Supplementary Material

Please refer to website
`https://github.com/MarcelaCespedes/Bayesian_inference_on_neuroimaging`
for R code (including code for plots) used in analysis outlined in the manuscript *Comparisons of neurodegeneration over time between healthy ageing and Alzheimer's disease cohorts via Bayesian inference*.

1 Derivation of model

In this section, we compare competing models for neurodegeneration with respect to age. Competing models include; linear, quadratic, cubir and quartic configurations, see equations (1) to (4), where $(.)$ denotes the the linear predictor in model (1).

$$\begin{aligned} Y_{ij}|\mu_{ij},\sigma^2 &\sim N(\mu_{ij},\sigma^2) \\ \mu_{ij} &= \beta_{0i} + \beta_1 x_{MCI,ij} + \beta_2 x_{AD,ij} + \beta_{3i} StndAge_{ij} + \beta_{4i} StndAge_{ij} x_{MCI,ij} + \beta_{5i} StndAge_{ij} x_{AD,ij} \\ \beta_{ki} &= \beta_k + b_{ki} \qquad \text{for } k = 0,3,4,5 \\ \boldsymbol{b_i} &\sim MVN(\mathbf{0},\Sigma) \end{aligned} \tag{1}$$

$$\mu_{ij} = (.) + \beta_6 StndAge_{ij}^2 + \beta_7 StndAge_{ij}^2 x_{MCI,ij} + \beta_8 StndAge_{ij}^2 x_{AD,ij} \tag{2}$$

$$\begin{aligned} \mu_{ij} = (.) &+ \beta_6 StndAge_{ij}^2 + \beta_7 StndAge_{ij}^2 x_{MCI,ij} + \beta_8 StndAge_{ij}^2 x_{AD,ij} + \\ &\beta_9 StndAge_{ij}^3 + \beta_{10} StndAge_{ij}^3 + \beta_{11} StndAge_{ij}^3 x_{AD,ij} \end{aligned} \tag{3}$$

$$\begin{aligned} \mu_{ij} = (.) &+ \beta_6 StndAge_{ij}^2 + \beta_7 StndAge_{ij}^2 x_{MCI,ij} + \beta_8 StndAge_{ij}^2 x_{AD,ij} + \\ &\beta_9 StndAge_{ij}^3 + \beta_{10} StndAge_{ij}^3 + \beta_{11} StndAge_{ij}^3 x_{AD,ij} + \\ &\beta_{12} StndAge^4 + \beta_{13} StndAge^4 x_{MCI,ij} + \beta_{14} StndAge_{ij}^4 x_{AD,ij} \end{aligned} \tag{4}$$

In Bayesian statistics, model choice can be handled via the posterior model probabilities. These probabilities can be estimated straightforwardly via normalising the model evidences, and, as the model evidence provides an inbuilt penalty for model complexity, there is a preference for the model with the largest value (MacKay, 2003). In this work, the integrated nested Laplace approximation (INLA) (Rue et al., 2009) was used to approximate the model evidences, and the logarithm of these values are shown in Table S1.

2 Assessment of normality assumption for models

Below are histograms of the residuals, scatter and quantile-quantile plots for the Ventricle and Hippocampus models presented in the manuscript in expression (3). Refer to Section 3 of the manuscript for assumptions of linear mixed effects models. Figure S1 shows residuals from both models are approximately normal, despite our response values being in (0, 1) range. Refer to Section 2 of the manuscript, for further discussion.

Model fitted, linear predictor with	**Ventricle**	**Hippocampus**
Age (1)	-1231	-1510
$Age + Age^2$ (2)	-1206	-1129
$Age + Age^2 + Age^3$ (3)	-1178	-1101
$Age + Age^2 + Age^3 + Age^4$ (4)	-1150	-1073

Table S1: Log evidence for competing models with non-linear terms with respect to age. We can see that the model with the highest log-evidence for both ventricle and hippocampus models consists of Age as a linear term, which is expression (1) as this has the highest log-evidence values for both regions.

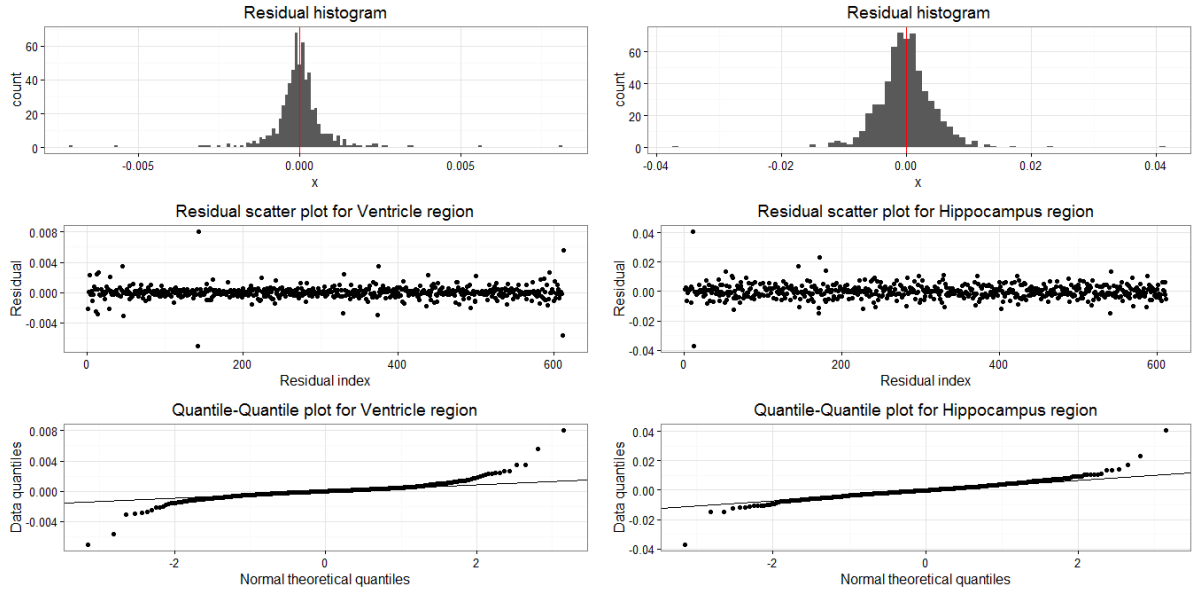


Figure S1: Ventricle (left) and hippocampus (right) assess linear mixed effect model for violation of normality assumptions.

3 Cross validation

We performed cross validation on the hippocampus and ventricle models described in the manuscript, to assess the predictive capability of a new observation. As discussed in Wang and Gelman (2014), there are no clear protocols for cross validation methods for multilevel models and out of sample validation methods for hierarchical models are not as straightforward as a random sample of the data for a holdout set. To that end, we carried out two approaches for leave-one-out cross validation (LOOCV) on the ventricle and hippocampus models presented in the manuscript.

Firstly, all observations for an individual were removed and the model was estimated with the remaining data, and was used to predict the observations for the missing individual. The posterior mean for each predicted value was subtracted from the known observation to attain a residual value. The residuals and predictive means were assessed for the predictive capability of an individual. This was performed on all individuals, including those with single observations and converters.

The second LOOCV technique involved randomly omitting one observation on individuals with repeated measures (199 participants in our data set). As our longitudinal data are unbalanced, this method assessed predicting values across all time points. Predicted values were computed as described above and the residual and predictive values were assessed for predictive capability within clustered groups.

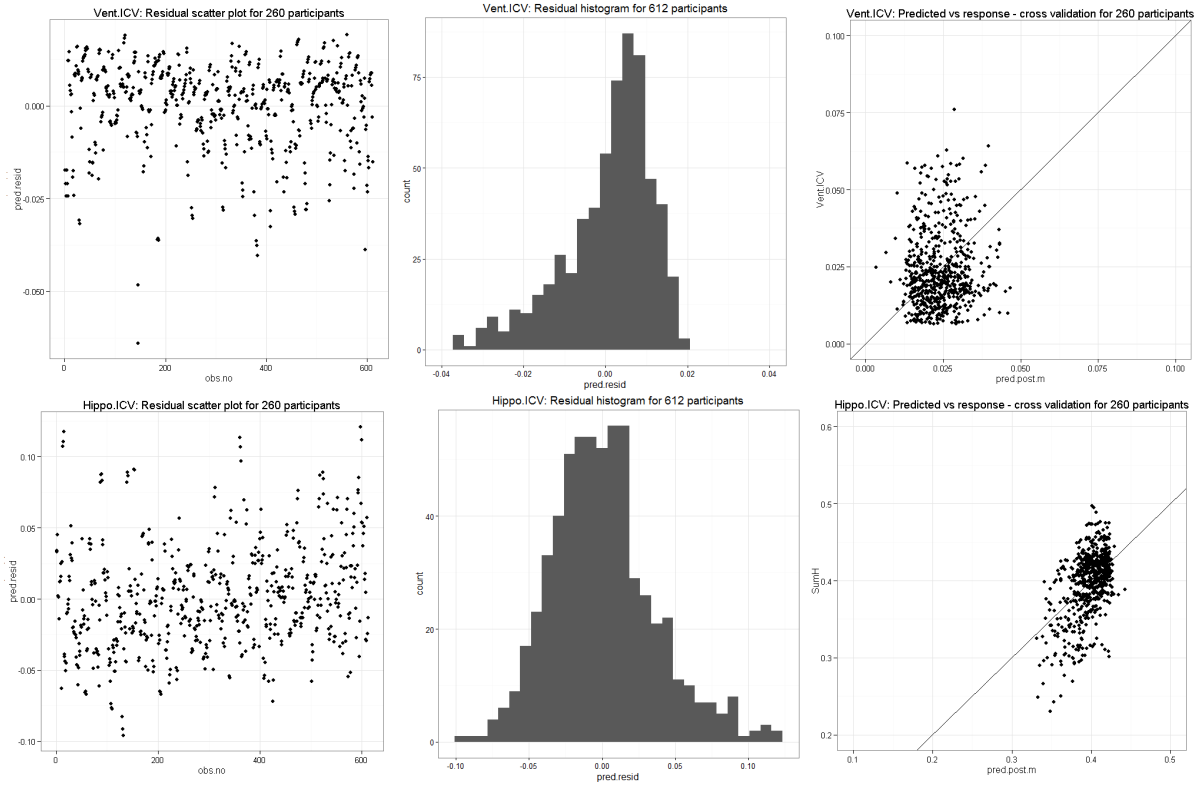


Figure S2: LOOCV-individual, residual (left), residual histogram (middle) and predicted versus response plots (right), for the ventricle (top) and hippocampus (bottom) models.

Left top and bottom plots in Figures S2 and S3 show that the variability in the predictions for out of sample data preserved the overall linear trend and there are no general pattern in the residual scatter plots. As the MSE is the sum of the variance, bias squared and irreducible error, in practise we seek to reduce both bias and variance, hence low MSE values are preferred. Nonetheless there MSE values are relatively low for both models, hence we are satisfied the models stated in (3) in the manuscript do not over-fit the data and provide adequate predictions of new data.

	Ventricle		Hippocampus	
	LOO-individual	LOO-within-a-cluster	LOO-individual	LOO-within-a-cluster
MSE	$1.86e^{-4}$	$1.42e^{-4}$	$3.91e^{-4}$	$1.66e^{-3}$

Table S2: Mean squared root error (MSE) for leave-one-out (LOO) on an individuals set of observations and within a cluster cross validation. The posterior mean for predictive value is $\tilde{y}_i$ and observed response is y_i for n observations computed as MSE$= \sum(\tilde{y}_i - y_i)^2/n$ as described in Timm (1980). Note both MSE values are relatively similar for the hippocampus and ventricle models.

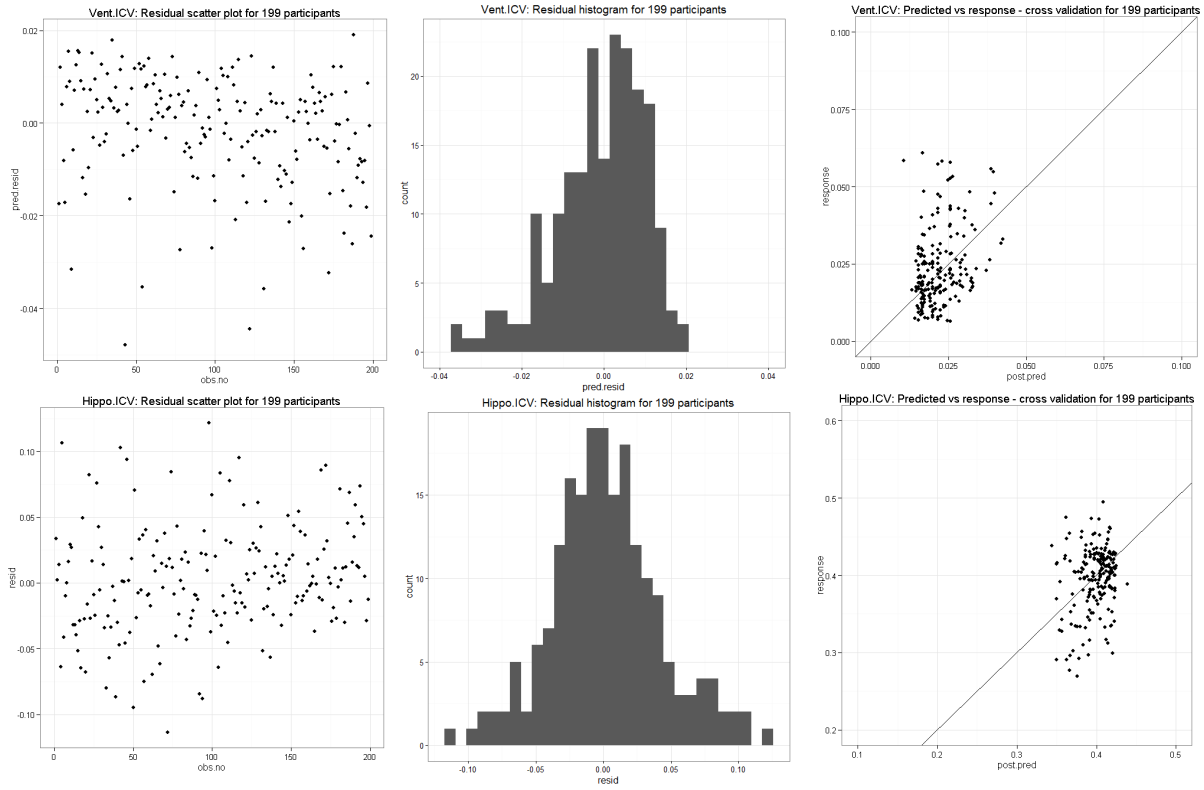


Figure S3: LOOCV-within-a-cluster, residual (left), residual histogram (middle) and predicted versus response plots (right), for the ventricle (top) and hippocampus (bottom) models.

4 Classical linear mixed effects model

In an effort to compare model (3) in the manuscript with popular conventional longitudinal methods, the model was also fitted by a classical linear mixed effects (LME) model. Recall the general LME model is of the following form,

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{b} + \varepsilon. \tag{5}$$

Design matrices are X and Z, and vectors $\boldsymbol{\beta}$ and $\mathbf{b}$ are the fixed and random effects respectively for p fixed and m random effects. We assume residuals $\varepsilon \sim N(0, \sigma^2 I)$, where I is the identity matrix. The random effects vector $\mathbf{b}$, assume $\mathbf{b} \sim N(\mathbf{0}, \sigma^2 D(\theta))$, where $D(\theta)$ is a symmetric and positive semi-definite matrix, parametrized by a variance component θ. For further details including methods to estimate the maximisation of the likelihood, see (Pinheiro, 1994; Corbeil and Searle, 1976).

The classical LME model for this case study is denoted as follows

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} StndAge_{ij} + \beta_2 x_{MCI} + \beta_3 x_{AD} + \beta_{4i} StndAge_{ij} x_{MCI,ij} + \beta_{5i} StndAge_{ij} x_{AD,ij} + \varepsilon_{ij} \\ \beta_{ki} &= \beta_k + b_{ki} \quad for \quad k = 0, 1, 4, 5 \\ \boldsymbol{b_i} &\sim MVN(\mathbf{0}, \Sigma). \end{aligned} \tag{6}$$

Random effects $\boldsymbol{b_i} = [b_{0i}, b_{1i}, b_{4i}, b_{5i}]$ denotes the i^{th} individual deviation at time point j, and covariance structure Σ is 4×4 diagonal matrix. Residuals ε are $i.i.d$ with distribution $N(0, \sigma^2 I_n)$, for n total observations.

R package `lme4()` (Bates et al., 2014b), was used to estimate model (6) (R Core Team, 2013).

	Parameter	Regions: units ICV volume/StndAge	
		Ventricle	Hippocampus
HC	β_3	$5.99e^{-3}$ ($3.29e^{-4}$)	-$1.18e^{-2}$ ($1.54e^{-3}$)
MCI	$\beta_3 - \beta_4$	$6.25e^{-3}$ ($6.69e^{-4}$)	-$2.10e^{2}$ ($5.61e^{-3}$)
AD	$\beta_3 - \beta_4$	$1.10e^{-3}$ ($5.36e^{-3}$)	-$2.82e^{-2}$ ($7.12e^{-3}$)
	Estimated difference of volumetric change among diagnosis groups		
HC - MCI	β_4	$2.68e^{-3}$ ($3.40e^{-4}$)	-$9.27e^{-3}$ ($4.07e^{-3}$)
HC - AD	β_5	$5.03e^{-3}$ ($1.49e^{-3}$)	-$1.64e^{-2}$($5.59e^{-3}$)

Table S3: Parameter estimates for model (6) for both regions, standard error in parenthesis. Similar to Table 1 in the manuscript, the fixed effect values are similar to the BLME model.

4.1 How do HC, MCI and AD participants degenerate over time?

Hypothesis tests were conducted in a similar manner to Bernal-Rusiel et al. (2013) for the ventricle and hippocampus models, refer to their supplementary material for their full model expressions, contrast matrices and null hypothesis statements. In a similar manner, we also state our null hypothesis to compare the rate of volumetric rate of change on diagnosis groups HC, MCI and AD for both brain regions, and set the significance level to $\alpha = 0.05$.

Test 1: **Are there any differences in the rate of change among the three groups?** The null hypothesis is $H_0 : \beta_4 = \beta_5 = 0$. The contrast matrix for both the ventricle and hippocampus models are of the form

$$T_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

	Ventricle	**Hippocampus**
F-statistic	5.850	6.10
p-value	0.0048	0.0058

Table S4: Results for the hypothesis Test 1 for Ventricle and Hippocampus models. As the p-value is less than α for both tests, we reject the null hypothesis and conclude there is a β_4 and β_5 are not equal to zero in both models.

Test 2:**Is there any difference in the rate of volumetric change between HC and MCI?** The null hypothesis is $H_0 : \beta_4 = 0$, with a contrast matrix for both regions as

$$T_2 = [0\ 0\ 0\ 0\ 1\ 0]$$

Test 3: Is there a difference in the rate of change between HC and AD? Similar to above the null hypothesis is $H_0 : \beta_5 = 0$, with a contrast matrix for both regions as

$$T_3 = [0\ 0\ 0\ 0\ 0\ 1]$$

	Ventricle	**Hippocampus**
F-statistic	0.623	5.19
p-value	0.431	0.031

Table S5: Results for the hypothesis Test 2 for Ventricle and Hippocampus models. The Ventricle model indicates we have insufficient evidence to reject the null hypothesis, and conclude the rate of change for MCI is not significantly different from baseline. Unlike the hippocampus model, there is sufficient evidence to suggest the rate of change for MCI is significanly different from baseline.

	Ventricle	**Hippocampus**
F-statistic	11.39	8.65
p-value	0.0018	0.0061

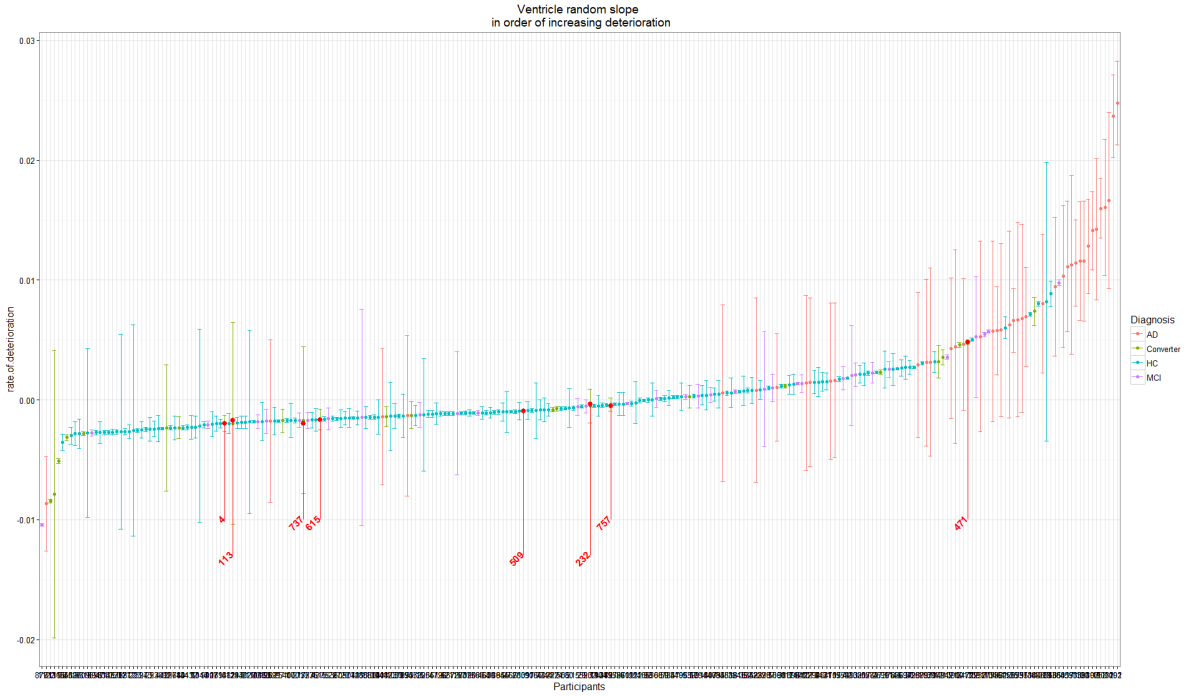
Table S6: Results for the hypothesis Test 3 for Ventricle and Hippocampus models. As both p-values are less than α, we reject the null hypothesis and conclude the rate of change for AD diagnosis is significanlty different from baseline.

4.2 How to identify individuals with high levels of neurodegeneration?

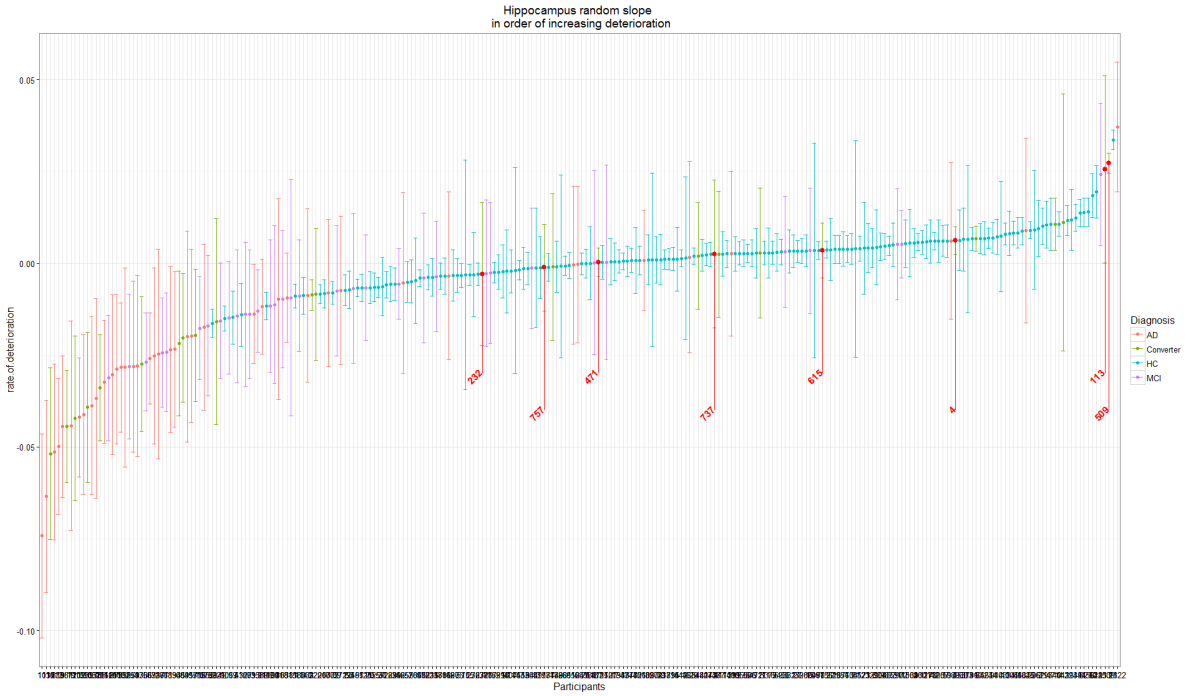
The second analysis presented in the manuscript in Sections 3.4 and 4.2 is presented here where possible, to assess individual participants rate of change. The caterpillar plots for each region are in Figure S4a and S4b which were derived from expression (6) above.

Caterpillar plots for classical mixed effects models are often used for analysis of the random effects (Bates et al., 2014a). The uncertainty for each individual is estimated by conditional variances of the random effects from the output of `lmer()` . The general pattern of HC to MCI converters is similar to those in the manuscript. The ventricle model shows ID's 4, 113, 737 and 509 and the hippocampus participant ID's 757, 232, 471 are in the lower half of the ranks similar to the Bayesian model Figure 4 in the manuscript.

Unfortunately, under this framework we cannot determine the probability of participants ranking in the highest or lowest degeneration extremes. As the ranking of participants relies on point estimates of the random effects and conditional variances, which does not include probability distribution to account for uncertainty between and within observations within clusters.



(a)



(b)

Figure S4: Individuals ranked by order of estimated ventricle (a) and hippocampus (b) rate of change.

	Ranking	AIBL.ID	Diagnosis	Posterior mean rate of deterioration for individuals (standard error)
Ventricle	1	877	AD	-1.04×10^{-2} (8.81×10^{-5})
	2	1122	HC	$-8.66 \times 10^{-3}(3.94 \times 10^{-3})$
	3	911	HC	$-8.45 \times 10^{-3}(1.66 \times 10^{-4})$
	4	111	HC	-7.86×10^{-3} (1.20×10^{-2})
	5	365	HC	-5.11×10^{-3} (2.16×10^{-4})
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	256	658	AD	1.60×10^{-2} (2.48×10^{-3})
	257	1032	AD	1.61×10^{-2} (5.69×10^{-3})
	258	10	AD	1.67×10^{-2}(7.34×10^{-3})
	259	1102	AD	$2.37 \times 10^{-2}(3.45 \times 10^{-3})$
	260	91	AD	2.478×10^{-2} (3.35×10^{-3})
Hippocampus	1	10	AD	-7.42×10^{-2} (2.78×10^{-2})
	2	1135	AD	-6.35×10^{-2} (2.62×10^{-2})
	3	12	Converter	$-5.19 \times 10^{-2}(2.34 \times 10^{-2})$
	4	1013	AD	-5.13×10^{-2} (2.39×10^{-2})
	5	819	AD	-4.98×10^{-2} (1.86×10^{-2})
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	256	483	MCI	$2.42 \times 10^{-2}(1.94 \times 10^{-3})$
	257	113	Converter	$2.56 \times 10^{-2}(2.55 \times 10^{-3})$
	258	509	Converter	2.72×10^{-2} (2.75×10^{-3})
	259	28	HC	3.36×10^{-2} (2.68×10^{-3})
	260	1122	AD	$3.71 \times 10^{-2}(1.78 \times 10^{-2})$

Table S7: Similar to Table 3 in the manuscript, participants by order of estimated rate of change, standard error in parenthesis. Snippet of table shows first and last five individuals for the ventricle and hippocampus volumes.

4.3 How do diagnosis trajectories vary over age?

As described in the manuscript, in this analysis $P(Diagnosis|\tilde{y}, age)$ is estimated for $Diagnosis = HC, MCI$ and AD, as shown in expression (5) in the manuscript. We note that in order to find these probabilities, for a given range in volume $\tilde{y}$ we need the probability of this range given a diagnosis classification and age, i.e. $P(\tilde{y}|Diagnosis, age)$.

A similar analysis cannot be performed with a classical LME model, as the method of maximisation of the likelihood does not allow for the straightforward computation of probabilities $P(Diagnosis|\tilde{y}, age)$. Another drawback of the classical approach is that it does not lend itself to the incorporation of relevant external data, to further extend statistical inference.

5 Posterior Predictive checks and parameter estimates

Posterior predictive checks were carried out to assess goodness-of-fit of our models in expression (3) of the manuscript, as predicted values were simulated from the joint posterior distribution. After burn-in and thinning, as specified in Section 3.2 of the manuscript, each predicted value consists of 8,000 simulations from which we compute the 95% credible intervals. Posterior predictive plots are shown in Figure S5. MCMC chain diagnostics such as trace, density and auto-correlation plots as well as the Gelman and Rubin convergence measures are available

upon request.

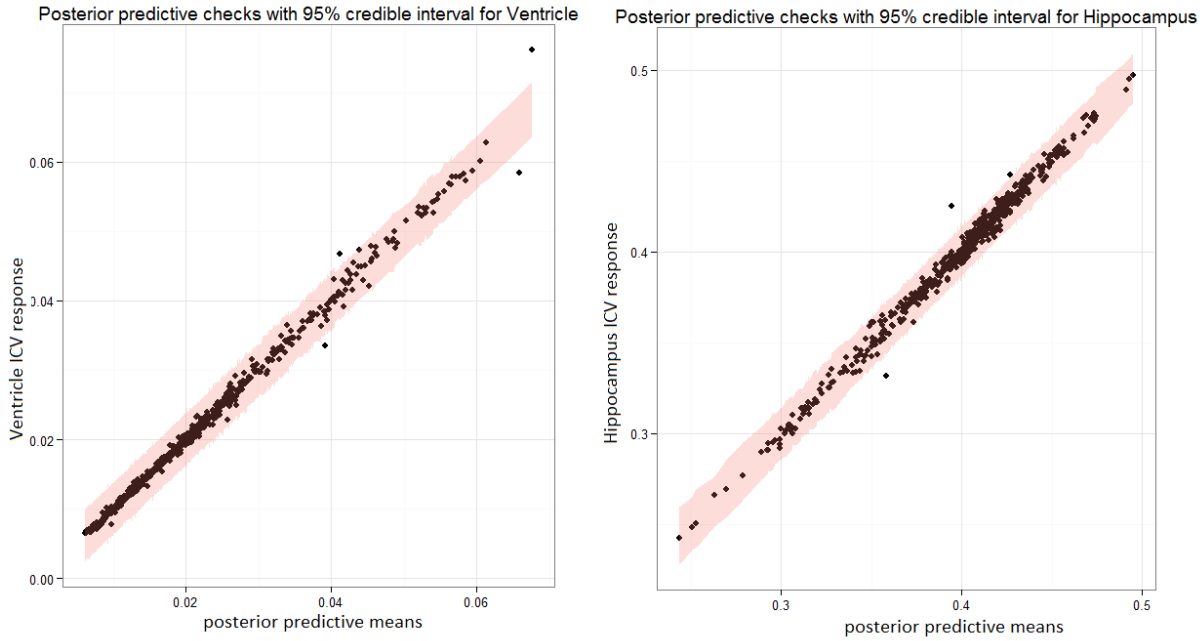


Figure S5: Posterior predictive means versus response values with the 95% credible interval. The tight bandwidth on all responses shows we have adequately captured the variability. As both the plots show a general diagonal pattern of $x = y$ for majority of the values (with the exception of a few cases), this provides evidence of accurate predicted values from our model.

	Ventricle ESS	**Hippocampus** ESS
β_{0i}	6895	5625
β_{1i}	7231	2830
β_2	5585	3832
β_3	3463	2779
β_{4i}	6657	1313
β_{5i}	5693	544
σ^2	7662	4810
σ_0^2	7351	3023
σ_1^2	7753	1824
σ_4^2	6670	977
σ_5^2	5352	1021
P.P.	0.993	0.995

Table S8: Posterior proportion of response (P.P), is a proportion of predicted values which lie within 95% credible interval of prediction values as seen in Figure S5. Effective sample size (ESS) denotes the estimated number of independent samples (no auto-correlation) obtained in our estimated parameters. As per our burn-in and thinning specifications stated in Section 3.2 of the manuscript, the ESS will be at most a value up to 8,000.

6 Distribution of ranks for converters

As described in Section 4.2 of the manuscript, distribution of ranks were performed on all (27) converters of the AIBL study, first on a subset of the first three time points; for ventricle model see Figure S6, hippocampus see Figure S7. Similarly the distribution ranks were estimated on the whole data set, Figure S8 shows the results for the ventricle model, and Figure S9 correspond to the hippocampus model.

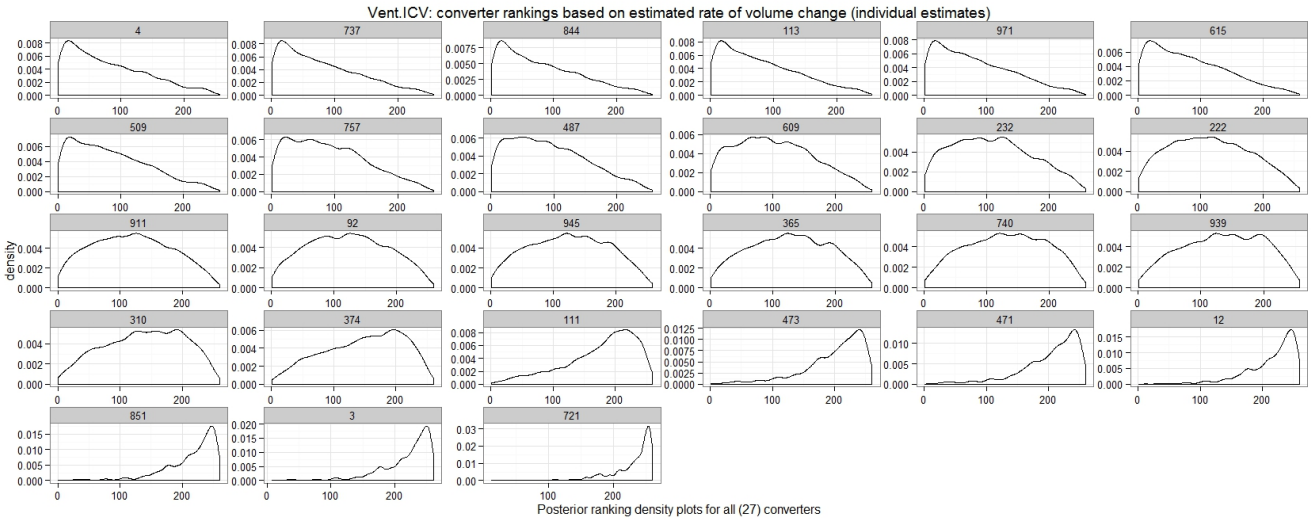


Figure S6: Ventricle converters posterior distribution of ranks for the first three time points.

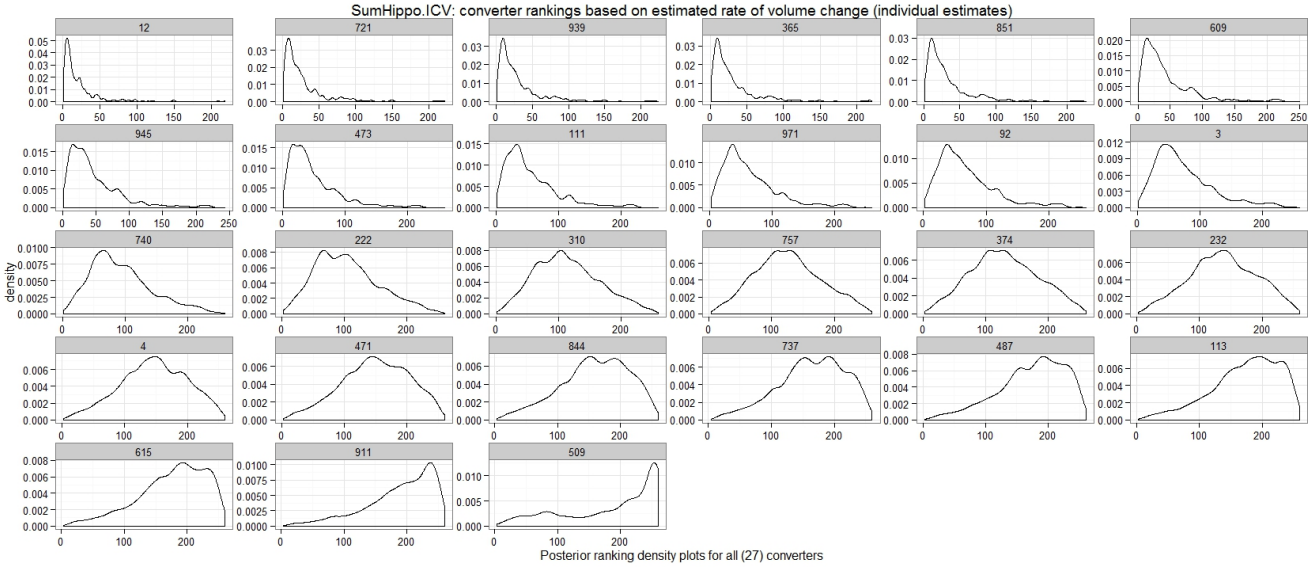


Figure S7: Hippocampus converters posterior distribution of ranks for the first three time points.

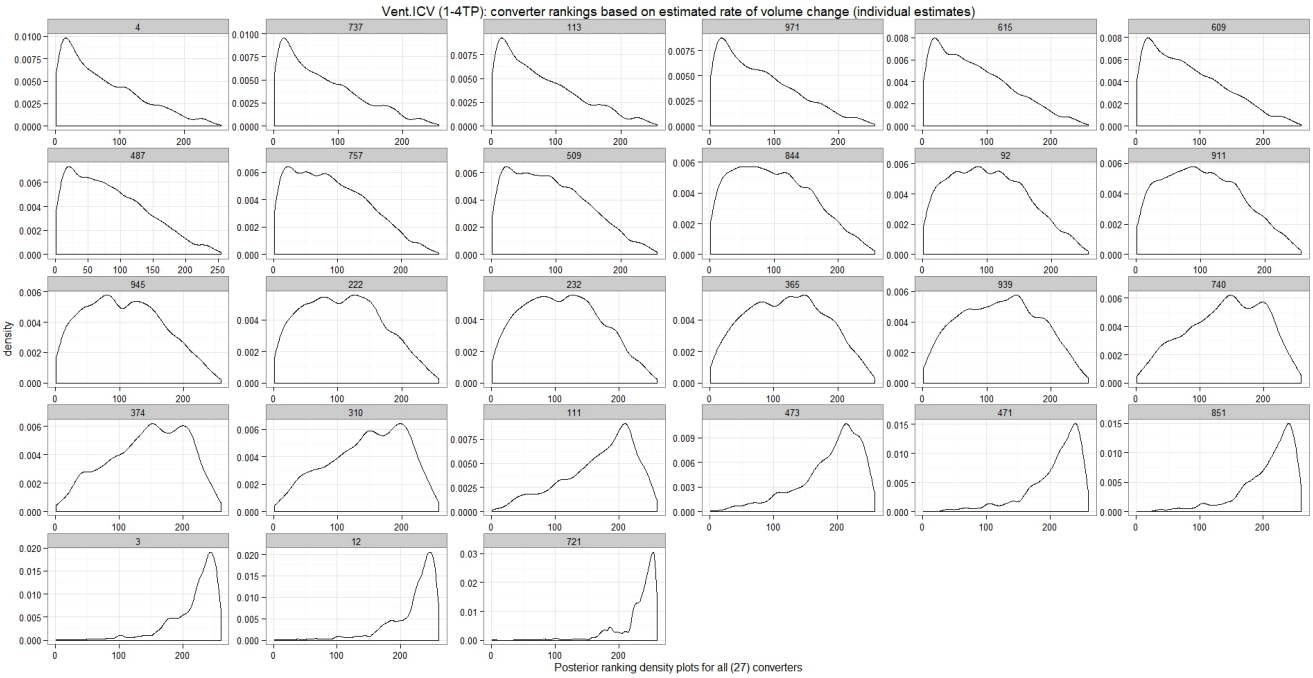


Figure S8: Ventricle converters posterior distribution of ranks for full data (4 timepoints).

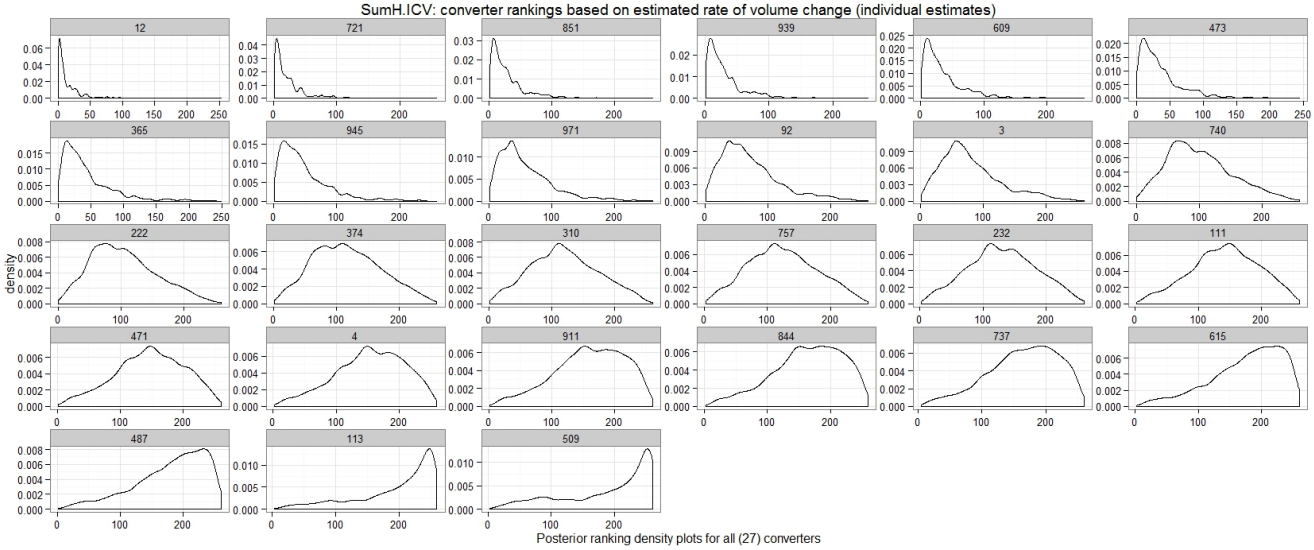


Figure S9: Hippocampus converters posterior distribution of ranks for full data (4 timepoints).

7 APOE and Gender diagnosis trajectories over age

As mentioned in Section 4.3 of the manuscript, initial exploration of diagnosis trajectories over groups; male, female, apolipoprotein $\varepsilon 4$ (APOE $\varepsilon 4$) carriers and non-carriers were also investigated for the ventricle and hippocampus models.

The broad prevalence rates utelised for Inference 3 were derived from Ward et al. (2012); Refshauge and Kalisch (2012) and is summarised in Table S9. Again the reader is cautioned that these are very broad estimates of prevalence rates and are generalised over many factors including lifestyle, genetic and demographic. These prevalence rates also do not take into account participants who develop other forms of dementia or any other neuropsychological disorders. The authors acknowledge there are several factors which the models presented in the manuscript do not account for. As the BLME models and inference derivation presented in this paper are the first of its kind, the objective of Inference 3 is to demonstrate probable diagnosis trajectories conditional on very broad, non-group specific prevalence rates. In order to account for gender and APOE $\varepsilon 4$ status and develop diagnosis trajectories specific to these groups, prevalence rates across ages 65-85 specific to these groups is required, which unfortunately is difficult to find in literature. Figure S10 are the disease trajectories for models (3) in the manuscript applied on male, female, APOE $\varepsilon 4$ carriers and non carriers groups separately, for the ventricle and hippocampus models. We assumed the same prevalence rates as in the manuscript.

Age	HC	MCI	AD
60	0.945	0.037	0.018
65	0.917	0.055	0.028
70	0.859	0.096	0.045
75	0.592	0.333	0.075
80	0.518	0.357	0.125
85	0.466	0.301	0.203

Table S9: Broad prevalence rates for healthy control (HC), mild cognitive impaired (MCI) and Alzheimer's disease taken from Ward et al. (2012); Refshauge and Kalisch (2012). These rates do not account for any lifestyle, demographic and genetic factors as well as other forms of dementia and neuropsychological disorders which are known to affect prevalence rates.

References

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014a.

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014b. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-7.

J. Bernal-Rusiel, D. N. Greve, M. Reuter, B. Fischl, and M. R. Sabuncu. Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *NeuroImage*, 66:249–60, February 2013. doi: 10.1016/j.neuroimage.2012.10.065.

R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.

D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

J. C. Pinheiro. *Topics in Mixed Effects Models*. PhD thesis, UNIVERSITY OF WISCONSIN–MADISON, 1994.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

A. Refshauge and D. Kalisch. Dementia in Australia, 2012. URL `http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=10737422943`. Cat. no. AGE70.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2): 319–392, 2009.

N. H. Timm. 2 multivariate analysis of variance of repeated measurements. *Handbook of statistics*, 1:41–87, 1980.

W. Wang and A. Gelman. Difficulty of selecting among multilevel models using predictive accuracy. *Statistics at its Interface*, 7(1):1–88, 2014.

A. Ward, H. M. Arrighi, S. Michels, and J. M. Cedarbaum. Mild cognitive impairment: disparity of incidence and prevalence estimates. *Alzheimer's & Dementia*, 8(1):14–21, 2012.

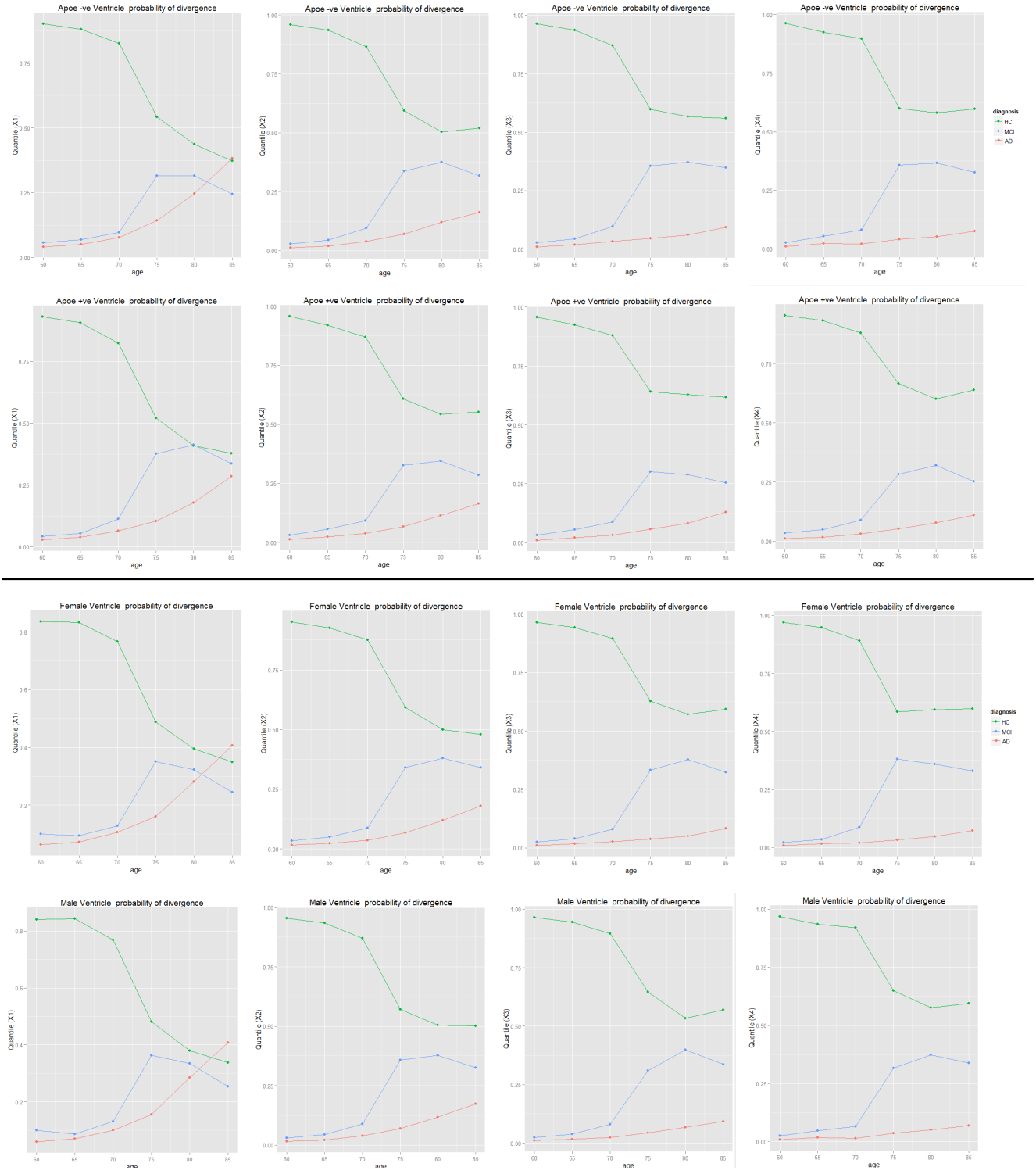


Figure S10: Male, female, APOE $\varepsilon 4$ carriers and non-carriers diagnosis trajectories for ventricle (top) and hippocampus (bottom) model. Volume quantiles X1, X2, X3 and X4 denote 75-100^{th}, 50-75^{th}, 25-50^{th} and 15-25^{th} quantiles respectively.