1. Supplementary multi-state model statistical methods

Most of the standard methods for analyzing time-to-event data (e.g., survival analysis, Cox models, and person-years methods) assume that the event times are known and thus are not applicable to data that were ascertained at scheduled intervals such as we have in the Mayo Clinic Study of Aging (MCSA). In our study we do not know the exact age at which a participant actually transitions from A−N− to A+N−, for example, only that the transition occurred sometime in the interval between their last A−N− visit and first A+N− one. For such data we use the multi-state Markov (MSM) model described below. The Markov assumption states that the rate of transition to a new state depends only on the current state and age. The Markov assumption — or "memoryless" property — is restrictive, but necessary in general to compute a likelihood for intermittently observed data.[1] A natural alternative to making this memoryless assumption would be to include time in the current state as an additional predictor in the model. Unfortunately, with intermittently observed states, we do not actually know how long a participant has been in their current state, but only a range of possible values. However, the apparent restrictiveness of the Markov assumption is ameliorated to some measure in the current study by the presence of age as a central covariate: rates rise during the state due to aging.

We label the six states in our study as 1=A−N−, 2=A+N−, 3=A−N+, 4=A+N+ (all non-demented), 5=dementia, and 6=death and let v be a six-element vector representing the initial frequency distribution over the states at age 50. Since being alive and not demented are enrollment criteria for the MCSA, we know that $v[5] = v[6] = 0$. The rates matrix, or transition intensity matrix, $R_{a,s}$ for each integer age a and sex s is a 6 by 6 matrix describing the transition rates that apply to participants over that year. We note that the R matrix depends on sex only because of sex-specific death rates. All other transitions rates are not different by sex. Our rates matrix was of the form

$$R_{a,s} = \begin{pmatrix} & r_{12} & r_{13} & 0 & 0 & d_1 M_{a,s} \\ m & & 0 & r_{24} & 0 & d_1 M_{a,s} \\ m & 0 & & r_{34} & r_{35} & d_1 M_{a,s} \\ 0 & m & m & & r_{45} & d_1 M_{a,s} \\ 0 & 0 & 0 & 0 & & d_2 M_{a,s} \\ 0 & 0 & 0 & 0 & 0 & \end{pmatrix}$$

Here the ijth element of $R_{a,s}$ represents the rate for the transition from state i to state j. The rates for the transition from states 1–4 to death are denoted by $d_1 M_{a,s}$ and the rates for dementia to death are denoted by $d_2 M_{a,s}$, where $M_{a,s}$ is the tabulated Minnesota death rate for that age and sex.[2]

As shown in the rate matrix above, four "backward transitions" are permitted in the model. These transitions include A+N− to A−N− (state 2 to 1), A−N+ to A−N− (state 3 to 1), A+N+ to A+N− (4 to 2), and A+N+ to A−N+ (4 to 3). We assume a common misclassification rate parameter denoted by m; this is assumed to be a technical misclassification error and so this rate does not depend on age.

Rates of 0 in the above rate matrix correspond to transitions that are not allowed in our model. State 6 (death) is an absorbing state and "backward" transitions from death are clearly not possible. Nor do we allow in our model transitions from state 5 (dementia) to any of the biomarker states 1–4. We note that direct transitions between biomarker states where both biomarkers change simultaneously (i.e., state 1 to 4, 2 to 3, 3 to 2, or 4 to 1) are not allowed as the odds of both happening at the exact same time are infinitesimally small. For example, any participant who was observed to be A−N− at one visit and A+N+ at the next is assumed to have passed through an intermediate state of either A+N− or A−N+.

Rates denoted by r_{ij} in the matrix above are of the form $\exp(\beta_{ij,0} + \beta_{ij,1}\text{age})$ with the exception that $r_{12} = \exp(\beta_{12,0} + \beta_{12,1}\text{age} + \beta_{12,2}\text{age}' + \beta_{12,3}\text{age}'')$. Here the single and double prime symbols denote functions of age required for a cubic spline model with knots at 55, 65, 75, and 90 years.[3] The diagonal elements of $R_{a,s}$ are chosen so that the rows sum to 0, a technical requirement of rate matrices. The values of m, d_1, d_2, and the 14-element vector β are the parameters of the model.

In a multi-state model, the state transition matrix gives the probability of transitioning between any two states and is a function of the rate matrix. The sex-specific state transition matrix $T_s(a, a+1)$ for age a to $a+1$ is $\text{expm}(R_{as})$, where expm( ) is the exponential of a matrix. (Evaluation of the expm function requires special software as it can be numerically difficult to evaluate; see for instance the documentation and references for the expm package within R.) The T_s matrix depends on sex only because transition rates to death vary by sex. The transition matrix between any two ages is a product, e.g. $T_s(67.4,70.1) = \text{expm}(0.6R_{67,s})T_s(68,69)T_s(69,70)\text{expm}(0.1R_{70,s})$. The state distribution vector at age a will then be $p_{a,s} = vT_s(50, a)$. A participant of sex s who was enrolled in state 1 at age 72.1 with two follow-up observations of state 3 at age 73.6 and state 3 again at age 74.8 contributes three terms to the likelihood:

- entry: $p_{72.1,s}[1]/(p_{72.1,s}[1] + p_{72.1,s}[2] + p_{72.1,s}[3] + p_{72.1,s}[4])$
- first transition: $T_s(72.1, 73.6)[1, 3]$ (the 1,3 element of the transition matrix)
- second transtion: $T_s(73.6, 74.8)[3, 3]$

The second two terms are identical to the calculations done by the R package msm.[1] The manual for that package provides much more information about the formulas above, along with details about special handling of the death state. The contribution of the first entry term in our model reflects that the MCSA is a population-based sample and hence the initial enrollment state of a participant provides direct information about the population prevalence.

Implicit in the discussion above, yet still worth pointing out directly, is that the transition matrix can be calculated between any two (unrounded) ages and therefore accommodates variation in the time between participant visits.

The table below shows the set of transitions for all 12,160 paired visits in the data set. A participant with four visits, for instance, would contribute three pairs to the table. Seventy percent of the pairs (8,536/12,160) are on the diagonal, i.e., at their subsequent visit participants were in the same state as they had been at their prior visit. Non-demented participants who have a clinical visit without imaging are only known to be in the non-demented state, that is, one of {A−N−, A+N−, A−N+, A+N+}. In this case the likelihood contribution involves a sum over states 1–4; for mathematical details we again refer the reader to the msm package[1]. A participant who had a non-imaging visit between two imaging visits would contribute to both column 1 and row 1 of the table. Participants who had only one visit are shown in the last column of the table.

Prior visit state	Subsequent visit state							
	Non-demented	A−N−	A+N−	A−N+	A+N+	Dementia	Death	Participants with a single visit
Non-demented	8082	375	213	210	314	217	780	0
A−N−	508	58	18	16	1	1	4	177
A+N−	232	3	50	0	7	0	8	22
A−N+	221	4	0	42	10	10	5	7
A+N+	256	0	2	3	87	19	24	3
Dementia	0	0	0	0	0	217	163	0

Supplementary Table 1. Prior visit state versus subsequent visit state for visit pairs

Because the rows of the T matrix must sum to 1 (everyone has to go somewhere) all of the rates are interlinked, and every observation in the data set contributes to the estimate of all of the rates. The amount and focus of the contribution can nevertheless be very disparate. For instance, a pair of visits that were a very short interval apart without a change in state is nearly uninformative since that is exactly what we would expect for any but extreme rate values. A participant who transitions from non-demented with an unknown biomarker state to demented gives direct information about dementia rates but has only minimal influence on estimated rates within states 1–4. A substantial number of such transitions however, as seen in this data set, will have a definite impact and a joint fit that uses all of the data is essential. A non-demented participant who has only a single visit without imaging provides no information since they provide no information on transition rates nor cross-sectional prevalences; these were the only observations omitted from the analysis.

The model parameters were estimated via maximum likelihood using the optim function in R. The estimated variance-covariance matrix was obtained from the inverse of the negative of the returned Hessian matrix. Confidence intervals for derived parameters such as the transition rates, differences in rates, and predicted frequencies were based on a parametric bootstrap, which drew 10,000 coefficient sets from a multivariate normal distribution, centered at the maximum likelihood estimate with variance equal to the estimated variance-covariance matrix.

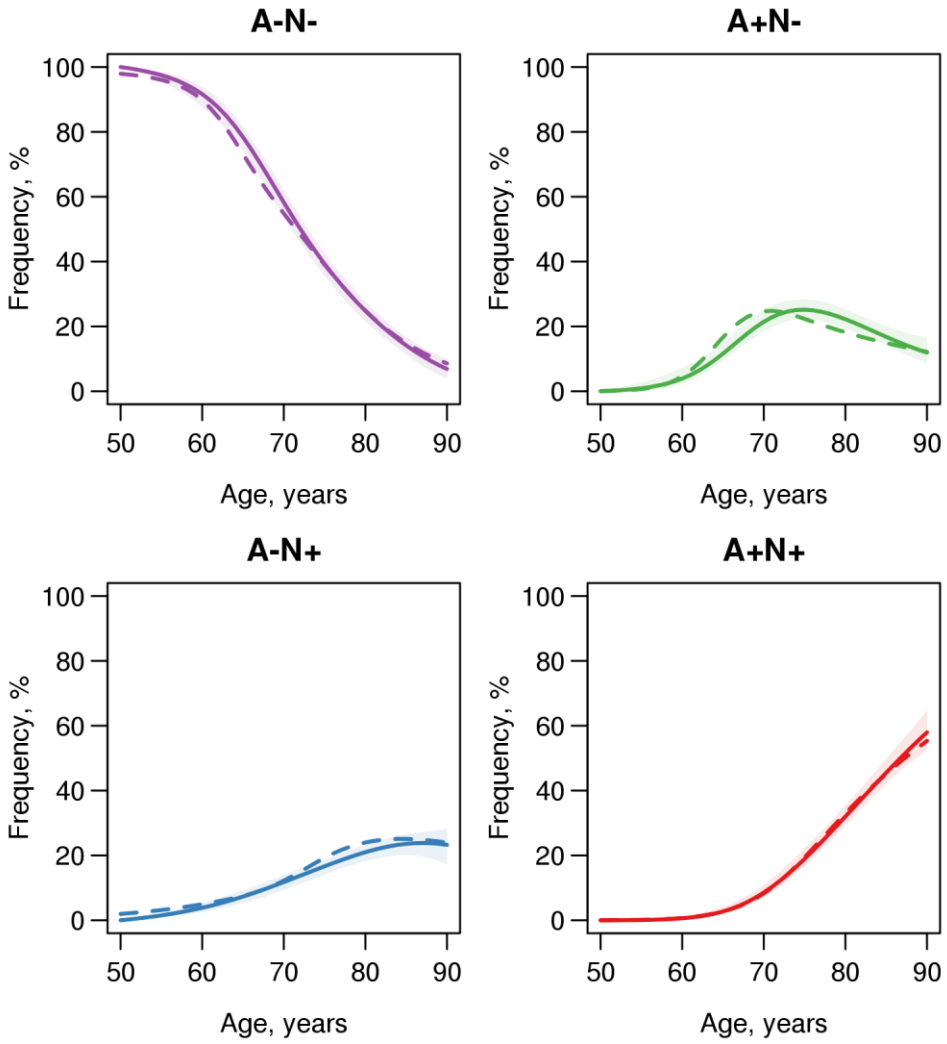
The transition rates summarized in Figures 2-5 are obtained from elements of the rates matrix $R_{a,s}$ while the frequencies in Figures 6 are obtained from $p_{a,s}$. The frequencies shown in supplementary figure 1 below are from $p_{a,s} = \nu T_s(50, a)$ and are compared to estimates from a maximum likelihood fit of a multinomial model using baseline data from the current study with knots at 55, 65, 75, and 90, a method as previously described.[4]

2. Supplementary analyses

All participants included in the study were randomly drawn from the Olmsted County population but only a subset agreed to participate in imaging. To assess whether those who agreed to imaging had a different overall trajectory, an important potential bias, we performed a sensitivity analysis. We fit a Cox proportional hazards model to predict dementia by imaging status (imaged versus not imaged) among 3632 participants with clinical follow-up. Age was used for the time scale with the follow-up period defined as age at baseline visit until age at first visit with dementia or last follow-up. Imaged vs. non-imaged was included in the model as a time-dependent covariate. That is, participants with their first imaging visit occurring at some visit other than baseline were classified as non-imaged until their first imaging visit, and then classified as imaged thereafter. Participants who were never imaged or imaged at baseline were classified as non-imaged and imaged, respectively, for the entire follow-up period.

The risk of dementia was found to be very similar for those who completed imaging studies compared to those who did not. Specifically, participation in imaging was associated with an 11% increase in hazard of dementia (hazard ratio: 1.11, 95% confidence interval: 0.81- 1.52, p=0.52). That imaging participants have similar rates of progression to dementia, a key late-stage outcome in our model of disease progression, provides supporting evidence in favor of the representativeness of the overall biomarker course of the MCSA imaging sample.

3. Supplementary Figure



Supplementary Figure 1. Estimated biomarker frequency from the multi-state Markov model using all available data (solid lines) with 95% confidence intervals (shaded regions) compared to the estimated frequency from a cross-sectional multinomial model using only baseline data (dotted lines). Estimates of frequencies for the multi-state model were calculated from the transition rates among a cohort of non-demented participants all assumed to be A−N− at age 50. Estimates were re-scaled and plotted among participants who remain alive and non-demented. Confidence intervals for the estimates were obtained by first randomly generating 10,000 multivariate normal variates centered at the maximum likelihood estimates with the variance-covariance matrix equal to the inverse of the negative of the Hessian matrix. Age-specific rates, and then frequencies, were calculated for each of the 10,000 variates. The 95% pointwise CIs were calculated as the 2.5th and 97.5th quantiles of these simulated frequencies.

References

1 Jackson C. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*. 2011;**38**(8):1–28.

2 Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000.

3 Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

4 Jack CR, Wiste HJ, Weigand SD, et al. Age-specific population frequencies of cerebral β-amyloidosis and neurodegeneration among people with normal cognitive function aged 50-89 years: a cross-sectional study. *Lancet Neurol*. 2014 Oct;**13**(10):997–1005.