Review, Replication, and Analysis of "A Bayesian Approach to Multistate Hidden Markov Models: Application to Dementia Progression" Williams et al. (2020)

J. Arturo Esquivel
a.esquivel@mail.utoronto.ca

Supervisor(s): Dr. Vianey Leos Barajas

Department of Statistical Sciences
University of Toronto

October 2023

Abstract

In this paper we review, replicate, and extend a novel application of continuous-time hidden Markov models (HMMs) in understanding the complex trajectory of dementia progression. The HMM developed incorporates corrections for two sources of bias common to population studies: the delayed enrollment bias and the death rate bias. We provide a detailed explanation of the methodologies used and developed in the original paper. We replicate the most relevant results from the paper and provide an implementation of their models using the popular software for Bayesian inference, Stan. We also discuss the choice of parametric state-dependent distributions in the paper and its limitations. Finally, we propose a different way to formulate state-dependent distributions, through the use of P-splines, making the models more flexible.

Contents

1 Introduction

Currently, one major demographic issue in the world is that of longevity. People live longer than before, and aging has presented new situations and challenges for humanity, such as dementia. Attaining a better understanding of dementia development, and about the role aging has in this process, is a pressing concern for society. Such is the aim of the Mayo Clinic Study of Aging (MCSA), a prospective study which purpose is to understand the processes that lead to dementia, particularly the type of dementia characterized as Alzheimer's disease (MCSA, 2023).

In this review, we consider the work of Williams et al. (2020). They developed a method to use the resulting MCSA data and draw inference on the role of aging in the development of dementia. The purpose of this report is to provide an extensive review and analysis of the work done by Williams et al. (2020). We provide a detailed description of the methods developed by the authors (Williams et al., 2020), along with relevant analysis of underlying theoretical considerations. Additionally, we present the results of an implementation of such methods on data closely resembling the MCSA data, and assess their relationship with those shown in the original paper. We carry out an implementation using a different MCMC approach as that used in the original paper, and discuss the outcomes it lead to. Finally, we propose a way to further extend the original model, make it more flexible, so that it can accommodate a wider variety of observations from similar, more complex, processes.

The authors built and estimated a continuous-time hidden Markov model (HMM), using a Bayesian framework, to sensibly accommodate: (1) age variation; (2) time variations between observations, (3) changing amount of information available per subject, and (4) subject specific covariates. Additionally, they demonstrate and offer solutions for multiple methodological gaps that can arise in disease progression analyses for studies similar to the MCSA. The model incorporates a correction for a common bias in *delayed enrollment* studies, such as the MCSA. It also includes a structure for estimating a separate bias, the *death rate bias*, which may arise in studies dependent on subject enrollment.

The authors extend previous work made by Jack et al. (2016), which conducted inference using a similar state space (see Figure 1). Jack et al. (2016) used a Markov model on the MCSA data to estimate age-specific biomarker state transition rates. They established that most rates are log-linear, and that at baseline (age 50) almost everyone is in a state of low biomarker burden. Williams et al. (2020) include sex, number of years of education, and presence of an APOE-ϵ4 allele as transition rate covariates, in addition to age.

In both studies the authors discretize the continuous space of biological measurements into high/low burden biomarker states. However, Jack et al. (2016) consider the states as known, using practitioner chosen hard biomarker cut-points to identify them. These biomarker measurements, associated with dementia, are proxies to the state active at the time, rather than an observation of the actual state. Therefore, as proposed by Williams et al. (2020), it may be more appropriate to consider the states as hidden and the data as emissions whose distributions depend on the actual (unobserved) state, making HMMs a natural approach. Time is considered as continuous because the process evolves in continuous time and subjects are observed at irregular times. Moreover, within the Bayesian framework, practitioner input can be included for parameter identifiability through prior distributions.

In this report we describe in detail the model developed by Williams et al. (2020) and show the results of a simulation study. We also assess the possibility of extending the model, making

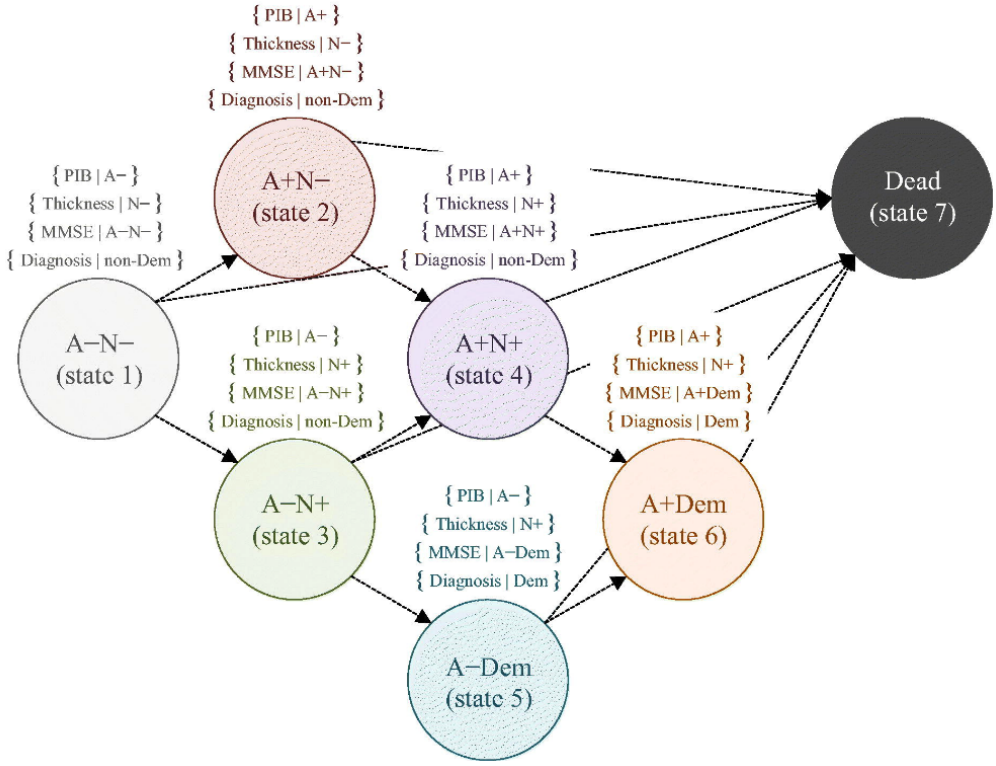


Figure 1: Dementia progression state space. Emitted response variables are displayed in brackets above the respective hidden state. A+ corresponds to high amyloid burden, and N+ corresponds to high neurodegenerative burden. States 1–4 are all non-demented (Williams et al., 2020).

it more flexible by using P-spline-based emission distributions. The structure of the document is the following: In §2 we provide a description of the MCSA data and of the alternative (simpler) data-set used for part of the analysis. In §3 we provide a description of the main methods used and detail the model developed by the authors. In §4 we describe the simulation analyses carried out and present the results obtained. In §5 we discuss the ways in which we extend the work of Williams et al. (2020) and carry out a small simulation study to show the impact using a semi-parametric approach can have. Lastly, §6 contains a discussion on the results of this study and its implications for associated future research.

2 Data

2.1 MCSA Data

The MCSA started in 2004, is still ongoing, and as of 2019 there were 4742 subjects in the study. It has enrolled a random sample (stratified by age and sex) of adults from Olmsted County in the American state of Minnesota. Adults are enrolled into the study at the (approximate) age of 50. Only those subjects diagnosed not to have dementia during their first observation are accepted into the study. Once enrolled, through clinical visits where information relevant to dementia is collected, subjects are followed up approximately every 15 months.

It has been widely established that amyloid buildup in the brain and considerable neurodegeneration are significantly related to dementia (see, for example, Dubois et al. (2014)). Thus,

approximately 50% of the subjects were chosen to undergo brain scans at every clinical visit (subjects not chosen have less information available). For those subjects, the Pittsburgh compound B (PIB) level from a positron emission tomography (PET) is recorded, as a measure of amyloid buildup. Similarly, cortical thickness measured from magnetic resonance imaging (MRI) is used as a proxy for the level of neurodegeneration. Both measures are continuous-valued.

During each visit, all subjects take the Mini Mental State Exam (MMSE). The MMSE is a questionnaire administered by medical professionals that is used to assess cognitive impairment, using an integer-valued 30 points system (see Folstein et al. (1975) for further details).

Lastly, at the end of each visit (observation), subjects are determined to be either healthy or cognitively impaired (demented) by professionals. Baseline data, such as age, sex, number of years of education, and the presence (or absence) of an APOE-ϵ4 allele, is also available for all subjects enrolled. Hence, for each subject, a number of variables are available at various (4 on average) points in time (distinct ages).

2.2 CAV Data

For some of the arguments (e.g., demonstrating the bias effect of *delayed enrollment*) made in Williams et al. (2020), the cardiac allograf vasculopathy (CAV) data-set is used. It contains data, available in the R package *msm*, collected by Sharples et al. (2003) to study the progression of CAV in survivors of heart transplants.

All patients in the study were observed at baseline (time of heart transplant) and had (approximately) yearly angiographic examinations. The response variable recorded at each follow up observation is the grade of CAV (state) diagnosed by the examiner. There are three possible states to be diagnosed: *no CAV* (state 1), *mild CAV* (state 2), and *severe CAV* (state 3). *Death* is considered to be the fourth sate and its occurrence is observed without error. The data also contain the time (in years) since the transplant and the sex of the subject as covariates.

Unfortunately, due to health information privacy restrictions, the actual MCSA data used in Williams et al. (2020) was not available for this study. For that reason, all the analysis regarding the model from the original paper is made using simulated data, simulated using a script made publicly available by the authors. The actual CAV data is available and simulated (synthetic) versions of it are used (both here and in Williams et al. (2020)) to carry out analyses. The CAV data-set is much simpler than the MCSA, but is similar in many of the MCSA particularities (such as the appropriateness of assuming an underlying continuous-time Markov process). For that reason, it is useful to impose to it conditions similar to those in the MCSA data (e.g., delayed enrollment) and carry out simulation analyses under this (simpler) setting. In §4.1 and §4.4 we provide a description on how both, CAV and MCSA, synthetic data-sets were built.

3 Methods

Here, we describe the basic structure of a hidden Markov model (HMM) and detail the continuous-time hidden Markov model developed by Williams et al. (2020) to understand the process of dementia progression.

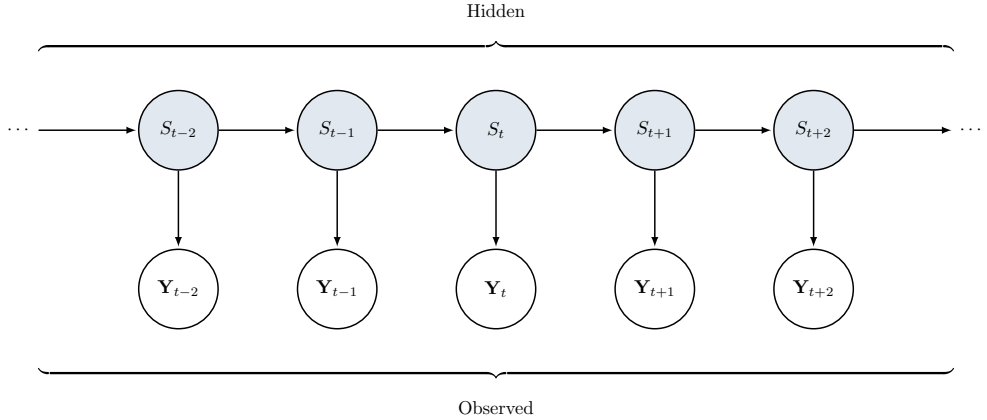


Figure 2: Basic dependence structure of an HMM. $S_1, \ldots, S_T$ is the state process, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ correspond to the state-dependent observed process.

3.1 Hidden Markov Model

A hidden Markov model (HMM) is a doubly stochastic process. It consists of an underlying, often non-observable (hidden), process, the state process; and an observable process, the state-dependent process (Rabiner and Juang, 1986). Figure 2 shows the basic dependence structure of an HMM. In a basic discrete-time setting, the state process is assumed to occur at homogeneous time-steps, and to be a first-order Markov chain. i.e., it is a stochastic process transitioning within $\{1, \ldots, K\}$, for K finite, satisfying the Markov property ($\mathbb{P}(S_{t+1} \mid S_t, \ldots, S_1) = \mathbb{P}(S_{t+1} \mid S_t)$). The state-dependent process' random mechanism is given by the state process. Meaning, the observations $\{\mathbf{Y}_t\}_{t=1}^T$ are generated from what are known as the state-dependent emission distributions $\{f_s\}_{s=1}^K$, determined by S_t, the state active at each time $t = 1, \ldots, T$. $\{\mathbf{Y}_t\}_{t=1}^T$ can be multivariate and are considered to be conditionally independent given the states $\{S_t\}_{t=1}^T$.

An HMM can be fully characterized by: K, the number of states; $\boldsymbol{\pi}$, the initial distribution of the state process at baseline ($t = 1$), with entries $\pi_k = \mathbb{P}(S_1 = k)$, for $k = 1, \ldots, K$; $\mathbf{P}(h, t)$, the transition probability matrix, with entries $P_{rs}(h, t) = \mathbb{P}(S_{h+t} = s \mid S_h = r)$, for $h = 1, \ldots, T$ and $t = 1, \ldots, T - h$; and the state-dependent distributions $f_s(\mathbf{Y}_t) = f(\mathbf{Y}_t \mid S_t = s)$, for $s = 1, \ldots, K$ and $t = 1, \ldots, T$.

3.2 HMM Likelihood Function

The joint distribution of the observations determines the form of what is commonly referred to as the marginal likelihood function. For one time series, it is given by a summation over all possible state sequences and the probabilities involved:

$$f\left(\{\mathbf{y}_t\}_{t=1}^T\right) = \sum_{s_1=1}^{K} f_{s_1}(\mathbf{y}_1) \cdot \sum_{s_2=1}^{K} P_{s_1 s_2}(1,1) f_{s_2}(\mathbf{y}_2) \cdots \sum_{s_T=1}^{K} P_{s_{T-1} s_T}(T-1,1) f_{s_T}(\mathbf{y}_T).$$

And, as shown by Zucchini et al. (2016), the marginal likelihood can be written as

$$f\left(\{\mathbf{y}_t\}_{t=1}^T\right) = \boldsymbol{\pi}' \mathbf{D}(\mathbf{y}_1) \mathbf{P}(1,1) \mathbf{D}(\mathbf{y}_2) \cdots \mathbf{P}(T-1,1) \mathbf{D}(\mathbf{y}_T) \mathbf{1}, \tag{1}$$

for $\mathbf{D}(\mathbf{y}_t)$ a $K \times K$ diagonal matrix with diagonal $f_1(\mathbf{y}_t), \ldots, f_K(\mathbf{y}_t)$, and $\mathbf{1}$ a column vector of ones.

3.3 Continuous-time HMM

When the state process (or its emissions) is observed at irregular times, a continuous-time setting can better accommodate its dynamics. When that is the case, the possibility of transitions occurring at any (continuous) point in time should be accounted for. Karlin and Taylor (1981) describe that for such a process, when it is time-homogeneous (i.e. transition probabilities are constant through time) and the probabilities $P_{rs}(t) = P_{rs}(h,t) = \mathbb{P}(S_{h+t} = s \mid S_h = r)$ are differentiable, the transition probability matrix $\mathbf{P}(t)$ satisfies the Kolmogorov backward and forward equations, which can be written as:

$$\frac{\partial \mathbf{P}(t)}{\partial t} = \mathbf{P}'(t) = \mathbf{Q}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q}.$$

Where $\mathbf{Q}$ is a $K \times K$ matrix, called the transition rate matrix or the intensity matrix, with entries:

$$q_{rs} = \begin{cases} \lim_{t \downarrow 0} \frac{P_{rs}(t)}{t}, & r \neq s, \\ -\sum_{s \neq r} q_{rs}, & r = s. \end{cases}$$

The off-diagonal elements of $\mathbf{Q}$ indicate the rate of change in the transition probabilities. When the process enters the r-th state, the time it takes for it to leave that state follows an exponential distribution with parameter $-q_{rr}$. And, once it leaves that state, the conditional probability of it transitioning to state s is $-q_{rs}/q_{rr}$ (Karlin and Taylor, 1981).

If the Kolmogorov equations are satisfied, then solving them provides an expression for $\mathbf{P}(t)$. Having such expression makes it possible to maintain the marginal likelihood characterization in (1). The Kolmogorov equations have a solution (Shelton and Ciardo, 2014) given by

$$\mathbf{P}(t) = e^{t\mathbf{Q}}, \tag{2}$$

where the exponential is the matrix exponential defined as

$$e^{t\mathbf{Q}} = \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbf{Q}^j.$$

Then, the expression in (1) becomes

$$f\left(\{\mathbf{y}_n\}_{n=1}^T\right) = \boldsymbol{\pi}' \mathbf{D}(\mathbf{y}_1)\mathbf{P}(t_2 - t_1)\mathbf{D}(\mathbf{y}_2) \cdots \mathbf{P}(t_T - t_{T-1})\mathbf{D}(\mathbf{y}_T)\mathbf{1}, \tag{3}$$

where now $\mathbf{y}_n$ corresponds to the n-th observation, observed at time t_n, continuous. The transition probability matrices are as in (2) and $t_T > t_{T-1} > \cdots > t_1 \geq 0$.

3.4 MCSA Dementia Progression Model

Williams et al. (2020) characterize the biology of dementia progression and its relationship with cognitive health through a seven-state model. The hidden states (i.e., all states but death, state 7) are characterized depending on the level of amyloid burden (A+, A-), the level of neurodegeneration (N+, N-) and the presence (or not) of dementia. The states are hidden because, even when PIB and cortical thickness measurements are available, they are only proxies to the actual levels of amyloid and neurodegeneration burdens, which determine the actual state of the process. The state space they consider is shown in Figure 1, where the arrows indicate the direction of transitions allowed. Note that only forward transitions are allowed. i.e., subjects are assumed to either stay the

same, or to experience cognitive decline, but never to improve. Another relevant feature is that high neurodegeneration (i.e. N+) is necessary for the development of dementia, whereas amyloid burden can either be high or low (i.e., A+ or A-) and dementia still be present. A transition from A+N+ (state 4) to A+N+ and the presence of dementia (state 6) characterizes Alzheimer's disease. Thus, this model enables the study of dementia specifically associated with Alzheimer's disease.

The model considers emitted responses consisting of four variables. i.e. for subject i at visit n, the emitted response is given by $\mathbf{y}_{i,n} = (y_{(i,n,1)}, y_{(i,n,2)}, y_{(i,n,3)}, y_{(i,n,4)})$. Where $y_{(i,n,1)} = \log(PIB-1)$ is the proxy for amyloid buildup level; $y_{(i,n,2)} = thickness$, is the proxy for the level of neurodegeneration; $y_{(i,n,1)} = MMSE$, is the cognitive impairment score; and $y_{(i,n,4)} = Diagnosis$, is the dementia diagnosis carried out. All four responses are assumed conditionally independent, given the state active (PI) and subject specific covariates. Missing responses are assumed missing at random conditional on the underlying active state. There are no emissions observed at death, and the time of death is assumed to be observed without error.

Gaussian emission distributions are assumed for $\log(PIB-1)$. The authors argue that such transformation is reasonable since there is evidence suggesting that the error from PIB measurements approaches a constant coefficient of variation. Two different means are assumed for the transformed PIB measurements depending on the level of amyloid buildup. i.e., μ_{A-} is the mean of the emission distribution when states of low amyloid burden (states 1,3, or 5) are active, and μ_{A+} when the level is high. Both groups are assumed to have the same variance and $\mu_{A-} \leq \mu_{A+}$, thus

$$f(y_{i,n,1} \mid s_{i,n}) = N(y_{i,n,1} \mid \mu_{A-}, \sigma_{pib}) \cdot I_{\{s_{i,n} \in \{1,3,5\}\}} + N(y_{i,n,1} \mid \mu_{A+}, \sigma_{pib}) \cdot I_{\{s_{i,n} \in \{2,4,6\}\}}. \quad (4)$$

Similarly, Gaussian distributions are assumed for thickness measurements. They have different means $\mu_{N-} \geq \mu_{N+}$ (higher measures of cortical thickness are associated with lower neurodegeneration) depending on the level of neurodegeneration of the active state, and same variance:

$$f(y_{i,n,2} \mid s_{i,n}) = N(y_{i,n,2} \mid \mu_{N-}, \sigma_{thick}) \cdot I_{\{s_{i,n} \in \{1,2\}\}} + N(y_{i,n,1} \mid \mu_{N+}, \sigma_{thick}) \cdot I_{\{s_{i,n} \in \{3,4,5,6\}\}}. \quad (5)$$

For the MMSE response variable, Gaussian distributions with different means are used. In this case there is a different mean for each of the six states (except death), and the mean score is also determined by subject specific covariates. The variance is assumed the same for all cases:

$$f(y_{i,n,3} \mid s_{i,n}) = N(y_{i,n,3} \mid \mu, \sigma_{mmse}), \quad (6)$$

with

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

All available covariates are considered: *age*; *male* = 1 when the subject is a male; *educ*, the number of years of education; and *apoe4* = 1 when there is an APOE-ϵ4 allele is present. *ntests* refers to the number of MMSE tests a subject has taken before a visit. It is included because it has been previously observed that individuals can improve their scores as they acquire familiarity with the test. *ntests* aims to control for that effect.

Lastly, for the diagnosis of dementia, Bernoulli emission distributions are used. $y_{(i,n,4)} = 1$ if individual i is diagnosed with dementia during visit n. p_0 and p_1 correspond to the probabilities of a dementia misdiagnosis. i.e., a dementia diagnostic when subject in states 1,2,3, or 4, and a non-dementia diagnosis when in states 5 or 6:

$$f(y_{i,n,4} \mid s_{i,n}) = Ber(y_{i,n,4} \mid p_0) \cdot I_{\{s_{i,n} \in \{1,2,3,4\}\}} + Ber(1 - y_{i,n,4} \mid p_1) \cdot I_{\{s_{i,n} \in \{5,6\}\}}. \tag{7}$$

3.5 Transition Rates

The formulation of the state process (see Figure 1) allows for only 13 different state transitions. So, the intensity matrix dictating the dynamics of the state process is given by

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

The logarithms of the transition rates in $\mathbf{Q}$, are modelled as a function of the same covariates as in (6). i.e., for $l \in \{1, \ldots, 13\}$,

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{age} + \beta_2^{(l)} \cdot \text{male} + \beta_3^{(l)} \cdot \text{educ} + \beta_4^{(l)} \cdot \text{apoe4}. \tag{8}$$

Both Williams et al. (2020) and Jack et al. (2016) argue that the death rates are mostly log-linear as a function of age, and so this functional form is appropriate. However, Jack et al. (2016) found that the effect of age on the transition rate from state 1 to state 2 is better modelled through log-cubic splines. Accordingly, Williams et al. (2020) use cubic splines for estimating said effect for q_1, in place of $\beta_0^{(1)}$ and $\beta_1^{(1)}$. They use knots at the ages of 55, 65, 75, and 90, with boundary knots at 50 and 120.

Note that modelling the transition rates as a function of age means that $\mathbf{Q}$ changes with time, which violates the time-homogeneity assumption required in §3.3. One way to deal with this, is to discretize the effect of age and assume transition rates remain constant between integer ages. Given that dementia is a process that takes years to evolve, the authors argue that this is not a very strong assumption to make. Such modification implies that the transition probability matrix in (2) has to be modified to

$$\begin{aligned} \mathbf{P}(h,t) = \exp\left\{(\lfloor h+1 \rfloor - h)\, \mathbf{Q}(\lfloor h \rfloor)\right\} \cdot \exp\left\{\mathbf{Q}\left(\lfloor h+1 \rfloor\right)\right\} \cdots \\ \cdots \exp\left\{\mathbf{Q}\left(\lfloor h+t \rfloor - 1\right)\right\} \cdot \exp\left\{(h+t-\lfloor h+t \rfloor)\, \mathbf{Q}(\lfloor h+t \rfloor)\right\}, \end{aligned} \tag{9}$$

where $\mathbf{Q}(x)$ contains the transition intensities corresponding to the integer year (age) x, $\lfloor \cdot \rfloor$ is the floor function, and $\lfloor h \rfloor \neq \lfloor h+t \rfloor$. Note that (9) is analogous to (2), it simply accounts for the different transition rates active (and their duration) along the period between h and $h+t$. It is also worth noting that, the time of death is observed without error. Hence, for transitions to state 7 the likelihood can incorporate that more precise information. And it can be easily done

by substituting $\mathbf{P}(t_T - t_{T-1})$ in (3) with the product of the matrix formed by the first six entries of $\mathbf{P}(t_T - t_{T-1})$ times the first six entries of the last column of $\mathbf{Q}(\lfloor T \rfloor)$. i.e., by considering the probability of being in any of the other six states at time T, and immediately transitioning to dead.

3.6 The Death Rate Bias

A relevant issue to consider in population based studies, is that of *death rate bias*. This effect arises because it is often less likely for sick and/or dying individuals to enroll and actively participate in a study. This can lead to lower death rates in the study's population, as compared with the overall population. The authors ague that to make adequate inference about the overall population, the *death rate bias* needs to be accounted for. They propose to correct it by explicitly estimating the log-linear bias in the non-dementia to dead transition rates. They propose to estimate it to be $c \leq 0$ at the time of enrollment, and assume it linearly decreases with a slope of $d \geq 0$, as the number of integer years enrolled in the study increases. Therefore, for $l \in \{3, 5, 8, 10\}$, (8) is modified into

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{age} + \beta_2^{(l)} \cdot \text{male} + \beta_3^{(l)} \cdot \text{educ} + \beta_4^{(l)} \cdot \text{apoe4} + g(\text{iyears}), \tag{10}$$

with *iyears* being the integer (since time must be discretized) number of years enrolled, and $g(\text{iyears}) = min\{c + d \cdot \text{iyears}, 0\}$. Note that the way g is defined, it is assumed that the *death rate bias* can only decrease the death rates.

3.7 The Delayed Enrollment Bias

The authors also address the problem of *delayed enrollment bias*, which arises when some individuals are not necessarily first observed at baseline (age 50 in the MCSA). In such situations, the likelihood of the paths that could have occurred between baseline and the time of enrollment should be considered. Williams et al. (2020) propose to do it by substituting the initial distribution $\boldsymbol{\pi}$ of those individuals with

$$\boldsymbol{\pi}(t_{i,1}) = [v_1(t_{i,1}), v_2(t_{i,1}), v_3(t_{i,1}), v_4(t_{i,1}), 0, 0, 0]' \cdot \frac{1}{\sum_{j=1}^{4} v_j(t_{i,1})}. \tag{11}$$

With $\boldsymbol{v}(t_{i,1})' = \boldsymbol{\pi}' \mathbf{P}(50, t_{i,1} - 50)$, and $t_{i,1}$ the time (age) of the first observation for subject i. The initial distribution is set up that way because dead and demented subjects are not enrolled into the study. Hence, the probability of the subject i being in a specific state at time $t_{i,1}$ has to be conditional on them being in a non-dead/non-demented state (states 1-4). Failing to account for such condition could lead to underestimating the rate at which the population transitions to dementia or death. In §4.2 we provide evidence on the effect such an omission can have.

4 Simulation Study

We now consider the simulation study carried out in the paper by Williams et al. (2020). The authors implement their model within a hierarchical Bayesian framework and carry out estimation via Markov chain Monte Carlo (MCMC). In particular, they use a *Metropolis-within-Gibbs* sampling approach. As an implementation proof of concept, and also as an attempt to test (maybe even improve) the practical performance of fitting such a model, we replicate some of their analyses

using the statistical software *Stan*, through the R package *RStan* (Stan Development Team, 2023). Thus, any Bayesian results shown in this report corresponding to the CAV data were obtained using the Hamiltonian Monte Carlo sampling algorithm built in *Stan* (Hoffman and Gelman, 2014).

One desirable feature of using a Bayesian setting is that incorporating prior information into the analysis becomes fairly simple. Accordingly, Williams et al. (2020) carried out their analysis using priors reflecting previous expert knowledge on the parameters. e.g., the prior distributions for the parameters of the brain scan emission distributions were *"chosen to correspond to biomarker values which are consistent with the medical community's most up-to-date understanding of the biology."* All priors used for this study, both for CAV and MCSA analyses, are the same as the ones specified by Williams et al. (2020). Normal prior distributions are used for all parameters. For constrained parameters, such as variances and probabilities, log-normal and logit-transformed Normal distributions are used respectively. Tables 2 and 3 in Appendices A and B contain all hyper-parameters used in the priors for the CAV and MCSA analyses respectively.

4.1 Synthetic CAV Data

As mentioned in §2.2, the CAV data contains observations of four possible states. Similar to the MCSA, the state space is assumed to have forward-only transitions, with observed "remissions" to be a result of misclassification. All subjects start in state 1 at baseline (time of heart transplant), since CAV does not develop instantly. In this data-set the only response variable is the diagnosed state and it is assumed to be the emission of a categorical distribution. Specifically, when a patient is in states 1, 2, or 3, there is a non-zero probability of observing (diagnosing) adjacent states. The probability mass function is assumed to follow the structure shown in Table 1, with p_1, p_2, p_3, and p_4 corresponding to the probabilities of different misclassification cases.

		Observed State			
		No CAV	Mild	Severe	Dead
True State	No CAV	$1-p_1$	p_1	0	0
	Mild	p_2	$1-p_2-p_3$	p_3	0
	Severe	0	p_4	$1-p_4$	0
	Dead	0	0	0	1

Table 1: Misclassification emission distributions. Each row contains the categorical distribution probabilities corresponding to the active (true) state.

Using this setting, treating time as continuous, and discretizing the intensity matrix for integer years, Williams et al. (2020) estimated the model using R package *msm* (Jackson (2011)). Integer years since heart transplant and sex were included as covariates for modelling the transition rates:

$$\log(q_l) = \beta_0^{(l)} + \beta_1^{(l)} \cdot \text{iyears} + \beta_2^{(l)} \cdot \text{male}. \tag{12}$$

With $l \in \{1, 2, 3, 4, 5\}$, and transition rate matrix

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

The estimates coming from using *msm* were then used by the authors as ground truth to

produce new data-sets, process replicated for this study. Also, some changes were made to the data generating process so that the CAV synthetic data includes characteristics similar to those of the MCSA data. An initial distribution is induced so that it is possible for the subjects to be in states other than state 1 at baseline (immediately after the heart transplant). $\boldsymbol{\pi} = [0.95, 0.04, 0.01, 0]'$ is used as initial distribution for each subject. The only subject specific covariate in this case, *sex*, is randomly chosen in proportions consistent with those observed from the empirical distributions of the real data.

The underlying state sequence is generated through sampling waiting times from independent exponential distributions, using the transition rates in correspondence with their interpretation provided in §3.3. Then, the times of follow-up visits are generated by sampling the inter-visit time empirical distribution observed in the actual data-set and adding it to the previous time observed. The process is repeated until either a visit is sampled at a time after the underlying process transitions to dead, or the individual exceeds 20 years enrolled in the study (the maximum enrollment time found in the real CAV data). Transitions to death are observed at the exact time that they happen. Lastly, for each visit, a diagnosis is sampled from the categorical distribution of Table 1 corresponding to the active state during the visit.

4.2 Delayed Enrollment Bias Demonstration

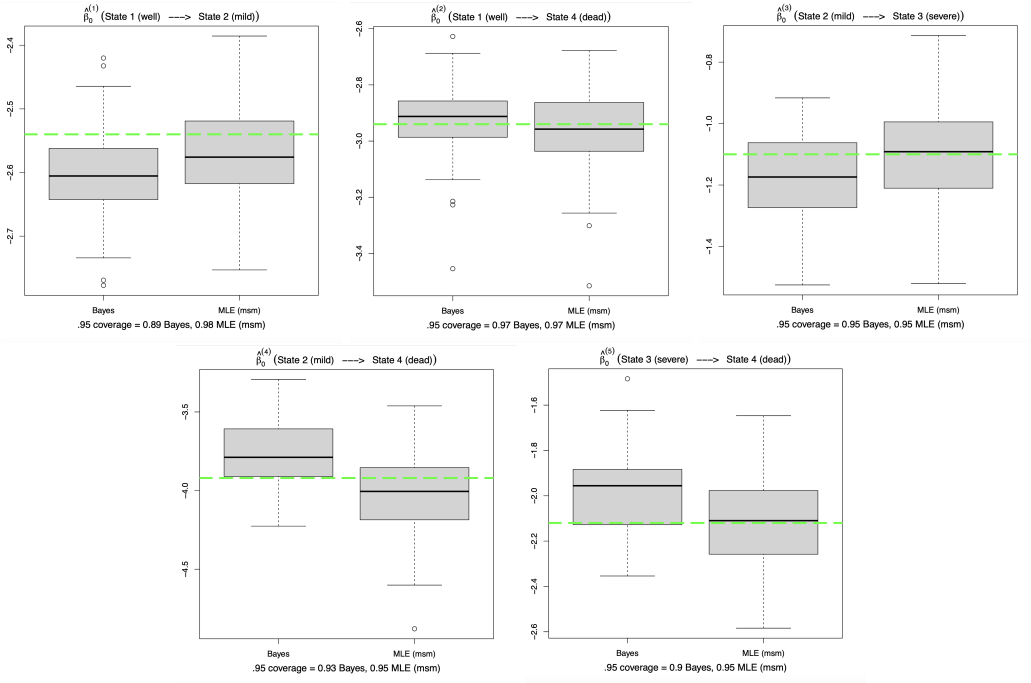


Figure 3: Synthetic CAV data (log) transition rates intercept estimates. All subjects are first observed at baseline. *Bayes* corresponds to the posterior mean estimates, and *MLE (msm)* to the maximum likelihood estimates obtained using *msm*. The true parameter values are indicated with the green dashed lines. The 95% coverage, i.e. the proportion of 95% credible (confidence for MLEs) intervals containing the true value, is indicated under each plot.

A sample of 700 (the real data-set contains 622) subjects was generated for 50 different data-sets as described above. Figure 3 shows the distribution of the estimates for the log-rate intercepts ($\hat{\beta}_0^{(l)}$) from (12) across the 50 data-sets. The Bayesian estimates, using our implementation of the

approach developed by the authors, correspond to the posterior means. And it is compared with the MLE results when using *msm*. As it can be seen in the plots, for such a simple context, both methods yield rather similar results. Note that these results are consistent with the implementation carried out by Williams et al. (2020) (Figure 5 in their paper).

A major difference between the CAV and MCSA data is that for CAV all individuals are observed at baseline. That is precisely the kind of study for which *msm* was designed, and it is how the synthetic data used to produce Figure 3 were simulated. However, there are population studies, such as the MCSA, where virtually no subject is observed at baseline, and where the *delayed enrollment bias* can arise. *msm* was not designed for such studies and thus can prove inappropriate to carry out inference. To inspect this issue, the CAV data structure was modified such that it has *delayed enrollment*. Hence, another 50 data-sets were generated using the same seeds but with the initial observations being at baseline with a probability of 0.25, and sampled from a $N(5, 1)$ with a probability of 0.75. Moreover, when a subject transitions to dead before the time of the initial observation, then the patient is not included in the sample. As Williams et al. (2020) argue, this is the crucial source of the bias, since the sample will less likely consider cases with adverse results shortly after the heart transplant. In such a scenario, a sample is not fully representative of the whole population.

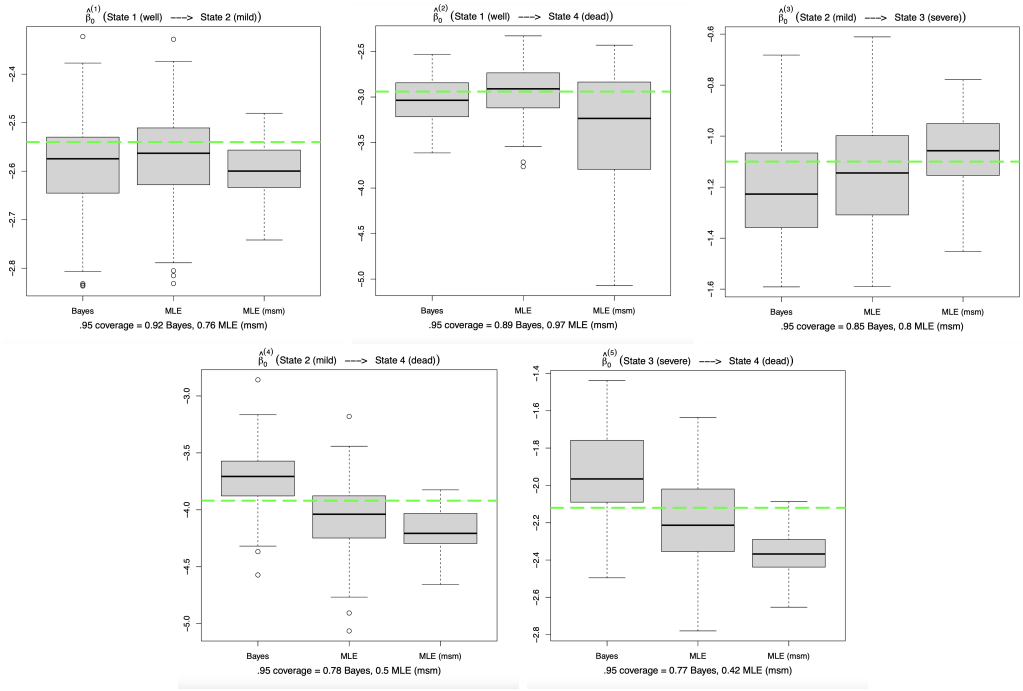


Figure 4: Synthetic CAV data (log) transition rates intercept estimates. Subjects can enroll in the study with delayment. *Bayes* and *MLE* correspond to estimates analogous to those shown in Figure 3, but accounting for *delayed enrollment*. *MLE (msm)* corresponds to the maximum likelihood estimates obtained using *msm* which does not consider any correction. The true parameter values are indicated with the green dashed lines. The 95% coverage is indicated under each plot for the Bayesian and *msm* approaches.

As mentioned before, one of the main contributions of Williams et al. (2020) is the development and implementation of an approach (see §3.7) to correct for this bias, correction that *msm* is not equipped to make. In Figure 4 the effect of conducting estimation without conditioning enrollment on being alive is showed. It contains the distributions of Bayes and MLE estimates, analogous to

those shown in Figure 3, with the only distinction that they include the initial distribution in (11). As it can be seen in the plots, not accounting for the bias can lead to estimates that indicate weaker baseline transition rates than they actually are. And not only the intercepts are affected, the effect the years since the transplant have on the transition rates is less accurately estimated when the bias is not accounted for (see Figures 9 and 10 in Appendix A). And in this simple simulation example, around 25% of the subjects are still observed at baseline. The bias can become even more significant when a higher proportion of participants (as is the case of the MCSA) is not observed at baseline. It seems clear from this analysis that methodology and implementations such as the ones carried out by Williams et al. (2020) are of great relevance for population studies with *delayed enrollment*.

4.3 Synthetic MCSA Data

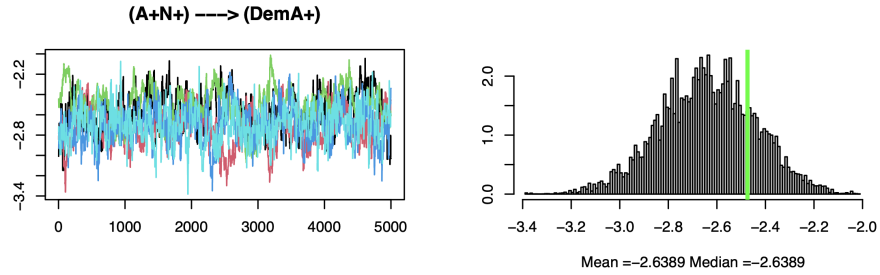
As mentioned before, the MCSA data contains HIPAA protected information and is not publicly available. Nevertheless, Williams et al. (2020) made publicly available a synthetic data-set which is supposedly very similar to the real MCSA data. We use that data-set to carry out a proof of concept of the authors' model.

The procedure used to generate the data is analogous to that in §4.1, with a few additional details. A sample of 4742 individuals (same as the real MCSA) was generated. Additional covariates (such as age or years of education) are also sampled randomly from the empirical distributions observed in the real data. For these data, baseline is the age of 50 and the maximum number of years subjects are simulated to remain in the study is 12. The age at the time of the first observation of each subject is sampled from the empirical distribution of all ages observed during the first visits in the actual data-set. For each clinical visit, subjects are simulated to remain in the study in accordance with a Bernoulli distribution with probability of 0.9125, which is consistent with the permanency rate observed in the real data. Also consistent with the conditions observed, 57% of the subjects are randomly chosen to undergo biomarker measurements. For these subjects chosen, during each visit, both PIB and cortical thickness are measured with a probability of 0.227; only thickness is measured with a probability of 0.231; only PIB is measured with a probability of 0.002; and no biomarker is measured with a probability of 0.54. The other two responses (i.e., MMSE and dementia diagnosis) are always observed. At each visit, all corresponding emitted responses are sampled independently from distributions determined by the active state (as described in §3.4), according with the model's estimates obtained by the authors from the real data.

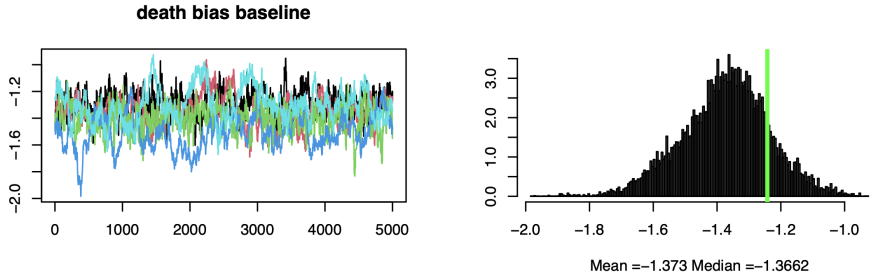
4.4 MCSA Data Implementation

As a final characterization in their model, the authors made multiple assumptions that they consider sensible for the biology of the process. First, they assume that non-dementia to death transition rates are all the same. Meaning that $q_3, q_5, q_8,$ and q_{10} are all the same with the exact same coefficients. Also, transitions to a state of high neurodegeneration (A-N+ or A+N+) from a state with high amyloid burden (A+N-) should occur at a rate at least as fast as those from states with a low burden (A-N-). Meaning that in every case the associated rate q_4 must be at least as large as q_2. Similarly, transitions to high amyloid burden should have higher rates when coming from a state of high neurodegeneration. Additionally, they assume that states with presence of dementia (states 5 and 6) transition to death at rates at least as high as those without dementia. They also assume that all the age coefficients ($\beta_1^{(l)}$) should be non-negative. Meaning that getting older can

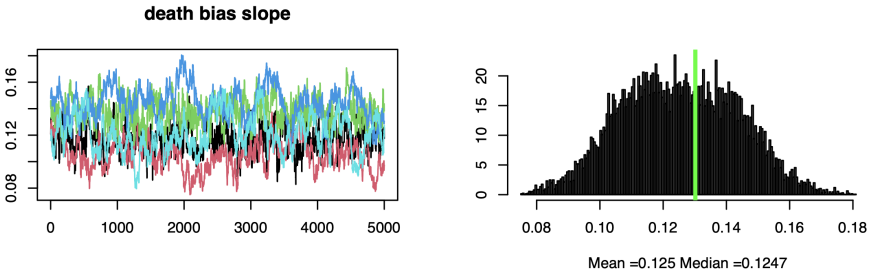
only lead to no change or an increase in the chances of dying and of dementia progressing. Lastly, the response distributions' means are constrained to ensure identifiability. The distribution means corresponding to the first two responses (amyloid buildup and thickness) are ordered depending on the level, as described in §3.4. The means for the MMSE score emission distributions are assumed to be monotonically non-increasing on the state number. i.e., the mean response in state 1 should be at least that of states 2 and 3, which at the same time should be at least that of state 4 and so on and so forth. One big benefit of using a Bayesian approach is that known information and constraints can be easily accommodated through the specification of priors.



(a) Alzheimer's log-transition rate intercept.



(b) Baseline death rate bias.



(c) Death rate bias slope.

Figure 5: MCSA model trace plots (left) and posterior samples (right). (a) contains the results for the (average) log-transition rate between state 4 and state 6. (b) those for the initial *death rate bias*. (c) shows the results for the *death rate bias* slope. Different colors are used for each chain. The true parameter values used for simulating the data are marked with the green line.

Using the MCSA synthetic data, we implemented the authors' model, using their MCMC algorithm to sample from 5 different chains. Figure 5 shows the trace plots and posterior distributions sampled from the model for 3 of the parameters of most interest: (a) the intercept of the (average) log-transition rate that characterizes Alzheimer's disease (state 4 to state 6); (b) the initial *death rate bias* (c from §3.6); and (c) the linear slope for the decrease of the *death rate bias* (d from §3.6). Analogous figures for other Alzheimer-related parameters are included in Appendix B. Overall, the

trace plots (both in Figure 5 and Appendix B) do not suggest any significant lack of convergence. Also, it is clear that most posterior samples are well located around the true parameter values.

Figure 6 shows how histograms of the observed brain scan responses compare to the state-dependent distributions and their mixture, for which the parameters are estimated using the posterior means from the model. We do not show a similar plot for $MMSE$ because the emission distributions depend on subject specific covariates and so a similar plot would not be all that meaningful. It seems the estimated emission distributions are consistent with the empirical distributions of the observed responses.

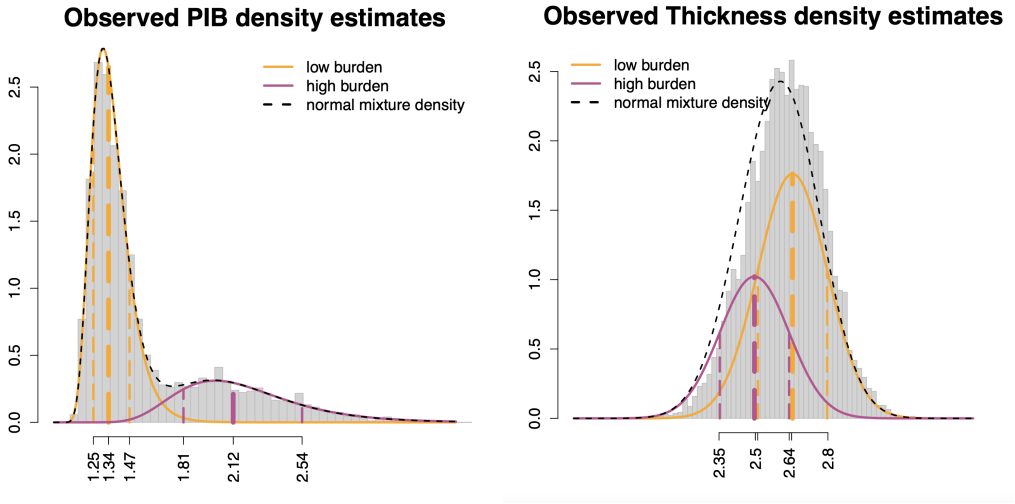


Figure 6: Observed response data for PIB (estimated using log(PIB - 1)) and (cortical) thickness. The densities correspond to the state-dependent distributions (with means depending on the biomarker level) estimated from the model. The black dashed lines represent the corresponding normal mixture density estimates. The vertical lines indicate the estimated means and a distance of 1 (estimated) standard deviations from them.

Given that the data used to obtain these results are synthetic, it is our opinion that there is not much value in analyzing the biological implications of the values obtained. However, looking at the results, it seems that the model's formulation offers reasonable insights into the dynamics underlying the data generating process. Even more complex effects such as the *delayed enrollment* or the *death rate* biases are captured by the model reasonably well. Suggesting that, if the simulation assumptions made are at all close to the real biology of the process, the authors' approach can yield meaningful and actionable insights regarding dementia progression. The results seem to be consistent with those obtained in Williams et al. (2020) and provide some reassurance that the model can indeed offer useful insights. e.g., Williams et al. (2020) argue the existence of a *death rate bias* by stating that when the bias indeed exists (as it is the case for the data used here) the model correctly estimates it and posterior samples do not contain the value zero. That is the case for this simulation study. Thus, given that the authors obtained similar results on the real data, evidence suggests that the *death rate bias* in fact exists and that the model is reasonably well equipped to estimate it.

5 Extensions

5.1 Stan Implementation of the Model

Our first extension to the work in Williams et al. (2020), is that of coding their methods into *Stan*. We deem it useful to program their models in *Stan* 1) as an alternative realization of their methodology, implementing their models through a different MCMC sampling mechanism; and 2) as an attempt to improve the practical use of similar HMM methods in further projects. *Stan* is an easily accessible software widely used for Bayesian inference. Having broadly applicable methods, such as a continuous-time HMM, implemented in sources like *Stan* make academic collaboration possible and more productive.

For both (CAV data and MCSA data) implementations, the authors use a Metropolis-within-Gibbs sampling approach. Meaning that they use a proposal distribution (specifically a multivariate normal) to update the HMM parameters. And that they are updated in groups built from parameters which exhibit strong correlations. i.e., in order to update a subset of the parameter vector, say $\boldsymbol{\theta}$, the proposal distribution for the next set of values is given by $N(\boldsymbol{\theta}^{(t)}, \tau\Sigma)$, where $\boldsymbol{\theta}^{(t)}$ is the current parameter vector in the MCMC chain, τ is a step size parameter, and Σ is the empirical covariance matrix of the parameter group from previous steps in the MCMC chain. A new parameter vector $\boldsymbol{\theta'}$ is accepted with probability given by the Metropolis ratio (the target distribution evaluated at $\boldsymbol{\theta'}$ divided by the target evaluated at $\boldsymbol{\theta}^{(t)}$). If the proposed vector is accepted, $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta'}$, and if it is rejected $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$. The process is carried out analogously to update every other subset of the parameter vector, using the most recent values of the other parameter groups to evaluate the target distribution. The piece-wise updating process is repeated until the desired number of sampled values is achieved.

τ is scaled down when the acceptance ratio gets too low (to propose smaller MCMC steps), and up when the acceptance ratio gets too high (to propose larger MCMC steps). Σ is updated at each step, but considers only a limited number of values from the MCMC chain history so as to exclude misrepresentative parameter vectors which the MCMC algorithm visited during the adaptation period. After a pre-specified number of (burn-in) steps, τ and Σ are held fixed at their most recent values and a traditional MCMC sampling with fixed tuning parameters is conducted.

Stan, on the other hand, uses a Hamiltonian Monte Carlo (HMC) sampling algorithm (Neal et al., 2011). It uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior distribution. It does this by simulating the Hamiltonian of a particle whose starting position is the current parameter values, and for which initial momentum values (kinetic energies) are generated randomly. Hamiltonian dynamics make it possible to produce distant proposals for the Metropolis algorithm, thereby avoiding the slow exploration of the state space, often afflicting random-walk (such as Metropolis-within-Gibbs) exploring procedures (Hoffman and Gelman, 2014). Thus, most complex problems with continuous state spaces can be more efficiently approached through HMC.

Note that the only discrete elements of the HMMs being considered are the states, and that such discrete nature is eliminated from the likelihood (and consequently from the target distribution) when we sum over all possible state sequences as in (1). Hence, HMMs, such as the ones considered in this work, are natural candidates to be modelled using HMC. For that reason, an additional motive for implementing the models in *Stan* is that of possibly achieving a more efficient implementation. One that translates into more acceptable running times and that makes replication and simulation studies practically useful. Computational efficiency aside, *Stan* provides

the practical advantage that parameter constraints can be easily imposed to the (constrained) parameter spaces without worrying about efficient proposal distributions, parameter sub-setting, or other major programming challenges.

Due to time constrictions regarding this project, we limit ourselves to fully implementing only CAV models in *Stan*. That is why (as mentioned throughout this paper) all CAV-related results were obtained using *Stan*, while the algorithm developed and made publicly available by Williams et al. (2020) was used for all MCSA-related results. Nevertheless, the MCSA dementia progression model has been almost fully programmed in *Stan*, producing promising results for small data-sets (800 entries, the synthetic MCSA data contains over 20000). All the code used for this study, including original and adapted from that made by Williams et al. (2020), can be found at the GitHub repository linked in Appendix C.

Overall, we consider the *Stan* implementation of the CAV model to be a success. As mentioned in §4.2, the results obtained are consistent with those obtained by Williams et al. (2020). Moreover, knowing the true values of the parameters, we can see (in §4.2 and Appendix A) that for the 50 different simulated data-sets, the parameter estimates are reasonably good. Trace plots of four different chains in Appendix A do not suggest a problem of lack of convergence and most posterior distributions indicate an adequate fit of the model. What is more, it seems our *Stan* implementation of the CAV model is in fact more efficient than the authors'. Due to time constraints, our models used 600 warm-up iterations and 600 samples, which proved to be enough to produce acceptable results (Williams et al. (2020) use 5000 and 10000 respectively). Four chains took, on average, around 5 hours to run. One chain takes around one day to run using the authors' algorithm (duration reported by them and corroborated by our own use of their code). It is worth noting that we use a significantly smaller sample size than them (700 subjects as opposed to 2000). However, even when we used larger samples (e.g., the implementations from the next section used a sample of 1500 subjects), two chains ran in less than 6 hours. It is also worth mentioning that this setting appears to be greatly dependent on the availability of strong and reliable prior information. If slightly less informative priors than those suggested by the authors were used, computational times increased significantly.

Given that the models used on the CAV data are in many ways analogous (only simpler) to the dementia progression model, it is our belief that future work dedicated to tuning and implementing the MCSA *Stan* development could prove fruitful.

5.2 Extending the Model: Using P-spline-based Emission Distributions

Sometimes simple parametric distributions are inadequate to fully capture the shape of the empirical state-dependent distributions. Langrock et al. (2018) argue that ignoring such possible lack of fit can lead to various undesirable consequences, such as: poor predictive performance; unreliable state decoding when emission distributions overlap for different states; invalid inference of possible covariate effects on the state-switching dynamics; and/or, invalid inference on the number of states. Even if said lack of fit does not meaningfully impact inference on other features of the process of interest, adequately estimating the state-dependent distributions and the information embedded in them might be of primordial interest. Being able to determine if a state-dependent distribution is, for example, skewed or has heavy tails can be of great value when studying biological processes. For that reason, we propose that the model developed by Williams et al. (2020) can be further refined with semi-parametric inference of the state-dependent distributions. In particular, with the

use of P-splines (i.e., penalized B-splines) for estimating the emission distributions.

For the specific case of the MCSA data, Figure 6 shows that the emission distributions overlap significantly (the same applies to what Williams et al. (2020) obtained for the real data). In such a case, choosing the right parametric distributions may prove challenging. Moreover, observing the empirical distributions for each of the states is simply not possible without first fitting the model if the states are not known. Even assessing if assuming that both groups share the same variance can prove difficult. To overcome this, as suggested in Langrock et al. (2015), one alternative is to use a polynomial spline written as a linear combination of (fixed) B-spline basis functions. Thus, we propose that Williams et al. (2020) model can be extended by formulating each of the 4 state-dependent distributions used for $\log(PIB-1)$ and *thickness* emissions using cubic B-splines. Note that we do not consider $MMSE$ emission distributions because the distributions depend on subject specific covariates, and thus it may be more appropriate to use a parametric setting. If access to the real data were possible, an assessment on whether or not to use a penalised B-splines formulation could prove valuable. For $\log(PIB-1)$ and *thickness* the proposition is to substitute the Gaussian emission distributions used, such that for subject i at visit n, the distribution is specified as

$$f(y_{i,n,r} \mid s_{i,n} \in g) = \sum_{\rho=1}^{m} \omega_{\rho,g}\phi_{\rho,g}(y_{i,n,r}),$$

with $r = 1, 2$, $\log(PIB-1)$ and *thickness* respectively; $g = g_1, g_2$, the groups of states for which the burden level is low and high respectively (e.g., when $r = 1$, $g_1 = \{1, 3, 5\}$); and $\phi_{1,g}, \ldots, \phi_{m,g}$, the set of m regularly spaced cubic B-splines.

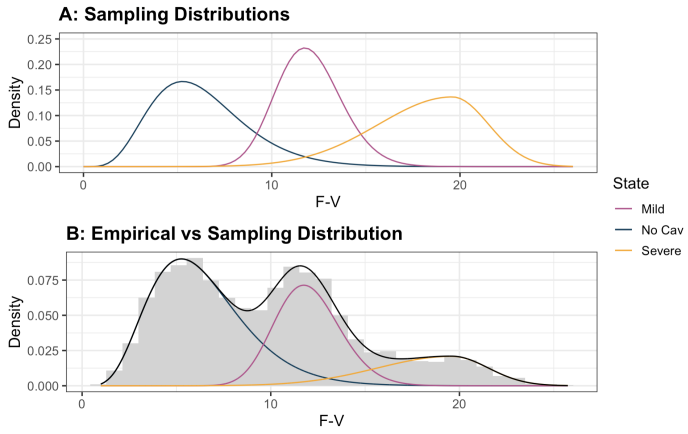


Figure 7: *F-V* state-dependent distributions and observed distribution. The color indicates which active state corresponds to the associated distribution emitting the responses. A) shows the densities of the state-dependent distributions used to simulate the data. B) contains a histogram of the *F-V* observed distribution, along with the mixture (in black) of the true state-dependent distributions weighted by the proportion of observations coming from each of the three states.

To demonstrate the effect using a spline-based formulation can have, we carried out a simple simulation study using the CAV data. We simulated one data-set with 1500 subjects the exact same way as described in §4.1, but with the assumption that a second fictitious response is observed. The assumption is that this second response variable (let us call it *F-V*) is some kind of biomarker measurement that somehow associates with the CAV condition of a patient (analogous to the biomarkers observed for dementia diagnosis). At each visit, a value of *F-V* is observed in accordance with the active state. When the state is *no CAV*, *F-V* is sampled from a $Gamma(6, 0.95)$; when

it is *mild CAV*, a $Gamma(48, 4)$ is used; and if *severe CAV*, values are simulated from a skewed normal distribution (Fernández and Steel, 1998) with mean equal to 18, standard deviation of 3 and skewness parameter of 0.7. Figure 7 shows these state-dependent distributions (top panel). It also shows (bottom panel) the empirical distribution for *F-V*, along with the sampling distributions (and their mixture) weighted by the proportion of the corresponding active states across all observations in the data.

Note that there is significant overlap between all three distributions, and that two of them (corresponding to states 1 and 3) are rather skewed). In fact, an *F-V* measurement of 13 has the same (considerably high) density when in state 1 than when in state 3. If something like this happened in reality, and if *F-V* was a measure used to determine the level of CAV in a patient, it could be paramount for physicians to adequately capture the fact that antagonistic states are almost equally likely to produce some of these rather likely values.

To analyze how a spline-based formulation can perform in a context like the one described above, we fitted the HMM described in §4.1 using two different formulations to model the added response variable. One is analogous to that used by Williams et al. (2020) on the MCSA data. i.e., the state-dependent distributions are assumed Gaussian, with shared variance and monotonically non-decreasing means ($\mu_1 \leq \mu_2 \leq \mu_3$). The second formulation utilizes P-splines as we propose, using 11 regularly spaced B-splines. Hence, for $s = 1, 2, 3$ the state-dependent distributions are of the form

$$f_s(y) = \sum_{\rho=1}^{11} \omega_{\rho,s} \phi_\rho(y).$$

As done in Langrock et al. (2018), in order to ensure that $f_s(y)$ is a proper probability density function, the splines are standardised so that they integrate to 1. Similarly, the weights $\omega_{1,s}, \ldots, \omega_{11,s}$ are parameterised using the softmax function:

$$\omega_{\rho,s} = \frac{\exp(\alpha_{\rho,s})}{\sum_{j=1}^{11} \exp(\alpha_{j,s})},$$

such that they are non-negative and sum to one, and the coefficients $\alpha_{\rho,s}$ are estimated without constraints. To make the number and position of the knots used less critical, penalized smoothing (hence the name P-splines) is enforced through the use of third order random walk priors on the coefficients (Fahrmeir et al., 2010). i.e., $\alpha_{\rho,s} \sim N(3\alpha_{\rho-1,s} - 3\alpha_{\rho-2,s} + \alpha_{\rho-3,s}, \sigma_a^2)$, for $\rho = 4, \ldots, 11$, and $p(\alpha_{1,s}, \alpha_{2,s}, \alpha_{3,s}) \propto const.$

Figure 8 compares the estimated distributions resulting from fitting the model using both approaches, along with the true sampling distributions and the empirical distribution. Looking at the plots, it is clear that the Gaussian approach (A) fails to capture appropriately the shape of the state-dependent distributions and so the correct probabilistic relationship between the values of *F-V* observed and the underlying active states. On the other hand, using P-splines (B) seems to attain a significantly better fit of the true state-dependent distributions. The estimated curves overlap almost completely with the true state-dependent distributions used to generate the data. Moreover, the estimated mixture using the spline-based approach follows much closer the empirical distribution of the data.

This is a simple artificial example. But, it serves to show that when the data generating process is of certain level of complexity, inadequate parametric models can fall short at fitting the true shape of emission mechanisms. Also, that using a spline-based model formulation can prove rather useful

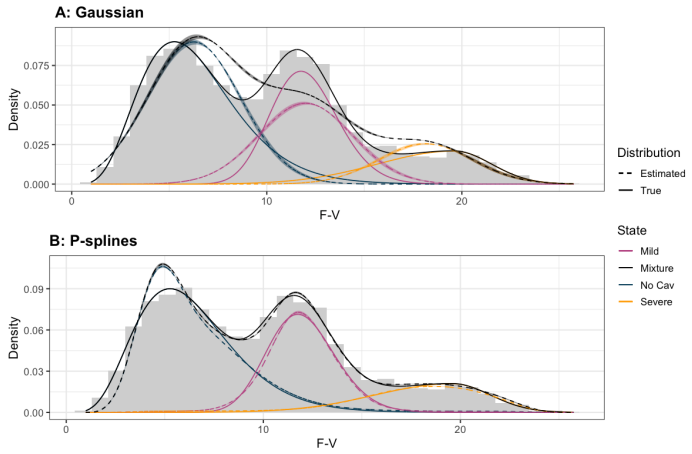


Figure 8: Estimated (dashed lines) and true (solid lines) F-V state-dependent distributions compared with the observed distribution, by model formulation. The estimates use the posterior means of the parameters that characterize them. 95% credible intervals are indicated by the shaded borders. State-dependent distributions are weighted by the true proportion of observations with the corresponding state active. A) shows the Gaussian formulation estimates. B) contains the distributions estimated using the P-splines approach proposed.

to better infer the underlying structure of the emission process. Even if a spline based formulation is not used for final inference and analysis, it can be used for model construction or validation. Knowing that semi-parametric approaches provide more flexibility, complex models (as the one Williams et al. (2020) use for dementia progression) can benefit from P-splines formulations, using them to compare/assess the results obtained with parametric settings. Looking at the emission distributions estimated using splines can also shred some light on which parametric setting to use (or not to use). e.g., in this case one could use the P-splines formulation to reckon that variance may not be the same for all three distributions, and to derive that a parametric model allowing for marked left and right skewness should be considered.

6 Discussion

We reviewed Williams et al. (2020), a paper where a continuous-time HMM was developed for the analysis of the MCSA data. We presented the fundamentals of the methodologies used, and detailed the steps taken to make the model as realistic of an abstraction of the actual data generating process as possible. This included handling issues inherent to enrollment population studies such as *delayed enrollment* and *death rate bias*. We also described the Bayesian computational framework developed to facilitate the use of critical prior information on many of the parameters. Via simulation studies we validated some of the most relevant arguments made by the authors, and replicated their results using synthetic data similar to the MCSA data. Additionally, using *Stan*, we provided and alternative MCMC implementation to that used by the authors. Lastly, we proposed an extension to make their model more flexible and accommodate complex emission mechanisms of the observed responses. Some relevant findings were that (1) evidence suggests that *delayed enrollment* and *death rate* biases are in fact elements that should be considered; (2) simulation results suggest the model developed by the authors is able to produce insightful results; (3) it appears that using *Stan* to fit their model could have major practical advantages; and (4) using semi-parametric emission distributions can lead to valuable cues about the true nature of

complex processes studied through HMMs.

We believe that a continuous-time HMM is a natural approach to analyse the MCSA data and model the phenomenon of dementia progression. The model Williams et al. (2020) developed encompasses a well thought framework, greatly informing many aspects of what is likely close to the true biology of the process. The Bayesian framework, as proposed by them, allows to easily carry out inference while accounting for such additional information. However, it is important to highlight that the various constraints placed within their model are rather stringent because they reflect expert domain knowledge. Choosing such impositions should always be a matter of caution, and result from reliable information. Similarly, prior specifications are, as noted in §5.1, heavily reliant on the availability of strong prior information. When in doubt, pertinent sensitivity tests should be carried out.

One drawback of this work is that, due to privacy restrictions, the real data is not available. For that reason, the scope of the study is limited to the kind of analyses that can be carried out on synthetic data, much of which entail results validation. The results obtained from this study are consistent with those reported by the authors, and mostly provide verification and credibility to their work. Nonetheless, if the real data became available, various and more comprehensive analyses could be carried out. Additional aspects, such as the group of covariates chosen by the authors, could be assessed, and the adequacy and optimal parsimony of the model could be more definitively established. Also, further validation steps, such as prior and posterior predictive checks, which are of great relevance when conducting Bayesian inference, and which the authors did not carry out profoundly (or at least did not report on), could be conducted. We believe this should be a subject of future work.

Finally, we extended the work made by the authors. We showed that using semi-parametric formulations, such as P-spline-based state-dependent distributions can yield better results than inadequate parametric settings. We believe that carrying out analyses under this framework, or at least incorporating it during the model formulation stages, can considerably help avoid the consequences of ill-fitted emission distributions. More importantly, our initial implementation of the models into *Stan* will facilitate collaboration and can likely lead to more efficient computations. Accordingly, a natural next step is to sharpen and conclude such efforts. They could facilitate deeper explorations of the particularities of dementia progression and the MCSA data.

References

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., DeKosky, S. T., Gauthier, S., Selkoe, D., Bateman, R., Cappa, S., Crutch, S., Engelborghs, S., Frisoni, G. B., Fox, N. C., Galasko, D., Habert, M.-O., Jicha, G. A., Nordberg, A., Pasquier, F., Rabinovici, G., Robert, P., Rowe, C., Salloway, S., Sarazin, M., Epelbaum, S., de Souza, L. C., Vellas, B., Visser, P. J., Schneider, L., Stern, Y., Scheltens, P., and Cummings, J. L. (2014). Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria. *The Lancet Neurology*, 13(6):614–629.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20:203–219.

Fernández, C. and Steel, M. F. (1998). On bayesian modeling of fat tails and skewness. *Journal of the american statistical association*, 93(441):359–371.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

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., and Petersen, R. C. (2016). Transition rates between amyloid and neurodegeneration biomarker states and to dementia: a population-based, longitudinal cohort study. *The Lancet Neurology*, 15(1):56–64.

Jackson, C. (2011). Multi-state models for panel data: the msm package for r. *Journal of statistical software*, 38:1–28.

Karlin, S. and Taylor, H. E. (1981). *A second course in stochastic processes*. Elsevier.

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D. L., and Papastamatiou, Y. P. (2018). Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, 72(3):179–200.

Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. L. (2015). Nonparametric inference in hidden markov models using p-splines. *Biometrics*, 71(2):520–528.

MCSA (2023). Mayo clinic study of aging. `https://www.mayo.edu/research/centers-programs/alzheimers-disease-research-center/research-activities/mayo-clinic-study-aging/overview` [Accessed: 2023-08-15].

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16.

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

Shelton, C. R. and Ciardo, G. (2014). Tutorial on structured continuous-time markov processes. *Journal of Artificial Intelligence Research*, 51:725–778.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.26.23.

Williams, J. P., Storlie, C. B., Therneau, T. M., Jr, C. R. J., and Hannig, J. (2020). A bayesian approach to multistate hidden markov models: Application to dementia progression. *Journal of the American Statistical Association*, 115(529):16–31.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series An Introduction Using R*. Chapman and Hall/CRC.

Appendix

A CAV Simulation Study Supplemental Material

Prior Means and Standard Deviations			
Transition rates parameters (see (12))			
Transition	$\beta_0^{(l)}$	$\beta_1^{(l)}$	$\beta_2^{(l)}$
all	–2 (1.1)	0 (0.11)	0(0.6)
Initial distribution			
π_1	π_2	π_3	
0.966 (0.1)	0.03 (0.1)	0.003 (0.1)	
Missclassification probabilities			
p_1	p_2	p_3	p_4
0.05 (0.1)	0.05 (0.1)	0.05 (0.1)	0.05 (0.1)
F-V response (§5.2)			
μ_1	μ_2	μ_3	σ
10 (5)	10 (5)	10 (5)	5 (2)
$\alpha_{1,s} - \alpha_{3,s}$	$\alpha_{4,s} - \alpha_{11,s}$	σ_a	
0 (0.5)	$3\alpha_{\rho-1,s} - 3\alpha_{\rho-2,s} + \alpha_{\rho-3,s}$ (σ_a)	0 (0.5)	

Table 2: Displayed are the means and standard deviations (in parentheses) of the (independent) normal priors placed on the CAV models' parameters. Note that probabilities and standard deviations are not re-parameterised. When using *Stan* parameter space constraints are indicated when the variables are defined, and are directly handled within *Stan*.

CAV data: log-rate iyears parameters

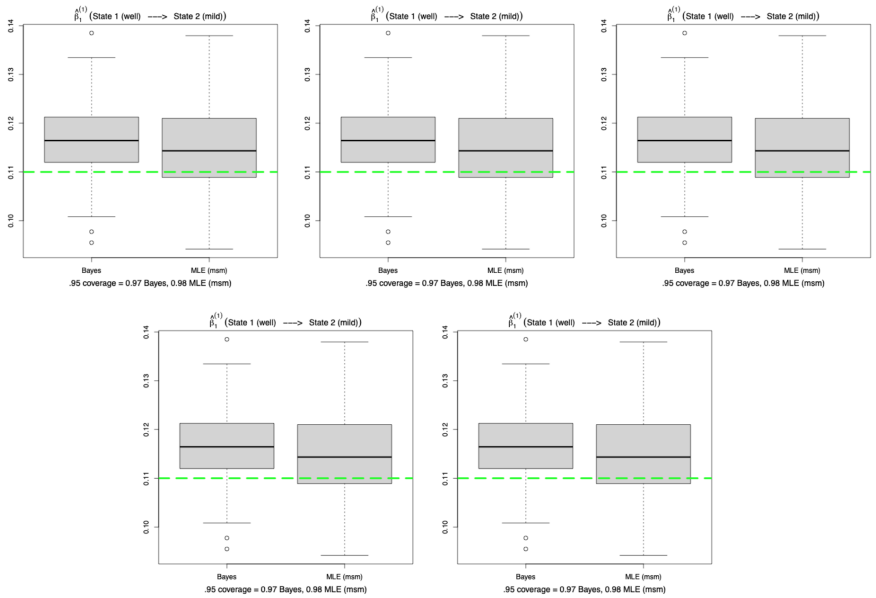


Figure 9: Synthetic CAV data (log) transition rates $\beta_1^{(l)}$ estimates. The plots are analogous to those in Figure 3, but for the estimates of the *iyears* coefficients.

CAV data *delayed enrollment*: log-rate iyears parameters

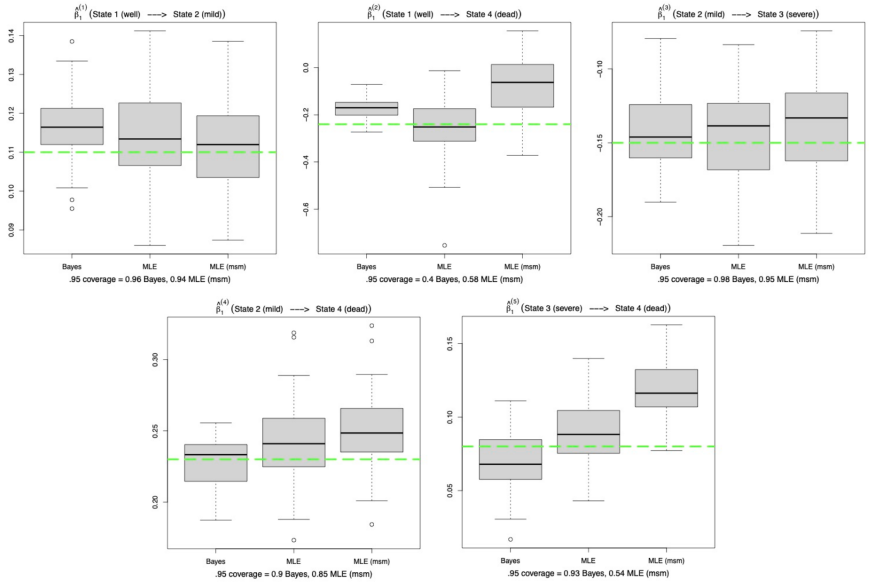


Figure 10: Synthetic CAV data (log) transition rates $\beta_1^{(l)}$ estimates. The plots are analogous to those in Figure 4, but for the estimates of the *iyears* coefficients.

CAV data: log-rate sex parameters

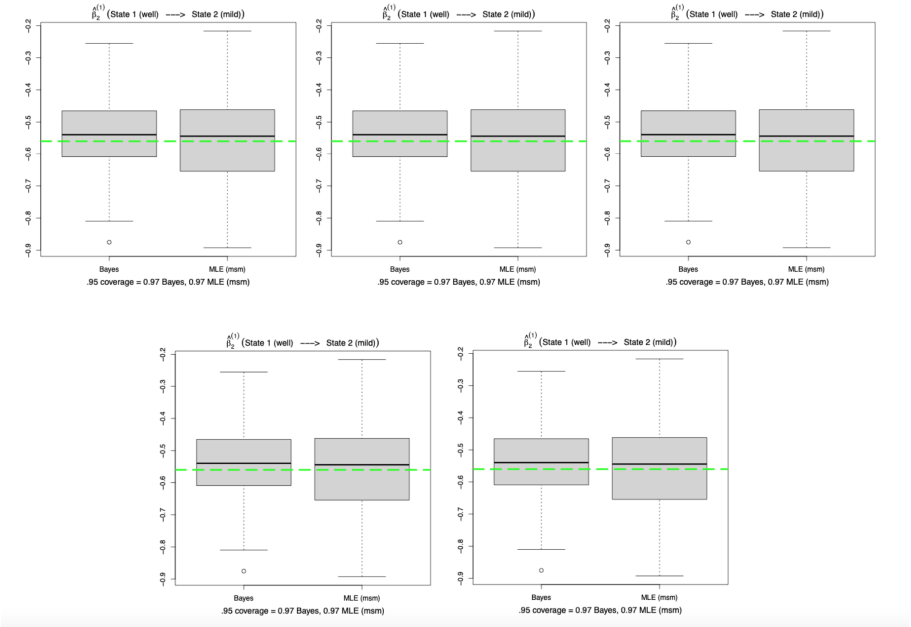


Figure 11: Synthetic CAV data (log) transition rates $\beta_2^{(l)}$ estimates. The plots are analogous to those in Figure 3, but for the estimates of the *sex* coefficients.

CAV data *delayed enrollment*: log-rate sex parameters

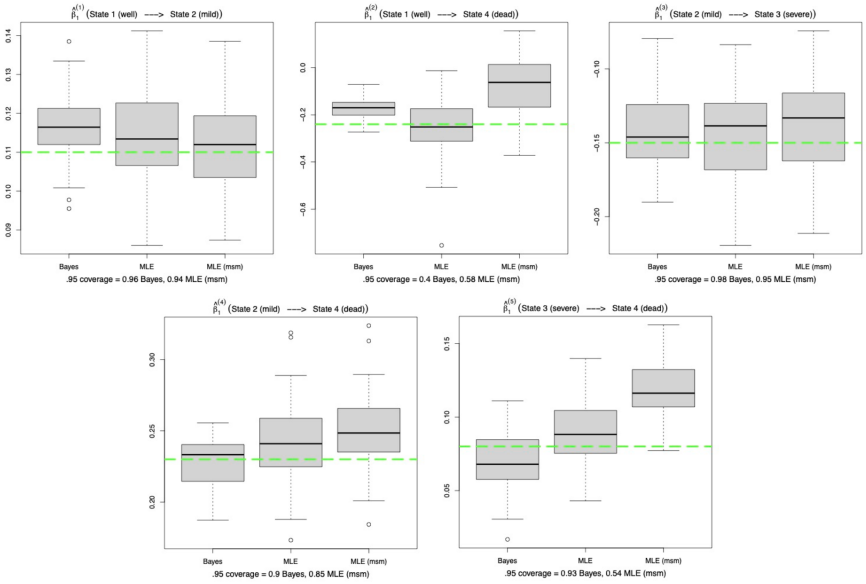


Figure 12: Synthetic CAV data (log) transition rates $\beta_2^{(l)}$ estimates. The plots are analogous to those in Figure 4, but for the estimates of the *sex* coefficients.

CAV data: misclassification probabilities

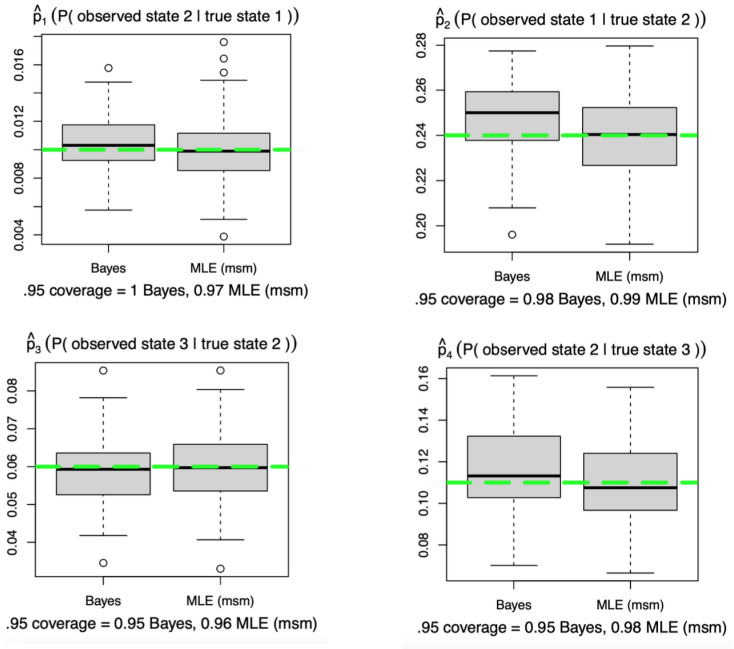


Figure 13: Synthetic CAV data misclassification probabilities estimates. The plots are analogous to those in Figure 3, but for the estimates of the state-dependent distributions' parameters.

CAV data *delayed enrollment*: misclassification probabilities

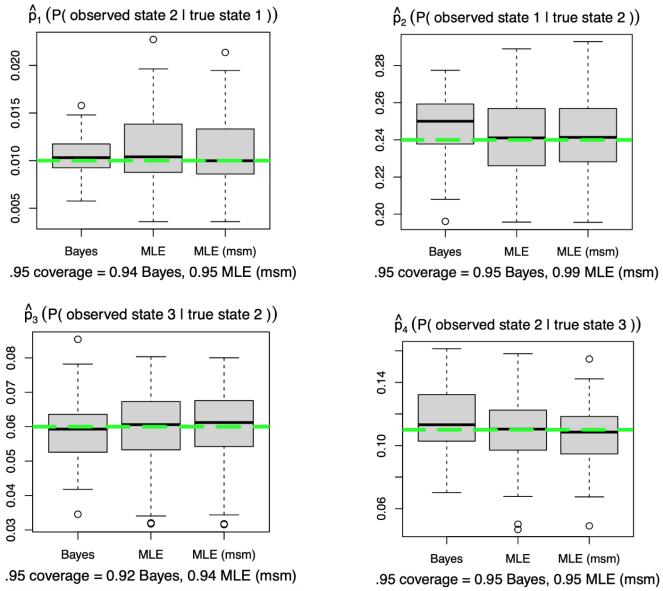
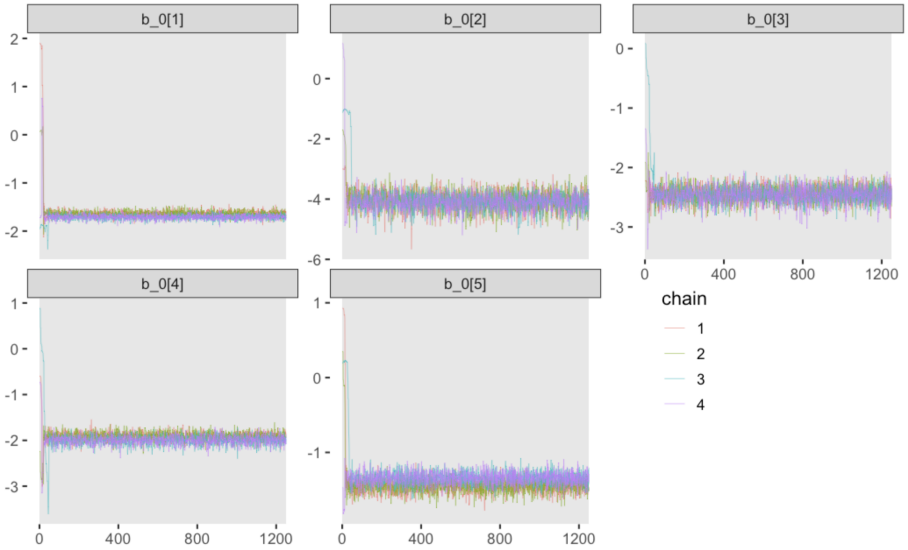


Figure 14: Synthetic CAV data misclassification probabilities estimates. The plots are analogous to those in Figure 4, but for the estimates of the state-dependent distributions' parameters.

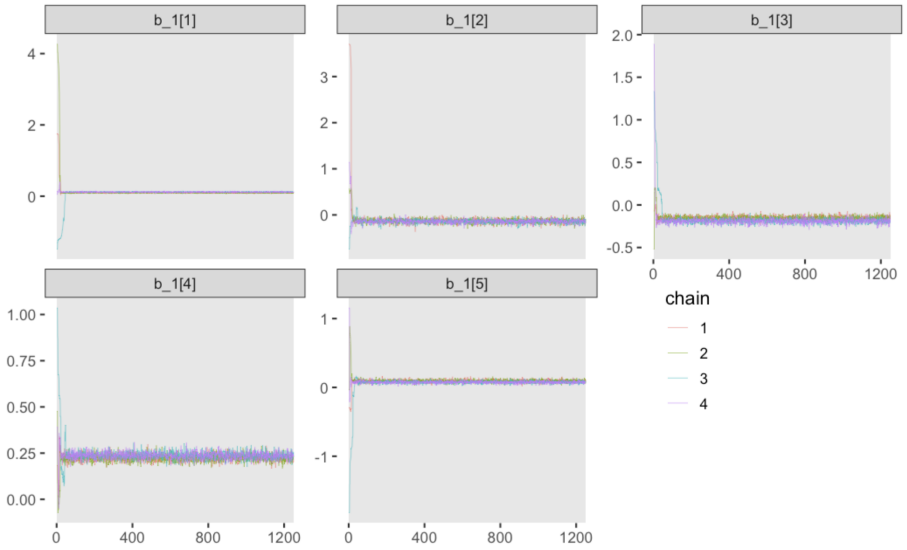
CAV data *delayed enrollment*: trace plots

The trace plots below were used to assess convergence of the HMC used for the CAV data. For conciseness sake, only the trace plots corresponding to the *delayed enrollment* implementation are shown. However, similar checks were performed for all models used in the study.

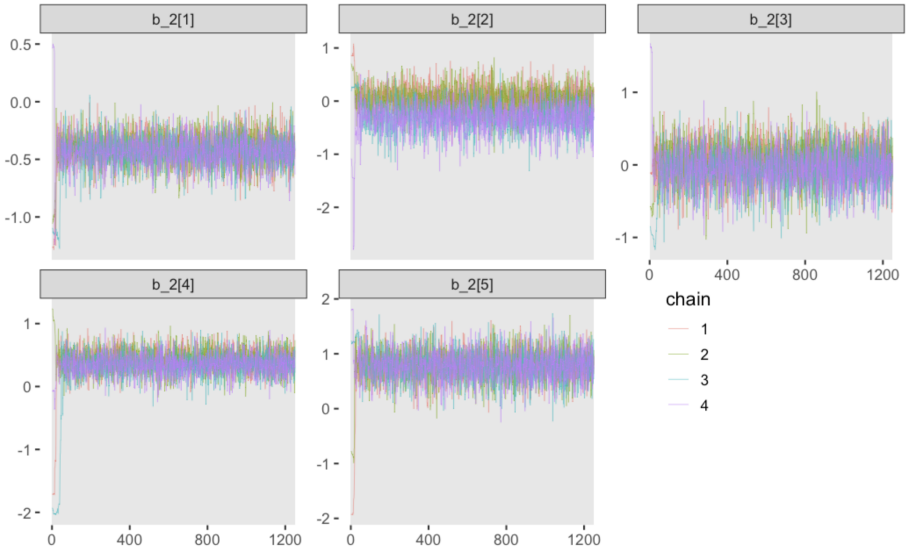
Intercepts ($\hat{\beta}_0^{(l)}$)



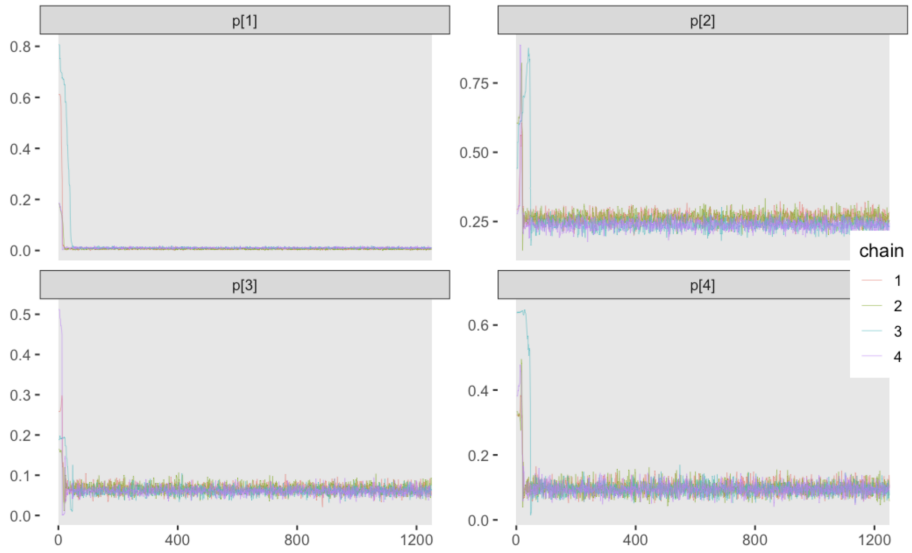
Log-rate iyears parameters ($\hat{\beta}_1^{(l)}$)



Log-rate sex parameters ($\hat{\beta}_2^{(l)}$)



Misclassification probabilities ($\hat{p}$)



B MCSA Simulation Study Supplemental Material

Prior Means and Standard Deviations

Transition rates parameters (see (8) and (10))

Transition	$\beta_0^{(l)}$	$\beta_1^{(l)}$	$\beta_2^{(l)}$	$\beta_3^{(l)}$	$\beta_4^{(l)}$	c	$c+d$
non-dem $\to$ 7	–4.41 (0.1)	0.094 (0.01)	0.47 (0.05)	0 (0.1)	0 (1)	–0.75 (0.375)	–0.60 (0.3)
5 $\to$ 7 and 6 $\to$ 7	–4 (1)	0.1 (0.05)	0 (1)	0 (0.1)	0 (1)		
all others	–3 (1)	0.1 (0.05)	0 (1)	0 (0.1)	0 (1)		

Cubic splines weights

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

μ_{A-}	μ_{A+}	$\log(\sigma^2_{pib})$
–1.3 (0.2)	–0.5 (0.2)	$\log((0.4/3)^2)$ (2)

Thickness response

μ_{N-}	μ_{N+}	$\log(\sigma^2_{thick})$
3.14 (0.2)	2.34 (0.2)	$\log((0.4/3)^2)$ (2)

MMSE response

$\alpha_1 - \alpha_4$	$\alpha_5 - \alpha_6$	α_7	α_8	α_9	α_{10}	α_{11}	$\log(\sigma^2_{mmse})$
–0.28 (0.75)	–7.3 (3)	0 (1)	0 (1)	0 (1)	0 (1)	0 (1)	–0.7 (2)

Dementia misclassification

logit(p_0)	logit(p_1)
–3 (1)	–3 (1)

Initial distribution

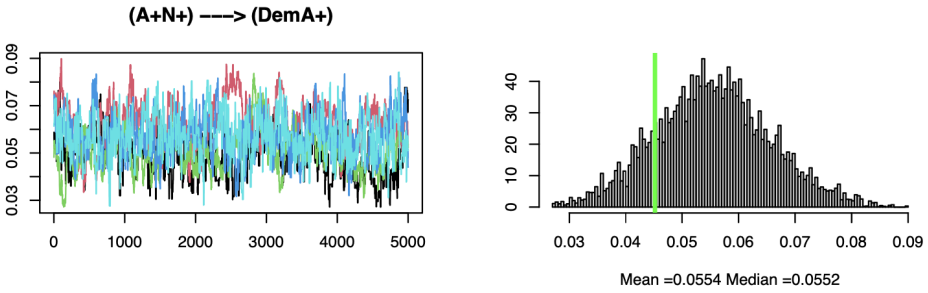
ξ_1	ξ_2	ξ_3
–3.5 (0.25)	–6 (1)	–6 (1)

Table 3: Displayed are the means and standard deviations (in parentheses) of the (independent) normal priors placed on the 81 model parameters. The parameters $c_1, \ldots, c_8$ correspond to the weights of the spline basis used to model the effect age has on the transition rate from state 1 to state 2 (see §3.5). ξ_1, ξ_2, ξ_3 are the logit parametrization of the initial distribution probabilities for states 2,3 and 4 respectively (1 is used for state 1).

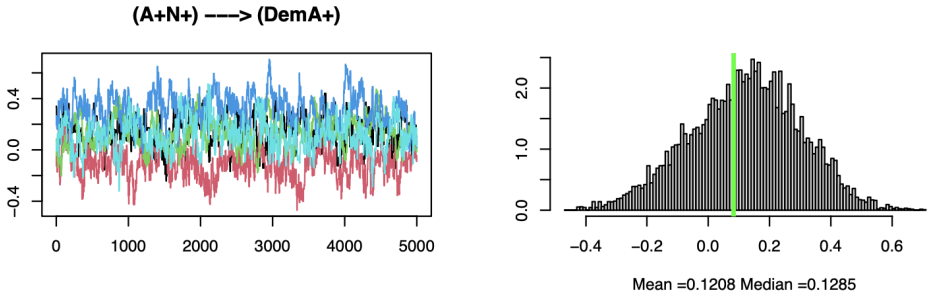
MCSA data: trace plots and posterior samples

The plots below are analogous to those shown in Figure 5, but for other Alzheimer-related parameters. They correspond to the trace plots and posterior samples from 5 different chains of the dementia progression model fitted to synthetic MCSA data. The model was fitted using the MCMC algorithm coded by Williams et al. (2020). For conciseness sake, similar plots are not included for the rest of the parameters, but they were produced, and can be found in the GitHub repository included in Appendix C, within the *Supplemental* folder.

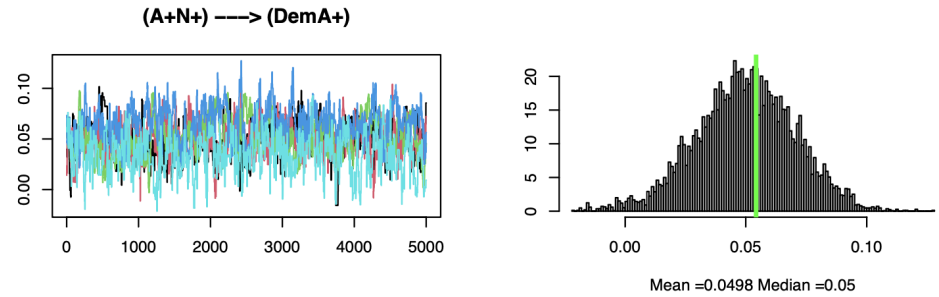
Log-rate age parameter ($\hat{\beta}_1^{(9)}$)



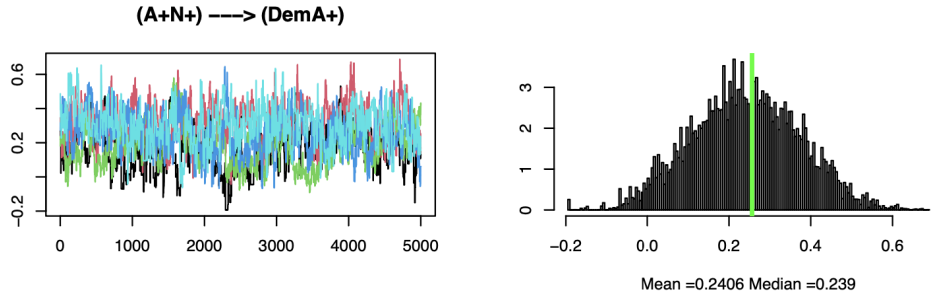
Log-rate sex parameter ($\hat{\beta}_2^{(9)}$)



Log-rate education parameter ($\hat{\beta}_3^{(9)}$)



Log-rate APOE-ϵ4 allele presence parameter ($\hat{\beta}_4^{(9)}$)



C Computational Work

All the computational work used for this report is publicly available in the following GitHub repository: `https://github.com/Esquivel-Arturo/Dementia_progression`.

Te repository is organized through three folders:

- *New_work* contains all of our original work made for this project. Some of the scripts contain code adapted from the work by Williams et al. (2020). When that is the case, the part of the script that was adapted is indicated.

- *Supplemental* contains results like those shown in Appendix B for every parameter in the dementia progression model.

- *TimeCapsule_Code* contains all the work done by Williams et al. (2020), as made publicly available in `https://jonathanpw.github.io/research.html`. All the material contained here is either unchanged, or minimally adapted to be used for this study.

The scripts were used for this study as follows:

- *Simulate_cav.r* (*New_work*) was used to generate all synthetic CAV data-sets.

- *fit_cav_models.r* (*New_work*) was used to fit all CAV models whose results are shown in Figures 4 and 3.

- *delayed_enrollment_plots.r* (*New_work*) was used to produce plots as the ones shown in Figures 4 and 3.

- *Simulate.r* (*TimeCapsule_Code*) was used to create the MCSA synthetic data-set used for §4.4.

- *RunFile.r* (*TimeCapsule_Code*) was used to implement the dementia progression model.

- *OutFile_figures.r* (*TimeCapsule_Code*) was used to produce the Figures in §4.4, and the material contained in *Supplemental*.

- *cav.stan* and *mcsa.stan* (*New_work*) contain the CAV and MCSA models respectively for implementation using *Stan*.

- *fit_ext_models.r* (*New_work*) was used to produce the results from §5.2.