Organization of supplementary material

Supplementary Material A

A.1 Table of MCSA HMM transition rate constraints

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).

A.2 Details about the synthetic data generation

The purpose of this section is to describe how the synthetic data was generated to simulate both the MCSA and modified CAV data sets. The same strategies were used to simulate both data sets, but the MCSA involved certain additional features to simulate.

Subject covariates such as 'sex/male', 'educ', and 'apoe4' are randomly chosen in proportions consistent with those of the empirical distributions from the actual data set. To simulate ages at first observation for the MCSA, for each subject we sample an age from the empirical distribution of all ages observed at first visit in the actual data set. The two scenarios for simulating years enrolled at first visit in the CAV data is described in the primary manuscript. The true state at initial observation is generated from the initial state probability vector.

Next, the true underlying state sequence at each instance 'dt' of time for each patient is generated by computing the row of the infinitesimal generator matrix $\boldsymbol{Q}$ corresponding to a subject's current state, and then sampling waiting times (i.e., realizations of independent exponential random variables), one for each nonzero rate parameter in the row of $\boldsymbol{Q}$. If the minimum of these sampled waiting times is smaller than 'dt' then the subject transitions to the state corresponding to the minimum waiting time, or else does not transition at that instant. This procedure is repeated until the subject transitions to dead. A caveat is that the defined instant of time, 'dt', must be much smaller than the scale on which the rates change as a function of age, or else the use of exponential waiting times is not appropriate. To generate the synthetic data sets for this paper dt $:= 1/365$, but we note that dt $:= 1/12$ results in very similar data sets. In fact, the less fine discretization (dt $:= 1/12$) was used for our preliminary analyses.

Once the true underlying state sequence has been generated the observed clinical visits are generated as follows. The simulated age a each subsequent clinical visit (after entry age) is generated by sampling an inter-observation time from the empirical distribution of all (non-death) inter-observation times from the actual data set, and adding the sampled inter-observation time to the current age. This process of generating clinical visits and corresponding ages is repeated until the subject transitions to dead or exceeds 12 years in the study for the MCSA (20 years for the CAV), whichever comes

first. The one exception is that transitions to dead are recorded using the exact time of transition to dead (from the simulated true state sequence described above). Additionally, prior to each sampled inter-observation time for the synthetic MCSA data there is a small Bernoulli probability that the subject leaves the study. Note that 12 years is approximately the maximum duration observed in the actual MCSA data. Details about simulating the various response functions are provided in the primary manuscript.

Lastly, note that the discretization of time, i.e. $\mathrm{dt} = 1/365$, only applies for simulating synthetic data. In fact, for our theorized HMM, and the estimation procedure, time is truly treated as continuous with no discretization. However, this means that we had to impose the assumption that the generator matrix, $\boldsymbol{Q}$, is constant over certain periods of continuous time. After some consideration both of the biology and of computational limitations, it was decided that assuming the transition rates as a function of age are constant between birthdays is reasonable. The buildup of amyloid plaques on the brain, the loss of cortical thickness, the development of dementia, and dying are processes that take decades/lifetimes to develop. Accordingly, rates of transition amongst our state space are unlikely to change by an estimable amount in less than a years' time.

A.3 Details about the MCMC proposal strategy

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 $\mathrm{N}(\boldsymbol{\theta}^{(t)}, \tau \cdot \Sigma)$, 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.