAS is supported by major awards from the Medical Research Council and the Department of Health (grant MRC/G99001400). The first and the fourth author were funded by the Medical Research Council grant US UC A030 0031.

References

Aalen, O.O., Farewell, V.T., De Angelis, D., Day, N.E., and Gill, O.N. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine* 16: 2191–2210. doi:10.1002/(SICI)1097-0258(19971015)16:19<2191::AID-SIM645>3.0.CO;2-5.

Aguirre-Hernandez, R. and Farewell, V.T. (2002). A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine* 21: 1899–1911. doi:10.1002/sim.1152.

Brayne, C., McCracken, C., and Matthews, E.F. (2006). Cohort profile: the Medical Research Council Cognitive Function and Ageing Study (CFAS). *International Journal of Epidemiology* 35: 1140–1145. doi:10.1093/ije/dyl199.

Folstein, M.F., Folstein, S.E., and McHugh, P.R. (1975). Mini-mental state. A practical method for grading cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12: 189–198. doi:10.1016/0022-3956(75)90026-6.

Himmelspach, J. and Uhrmacher, A.M. (2007). Plug 'n simulate. In: *ANSS: Proceedings of the 40th Annual Simulation Symposium*. Washington, DC, USA: IEEE Computer Society: 137–143. doi:10.1016/B978-0-12-378638-8.00023-3.

Izmirlian, G., Brock, D., Ferrucci, L., and Phillips, C. (2000). Active Life Expectancy from Annual Follow-Up Data with Missing Responses. *Biometrics* 56: 244–248. doi:10.1111/j.0006-341X.2000.00244.x.

Jackson, C.H. (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software* 38. http://www.jstatsoft.org/v38/i08.

Kalbfleisch, J. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 80: 863–871. doi:10.1080/01621459.1985.10478195.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* 42: 855–865. doi:10.2307/2530699.

Kulkarni, V.G. (2011). *Introduction to Modeling and Analysis of Stochastic Systems*. New York: Springer.

Le Couteur, D.G., Doust, J., Creasey, H., and Brayne, C. (2013). Political drive to screen for pre-dementia: not evidence based and ignores the harms of diagnosis. *British Medical Journal* 347: f5125. doi:10.1136/bmj.f5125.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley.

Matthews, F.E., Miller, L.L., Brayne, C., and Jagger, C. (2006). Regional differences in multidimensional aspects of health: findings from the MRC cognitive function and ageing study. *BMC Public Health* 6. doi:10.1186/1471-2458-6-90.

Putter, H., Fiocco, M., and Geskus, R.B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 26: 2389–2430. doi:10.1002/sim.2712.

R Development Core Team (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Reuser, M., Willekens, F.J., and Bonneux, L. (2011). Higher education delays and shortens cognitive impairment. A multistate life table analysis of the US Health and Retirement Study. *European Journal of Epidemiology* 26: 395–403. doi:10.1007/s10654-011-9553-x.

Satten, G.A. and Longini, I.M. (1996). Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease (with discussion). *Applied Statistics* 45: 275–309. doi:10.2307/2986089.

Titman, A.C. and Sharples, L.D. (2010). Model diagnostics for multi-state models. *Statistical Methods in Medical Research* 19: 621–651. doi:10.1177/0962280209105541.

Van den Hout, A., Jagger, C., and Matthews, F.E. (2009). Estimating life expectancy in health and ill health by using a hidden Markov model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58: 449–465. doi:10.1111/j.1467-9876.2008.00659.x.

Van den Hout, A. and Matthews, F.E. (2008). Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model. *Statistics in Medicine* 27: 5440–5455. doi:10.1002/sim.3360.

Van den Hout, A. and Matthews, F.E. (2010). Estimating stroke-free and total life expectancy in the presence of non-ignorable missing values. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 173: 331–349. doi:10.1111/j.1467-985X.2009.00610.x.

Willekens, F. (2005). Biographic forecasting: Bridging the micro-macro gap in population forecasting. *New Zealand Population Review* 33: 77–124.

Willekens, F. (2006). Description of the microsimulation model (continuous time microsimulation). The Hague: NIDI: 11–111. (Technical report).

Zinn, S., Himmelspach, J., Gampe, J., and Uhrmacher, A.M. (2009). MIC-CORE: a tool for microsimulation. *Proceedings of the 2009 Winter Simulation Conference* 992–1002. doi:10.1109/WSC.2009.5429424.