A Bayesian Approach to Multi-State Hidden Markov Models: Application to Dementia Progression

Jonathan P Williams[†‡], Curtis B Storlie[†], Terry M Therneau[†], Clifford R Jack Jr[†], Jan Hannig[‡]

[†]Mayo Clinic
[‡]University of North Carolina at Chapel Hill

Abstract

People are living longer than ever before, and with this arise new complications and challenges for humanity. Among the most pressing of these challenges is to understand the role of aging in the development of dementia. This paper is motivated by the Mayo Clinic Study of Aging data for 4742 subjects since 2004, and how it can be used to draw inference on the role of aging in the development of dementia. We construct a hidden Markov model (HMM) to represent progression of dementia from states associated with the buildup of amyloid plaque in the brain, and the loss of cortical thickness. A hierarchical Bayesian approach is taken to estimate the parameters of the HMM with a truly time-inhomogeneous infinitesimal generator matrix, and response functions of the continuous-valued biomarker measurements are cutoff point agnostic. A Bayesian approach with these features could be useful in many disease progression models. Additionally, an approach is illustrated for correcting a common bias in delayed enrollment studies, in which some or all subjects are not observed at baseline. Standard software is incapable of accounting for this critical feature, so code to perform the estimation of the model described below is made available online.

Corresponding Author: Jonathan Williams, `williams.jonathan1@mayo.edu`

1   Introduction

People are living longer and with this arise new complications and challenges for humanity. Among the most pressing of these challenges is to understand the role of aging in the development of dementia. The Mayo Clinic Study of Aging (MCSA) is a large research initiative with the goal to understand the natural history of dementia and particularly Alzheimer's Disease. This study has enrolled a large age/sex stratified random sample of Olmsted County, Minnesota. Subjects are followed forward approximately every 15 months, and clinical visits collect information on three major aspects of the disease: amyloid burden in the brain (PIB scans), neuro-degeneration (MRI structural scans), and clinical status (MMSE scores, clinical diagnosis, etc.), along with baseline data (age, sex, clinical and genetic markers). The MCSA study began in 2004, currently has 4742 subjects, and is still ongoing (at the time of this writing) and enrolling new subjects.

This paper is motivated entirely by the MCSA and how the data can be used to draw inference on the role of aging in the development of dementia. The goal is to create a model of progression to dementia which can accommodate: (1) A wide variation in age (the dominant variable under consideration), (2) Significant fluctuation in the time between subject visits, (3) Different amount of information available for each subject (e.g., missing visits or PIB and/or MRI scans), and (4) Subject specific covariates.

An ideal solution would be one in which the rates of progression to various cognitive states leading to dementia could be quantified precisely, and even more basically, that various pre-dementia cognitive states could be identified systematically. "Does gender play a significant role in rates of progression?", or "Through which cognitive states does age have the most pronounced effect on progression?", are some of the key questions of interest. Precise estimates of quantities such as expected time in a given state, probability of ever entering a given state, probability of spending more than a year in a given state, or the fraction of the population that will pass through a high amyloid burden state on

the way to dementia, are all of clinical interest. One natural approach is to consider a Markov model of progression. This idea is explored in Jack et al. (2016). What they found is that almost all rates are log-linear, and at age 50 nearly everyone is in state A−N− (i.e., low Amyloid burden and low cortical thickness loss burden which is state 1 in Figure 1) but that soon begins to change.

It is known in the medical community that large amounts of amyloid protein buildup, and significant neuro-degeneration are strongly associated with dementia. Accordingly, a simplistic formalization of the biology is to theorize a seven-state model to describe cognitive health in relation to dementia. Figure 1 illustrates such states, and depicts the allowed transitions with directional arrows. A notable feature of the state space is that an individual must be in a high neuro-degeneration burden state (i.e., N+) to develop dementia, but not necessarily in a high amyloid burden state (i.e., A+). In fact, the transition from A+N+ (state 4) to A+Dem (state 6) is identified as Alzheimer's Disease, a particular type of dementia. Isolating this Alzheimer's Disease transition is not possible using the previous state space of Jack et al. (2016). Amyloid build-up and neuro-degeneration each develop on a continuum, but the time spent in any intermediate states, not explicitly represented by the state space in Figure 1, is believed to be relatively short and thus ignorable in these data.

Time must be treated as continuous because patient visit times are irregular, and the underlying sequence of states visited for an individual is hidden by uncertainty. The states of high amyloid burden and neuro-degenerative burden are not precisely defined and are best treated as hidden. High amyloid burden means a build-up of amyloid plaques in the brain significant enough to effect pathways and lead to neuro-degeneration, but precise measurements of the extent of amyloid protein would require autopsy. High neuro-degeneration burden simply represents a state of loss of neurons and synapses denoted by atrophy of the cerebral cortex in Alzheimer's-sensitive areas. Accordingly, biomarker

measurements (from PIB and MRI scans) associated with amyloid buildup and neuro-degeneration, respectively, are used as emitted response variables (in a traditional hidden Markov model sense) to make inference on an underlying sequence of states visited.

There is extensive statistical literature on the estimation of various types of discrete-state space Markov models, with numerous applications. Broadly speaking, these models can be treated as continuous-time or discrete-time, with a known or hidden state space. See the following works for a discussion of continuous-time Markov models with a known (observed) discrete-state space (Kalbfleisch & Lawless 1985, Gentleman et al. 1994, Hanks et al. 2015, Titman & Sharples 2008).
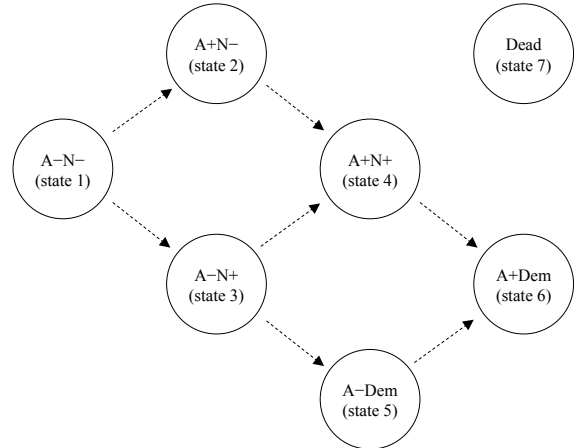

Figure 1: State space. A+ corresponds to high amyloid burden, and N+ corresponds to high neuro-degenerative burden. States 1-4 are all non-demented. Transitions to Dead are also allowed from all states.

Traditional applications of HMM in engineering and biological sequencing observed data over equally-spaced observations in time (Jackson et al. (2003)). With equally-spaced time points a discrete-time HMM is appropriate, and is less computationally expensive to implement because the transition probability matrix can be computed without having to compute a matrix exponential (as is the case in the continuous-time setting). See Rabiner (1989), Rydén (1994), Shirley et al. (2010), Scott (2002), Scott et al. (2005), and Robert et al. (2000) for implementations of discrete-time, discrete-state HMM. A good HMM tutorial for this setting is provided in Rabiner (1989).

In many settings, including much of medical research, subjects are often observed irregularly in time, which means that time must be treated as continuous. Accounts of continuous-time, discrete-state HMM are given by Lange & Minin (2013), Bureau et al. (2003), Jackson et al. (2003), Titman & Sharples (2008), Jack et al. (2016), and, within

a Bayesian framework, Zhao et al. (2016). Further, Satten & Longini (1996) is a very complete account of how to implement a continuous-time, discrete-state space HMM, and is recommended reading for anyone not familiar with the methodology.

The work of Jackson et al. (2003) uses an HMM to model the state misclassification error of a disease, and includes age as a covariate on the transition rates. However, they make a very restrictive assumption that the transition rates are constant between subject observation times. The work of Jack et al. (2016) is seemingly the first in the literature to estimate the transition rates as a function of age, in a truly continuous-time fashion for a multi-state model of dementia, however, Yu et al. (2010) also treated transition probabilities as a function of age, in a discrete-time fashion.

Wei & Kryscio (2016) take an alternative approach to fitting a discrete-state space model of dementia by using a semi-Markov model. In a semi-Markov model the transition intensity of each state depends on the amount of time spent in the current state (Wei & Kryscio (2016)). They discuss how to fit a semi-Markov model under the assumption of no misclassification error. See Titman & Sharples (2010) for additional methodology and application of semi-Markov models, as well as for a discussion on how to handle misclassification error within such models. Johnson & Willsky (2013) provides an account of estimating hidden semi-Markov models in a Bayesian framework. However, applications of semi-Markov models are more limited than Markov models. If it cannot be assumed that subjects make no more than one state transition between observations, then computing the likelihood function becomes infeasible with more than 3-4 states (Wei & Kryscio (2016), Titman & Sharples (2010)).

Most implementations of continuous-time, discrete-state HMMs estimate parameters in a maximum likelihood fashion. However, as mentioned in Jack et al. (2016) optimization becomes exceedingly difficult as more parameters are introduced in the model. Convergence time for standard methods may become impractical, and analyti-

4

cal gradient formulas for use in more efficient optimization procedures can become intractable. Additionally, it is often difficult and/or awkward to fit prior information into an optimization-based frequentist approach (e.g., via constrained optimization or penalty functions which require tuning parameters), and deriving confidence intervals becomes a challenge. Further, as the model becomes more complex to better capture reality in our application, prior information becomes necessary for practical identifiability of HMM parameters which makes Bayesian methodology the natural approach.

For the reasons above, we propose a hierarchical Bayesian framework with model estimation via Markov chain Monte Carlo (MCMC). Using MCMC for the estimation of complicated models requires creative proposal strategies, but is extremely flexible for a variety of model specifications. Moreover, credible regions become a convenient way to represent uncertainty. In the following we illustrate a general and effective framework for fitting continuous-time, discrete-state HMMs in a Bayesian manner.

This paper builds on the work of Jack et al. (2016) with more sophisticated modeling which allows for deeper insights. In Jack et al. (2016), the states similar to those in Figure 1 were actually defined by the biomarkers used to make inference on amyloid and neuro-degeneration. Specifically, hard cutoff points were chosen to distinguish between high/low burden biomarker states. However, hard cutoff points for discretizing continuous measurements of biological processes are practically and philosophically problematic, and have to be chosen more or less arbitrarily. Alternatively, we propose an approach which is cutoff point agnostic, and actually estimates the biomarker regions most associated with high/low burden states. Time is treated as continuous, and the infinitesimal generator matrix of the underlying Markov process is allowed to be truly time-inhomogeneous (as a function of an individual's age). Another contribution is that in addition to the effect of age, the effects of the covariates gender, number of years of education, and presence of an APOE-$\varepsilon 4$ allele on the infinitesimal transition rates are also

estimated. The importance of these variables has been well documented in the medical literature but their effect on aging has not been studied in this context (i.e., how they affect the transition rates of an HMM).

In addition to the new insights these features bring to the medical community, an approach is illustrated for correcting a common bias in *delayed enrollment* studies which has been overlooked in the literature. The term *delayed enrollment*, here, is used to describe a study with a given baseline (age 50 in the case of the MCSA) such that some or all subjects are *not* observed at baseline. We demonstrate empirically on a simple data set that the *delayed enrollment* can result in substantially biased estimates of parameters if it is not carefully addressed. We argue that the effects of this bias cannot be ignored, and existing software is not equipped to handle this feature. Finally, our work also provides a framework for estimating the strength and persistence of a separate *death bias* specific to death rates, which could be present in a study with or without *delayed enrollment*. Estimating multi-state models of this nature, where time is treated as continuous, is statistically and computationally non-trivial, and existing statistical software is incapable of handling many of the subtleties of our approach. Flexible software to fit the models described below is provided in a GitHub repository at `https://jonathanpw.github.io/software.html`.

The organization of the rest of the paper is as follows. Section 2 describes the methodology, the likelihood function, and the Bayesian computations. An empirical analysis of the *delayed enrollment bias* is presented in Section 3, followed by a simulation study of our Bayesian estimation procedure performed on synthetic data generated to resemble the MCSA data set. Finally, a detailed analysis of the MCSA data is presented in Section 4, accompanied by a discussion and interpretation of the biological findings. This paper also has accompanying supplementary material containing details of the MCMC algorithm and further MCSA analysis results.
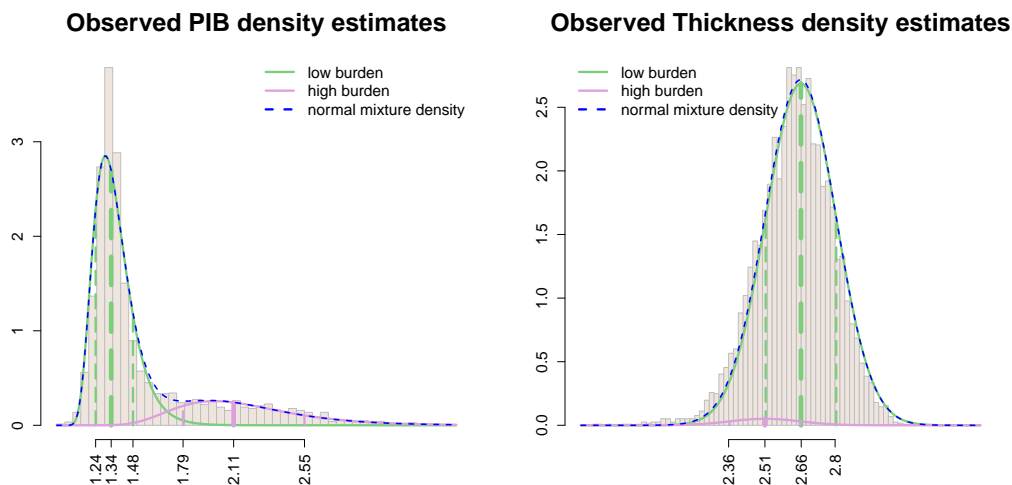
6

Figure 2: Observed response data for PIB which is a PET measure of amyloid, and (cortical) Thickness which is associated with neuro-degeneration. Note that the response densities for PIB correspond to the data transformation, $\log(\text{PIB} - 1)$. The component density estimates correspond to the posterior mean estimates from Section 4. The blue dashed lines represent the normal mixture density estimates.

2 Methodology

The specifics of the HMM constructed to make inference on the MCSA data are detailed in the subsections to follow, however, a brief overview is provided here. Amyloid PET (in this case Pittsburgh Compound B or PIB) and cortical thickness (Thickness) are assumed to be continuous outcomes associated with amyloid deposition and neuro-degeneration, respectively. States 2, 4, and 6 which correspond to increased amyloid burden will emit PIB values from a distribution corresponding to A+, while states 1, 3, and 5 will emit PIB from the distribution corresponding to $A-$, and similarly for neuro-degenerative burden (N+ or N−). The prior distributions for these response distribution parameters was chosen to correspond to biomarker values which are consistent with the medical community's most up-to-date understanding of the biology. Using a Gaussian distribution for $\log(\text{PIB} - 1)$ and for Thickness appears to be quite reasonable, and there is evidence that the error from the PIB scans follows more closely to a constant coefficient of variation than to a constant variance. Figure 2 displays histograms of the observed response data along with the respective normal mixture densities resulting from the posterior mean estimates from the full HMM model described below.

7

For subjects in the study ($\sim 50\%$) who were not chosen to undergo regular brain scans less information about their underlying states is available. To reduce the uncertainty, especially for these subjects, the Mini Mental State Exam (MMSE) is used as an additional response. The MMSE is a questionnaire-based test administered by a medical professional to assess cognitive impairment on an integer scale out of 30 points (Xu et al. 2015). This response is assumed a separate Gaussian emission distribution for each of the first six states (deceased subjects do not emit cognitive test scores). Figure 3 overlays the estimated six-component normal mixture density function on top of a histogram of the observed MMSE scores from the study subjects' visits.
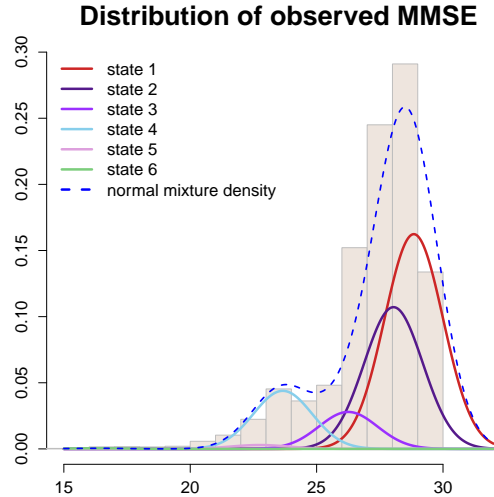

Figure 3: Observed response data for MMSE test scores associated with the six non-death states in the state space. The component density estimates here correspond to the posterior mean estimates from Section 4. The blue dashed line represents the normal mixture density estimate.

Finally, at the time of observation all subjects are determined to be either cognitively unimpaired, or to be demented. This represents a substantial amount of information for making inference on a subject's underlying state sequence. However, diagnosing dementia is not an exact science, and so the observed label is not without error. Accordingly, a simple misclassification response model is used to allow for a probability of a dementia diagnosis given the underlying state is a dementia state (i.e., states 5 or 6), and given it is not a dementia state. Death is the only state in the state space which is known without error, and in fact the exact time of death is known, as well.

2.1 Continuous-time transition probabilities

This section serves to specify the hidden Markov model in the context of the state space illustrated in Figure 1. For $r, s \in \{1, 2, 3, 4, 5, 6, 7\}$ and $h, t \geq 0$, the probability of

transitioning from state r at time h to state s at time $h + t$ is denoted by $P_{r,s}(h,t) = P(S(h+t) = s|S(h) = r)$. Assuming these probabilities are differentiable functions in t and that the Markov process is time-homogeneous, it can be shown that they satisfy the Kolmogorov forward equations (Karlin & Taylor (1981)),

$$\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}, \tag{1}$$

where $\boldsymbol{Q}$ is called the transition rate matrix, and $\boldsymbol{P}(t)$ is the matrix with components

$$P_{rs}(t) := P_{rs}(h = 0, t). \tag{2}$$

Note that h can be taken to be 0 in (2) because the probabilities are assumed for now to be time-homogeneous. The off-diagonal components of $\boldsymbol{Q}$ are interpreted as the change in transition probabilities for an infinitesimal amount of time into the future, i.e.,

$$q_{rs} = \lim_{t \downarrow 0} \frac{P(S(t) = s|S(0) = r)}{t}, \quad r \neq s, \tag{3}$$

with diagonal elements $q_{rr} = -\sum_{s \neq r} q_{rs}$.

The forward equations in (1) have the matrix exponential solution, $\boldsymbol{P}(t) = e^{t\boldsymbol{Q}}$. However, as discussed in Section 1, the transition rates will be expressed as a function of a subject's age at the time of transition. That is, $\boldsymbol{Q} = \boldsymbol{Q}(t)$ which violates the time-homogeneity of the Markov process. A simple work around is to discretize the effect of age and assume that the transition rates only change when a subject's integer age changes. Doing so implies that subjects' transition rates, $\boldsymbol{Q}$, remain constant between birthdays and yields

$$\boldsymbol{P}(h,t) = e^{(\lfloor h+1 \rfloor - h)\boldsymbol{Q}(\lfloor h \rfloor)} \cdot e^{\boldsymbol{Q}(\lfloor h+1 \rfloor)} \cdots e^{\boldsymbol{Q}(\lfloor h+t \rfloor - 1)} \cdot e^{(h+t-\lfloor h+t \rfloor)\boldsymbol{Q}(\lfloor h+t \rfloor)}, \tag{4}$$

for $\lfloor h \rfloor \neq \lfloor h + t \rfloor$, where h represents the subject's current age, t is the time (in years) into the future, and $\lfloor \cdot \rfloor$ is the floor function.

Observe that with expression (4) all transition probabilities can be computed, as long as the components of $\boldsymbol{Q}(t)$ are specified. As mentioned above, the transition rates will be modeled as a function of age, gender, presence of an APOE-$\varepsilon 4$ allele, and number of years

of education. Specifically, denoting each of the 13 nonzero transition rates illustrated in Figure 1 by q_l for $l \in \{1, \ldots, 13\}$,

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{age} + \beta_2^{(l)} \cdot \text{male} + \beta_3^{(l)} \cdot \text{educ} + \beta_4^{(l)} \cdot \text{apoe4}, \qquad (5)$$

where,

$$\boldsymbol{Q} = \begin{pmatrix} -q_1 - q_2 - q_3 & q_1 & q_2 & 0 & 0 & 0 & q_3 \\ 0 & -q_4 - q_5 & 0 & q_4 & 0 & 0 & q_5 \\ 0 & 0 & -q_6 - q_7 - q_8 & q_6 & q_7 & 0 & q_8 \\ 0 & 0 & 0 & -q_9 - q_{10} & 0 & q_9 & q_{10} \\ 0 & 0 & 0 & 0 & -q_{11} - q_{12} & q_{11} & q_{12} \\ 0 & 0 & 0 & 0 & 0 & -q_{13} & q_{13} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Overall death rates over age 50 are log-linear (see Figure 4) which makes the log-linear function of age a natural starting place. This functional form was also argued in Jack et al. (2016) to be reasonable for all of the rates except the rate from A−N− (state 1) to A+N− (state 2). They compared log-linear rate to that obtained from log-cubic splines. We came to the same conclusion and accordingly, a cubic spline is used for estimating only the rate of transition from state $1 \rightarrow 2$,

Log death rate

Figure 4: Annualized natural logarithm of Minnesota overall population death rates. Solid line corresponds to female, and dashed line corresponds to male.

with knots at ages 55, 65, 75, 90, and boundary knots at 50 and 120.

2.2 Likelihood function

If the states were known for each subject at each observation, then the contribution to the likelihood function from each subject consists of a product of matrix exponentials, i.e., apply (4) to get the probability of being in a given state for each observation time. However, the underlying state sequences are not observed. Within an HMM, rather, responses emitted from the underlying process (conditional on the true state of the process
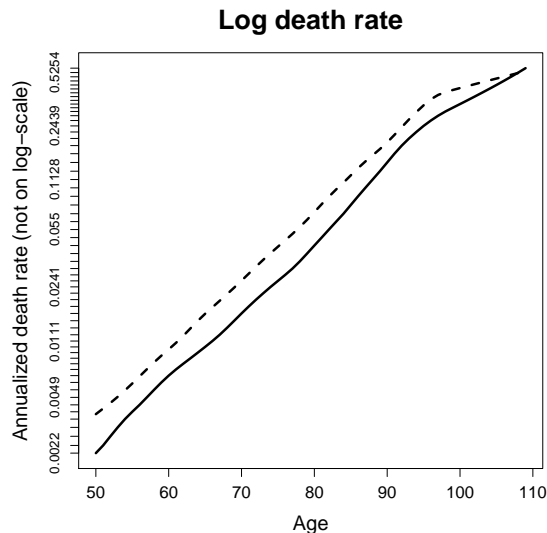
at a given point in time) are used to inform of the underlying state. In this application, there are four emitted responses (i) $\log(\text{PIB} - 1)$, (ii) Thickness, (iii) MMSE (see Figures 2 and 3), and (iv) dementia diagnosis (binary). Denote the observations for each of these four responses by $\boldsymbol{y}_{i,k} = [y_{i,k,1}, y_{i,k,2}, y_{i,k,3}, y_{i,k,4}]'$, respectively, for the ith subject's k^{th} clinical visit (observation). Further, the ith subject has n_i clinical visits, and an (unknown) underlying sequence of states $\{s_{i,k}\}$.

Thus, the likelihood contribution from the ith subject at the kth visit can be expressed as

$$f(\boldsymbol{y}_{i,k}) = \sum_{s_{i,k}=1}^{7} f(\boldsymbol{y}_{i,k}, s_{i,k}) = \sum_{s_{i,k}=1}^{7} P(s_{i,k}) \cdot f(\boldsymbol{y}_{i,k}|s_{i,k}),$$

where the sum is taken over all possible states (since the true state sequence is unknown). If a given response $y_{i,k,j}$ is missing (e.g., missing PIB scan), then the missing value is integrated out of the likelihood (i.e., the response density of the missing value contributes a 1 to the likelihood function).

Making the standard assumption that the responses are conditionally independent given an underlying state sequence $s_{i,1}, \ldots, s_{i,n_i}$, and applying the Markov property for the state sequence gives,

$$f(\boldsymbol{y}_{i,1}, \ldots, \boldsymbol{y}_{i,n_i}) = \sum P(s_{i,1}) P(s_{i,2}|s_{i,1}) \cdots P(s_{i,n_i}|s_{i,n_i-1}) \cdot \prod_{k=1}^{n_i} f(\boldsymbol{y}_{i,k}|s_{i,k})$$

$$= \sum_{s_{i,1}=1}^{7} P(s_{i,1}) f(\boldsymbol{y}_{i,1}|s_{i,1}) \cdot \sum_{s_{i,2}=1}^{7} P(s_{i,2}|s_{i,1}) f(\boldsymbol{y}_{i,2}|s_{i,2}) \cdots \sum_{s_{i,n_i}=1}^{7} P(s_{i,n_i}|s_{i,n_i-1}) f(\boldsymbol{y}_{i,n_i}|s_{i,n_i})$$

$$= \boldsymbol{\pi}' \boldsymbol{D}_{(i,1)} \cdot \boldsymbol{P}(t_{i,1}, t_{i,2} - t_{i,1}) \boldsymbol{D}_{(i,2)} \cdots \boldsymbol{P}(t_{i,n_i-1}, t_{i,n_i} - t_{i,n_i-1}) \boldsymbol{D}_{(i,n_i)} \cdot \mathbf{1}, \tag{6}$$

where $\boldsymbol{\pi}$ is the initial state probability vector, $\boldsymbol{D}_{(i,k)}$ is a diagonal matrix with diagonal components $f(\boldsymbol{y}_{i,k}|s_{i,k})$ for each $s_{i,k} \in \{1, \ldots, 7\}$ in the state space, $\boldsymbol{P}(t_{i,k-1}, t_{i,k} - t_{i,k-1})$ is the transition probability matrix given in (4) with $t_{i,k}$ denoting the (continuous) age of subject i at visit k, and $\mathbf{1}$ is a column vector of ones. There are three subtle, but important features of the MCSA data that need to be addressed which result in slight modifications of this likelihood function.

11

The first feature is that the first observation time for the ith subject, $t_{i,1}$, is not necessarily equal to the baseline age of 50 years old. In this case, the probability of transitioning from the (unknown) underlying state at baseline to the (unknown) underlying state at the first observation must be accounted for in the likelihood function. That is, the initial state probability vector, $\boldsymbol{\pi}$ in (6), is replaced with,

$$\boldsymbol{\pi}(t_{i,1}) = [v_1(t_{i,1}) \,,\, v_2(t_{i,1}) \,,\, v_3(t_{i,1}) \,,\, v_4(t_{i,1}) \,,\, 0 \,,\, 0 \,,\, 0]' \frac{1}{\sum_{j=1}^{4} v_j(t_{i,1})}, \qquad (7)$$

where $v(t_{i,1})' = \boldsymbol{\pi}_0' P(50, t_{i,1} - 50)$, and $\boldsymbol{\pi}_0$ is the initial state probability vector for a subject at baseline. The last three components are set equal to zero due to the fact that demented and dead subjects are not enrolled into the study. If the initial probabilities from baseline to enrollment age were not conditioned on the underlying states being non-demented and non-dead, then the transition rates (especially those to dementia and death) will exhibit strong downward bias (with respect to the true population parameter values) due to the enrollment scheme. We refer to this bias as the *delayed enrollment bias*, and it will be illustrated empirically in Section 3.1. For example, when a subject is enrolled in the study, e.g., at age 71, although 21 years worth of clinical visits were not observed, the response variables may indicate that this subject is likely to be in state 4 which informs all of the transition rates, each to a greater or lesser extent. Relevant standard software such as the `msm` package in R (Jackson 2011) does allow for specification of a common baseline in a *delayed enrollment* study (after manually adding censored observations in the data set at baseline), but unfortunately does not offer the ability to perform conditioning (rescaling) on the initial state probability vector, as in (7).

The second feature is that death is observed without error, and time of death is known exactly, so the likelihood must be modified to account for this more precise information (Satten & Longini 1996). Suppose that subject i transitions to death at time t_{i,n_i}. The final term in the subject's contribution to the likelihood can be re-expressed as follows. For a state s_{i,n_i-1} at time t_{i,n_i-1}, and $\varepsilon \in (0, t_{i,n_i} - t_{i,n_i-1})$, let,

$$B(\varepsilon) := \{s_{i,n_i}(t_{i,n_i}) = 7, s(t_{i,n_i} - \varepsilon) < 7\}.$$

Then $B := \bigcap_\varepsilon B(\varepsilon)$ is the event that the ith subject dies precisely at time t_{i,n_i}. Further,

$$P(B(\varepsilon)|s_{i,n_i-1}) = \sum_{s(t_{i,n_i}-\varepsilon)=1}^{6} P\big(s(t_{i,n_i} - \varepsilon)|s_{i,n_i-1}(t_{i,n_i-1})\big) \cdot P\big(s_{i,n_i}(t_{i,n_i}) = 7|s(t_{i,n_i} - \varepsilon)\big).$$

Thus, dividing both sides by ε and taking the limit as $\varepsilon \to 0$ gives a likelihood function value of

$$\sum_{s(t_{i,n_i})=1}^{6} P(s(t_{i,n_i})|s_{i,n_i-1}) \cdot Q_{s,7}(\lfloor t_{i,n_i} \rfloor) \tag{8}$$

evaluated at the event $\{B|s_{i,n_i-1}\}$. The quantity in (8) is interpreted as the average probability of being in each of the first six states the instant prior to death, each weighted by the probability of transitioning to death at the next instant (given by the instantaneous transition rates $Q_{s,7}(\lfloor t_{i,n_i} \rfloor)$). Note that the response functions are not needed/defined when $s_{i,n_i} = 7$ because death is assumed observed without error.

The third feature, a common feature of many human population based studies (*delayed enrollment* or not), is the following *death bias*. In addition to not representing those members of the population who would have enrolled if they were not already dead (which is tied into the *delayed enrollment bias*), it is reasonable to assume that people are much less likely to enroll in a study if they are very sick and/or dying. As a result, the death rate for the sub-population of individuals who are most likely to enroll is probably smaller than the overall population death rate, for at least the first few years after enrollment into the study. Consistent with the difference between this healthier sub-population and the overall population, this phenomenon will lead to a reduced estimate of the death rate and higher likelihoods of other paths in the state space. With respect to the true parameter values of the overall population, this represents a bias which we term the *death bias*. Note that the *death bias* is in fact distinct from the *delayed enrollment bias*.

Our proposed approach to correct for the *death bias* is to explicitly estimate the *death bias* on the non-dementia to death rate. This can be done in a linear fashion by including

two additional parameters in the log-linear rate equation (5). The first, say $c \leq 0$, will be for estimating the baseline effect of the *death bias*, and the second, say $d > 0$, will be for estimating the linear slope at which the *death bias* vanishes for every integer year in the study (since the time effect is discretized annually). That is, equation (5) for only the non-dementia to death rate ($l \in \{3, 5, 8, 10\}$) becomes,

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{age} + \beta_2^{(l)} \cdot \text{male} + \beta_3^{(l)} \cdot \text{educ} + \beta_4^{(l)} \cdot \text{apoe4} + g(\text{iyears}), \quad (9)$$

where 'iyears' is integer years enrolled, and $g(\text{iyears}) := \min\{c + d \cdot \text{iyears}, 0\}$. The *death bias* term is only allowed to decrease the non-dementia to death rate. Further, the smallest root of g is the duration for which the *death bias* persisted in the study. The coefficients in equation (9) become identifiable due to strong prior information available for the overall population death rate.

Finally, in order to complete the specification of the likelihood, the form of the response density $f(\boldsymbol{y}_{i,k})$ and the baseline state probability vector $\boldsymbol{\pi}_0$ must be specified. It is assumed that the four responses, $y_{i,k,1} = \log(\text{PIB} - 1)$, $y_{i,k,2} = \text{Thickness}$, $y_{i,k,3} = \text{MMSE}$, and $y_{i,k,4} = \text{Dementia}$, are conditionally independent given state $s_{i,k}$ and subject specific covariates. As illustrated in Figure 2, the transformed PIB measurements are assumed to be generated according to two normal random variables with different means. One of the means, say μ_{A-}, corresponds to the distribution of transformed PIB measurements for an individual in a low amyloid burden state, and the other, say μ_{A+}, corresponds to a high amyloid burden state. Specifically,

$$f(y_{i,k,1}|s_{i,k}) = \text{N}(y_{i,k,1}|\mu_{A-}, \sigma_{\text{pib}}) \cdot I_{\left\{s_{i,k} \in \{1,3,5\}\right\}} + \text{N}(y_{i,k,1}|\mu_{A+}, \sigma_{\text{pib}}) \cdot I_{\left\{s_{i,k} \in \{2,4,6\}\right\}}, \quad (10)$$

where I_A is the indicator function equal to 1 if A and 0 otherwise. The variance of both Gaussians are assumed to be equal to aid in identifiability of the two groups. The density function for Thickness and MMSE are defined analogously.

For the Thickness response variable,

$$f(y_{i,k,2}|s_{i,k}) = \mathrm{N}(y_{i,k,2}|\mu_{N-}, \sigma_{\text{thick}}) \cdot I_{\{s_{i,k}\in\{1,2\}\}} + \mathrm{N}(y_{i,k,2}|\mu_{N+}, \sigma_{\text{thick}}) \cdot I_{\{s_{i,k}\in\{3,4,5,6\}\}}, \quad (11)$$

and for the MMSE response variable,

$$f(y_{i,k,3}|s_{i,k}) = \mathrm{N}(y_{i,k,3}|\mu, \sigma_{\text{mmse}}), \quad (12)$$

with

$$\mu = \sum_{j=1}^{6} \alpha_j \cdot I_{\{s_{i,k}=j\}} + \alpha_7 \cdot \text{age} + \alpha_8 \cdot \text{male} + \alpha_9 \cdot \text{educ} + \alpha_{10} \cdot \text{apoe4} + \alpha_{11} \cdot \text{ntests}.$$

The first four covariates are the same as those in Section 2.1, and 'ntests' is the number of times a subject has taken the MMSE by a given clinical visit. It is observed in the medical practice that scores on the MMSE may improve as an individual becomes familiar with the exam, and so the 'ntests' covariate is included to control for this effect.

The probability mass function for misdiagnosis of dementia is

$$P(y_{i,k,4}|s_{i,k}) = \text{Bernoulli}(y_{i,k,4}|p_0) \cdot I_{\{s_{i,k}\in\{1,2,3,4\}\}} + \text{Bernoulli}(1-y_{i,k,4}|p_1) \cdot I_{\{s_{i,k}\in\{5,6\}\}}, \quad (13)$$

where $y_{i,k,4} = 1$ if the subject was diagnosed with dementia, and $y_{i,k,4} = 0$ if not. Accordingly, p_0 and p_1 are misclassification probabilities, with p_0 the probability of an incorrect diagnosis of dementia, and p_1 the probability of an incorrect non-diagnosis of dementia.

Lastly, the baseline state probability vector is $\boldsymbol{\pi}_0 = [\pi_{0,1}, \pi_{0,2}, \pi_{0,3}, \pi_{0,4}, 0, 0, 0]'$ where $\pi_{0,j} = P(s_{i,1} = j)$ for $j \in \{1, 2, 3, 4\}$, for all subjects, i. As the MCSA did not enroll demented or deceased individuals, subjects would necessarily have been in states 1-4 at baseline. Accordingly, $\sum_{j=1}^{4} \pi_{0,j} = 1$, and the last three components of $\boldsymbol{\pi}_0$ must be zero.

2.3 Bayesian Computation

The HMM above has been specified independent of a choice of estimation procedure. A hierarchical Bayesian model representation will be specified below, along with important details about the MCMC estimation. There are certain advantages to implementing Bayesian procedures for an HMM, particularly for one with many nuances and features. It is natural to use prior information for the parameters in this problem, e.g., the known

overall population death rate, and dementia diagnosis misclassification rates from previous studies based on autopsy data. In a frequentist approach these parameters would need to be simply fixed at the prior beliefs, or coerced with penalty functions which add to the optimization complexity and necessitate the choice of tuning parameters. It is also relatively simple in a Bayesian framework to place constraints on the HMM parameters through prior specifications, and credible regions provide a natural means for which to represent uncertainty. Obtaining an MLE and computing confidence intervals in frequentist based approach becomes much more difficult for complex HMMs, particularly under the presence of constraints with estimates on or near the boundaries.

The standard R package for estimating an HMM with arbitrary continuous-time observations is `msm` (Jackson (2011)). While `msm` is a well written and powerful package, it does not offer Bayesian estimation options, and it cannot correct for the biases on the transition rates that arise from a *delayed enrollment* sample. The implications of this latter issue will be documented in Section 3.1.

The code for the Bayesian implementations used here is part of a broader project to write a comprehensive R package with both frequentist and Bayesian capabilities for estimating an HMM, and with the ability to control for *delayed enrollment bias*. Thus, for ease of generalizability, a multivariate normal prior is placed on the vector of HMM parameters, and multivariate normal proposals are used. For parameters which are non-negative, such as the variance parameters for the response functions, the Gaussian priors are placed (and Gaussian proposals are made) on the natural logarithm of the parameter. For parameters which are constrained to be between 0 and 1, such as the dementia misclassification parameters and the initial state probability parameters, the logit transformation of the parameters is used. For example, the initial state probability vector is re-expressed as

$$\boldsymbol{\pi}_0 = \left[1 \,,\, e^{\xi_1} \,,\, e^{\xi_2} \,,\, e^{\xi_3} \,,\, 0 \,,\, 0 \,,\, 0 \right]' \frac{1}{1 + e^{\xi_1} + e^{\xi_2} + e^{\xi_3}} \tag{14}$$

16

where ξ_1, ξ_2, and ξ_3 are assigned Gaussian priors and proposals.

Due to the complexity of the posterior density function pure Gibbs sampling is not possible, and a Metropolis-within-Gibbs sampling approach is used. For more efficient mixing of the MCMC sampling, the HMM parameters are updated in groups, and the proposal scheme is adaptive during the burnin period. The update groups are chosen based on parameters which exhibit strong correlations, and the number of groups is chosen to strike a balance between good mixing and computation time for each iteration. Unlike the adaptive proposal scheme, the number and composition of the groups is chosen prior to the MCMC implementation.

The logic behind the adaptive proposal scheme is most easily illustrated with an example. In order to update a subset of the parameter vector, say $\boldsymbol{\theta}$, the proposal distribution is specified as $N(\boldsymbol{\theta}^{(t)}, \tau \cdot \Sigma)$, where $\boldsymbol{\theta}^{(t)}$ is the current parameter vector in the MCMC chain, τ is a scale parameter, and Σ is the empirical covariance matrix for some specified number of previous steps in the MCMC chain. A desirable acceptance ratio is targeted by scaling down τ when the acceptance ratio gets too small (to propose smaller MCMC steps), and by scaling up τ when the acceptance ratio gets too large (to propose larger MCMC steps). The empirical covariance, Σ, is updated at each step, but observes a limited history of the MCMC chain so as to forget misrepresentative parameter vectors which the MCMC algorithm visits early on during the adaptation period as it settles in closer to the posterior distribution. At a pre-specified number of steps, τ and Σ are held fixed at their most recent values and a traditional MCMC sampling with fixed tuning parameters is conducted.

The specific values for the prior means and variances are provided in Table 2 in Section 3.2, where the data from the MCSA is discussed in more detail. The specific HMM parameter constraints are enumerated there, as well, along with a discussion of other methodology designed particularly for this data set.

17

		Observed state		
		no CAV	mild	severe	death
	no CAV	$1 - p_1$	p_1	0	0
True state	mild	p_2	$1 - p_2 - p_3$	p_3	0
	severe	0	p_4	$1 - p_4$	0
	death	0	0	0	1

Table 1: Misclassification response function.

3 Simulation results

Before proceeding directly to estimating the HMM parameters for the MCSA data, the following two subsections serve to validate the approach presented in this work. First, the potential bias resulting from a *delayed enrollment* study, where some or all subjects are not observed at baseline, is illustrated on synthetic data in Section 3.1. In Section 3.2 the MCSA data is replicated with simulated data sets, and the estimation procedure described in the previous sections is implemented on these synthetic data to assess its performance in a setting that closely resembles the intended application.

3.1 The effect of a *delayed enrollment* study

A toy example is analyzed in order to illustrate the bias which can result from a *delayed enrollment* study. The cardiac allograft vasculopathy (CAV) data set from the `msm` package is used for this purpose. The data are from a study of the progression of CAV which is a common cause of death after heart transplant (Sharples et al. 2003). The state space as described in Jackson (2011) includes four states labeled 'no CAV' (state 1), 'mild/moderate CAV' (state 2), 'severe CAV' (state 3), and 'death' (state 4), and forward-only transitions are assumed (patients only get worse). Observed remissions in the state of a patient is considered a result of misclassification error.

In this simple data set, the only response variable is the observed state at each visit which is assumed to follow a categorical distribution. In particular, given a patient is in state 1, 2, or 3 there is a nonzero probability of observing an adjacent state. Additionally, as is the case for the MCSA, death and time of death are known without error. Formally,

given a true state (row) the response probability mass function takes the form given in Table 1, where the probabilities p_1, p_2, p_3, and p_4 are each interpreted as the probability a of particular misclassification. Note that the rows must sum to one.

Time is treated as continuous and discretized annually. Integer years since heart transplant, and gender are included as covariates on the transitions rates for the CAV data,

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{iyears} + \beta_2^{(l)} \cdot \text{sex}, \tag{15}$$

where,

$$\boldsymbol{Q} = \begin{pmatrix} -q_1 - q_2 & q_1 & 0 & q_2 \\ 0 & -q_3 - q_4 & q_3 & q_4 \\ 0 & 0 & -q_5 & q_5 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

To replicate the CAV data in a truth known simulation, first the HMM parameter estimates were obtained using the `msm` package on the original CAV data. As suggested in Jackson (2011), the 'BFGS' quasi-Newton optimization algorithm was specified to estimate the HMM parameters with the 'msm' function from `msm`. The parameter estimates from the CAV data set are then used as the 'true' values of the parameters to replicate the data set. However, certain features of the CAV data are adjusted to better facilitate the analysis here.

In the CAV data, all patients are observed at baseline, and in fact at baseline all patients are in state 1 because CAV does not develop immediately. However, important features of the MCSA are that not all subjects are in state 1 at baseline, and virtually none of the subjects are observed at baseline. Thus, each patient in the simulated CAV data is generated with a baseline state according to the following distribution,

$s_{i,1}$	no CAV	mild	severe	death
$P(s_{i,1})$	0.95	0.04	0.01	0

Sex of the patient is randomly chosen with probability consistent with that of the actual CAV data proportions of each sex. The true state sequence for each patient is generated by first computing the row of the infinitesimal generator matrix $\boldsymbol{Q}(\text{iyears}, \text{sex})$ corresponding to the patient's current state, and then sampling from exponential dis-

tributions (waiting times), one for each nonzero rate parameter in the row vector. If the minimum of these sampled waiting times is smaller than an 'infinitesimal' amount of time then the patient transitions to the state corresponding to the minimum waiting time, or else does not transition at that 'instant'. For simulating the data, an 'instant' or 'infinitesimal' amount of time is discretized to be 1/365. This procedure for generating the state sequence of each patient is carried out until the patient transitions to state 4 (death). These state sequences are taken to be the true underlying sequence of states.

Once the true sequence of states is generated for each patient, inter-observation times are sampled from the non-death inter-observation times from the actual CAV data set. A sequence of observed visit times is then generated in this fashion, and the true state for the patient (in addition to the covariate values) corresponding to each visit time is recorded from the underlying state sequence. The one exception is that transitions to death are recorded using the exact time of transition for death. To introduce misclassification error in the underlying state for each visit, observed states are generated from the misclassification response distribution (Table 1), with the parameters estimated via msm.

A sample size of 2000 was generated for 100 simulated data sets. Figure 5 shows estimates of the log-rate intercept coefficients from equation (15) using the posterior mean from the proposed Bayesian approach, and the MLE obtained via msm. The box plots are over the 100 estimates of each parameter and demonstrate that in a simple idealized setting the Bayesian and MLE estimates are very similar.

In the synthetic data example above, all patients are observed at baseline; this is the type of study for which the msm package was designed. To illustrate the bias that ensues when subjects are not observed at baseline, the 100 data sets are generated once more from the same random generator seeds. However, instead of beginning with the initial observations at baseline (zero years after heart transplant), the time of initial observation is generated, with probability 0.75, from a Gaussian distribution with a mean of 5 integer
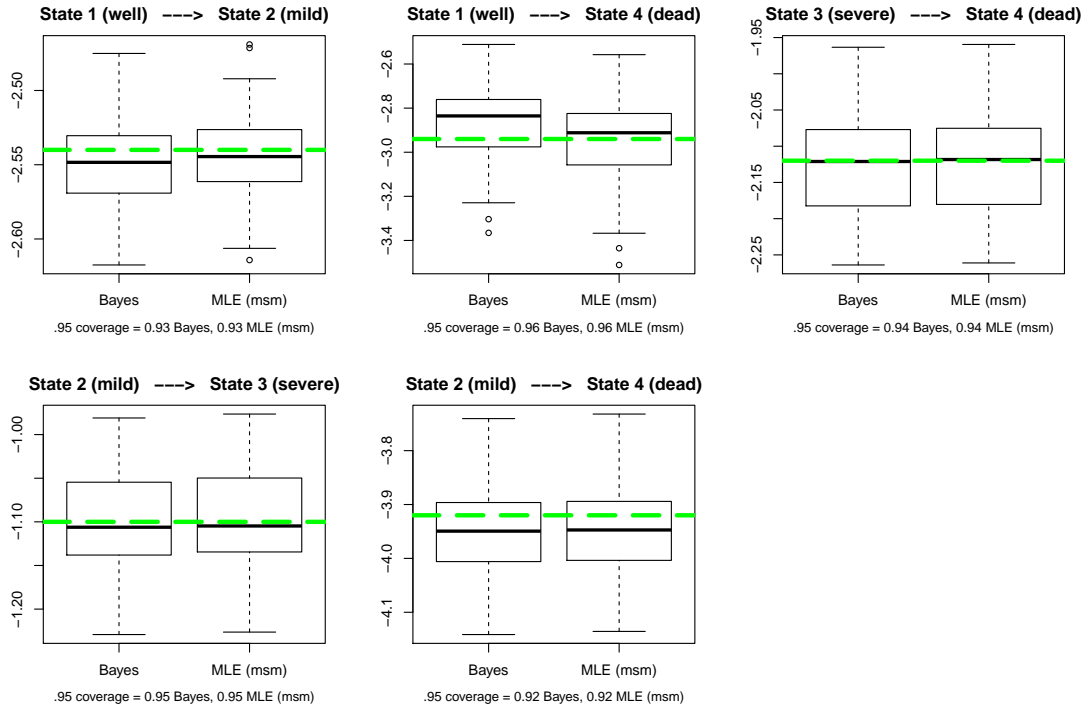
20

Figure 5: Intercept coefficient estimates for a traditional study in which all subjects are enrolled at a common baseline time. 'Bayes' corresponds to the Bayesian posterior mean estimates, and 'MLE (msm)' corresponds to maximum likelihood estimates computed from the 'msm' function. Green dashed lines represent the true values. Coverage is the proportion of .95 probability credible intervals (confidence intervals for the MLE) which contain the true parameter value.

years and a standard deviation of 1, i.e., the initial observations remain at baseline with probability 0.25. Additionally, if a patient transitions to death prior to the generated initial observation time, then the patient is not included in this *delayed enrollment* study. This is a critical point because it is the cause of the bias: the *delayed enrollment* study is less likely to include patients with immediate adverse reactions to heart transplant. If the sample were truly representative, then all 2000 patients would be represented. However, studies typically only sample from the living. The average sample size of the 100 synthetic data sets for the *delayed enrollment* study is 1686.

In Section 2.2 the procedure for accounting for the *delayed enrollment bias* in the likelihood function was described. It amounts to evaluating the transition rate matrix, $\boldsymbol{Q}$, (here, annually) from baseline to initial observation, and then computing the conditional probabilities for the initial states of enrollment. This conditioning feature is not
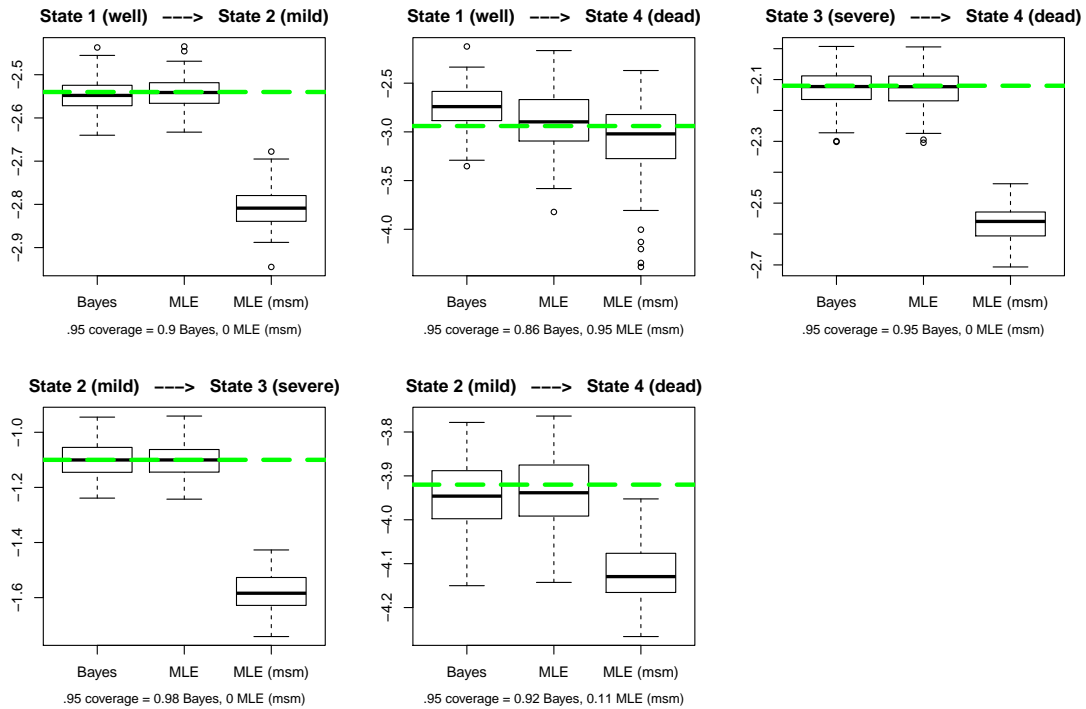
Figure 6: Intercept coefficient estimates for a *delayed enrollment* study. 'Bayes' corresponds to the Bayesian posterior mean estimates, MLE corresponds to the MLE obtained via optimizing the likelihood function from Section 2.2 using the 'optim' function in R, and 'MLE (msm)' corresponds to maximum likelihood estimates computed from the 'msm' function (assuming all first observations happen at baseline). Green dashed lines represent the true values. Coverage is the proportion of .95 probability credible intervals (confidence intervals for the MLE) which contain the true parameter value.

available in the `msm` package which was not designed for a *delayed enrollment* study. Figure 6 illustrates the effect of ignoring this feature and estimating as though enrollment is *not* conditional on being alive, i.e., MLE(msm). The estimates from Bayes and MLE are analogous to those in Figure 5, however, they do explicitly account for the initial probability according to (7). Without accounting for the *delayed enrollment* effect in the likelihood function certain estimates are significantly biased downward, suggesting slower rates of transition. This is because, of the 2000 patients, those which happened to transition quickly through the state space are no longer observed in the data set. The biases become more extreme as fewer patients are observed at baseline; recall that about 25% of the patients are still observed at baseline.

The bias also filters into other HMM parameter estimates; see the Supplementary Material for the full results of these two simulation setups. The objective of this synthetic

delayed enrollment study example is to demonstrate one of the crucial reasons why the analysis of the MCSA data requires methodological developments which are not readily available in standard software. The `msm` package, as well as other similar software, are not flawed, rather they were simply not designed for this type of application.

3.2 Synthetic Data Based on Mayo Clinic Study of Aging

Section 3.1 presented the results of the proposed estimation procedure in a simple idealized simulation example. In this section a more realistic simulation is presented which is intended to closely replicate the MCSA data generating process with respect to sample size, the frequency of clinical visits, and the proportion of biomarker measurements available. The three objectives are to (i) provide evidence that the synthetic data reasonably resembles the real data in an effort to verify that the data generating mechanisms of the real data are sufficiently understood, (ii) validate the estimation procedure by demonstrating that credible regions concentrate around the true parameter values, and (iii) demonstrate the reliability of the estimates with respect to the true parameter values.

The same techniques used to generate data resembling the CAV data set are applied here as well, just with more features to simulate such as the additional response functions and a *death bias*. Before diving right into the estimation results, a few estimation details are discussed, mainly relating to prior information. In the state space of Figure 1, there are a number of assumptions which can reasonably be made about the biology of this process. For instance, the rate of transition to the N+ state with high amyloid burden should be at least as fast as with low amyloid burden. That is, the rate parameter from Q for transitioning from A+N− to A+N+ should be at least as big as the rate parameter from Q for transitioning from A−N− to A−N+. Similar constraints should be true for transitioning from low to high amyloid burden with respect to high/low neurodegeneration burden (i.e., having N+ cannot lower the rate of transitioning to A+), and those with dementia die at a rate no less than those without dementia. These constraints

are a mathematical formulation of the assumption that having a larger burden may not escalate disease progression, but it certainly cannot help. The full list of constraints on the rates is given in the following table,

The rate from ...	*... should be at least that from ...*
A+N− (state 2) *to* A+N+ (state 4)	A−N− (state 1) *to* A−N+ (state 3)
A+N+ (state 4) *to* A+Dem (state 6)	A−N+ (state 3) *to* A−Dem (state 5)
A−N+ (state 3) *to* A+N+ (state 4)	A−N− (state 1) *to* A+N− (state 2)
A−Dem (state 5) *to* A+Dem (state 6)	A−N+ (state 3) *to* A+N+ (state 4)
A−Dem (state 5) *to* Dead (state 7)	non-Dem (states 1-4) *to* Dead (state 7)
A+Dem (state 6) *to* Dead (state 7)	A−Dem (state 5) *to* Dead (state 7).

In addition to these constraints, it is reasonable to assume that all of the age coefficients on the rate parameters (see equation (5)) are nonnegative. That is, becoming older will not slow down an individual's rates of progression through the state space. Lastly, constraints are placed on the response variable parameters to help with identifiability. The mean $\log(\text{PIB} - 1)$ measurement associated with the low amyloid burden state should not exceed that of the high amyloid burden state. Analogously, the mean Thickness measurement associated with high neuro-degeneration burden should be less than or equal to that for low neuro-degeneration burden (low Thickness is associated with high burden). Finally, the mean MMSE scores should be monotone non-increasing in the state ordering $\{1, 2, 3, 4, 5, 6\}$, i.e., this implies that being in an N+ state should generally be associated with lower MMSE scores than being in an A+ state.

One of the advantages of the Bayesian approach of estimation is that imposing this long list of important constraints in the model can be easily handled through specification of the priors, and the priors have a natural ability to accommodate other known information about the model parameters. In particular, the MCSA sampled subjects are from the greater Rochester, MN area, and Minnesota death rates for women and men of all ages (from 1970 to 2004) are made available in the US Decennial Census, which are captured in the "survexp.mn" data set in the `survival` package in R (Therneau 2015).

It is assumed that all non-dementia to death rates are the same (and hence share the same coefficients in equation (9)). The fact that the A and N states do not have obvious

Prior means and standard deviations

State transition parameters (see (5) and (9))

Transition	$\beta_0^{(l)}$	$\beta_1^{(l)}$	$\beta_2^{(l)}$	$\beta_3^{(l)}$	$\beta_4^{(l)}$	c	$c+d$
non-dem→7	-4.41 (.1)	.094 (.01)	.47 (.05)	0 (.1)	0 (1)	-.75 (.375)	-.60 (.3)
5→7 and 6→7	-4 (1)	.1 (.05)	0 (1)	0 (.1)	0 (1)		
all others	-3 (1)	.1 (.05)	0 (1)	0 (.1)	0 (1)		

Cubic spline parameters for state 1 to state 2 as a function of age

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
-5 (1)	-4 (2)	-3 (2)	-2 (2)	-1 (2)	0 (3)	1 (3)	2 (3)

log(PIB − 1) response (see (10))

μ_{A-}	μ_{A+}	$\log(\sigma^2_{\text{pib}})$
-1.3 (.2)	-.5 (.2)	$\log\left((.4/3)^2\right)$ (2)

Thickness response (see (11))

μ_{N-}	μ_{N+}	$\log(\sigma^2_{\text{thick}})$
3.14 (.2)	2.34 (.2)	$\log\left((.4/3)^2\right)$ (2)

MMSE response (see (12))

α_1-α_4	α_5-α_6	α_7	α_8	α_9	α_{10}	α_{11}	$\log(\sigma^2_{\text{mmse}})$
-.28 (.75)	-7.3 (3)	0 (1)	0 (1)	0 (1)	0 (1)	0 (1)	-.7 (2)

Dem misclass (see (13))

$\text{logit}(p_0)$	$\text{logit}(p_1)$
-3 (1)	-3 (1)

Initial probabilities (see (14))

ξ_1	ξ_2	ξ_3
-3.5 (.25)	-6 (1)	-6 (1)

Table 2: Displayed are the means and standard deviations of the (independent) normal priors placed on the 81 model parameters. Standard deviations are in parentheses, and these priors assume that the data have been centered. The parameters $c_1, \ldots, c_8$ are the control points used to estimate the cubic spline for the state 1 to state 2 transition (for baseline and age), as discussed at the end of Section 2.1.

external manifestations makes this assumption reasonable. With so many paths for which an individual can transition through the state space, the strong prior information on the death rates is crucial for practical identifiability/reducing uncertainty in the estimated transition rates. For example, an individual who is observed first in state 2 and is next observed dead could have transitioned directly from state 2 to 7 (death), or state $2 \to 4 \to 6 \to 7$, for example. Since the death rate is known to not deviate far from the overall Minnesota rate, it serves to untangle this ambiguity in a broad sense, or "on average". The full table of prior distribution specifications for each of the 81 parameters in the model is provided in Table 2.

The simulation study consisted of simulating 50 synthetic data sets resembling the MCSA data. The synthetic data sets were simulated to contain similar amounts of
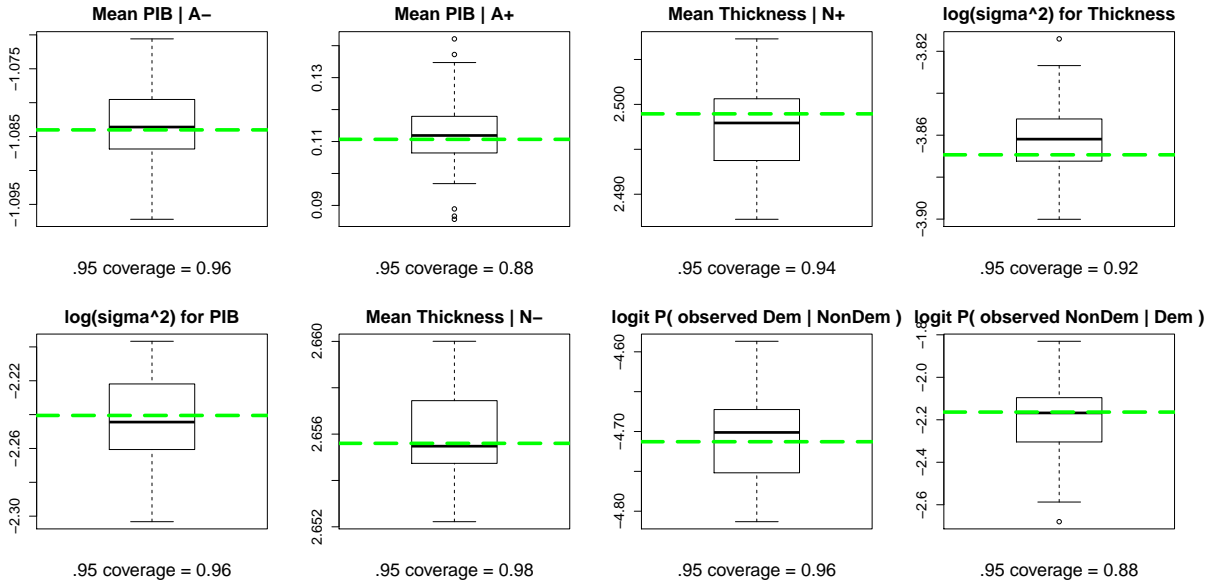
Figure 7: Various response coefficient estimates for the synthetic MCSA data. Presented are box plots of posterior means of labeled parameters, for 50 synthetic MCSA data sets. Green dashed lines represent the true values. Coverage is the proportion of .95 probability credible intervals which contain the true parameter value.

information to the real data set. That is, 4742 subjects were simulated starting from random ages and assigned other covariates randomly from the empirical distribution of the MCSA data. The simulated subjects are "observed" at times which are determined by sampling from the actual inter-observations times in the MCSA data.

The maximum length of time in the study for subjects in the MCSA is not much longer than 12 years. Thus, for reasonable comparison, subjects in the synthetic data sets are observed for 12 years or until time of death, whichever comes first. Approximately 57% of subjects in the MCSA data set have at least one observed biomarker measurement. Of that 57%, the proportions of observed biomarkers over all visits is presented in the following table,

		PIB	
		measured	not measured
Thickness	measured	0.231	0.234
	not measured	0.002	0.533

.

To keep consistent with this feature, ∼53% of the synthetic subjects were given no biomarker data for any of their visits, while the remaining ∼47% were given observed PIB or Thickness measurements according to the above distribution. Death was observed
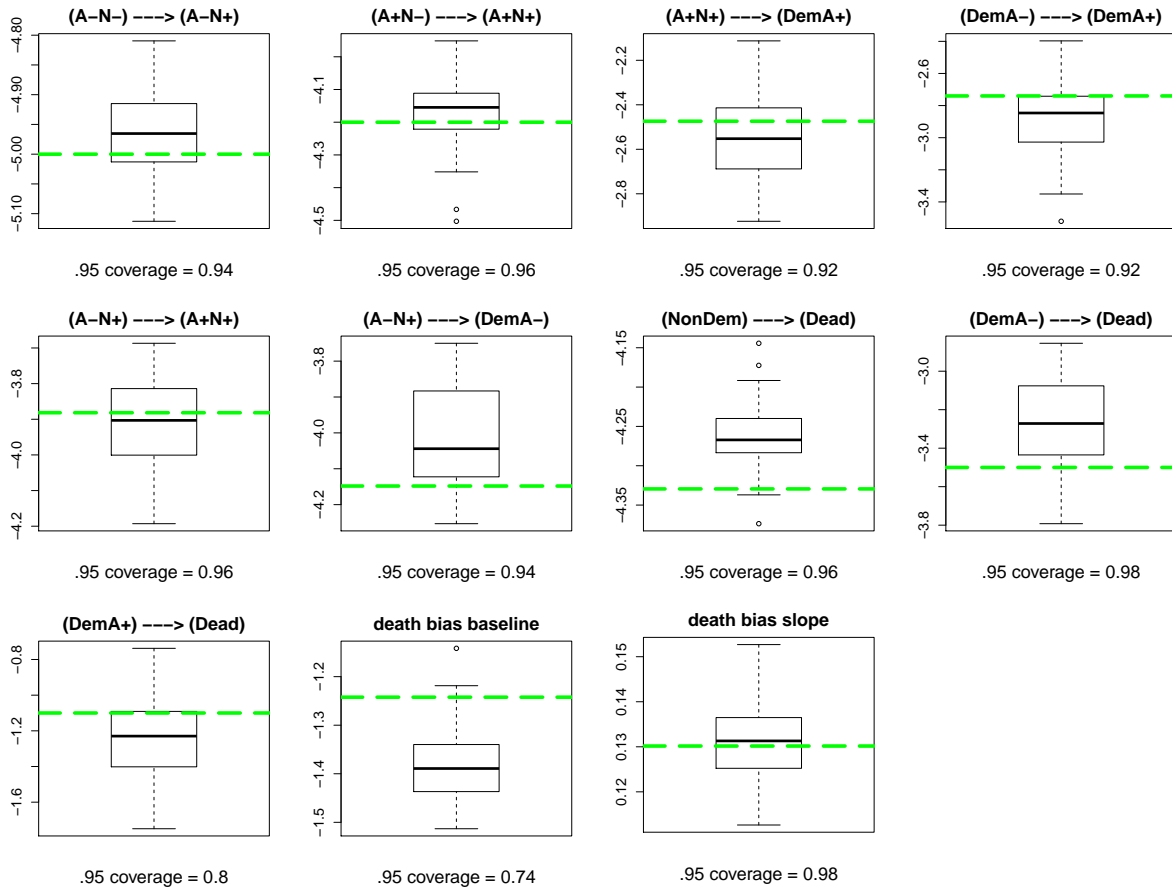
Figure 8: Intercept coefficient estimates for the synthetic MCSA data, see (5) and (9). Note that the covariates in the data are centered. Presented are box plots of posterior means of labeled parameters, for 50 synthetic MCSA data sets. Green dashed lines represent the true values. Coverage is the proportion of .95 probability credible intervals which contain the true parameter value.

for just over 28% of the actual MCSA study subjects, and the number of clinical visits for study subjects varied between 1 and 10 visits, with a median of 4. The synthetic data sets observe death for ∼31% of subjects on average, and the number of clinical visits for synthetic study subjects varies between 1 and 11 visits, with a median of around 6.

Figures 7 and 8 provide a summary of the results in the form of box plots of the posterior mean estimates, and coverage for 95 percent credible intervals, for 19 of the more interesting model parameters. The Supplementary Material contains a more detailed summary of these results, including box plots, coverages, histograms, and MCMC trace plots for all 81 model parameters. Overall, this simulation exercise lends some confidence to results of the actual MCSA analysis presented next.

4 Analysis of the Mayo Clinic Study of Aging Data

This section presents analysis and interpretation of the HMM parameter estimates of the actual MCSA data. Since there are 81 HMM parameters, some playing very different roles, presenting the output in a concise manner is challenging and depends on the question being asked. We present a targeted summary of a few important questions here. See the Supplementary Material for estimates of all 81 HMM parameters, including the MCMC trace plots, histograms, and 95 percent credible intervals.

The state space as it is defined in Figure 1 allows for the computation of transition probabilities broken down for particular types of dementia. Most notably, the devel-
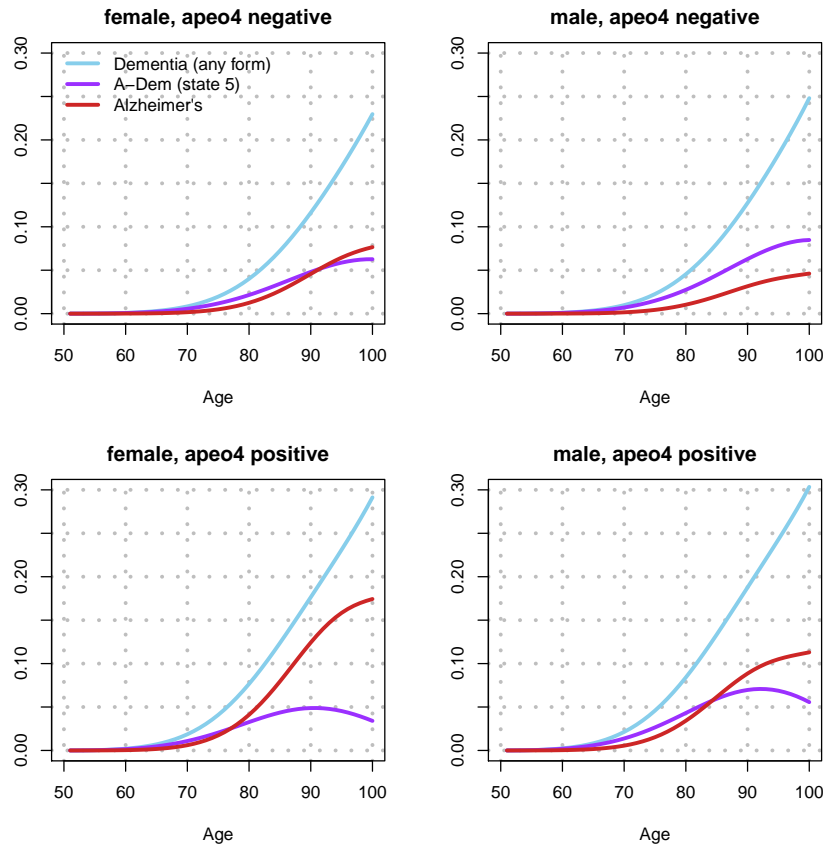

Figure 9: Evolution of transition probabilities. The curves represent the probability of transitioning to the respective state for each given age, computed using the posterior mean estimates of the HMM parameters. The probabilities are conditional on not transitioning to state 7 (Dead), and correspond to an individual in state 1 (A−N−) at the baseline age of 50. The label 'apoe4 neg' corresponds to an individual with no APOE-ε4 alleles, and 'apoe4 pos' corresponds to an individual with at least one APOE-ε4 allele. The curve labeled 'Alzheimer's' depicts the probability of making the transition from state 4 (A+N+) to state 6 (A+Dem), given not dead.

opment of Alzheimer's Disease as defined here corresponds to a transition from state 4 (A+N+) to state 6 (A+Dem), i.e., there was Amyloid build up, prior to the neuro-degeneration leading to Dementia. See Figure 9 for the estimates of how these transition probabilities evolve over time. While Alzheimer's Disease is slightly prominent among females of a given age versus males of the same age, it is interesting that the likelihood of dementia (of any kind) is nearly the same, i.e., males are more likely to develop non-Alzheimer's related dementia.

The estimated *death bias* for this study is also of interest, particularly due to the novel approach taken to account for it. It is observed that individuals just enrolled in the study experience a death rate which is $\sim$31 percent (posterior mean) of the population death rate and it remains lower than the rest of the population for several years after enrollment; See Figure 10. This suggests that the *death bias* cannot be ignored.

Another feature of the analysis is that it is cutoff point agnostic, by design. Instead of hardwiring cutoff points, the suggested
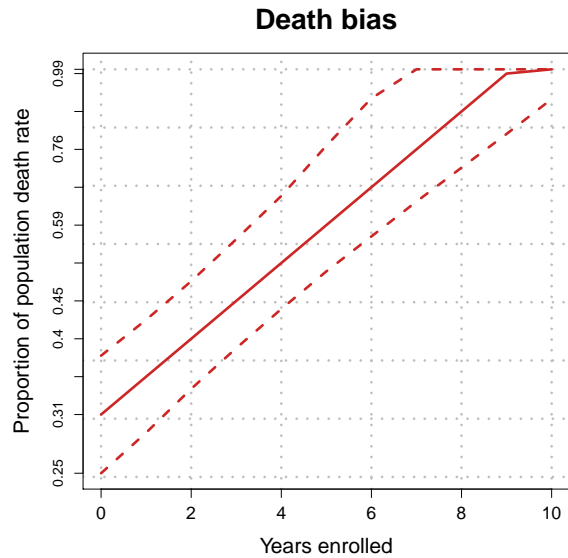

Figure 10: Posterior mean estimate of the *death bias* (9). Values are interpreted as the proportion of the population death rate that is experienced by subjects in the MCSA, for each integer year a subject is enrolled in the study. For example, subjects just enrolled in the study experience a death rate which is 31 percent of the population death rate. The dotted lines are 95% credible bands.

cutoff points in the medical literature have been used as prior information for the response distributions from high/low amyloid burden and high/low cortical thickness loss burden in the $\log(\text{PIB}-1)$ and Thickness measurements, respectively. Recall, Figure 2 displayed the estimated distribution of these response biomarkers for high/low burden states.

Table 3 shows the estimated dementia misclassification probabilities. These estimates indicate that physicians tend to be conservative in diagnosing dementia. That is, they

		Observed status	
		Diagnosed not demented	Diagnosed demented
True status	Not demented	0.992	0.008 [0.007, 0.009]
	Demented	0.107 [0.070, 0.151]	0.893

Table 3: Posterior mean estimates of the dementia diagnosis response parameters. The components are probabilities. Note that each row corresponds to only one parameter, but both columns have been filled in for ease of interpretability (rows must sum to one). The brackets represent 95 percent credible intervals (for the components directly corresponding to the parameters which were estimated).

very seldom diagnosis an individual without dementia as demented, but about 1 out of 10 individuals that truly have dementia is not diagnosed as such.

Finally, presented in Figures 11, 12, and 13 are heat maps of the state space corresponding to individual specific estimated transition intensities (posterior mean estimates). These depictions of the state space in relation to the estimated parameters captures a holist picture of the model as a physical system. With these plots, it is easy to identify which transition rates are most intense at various years of age, and to get a general sense of when rates begin to 'heat up'. In fact, these plots can be created for every integer age greater than or equal to 50, and they are presented as a movie in the Supplementary Material.

There are many variables whose effect is known medically that could be used to validate some of the results. For example, it is observed that the rates are relatively dormant for the first two decades starting from 50 years old. After that, transition rates among the first four states seem to increase slower than the transition rates to more advanced states. Transitions particularly to A+Dem from A+N+ or from A−Dem are the most intense over almost all ages. APOE-ε4 is known to increase risk of A+ by a factor of 2-3. Based on this, the model should yield a higher rate estimate of A−N− (state 1) to A+N− (state 2) and A−N+ (state 3) to A+N+ (state 4) among APOE-ε4 carriers than non-carriers. Examination of Figures 11-13 illustrates that at every age and for both men and women, the relationship between both of these transition rates and APOE-ε4 in the model output is exactly as would be expected. Additionally, the rate of A+N+ to A+Dem should be greater in APOE-ε4 carriers than non-carriers. This is
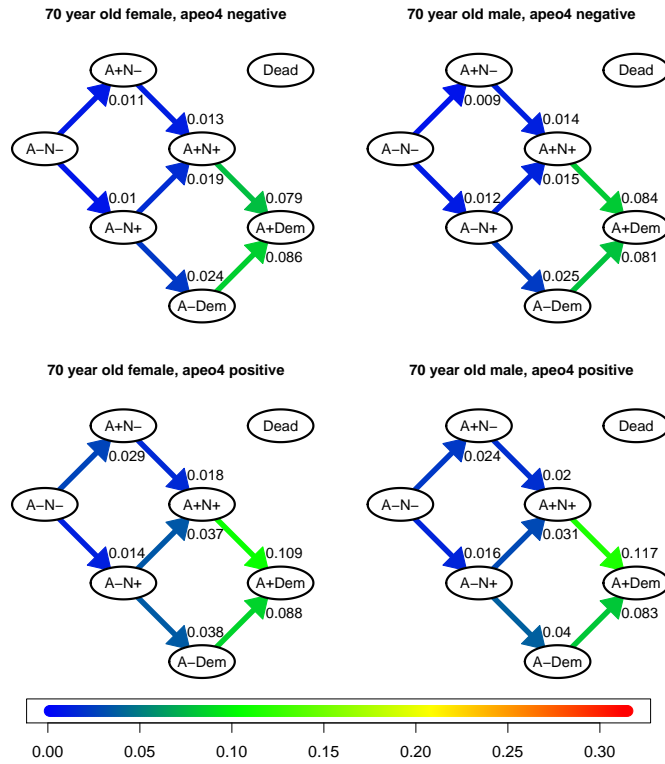
Figure 11: Posterior mean estimates of the transition rate components of Q at age 70. Each component can be interpreted as the rate parameter for a univariate exponential random variable. These plots correspond to an individual with a college degree. Transitions to death are omitted for clarity.
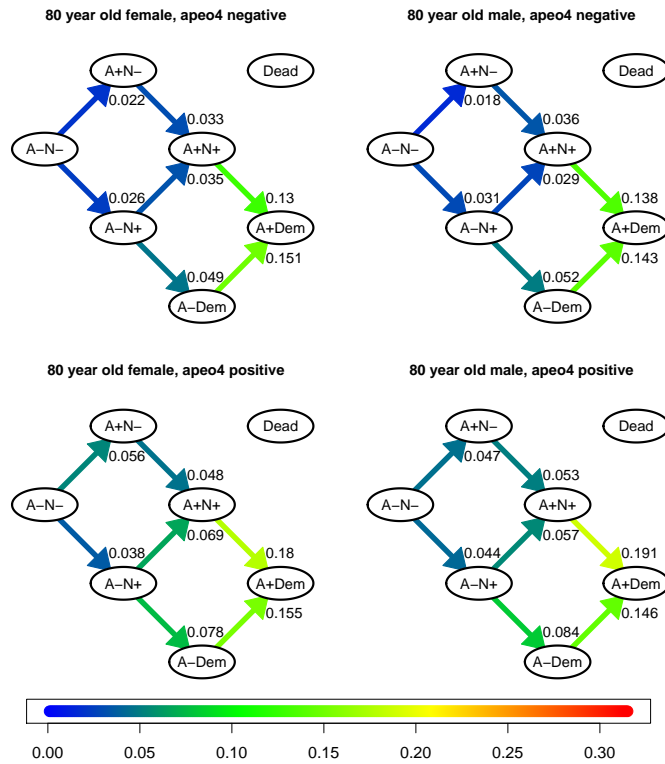

Figure 12: Posterior mean estimates of the transition rate components of Q at age 80.
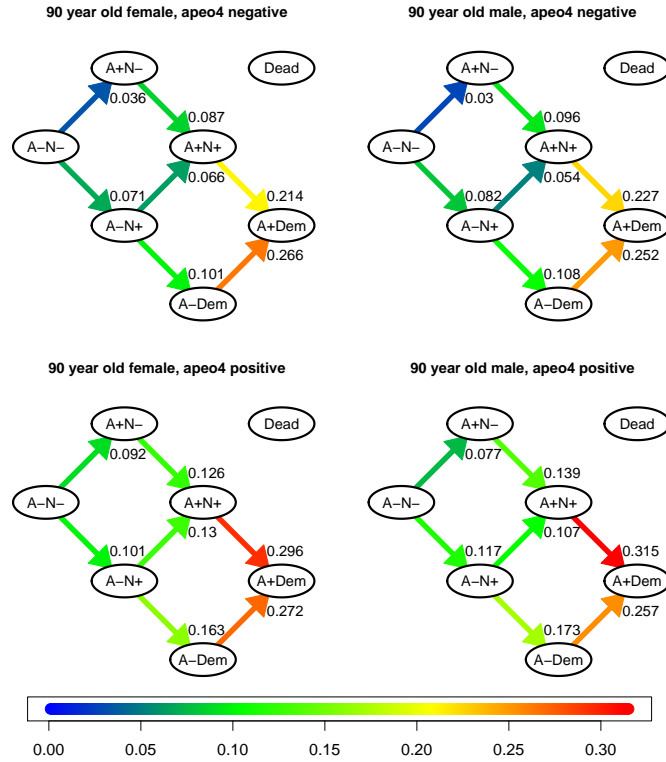
31

Figure 13: Posterior mean estimates of the transition rate components of Q at age 90.

the case in the model for all three ages shown and for both men and women as would be predicted from the known biology.

5 Conclusions & Further Work

A continuous time HMM was developed for the analysis of the MCSA data. Much care was taken to make this model as realistic of an approximation to the actual data generating process as possible, including the treatment of important features such as *delayed enrollment* and *death bias.* A Bayesian computational framework was developed in order to facilitate computation and quantification of uncertainty, as well as allow for essential prior information on many of the parameters. The model and its estimation performance was validated via several simulation studies, prior to presenting the results of its application to the MCSA data. Several important findings were that (i) the *delayed enrollment* and *death bias* play a significant role in this study, (ii) females of a certain age are more prone to Alzheimer's related dementia than male counterparts, but they are

not more prone to dementia, in general, and (iii) individuals with at least one APOE-$\varepsilon4$ allele are more than twice as likely to develop Alzheimer's than those with no alleles.

One limitation of the proposed approach is that it does not allow for rates to vary depending on the extent of Amyloid (or cortical thickness) burden. It is known that once there is sufficient Amyloid build up, then the rate of neuro-degeneration is elevated, however, this effect may not be constant across all levels of Amyloid build up. This could be tested by allowing for multiple discrete Amyloid states (e.g., low, medium, high). However, this feature would be best addressed with a continuous state space for both Amyloid and cortical thickness burden. This is a subject of future work.

References

Bureau, A., Shiboski, S. & Hughes, J. P. (2003), 'Applications of continuous time hidden markov models to the study of misclassified disease outcomes', *Statistics in Medicine* **22**(3), 441–462.

Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. (2012), *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C. & Yan, P. (1994), 'Multi-state markov models for analysing incomplete disease history data with illustrations for hiv disease', *Statistics in Medicine* **13**(8), 805–821.

Ghahramani, Z. (2001), 'An introduction to hidden markov models and bayesian networks', *International journal of pattern recognition and artificial intelligence* **15**(01), 9–42.

Hanks, E. M., Hooten, M. B. & Alldredge, M. W. (2015), 'Continuous-time discrete-space models for animal movement', *The Annals of Applied Statistics* **9**(1), 145–165.

Himmelmann, S. S. D. D. L. & www.linhi.com (2010), *HMM: HMM - Hidden Markov Models*. R package version 1.0.
URL: *https://CRAN.R-project.org/package=HMM*

Jack, C. R., Therneau, T. M., Wiste, H. J., Weigand, S. D., Knopman, D. S., Lowe, V. J., Mielke, M. M., Vemuri, P., Roberts, R. O., Machulda, M. M., Senjem, M. L., Gunter, J. L., Rocca, W. A. & Petersen, R. C. (2016), 'Transition rates between amyloid and neurodegeneration biomarker states and to dementia: a population-based, longitudinal cohort study', *Lancet Neurol* **15**, 56–64.

Jackson, C. H. (2011), 'Multi-state models for panel data: The **msm** package for **r**', *Journal of Statistical Software* **38**(8).

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W. & Couto, E. (2003), 'Multistate markov models for disease progression with classification error', *The Statistician* **52**(2), 193–209.

Johnson, M. J. & Willsky, A. S. (2013), 'Bayesian nonparametric hidden semi-markov models', *Journal of Machine Learning Research* **14**, 673–701.

Kalbfleisch, J. D. & Lawless, J. F. (1985), 'The analysis of panel data under a markov assumption', *Journal of the American Statistical Association* **80**(392), 863–871.

Karlin, S. & Taylor, H. M. (1981), *A second course in stochastic processes*, Academic Press, New York, NY.

Lange, J. M. & Minin, V. N. (2013), 'Fitting and interpreting continuous-time latent markov models for panel data', *Statistics in medicine* **32**(26), 4581–4595.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rabiner, L. R. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**(2), 257–286.

Robert, C. P., Rydn, T. & Titterington, D. M. (2000), 'Bayesian inference in hidden markov models through the reversible jump markov chain monte carlo method', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(1), 57–75.

Roberts, R. O., Geda, Y. E., Knopman, D. S., Cha, R. H., Pankratz, V. S., Boeve, B. F., Ivnik, R. J., Tangalos, E. G., Petersen, R. C. & Rocca, W. A. (2008), 'The mayo clinic study of aging: Design and sampling, participation, baseline measures and sample characteristics', *Neuroepidemiology* **30**, 58–69.

Rydén, T. (1994), 'Consistent and asymptotically normal parameter estimates for hidden markov models', *The Annals of Statistics* pp. 1884–1895.

Satten, G. A. & Longini, I. M. (1996), 'Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **45**(3), 275–309.

Scott, S. L. (2002), 'Bayesian methods for hidden markov models: Recursive computing in the 21st century', *Journal of the American Statistical Association* **97**(457), 337–351.

Scott, S. L., James, G. M. & Sugar, C. A. (2005), 'Hidden markov models for longitudinal comparisons', *Journal of the American Statistical Association* **100**(470), 359–369.

Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J. & Large, S. R. (2003), 'Diagnostic accuracy of coronary angiography and risk factors for post–heart-transplant cardiac allograft vasculopathy', *Transplantation* **76**(4), 679–682.

Shirley, K. E., Small, D. S., Lynch, K. G., Maisto, S. A. & Oslin, D. W. (2010), 'Hidden markov models for alcoholism treatment trial data', *The Annals of Applied Statistics* **4**(1), 366–395.

Therneau, T. M. (2015), *A Package for Survival Analysis in S.* version 2.38.
  URL: *https://CRAN.R-project.org/package=survival*

Titman, A. C. & Sharples, L. D. (2008), 'A general goodness-of-fit test for markov and hidden markov models', *Statistics in medicine* **27**(12), 2177–2195.

Titman, A. C. & Sharples, L. D. (2010), 'Semi-markov models with phase-type sojourn distributions', *Biometrics* **66**, 742–752.

Tombaugh, T. N. & McIntyre, N. J. (1992), 'The mini-mental state examination: a comprehensive review', *Journal of the American Geriatrics Society* **40**(9), 922–935.

Wei, S. & Kryscio, R. J. (2016), 'Semi-markov models for interval censored transient cognitive states with back transitions and a competing risk', *Statistical Methods in Medical Research* **25**(6), 2909–2924.

Xu, X., Chong, E., Hilal, S., Ikram, M. K., Venketasubramanian, N. & Chen, C. (2015), 'Beyond screening: Can the mini-mental state examination be used as an exclusion tool in a memory clinic?', *Diagnostics* **5**(4), 475–486.

Yu, L., Griffith, W. S., Tyas, S. L., Snowdon, D. A. & Kryscio, R. J. (2010), 'A nonstationary markov transition model for computing the relative risk of dementia before death', *Statistics in Medicine* **29**, 639–648.

Zhao, T., Wang, Z., Cumberworth, A., Gsponer, J., de Freitas, N. & Bouchard-Côté, A. (2016), 'Bayesian analysis of continuous time markov chains with application to phylogenetic modeling', *Bayesian Analysis* **11**(4), 1203–1237.