Research Article

Statistics in Medicine

Received 4 January 2011, Accepted 22 February 2012 Published online 20 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5370

Associations between variability of risk factors and health outcomes in longitudinal studies

Michael R. Elliott,[a,b*†] Mary D. Sammel[c] and Jessica Faul[d]

Many statistical methods have been developed that treat within-subject correlation that accompanies the clustering of subjects in longitudinal data settings as a nuisance parameter, with the focus of analytic interest being on mean outcome or profiles over time. However, there is evidence that in certain settings, underlying variability in subject measures may also be important in predicting future health outcomes of interest. Here, we develop a method for combining information from mean profiles and residual variance to assess associations with categorical outcomes in a joint modeling framework. We consider an application to relating word recall measures obtained over time to dementia onset from the Health and Retirement Survey.

Keywords: differential measurement error; Markov chain Monte Carlo; total recall; dementia; Health and Retirement Survey

1. Introduction

Summary statistics such as the sample mean often describe central tendencies in observed risk factors measured at a particular point in time. However, there are many applications in which characteristics of the risk factors over time are of primary scientific interest for predicting disease. These characteristics can include both changes in the mean function, such as a slope, and measures of variability about the mean. Whereas variances are sometimes modeled to accommodate heteroscedaticity or a hierarchical covariance structure [1], methods that treat variances as being of primary interest and the mean or trend as a nuisance are far less common than the converse [2]. Examples of the latter include Harlow *et al.* [3], where within-woman variability in menstrual cycle length at earlier ages was demonstrated to be an important predictor of abnormal uterine bleeding at later ages; Sammel *et al.* [4], who used a two-stage model to show that high levels of variability in reproductive hormone levels were associated with increased prevalence of menopausal symptoms such as severe hot flashes; Elliott [5], who used a penalized spline model to detrend affect data in a sample of recovering myocardial infarction patients and showed that both low levels and high levels of the residual variance were associated with increased risk of depression; and Kikuya *et al.* [6], who found that day-to-day variability in blood pressure was associated with increased cardiovascular and stroke mortality risk whereas day-to-day variability in heart rate was associated with increased cardiac and stroke mortality risk. Outside of the medical and public health arenas, economists have considered price volatility as an important predictor of risk and returns

[a]*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
[b]*Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106, U.S.A.*
[c]*Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr., Philadelphia, PA 19104, U.S.A.*
[d]*Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106, U.S.A.*
**Correspondence to: Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
[†]*E-mail: mrelliot@umich.edu*

in financial markets [7], but in general, methods to assess associations between variability and risk are underdeveloped in the biostatistics literature.

Methods for longitudinal data have usually treated within-subject variability as a nuisance parameter, with the focus being on central tendency measures to relate the risk factor to an outcome of interest. For example, Muthén *et al.* [8] used growth mixture models to classify subjects on the basis of underlying mean trends in childhood aggression measures and related these classes to juvenile delinquency risk. Ye *et al.* [9] used a measurement error model to relate profiles of prostate-specific antigen levels to time to prostate cancer recurrence. These methods focus on error that is nondifferential, such that it does not influence the outcome of interest directly. This manuscript develops methods to model the joint distribution of within-subject mean trends *and* variability in repeated measurements of risk factors as predictors of categorical health outcomes, with the goal of maximizing the predictive power that can be gleaned from longitudinal risk factor data. We use these methods to determine how trends and variability in memory tests are associated transition to dementia using data from the Health and Retirement Study (HRS). In particular, we jointly model subject-level slope-intercepts and residual variances of a memory recall test as predictors of onset of dementia during the follow-up period. Although we use a fully Bayesian approach, we provide a simulation study to consider the repeated sampling properties of our proposed method.

1.1. Memory and cognition testing in the Health and Retirement Study

Many studies have shown a positive relationship between older age and variability in performance on sensory, motor, and cognitive tasks [10, 11]. There is evidence that intra-individual cognitive variability is a significant source of this variability in performance between groups, especially among older adults [12, 13]. Until recently, age-related increases in intra-individual variability have usually been attributed to lack of instrument reliability [14]. Although variability arising from this type of measurement error may not be meaningful in and of itself, in many cases intra-individual variability may arise from origins other than measurement error and may provide insight into underlying psychological processes and lead to several possible theoretical interpretations [14]. Intra-individual variability may provide support to existing theories of cognitive aging including the differentiation-dedifferentiation theory and the 'common cause' hypothesis [11]. Intra-individual variability may also be an early marker of cognitive deficits [15]. More specifically, this type of variability might reflect an adaptive response to cognitive decline. For example, it may require high levels of attentional capacity to maintain low levels of variability over repeated trials of a task or across cognitive tasks [14].

Current methods of detection of dementia often rely on information on a person's level of performance from a single assessment. A main limitation of this approach is the inability to distinguish between poor performance due to low baseline intellectual ability, preclinical stages of dementia, or fluctuations in cognitive performance due to mood, sensory stimulation, or other state-based differences [16, 17]. Because performance on tests such as the Mini-Mental State Examination is known to vary even over relatively short time intervals, our ability to detect early stages of dementia using these methods is not reliable [18]. However, interpreting lack of reliability in this case is difficult, as it may reflect characteristics of both the specific cognitive test and the individual being tested; that is, inconsistent classification arises from variability due to classification error as well as intra-individual variability in performance. Despite the difficulties in measurement, investigators have hypothesized that variable cognitive performance would precede consistently poor performance in individuals with mild cognitive impairment and those in the very early stages of dementia and thus provide information that predicts over and above what can be achieved by predicting from level information alone [13, 17]. We assess this hypothesis using data from the HRS.

The HRS is a nationally representative, prospective panel study of community-dwelling US adults born between 1890 and 1959 with oversampling of minorities and Florida residents [19]. The HRS include five cohorts: the Asset and Health Dynamics Among the Oldest Old Study (AHEAD) cohort of persons born between 1890 and 1923; the Children of the Depression Age cohort of those born between 1924 and 1930; the original HRS cohort of those born between 1931 and 1941; the War Babies cohort of those born between 1942 and 1947; the Early Baby Boomer cohort of those born between 1948 and 1953 and the Middle Baby Boomer cohort of those born between 1954 and 1959 [19]. Our focus will be on the AHEAD cohort, consisting of 8222 subjects born between 1890 and 1923, who have had data collected in 1993, and, if they survived, 1995, 1998, 2000, 2002, 2004, 2006, and 2008. Interviews are conducted by telephone for most respondents under 80 years of age and face-to-face for persons 80 years

of age or older. Baseline and re-interview rates have been consistently high, with baseline rates ranging from 70% to 81% across the cohorts. Follow-up response rates are on average in the low to mid-90% range; subjects who fail to respond at a given wave are attempted at the next wave unless they have died; hence, some missingness is intermittent.

In the HRS, we assessed cognitive function through several questions asked at every wave. For these analyses, we only considered performance on the episodic memory tasks. In particular, we focus on 'total recall', a measure of episodic memory that consists of immediate and delayed recall of a 10-word list that is asked in the HRS survey. Data on dementia diagnosis come from Medicare claims records linked to HRS respondents. We matched longitudinal HRS survey data to administrative Medicare records for HRS respondents who have previously consented to have their Medicare data released; over 80% consent to do so. Of those who provided an identification number, 98% have been successfully matched to Medicare files. Dementia is defined using the ICD-9 codes listed in the Chronic Condition Data Warehouse definition of Alzheimer's Disease and Related Disorders or Senile Dementia at

`http://www.ccwdata.org/cs/groups/public/documents/document/ccw_conditioncategories.pdf.`

We classified respondents as having received a dementia diagnosis if they had at least one dementia diagnosis code in any of the Medicare claims files, including inpatient, outpatient, part B physician supplier, Skilled Nursing Facility, hospice, and durable medical equipment files. This claims-based diagnostic measure has reasonable sensitivity and specificity for dementia (0.85 and 0.89; see Taylor *et al.* [20]). The key question of interest is the degree to which trends and variability in cognitive and memory tests are associated with transition to dementia, after adjustment for educational level, race/ethnicity, and gender.

2. A model to relate the first two moments of subject-level risk factors to a health outcome

Sammel *et al.* [4] obtained subject-level growth curves and residual variances and used them to predict outcomes in a two-stage model. Here, we extend this idea as a shared parameter model linking mean and variance parameters governing the continuous subject-level longitudinal risk factor measures Y with the binary outcome of interest W:

$$Y_{it} \mid \boldsymbol{\beta}_i, \sigma_i^2 \sim N(f(\boldsymbol{\beta}_i; t), \sigma_i^2) \tag{1}$$

$$W_i \mid \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{Z}_i, \boldsymbol{\gamma} \sim BER(\pi_i), \log\left(\frac{\pi_i}{1-\pi_i}\right) = g(\boldsymbol{\gamma}; \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{Z}_i)$$

$$\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \Sigma \sim N(\boldsymbol{\beta}, \Sigma)$$

$$log(\sigma_i^2) \mid \sigma, \Psi^2 \sim N(\sigma, \Psi^2).$$

We assume that the longitudinal risk factor measures are normally distributed with mean $f(\boldsymbol{\beta}_i; t)$ that may be a linear or nonlinear (polynomial or spline) function of t parameterized by the subject-level parameters $\boldsymbol{\beta}_i$ and subject-level residual variance σ_i^2. Similarly, we assume that the log-odds of the outcome is given by a function $g(\boldsymbol{\gamma}; \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{Z}_i)$ that allows for linear or nonlinear relationships between the subject-level mean profile and residual variance parameters as well as other subject-level covariates, parameterized by the population-level parameters $\boldsymbol{\gamma}$. For a fully Bayesian model, we ensure a proper posterior by proposing the following conjugate independent hyperpriors:

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, V_{\beta_0})$$

$$\sigma \sim N(\boldsymbol{\sigma}_0, V_{\sigma_0})$$

$$\Sigma^{-1} \sim Wishart(k, S_0)$$

$$\Psi^{-2} \sim \Gamma(a_\sigma, b_\sigma)$$

$$\gamma \sim N(\gamma_0, V_\gamma).$$

Very weakly informative hyperprior parameters were used to avoid unduely influencing the information provided by the data.

The posterior distribution is obtained using a Markov chain Monte Carlo (MCMC) approach that combines Gibbs sampling with Metropolois–Hastings draws [21,22]. In brief, Gibbs sampling obtains draws from a joint distribution of $p(\theta \mid \text{data})$ for $\theta = \{\theta_1, \ldots, \theta_q\}$ by initializing θ at some reasonable $\theta^{(0)}$ and drawing $\theta_1^{(1)}$ from $p(\theta_1 \mid \theta_2^{(0)}, \ldots, \theta_q^{(0)}, \text{data})$, $\theta_2^{(1)}$ from $p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_q^{(0)}, \text{data})$, and so forth. As $T \to \infty$, $\theta^{(T)} \dot{\sim} p(\theta_1, \ldots, \theta_n \mid \text{data})$. The conditional draws are obtained using adaptive rejection sampling [23] in WinBugs software (WinBugs V1.4.3, Imperial College and MRC, UK, 2007).

3. Associations between dementia and baseline level, change, and variability in total recall

3.1. Preliminary analyses

To obtain accurate information about dementia onset, we restrict our analysis to the AHEAD cohort subjects who are fee-for-service Medicare beneficiaries and were not diagnosed with dementia at the time of the baseline (1993) interview (4983 of 8222 subjects). To estimate stable subject-level intercepts, slopes, and variances, we further restrict the analysis to the 2372 subjects who had at least four interviews during the follow-up period; an additional 20 were excluded for lacking age data, yielding a total of 2352 for analysis.

Table I shows the distribution of total recall by year of follow-up, along with age, gender, education, and race/ethnicity of participating survivors. An interview was completed at a given follow-up with 69–99% of survivors with four or more total interviews. Mean recall in 1993 was 8.5 words out of 20 (10 for immediate recall and 10 for delayed recall), declining to 6.5 in 2008. Subjects' mean age at baseline was 75.4 years, increasing to 88.4 years among participating survivors in 2008. At baseline, 65% of subjects were female; 33% had less than a high school education, and 15% had more than a high school education; and 83% were white, 12% African–American, and 4% were Hispanic. Participating survivors became increasingly female and more highly educated through the follow-up period.

A dementia diagnosis was obtained among 605 subjects (25.7%) by 2008. Figure 1 plots total recall by age among a subsample of subjects who did not develop dementia and a subsample of subjects who did. There are no clear associations between the observed recall trends and the development of dementia.

Table I. Total recall, age, gender, education, and race/ethnicity by year of follow-up among participating survivors.

Year	1993	1995	1998	2000	2002	2004	2006	2008
n survived	2352	2318	2317	2312	2261	2042	1657	1000
n interviewed	2292	2286	2269	2247	1843	1481	1146	852
Recall	8.5(3.7)	8.8(3.6)	8.2(3.6)	7.5(3.5)	7.2(3.5)	6.8(3.3)	6.5(3.3)	6.5(3.3)
Age	75.4(4.6)	77.4(4.6)	79.7(4.6)	81.8(4.6)	84.0(4.6)	85.7(4.4)	87.3(4.1)	88.4(3.5)
% Female	65.2	64.9	64.6	64.4	66.0	65.8	67.2	69.0
% <HS	33.2	33.0	32.8	30.7	29.4	29.0	29.0	27.8
% HS	52.0	52.2	52.4	52.6	53.7	54.2	53.7	53.6
% >HS	14.7	14.7	14.8	14.8	15.6	16.3	17.3	18.5
% White	83.1	83.2	83.4	83.5	83.9	83.7	83.2	83.3
% Black	12.1	11.8	11.9	11.7	11.6	11.7	11.8	11.6
% Hispanic	3.8	4.0	3.7	3.8	3.7	4.0	4.2	4.5
% Other	0.9	1.0	1.1	1.1	0.8	0.7	0.8	0.6

HS = High school. Standard deviations in parentheses.

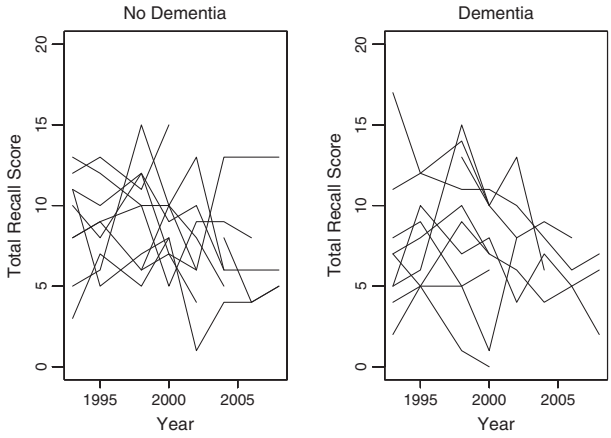

Figure 1. Total recall by age among (a) 10 subjects without a dementia diagnosis and (b) 10 subjects with a dementia diagnosis.

3.2. Joint modeling of recall mean and variance to predict onset of dementia during follow-up

We fit the model proposed in Section 2 to the total recall and dementia outcome data, letting the risk factor measures y_{it} be the total recall score raised to the power 0.7325 to improve the approximation to normality. The outcome w_{it} is an indicator for whether or not dementia was diagnosed at any time during the follow-up period. Recall measures are obtained from all subjects regardless of dementia diagnosis. Because of the small number of recall observations per subject (4–8), we only considered low-degree polynomials for $f(\boldsymbol{\beta}_i, t)$; on the basis of preliminary mixed-model analysis, we chose a quadratic trend: $f(\boldsymbol{\beta}_i, t) = \beta_{0i} + \beta_{1i}\tilde{a}_{it} + \beta_{2i}\tilde{a}_{it}^2$, where $\tilde{a}_{it}$ is the age of the ith person at the t follow-up visit standardized for numerical stability reasons by subtracting the global mean age and dividing by the global standard deviation of age: $\tilde{a}_{it} = (a_{it} - 81.0496)/5.8487$ for age a_{it}. Thus, β_{0i} corresponds to a subject-level mean, β_{1i} to a subject-level slope, and β_{2i} to a subject-level curvature in (transformed) recall scores. For the dementia model, we assume $g(\boldsymbol{\gamma}, \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{z}_i) = \gamma_0 + \gamma_1\beta_{0i} + 10\gamma_2\beta_{1i} + 100\gamma_3\beta_{2i} + \gamma_4\sigma_i^2 + \gamma_5 S(\sigma_i^2) + \boldsymbol{\gamma}_6^T \boldsymbol{z}_i$. (We inflated the random effects associated with slope and curvature to bring the components of γ to be on the same scale and thus improve convergence of the MCMC algorithm.) The function $S(x) = (x-x_1)_+^3 - ((x_3-x_1)/(x_3-x_2))(x-x_2)_+^3 + ((x_2-x_1)/(x_3-x_2))(x-x_3)_+^3$ contains the nonlinear component of a restricted cubic spline [24] with knots at x_1, x_2, and x_3, termed 'restricted' because $S(x)$ is constrained to be linear in its tails ($x < x_1$ and $x > x_3$), thus avoiding overfitting while still accommodating any nonlinearities in the relationship between risk of dementia onset and subject-level residual variances. We chose the knot values on the basis of a visual inspection of histograms of the posterior means of the residual variances of a transformed recall scores-only model and fixed them as (1,3,5). The covariate vector $\boldsymbol{z}_i$ includes dummy variables for education, race/ethnicity, gender, and baseline age categories (65–70, 71–75, 76–80, and 80+). Finally, we assume relatively flat hyperpriors $\boldsymbol{\beta}_0 \equiv 0$, $V_{\beta_0} = diag(1000)$, $\sigma_0 = 0$, $V_{\sigma_0} = 100$, $k = 3$, $S_0 = diag(0.1)$, $a_\sigma = b_\sigma = 0.01$. Having brought the predictors of dementia to approximately a unit scale, we use priors of the form $\boldsymbol{\gamma}_0 \equiv 0$ and $V_\gamma = diag(100)$. Four chains of 20,000 draws were obtained after a burn-in of of 1000 draws. Convergence of the MCMC algorithm was assessed for each parameter using the Gelman–Rubin statistic $\hat{\sqrt{R}}$ [25], which is (approximately) the square root of the total variance of the draws of the parameter divided by the within-chain variance. The maximum value was 1.03 across all population parameters, considered sufficient for convergence.

The posterior mean of $\boldsymbol{\beta}$ and Σ were

$$\begin{pmatrix} 4.38 \\ -0.59 \\ -0.09 \end{pmatrix} \text{and} \begin{pmatrix} 1.190 & 0.078 & -0.091 \\ 0.078 & 0.185 & -0.012 \\ -0.091 & -0.012 & 0.015 \end{pmatrix},$$

respectively, indicating an overall accelerating decline in recall, with considerable between-subject variability. Figure 2 shows the observed and predicted values of y_{it} for four randomly chosen subjects

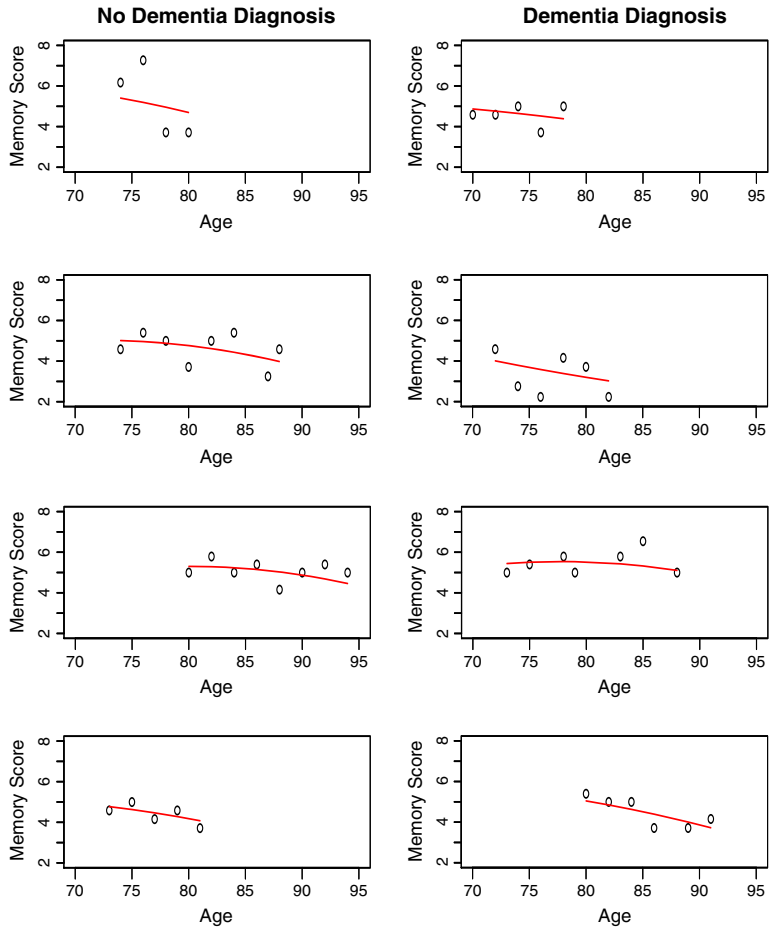

Figure 2. Observed and predicted recall scores with 0.7325 power transformation, among a subsample of those without a dementia diagnosis and with a dementia diagnosis.

without and with dementia diagnosis, where the predicted values are given by $\hat{y}_{it} = \hat{\beta}_{0i} + 10\hat{\beta}_{1i}\tilde{a}_{it} + 100\hat{\beta}_{2i}\tilde{a}_{it}^2$ for $\hat{\beta}_{pi} = E(\beta_{pi} \mid y)$. Some subjects had approximately flat trends, whereas others decreased with varying degrees of rapidity, with no obvious differences between those who developed dementia and those who did not.

Table II provides the posterior means and 95% credible intervals for the predictors of dementia onset during follow-up. There is moderate evidence that higher intercepts and positive or less pronounced negative curvatures are associated with increased risk of dementia and somewhat stronger evidence that increased variability in cognitive scores is positively associated with risk of dementia, up to a threshold level in the upper tail of the individual recall score variability distribution. To better visualize these relationship, we plot in Figure 3 the log-odds of dementia for a given subject-level mean, slope, curvature, and residual variance relative to the posterior mean of these population-level quantities (in the case of variance, this is $E\left(e^{\sigma+\Psi^2/2}/y\right) = 0.93$), holding baseline covariates constant. (The tick marks at the bottom of the plot denote random samples of the posterior means of β_{0i}, β_{1i}, β_{2i}, and σ_i^2 and provide a way to interpret the range of risks associated with the subject-level residual variance in the population.) Subjects with variances below the population mean are at reduced risk, whereas subjects above this mean are at increased risk up to a threshold of approximately 2.5–3. As a specific example, we predicted subjects with a variance of 2.5 to have an odds ratio of 2.48 (95% CI 1.14,7.85) for development of dementia relative to subjects with a variance of 0.5. Similarly, subjects with less negative curvature are at a higher risk of dementia than subjects with more negative curvature: for example, subjects with a curvature of 0 (linear trend) are predicted to have an odds ratio of 5.42 (95% CI 1.01,947.19) for development of dementia relative to those with a curvature of −0.15. Among the baseline covariates, only gender showed any evidence of being associated with risk of dementia onset, with males having an odds of 0.76 relative to females (95% CI 0.52–1.01).

Table II. Log OR of dementia as a function of trends and variances of total recall score with power transformation a, adjusted for education, race/ethnicity, and gender.

	Two-stage model	Joint model $a = 0.7325$	Joint model $a = 2/3$	Joint model $a = 3/4$
Subject-level intercept β_{i0}	$-0.003_{(-0.040,0.034)}$	$1.001_{(-0.028,2.957)}$	$0.607_{(-0.023,2.283)}$	$\mathbf{0.925_{(0.004,2.615)}}$
Subject-level slope β_{i1}	$-0.001_{(-0.025,0.025)}$	$0.400_{(-0.615,2.344)}$	$-0.077_{(-1.043,0.747)}$	$0.442_{(-0.398,2.057)}$
Subject-level curvature β_{i2}	$\mathbf{0.065_{(0.018,0.112)}}$	$\mathbf{14.81_{(0.05,45.69)}}$	$\mathbf{9.26_{(0.21,38.44)}}$	$\mathbf{13.53_{(0.44,39.53)}}$
Subject-level linear variance σ_i^2	$0.050_{(-0.081,0.182)}$	$\mathbf{0.639_{(0.084,1.524)}}$	$\mathbf{0.766_{(0.195,1.490)}}$	$\mathbf{0.478_{(0.007,1.088)}}$
Subject-level nonlinear variance $S(\sigma_i^2)$	$-0.002_{(-0.012,0.008)}$	$\mathbf{-0.092_{(-0.336,-0.008)}}$	$\mathbf{-0.777_{(-2.293,-0.061)}}$	$\mathbf{-0.042_{(-0.117,-0.001)}}$
< HS (versus >HS)	$-0.096_{(-0.314,0.122)}$	$-0.140_{(-0.478,0.156)}$	$-0.123_{(-0.397,0.139)}$	$-0.136_{(-0.465,0.152)}$
HS (versus >HS)	$0.027_{(-0.251,0.305)}$	$0.012_{(-0.375,0.407)}$	$0.017_{(-0.304,0.343)}$	$0.011_{(-0.383,0.377)}$
Black (versus White)	$0.003_{(-0.294,0.299)}$	$0.024_{(-0.394,0.452)}$	$0.011_{(-0.332,0.353)}$	$0.015_{(-0.370,0.412)}$
Hispanic(versus White)	$-0.036_{(-0.530,0.457)}$	$-0.095_{(-0.782,0.539)}$	$-0.084_{(-0.654,0.485)}$	$-0.082_{(-0.752,0.527)}$
Other (versus White)	$0.013_{(-0.921,0.946)}$	$0.041_{(-1.105,1.175)}$	$0.001_{(-1.042,0.947)}$	$0.003_{(-1.209,1.114)}$
Male (Versus Female)	$\mathbf{-0.217_{(-0.417,-0.017)}}$	$-0.273_{(-0.645,0.005)}$	$\mathbf{-0.234_{(-0.493,-0.006)}}$	$\mathbf{-0.263_{(-0.613,-0.002)}}$
Baseline age 70–74 (versus 65–70)	$-0.152_{(-0.468,0.163)}$	$-0.226_{(-0.715,0.191)}$	$-0.186_{(-0.556,0.154)}$	$-0.210_{(-0.699,0.189)}$
Baseline age 75–79 (versus 65–70)	$-0.033_{(-0.375,0.308)}$	$-0.081_{(-0.568,0.357)}$	$-0.064_{(-0.473,0.314)}$	$-0.069_{(0-.507,0.365)}$
Baseline age 80+ (versus 65–70)	$-0.028_{(-0.406,0.351)}$	$-0.128_{(-0.684,0.358)}$	$-0.096_{(-0.551,0.317)}$	$-0.126_{(-0.668,0.358)}$

SD = standard deviation (estimated empirically in two-stage model and parametrically in joint model); HS = High school. Boldface denotes CIs that exclude 0.

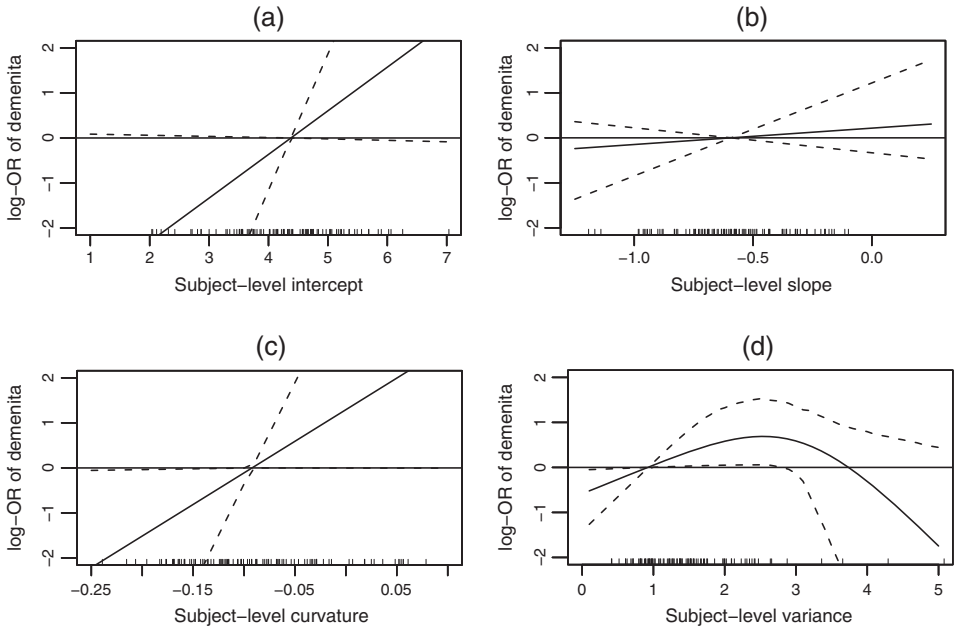

Figure 3. log-OR for dementia as a function of subject-level intercept, slope, curvature, and variance versus posterior mean of the population-level mean of intercept, slope, curvature, and variance. Tick marks show random sample of posterior means of 200 individual level variances σ_i^2.

For comparison, we fit a two-stage model to these data. First, a separate Gaussian regression model was used to relate recall data to a quadratic association of age (standardized to have a mean of 0 and variance of 1 across all subjects) for each subject. Next, we used the estimated intercepts, slopes, curvature, and residual variance obtained to predict the probability of dementia onset in a logistic regression model, adjusting for education, race/ethnicity, gender, and baseline age. Table II shows that as in the joint model, curvature and gender are associated with risk of dementia and in the same direction (positive or less pronounced negative curvatures associated with increased risk; males with decreased risk).

3.3. Model checking

We use posterior predictive distribution (PPD) model checking [25] to assess whether the proposed model provides a reasonable approximation to the true data. The PPD 'p-value' represents the probability that an observed statistic (which can be a function of both the data y and the parameter θ) is more extreme than replicated statistic, conditional on the observed data: $P(T(y^{\text{obs}}, \theta) \leq T(y^{\text{rep}}, \theta) \mid y)$, where y^{rep} is drawn from the PPD $f(y^{\text{rep}} \mid y) = \int f(y^{\text{rep}} \mid \theta, y) p(\theta \mid y) \text{d}\theta$. Although PPD p-values are not true p-values in that they do not have a uniform distribution, values close to 0 or 1 give evidence of poor model fit. For the predictor data Y_{it} (transformed recall scores), we computed for each subject a chi-square discrepancy statistics of the form $T_i(\boldsymbol{y}_i; \boldsymbol{\beta}_i, \boldsymbol{\sigma}^2) = \Sigma_t (y_{it} - f(\boldsymbol{\beta}_i, t))^2 / \boldsymbol{\sigma}_i^2$. We compute $P(T_i(\boldsymbol{y}_i^{\text{obs}}; \boldsymbol{\beta}_i, \boldsymbol{\sigma}^2) < T(\boldsymbol{y}_i^{\text{rep}}; \boldsymbol{\beta}_i, \sigma^2) | (\boldsymbol{y}_i^{\text{obs}}))$ by keeping $\boldsymbol{y}_i$ fixed at its observed values and computing 200 values of $T(\boldsymbol{y}_i^{\text{obs}}; \boldsymbol{\beta}_i, \boldsymbol{\sigma}_i^2)$ from 200 draws from the posterior of $\boldsymbol{\beta}_i, \sigma_i^2$ and comparing these with 200 draws from $T(\boldsymbol{y}_i^{\text{rep}}; \boldsymbol{\beta}_i, \sigma_i^2)$, which has a $\chi_{n_i}^2$ distribution. Figure 4 shows the resulting histogram of the 2352 PPD p-values for each subject's recall score trajectory. The median PPD value was 0.48; the range was 0.20 to 0.78, indicating a reasonable degree of model fit for all subjects. The largest p-value was for a subject whose recall scores was extremely low across all follow-ups, indicating 'floor effects' that were not entirely captured by the normality assumption.

Some preliminary transformations indicated a rather poor fit: Figure 4 shows the equivalent histograms for recall score power transformations of 2/3 and 3/4. Roughly speaking, power transformations less than 0.7325 led to overdispersed data, and power transformations greater than 0.7325 lead to underdispersed data, although the integer nature of the underlying recall scores plays a role as well. Because of this sensitivity, we report in Table II the association between dementia onset and the recall score mean profiles

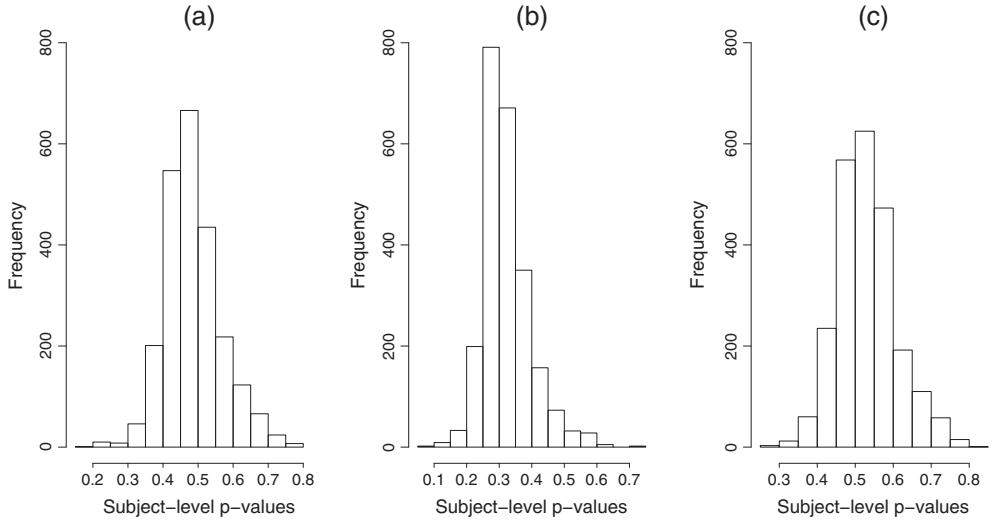

Figure 4. Histogram of posterior predictive distribution p-values for subject-level recall score trajectory discrepancy statistic: (a) power transformation of 0.7325, (b) power transformation of 2/3, and (c) power transformation of 3/4.

and residual variances for the two alternative power transformations. We found that there was a modest degree of sensitivity to the choice of transformation but that the basic finding remained that overall mean, curvature, residual variance, and gender are associated dementia onset, with the functional form of the residual variance association being similar. The 2/3 power transformation suggested a somewhat stronger quadratic association between risk of dementia and residual variance, as well as somewhat narrower intervals for the baseline covariate effects, possibly owing to somewhat less predictive information being captured from the recall scores.

We also considered the predictive distribution of the dementia outcome W_i. We considered the total count $T = \sum_i w_i$, which was $T^{\text{obs}} = 605$ in the observed data; comparing this with $T^{\text{rep}} = \sum_i w_i^{\text{rep}}$ yields a the posterior predictive 'p-value' of 0.50, where w_i^{rep} is drawn from a Bernoulli distribution with probability $\pi_i^{\text{rep}} = \exp^{g(\gamma^{\text{rep}}, \boldsymbol{\beta}_i^{\text{rep}}, \sigma^{2\,\text{rep}}_i, z_i)}/(1 + \exp^{g(\gamma^{\text{rep}}, \boldsymbol{\beta}_i^{\text{rep}}, \sigma^{2\,\text{rep}}_i, z_i)})$ and $\boldsymbol{\gamma}^{\text{rep}}, \boldsymbol{\beta}_i^{\text{rep}}, \boldsymbol{\sigma}^{2\,\text{rep}}_i$ and drawn from their posterior distributions in the MCMC chain. To assess whether the predicted probabilities are reasonable at the tails of their distribution as well as overall, we considered a Hosmer–Lemeshow-type fit statistic $T_k = \sum_{i \in k} w_i$, where $k = 1, \ldots, 10$ indexes the deciles of the probability of dementia π_i: we compute $T_k^{\text{obs}} = \sum_{i \in k} w_i$ where k is based on the posterior draws of π_i^{rep} and compare with the distribution $T_k^{\text{rep}} = \sum_{i \in k} w_i^{\text{rep}}$, on the basis of both the posterior draws of π_i^{rep} and the posterior predictive draws of w_i^{rep}. This yielded PPD p-values across the deciles of (0.32,0.38,0.34,0.47,0.44,0.43,0.56,0.52,0.52,0.55), indicating reasonable fit for the second stage of the model over the range of $\hat{\pi}_i$, with a very modest tendency to overestimate risk of dementia in the lowest-probability subjects and underestimate risk of dementia in the highest-probability subjects.

4. Simulation study

We conducted a simulation study as follows. We consider 1000 observations with five repeated measures per subject, with the continuous predictors generated under a Gaussian random effects linear model and the dichotomous outcome generated under a logistic model that is a function of the random effects that govern the predictors:

$$Y_{it} \mid \boldsymbol{\beta}_i, \sigma_i^2 \sim N(\beta_{0i} + \beta_{1i} t, \sigma_i^2),\ t = 0, \ldots, 4,\ i = 1, \ldots, 1000$$

$$W_i \mid \boldsymbol{\beta}_i, \sigma_i^2, \boldsymbol{\gamma} \sim BER(\pi_i), \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \gamma_0 + \gamma_1 \beta_{0i} + \gamma_2 \beta_{1i} + \gamma_3 \sigma_i^2$$

Table III. Simulation study results: population-level regression parameters. Bias for posterior mean; coverage for 95% credible intervals.

Parameter (True value)	Two-stage model Bias	Two-stage model Nominal 95% Coverage	Two-stage model Power	Joint model Bias	Joint model Nominal 95% Coverage	Joint model Power
β_0 (5)	—	—	—	0.00	94	—
β_1 (0.1)	—	—	—	−0.00	96	—
γ_0 (−4)	2.75	0	—	−0.20	99	—
γ_1 (0.75)	−0.34	1	100	0.01	94	99
γ_2 (0)	0.43	35	65	0.01	93	7
γ_3 (2)	−1.53	0	100	0.17	95	100

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \Sigma)$$

$$log(\sigma_i^2) \sim N(\sigma, \Psi^2)$$

We have $\boldsymbol{\beta} = (4\ -0.1)^T$, $\Sigma = \begin{pmatrix} 0.7 & 0.07 \\ 0.07 & 0.12 \end{pmatrix}$, $\sigma = -0.4$, $\Psi^2 = 0.3$, and $\boldsymbol{\gamma} = (-6\ 0.75\ 0\ 2)^T$. This corresponds to a 1 standard deviation increase in the subject-level intercept being associated with a 69% increase in the odds of the outcome and a 1 standard deviation increase in the subject-level variance being associated with a 310% increase in the odds of the outcome. The probability of the outcome is approximately 19% when the random effects are fixed at their population means.

Two chains of 2500 draws were obtained after a burn-in of 2500 draws. The bias (based on the posterior mean), nominal 95% credible interval coverage, and power for the population-level regression parameters estimated from 100 simulated datasets are given in Table III. Bias is negligible for the population slope and intercept for the predictors, and coverage is approximately correct. There is a modest degree of bias for the logistic regression parameters, on the order of 10–15%, but again, coverage is approximately correct. For comparison, we include the results of a two-stage analysis for the logistic regression parameters. Bias toward the null is severe, with nominal 95% coverage being essentially 0 for three of the four parameters.

5. Discussion

Despite the great proliferation of longitudinal health data during the past three decades, relatively little attention has been paid to the role that variability in such data might play in predicting outcomes of interest. This manuscript attempts to fill this gap by developing a method to combine information about both mean trends and variances in longitudinal data to predict categorical outcomes of interest. We applied the method to predict onset of dementia in elderly adults over a 14-year time period using recall data measured every 2 years. We found residual variability to be associated with dementia risk, with subjects with low variability being less likely to develop dementia by the end of the follow-up period of 14 years than subjects with moderate to high variability. Overall, mean level and curvature (quadratic trend of recall) were marginally associated with dementia risk, with increase mean level and increased quadratic trend found to be associated with increased risk of dementia onset. We found little predictive power for linear trend or in the baseline measures of education, race/ethnicity, or age. The cognitive performance trends had associations that were reversed from those hypothesized with the diagnostic outcome (higher intercepts associated with decreased risk, accelerating declines associated with increased risk), possibly owing to the fact that intercepts and curvatures had a strong negative correlation (posterior mean of −0.68), which would lead to some instability in estimation of associated effects. Increased within-person variability had a significant prognostic relationship in the hypothesized direction.

The importance of using a joint model to assess the relationship between the individual parameters governing the trends and variability of recall and risk of dementia is seen in the fact that simple two-stage models that used the results from individual linear regression fits of recall data either had had little relationship with dementia risk or was in the reverse direction from that hypothesized (negative curvature, i.e., an increasingly rapid reduction in memory performance, was associated with a decreased

risk of dementia onset). This is likely due to very substantial bias toward the null induced by measurement error, and/or spurious relationships induced by floor or ceiling effects, in the first-stage model. A two-stage procedure that used, for example, empirical Bayes estimates from random effects for both the means and variances to stabilize first-stage estimation of the mean trends and residual errors would likely have improved performance, but standard software for fitting random effects models does not typically allow for estimation of subject-level random effects for the variance components.

The authors also attempted to fit the shared parameter model in (1) using a fully likelihood-based method. However, integrating out the subject-level variance random effects proved virtually intractable using adaptive Gaussian quadrature methods available for the PROC NLMIXED procedure available SAS V 9.1 (SAS Institute, Cary, NC, USA). Our simulation study suggests that the Bayesian approach with weakly informative priors has reasonable repeated sampling properties and is vastly superior to estimates obtained from a two-stage method.

Use of a low-degree polynomial to model the longitudinal recall data was dictated by both the nature of the question at hand – specifically the desire to link slopes and curvatures of recall trends to dementia risk – and the relative paucity of follow-up visits for each individual. Even if sufficient longitudinal data were available, higher order terms in nonlinear growth profiles for the longitudinal risk factors yield predictors of outcome that are difficult to interpret. Elliott [5] used a penalized spline model to detrend subject-level daily affect data consisting of up to 35 follow-up measures, relating the risk of depression to latent clusters of subject-level variability. An extension that would incorporate information from both means and variances could assign subjects into latent classes of profiles and variances and link these classes to a categorical outcome of interest via a log-linear model [26, 27]. Such an approach requires some clustering of the profiles and the residual variability of longitudinal predictors; in the application of interest here, no such clustering was evident.

As Table I shows, most of the missing data in the longitudinal predictors was structurally missing (owing to death rather than loss to follow-up). Use of a linear mixed model for the longitudinal data makes a missing at random (MAR) assumption [28]. When the missingness is intermittent, the MAR assumption seems reasonable: violation would require that we systematically miss low-mean or high-mean or highly variable observations within a subject. The MAR assumption is stronger in the small fraction of dropout data, in that it requires that our modeling assumptions of the mean or variance be correct. Future work could consider selection models that would provide sensitivity analysis for violations of the MAR assumptions.

Finally, some discussion of the limitations of our analysis is in order. First, restricting our analysis to those subjects with four or more recall scores, although necessary to provide information about quadratic trends and residual variances, may also lead to selection bias, because subjects with three or fewer follow-ups may be at higher risk of death, which could have a variety of impacts on risk of dementia during follow-up. Indeed, subjects with three or fewer follow-ups were older (79.3 versus 75.4 years) and had lower recall scores (6.3 vs 8.5) at baseline and were more likely to be male (42.5% vs 35.3%), lack a high school degree (52.5% vs 33.3%), and be African–American (16.3% vs 12.1%) (all $p < .001$ by t-test or χ^2 test). This limitation might be addressed in part by treating the outcome as a time-to-event measure rather than a dichotomous outcome over the whole follow-up period. This would have the advantage of accounting for administrative or competing risk censoring in cases of dropout or death, as well as increased power and a more nuanced understanding of the associations of interest. However, the problem of insufficient information to estimate trends and residual variance in subjects with few follow-up measures will remain and highlights the requirement for sufficient follow-up data to implement the proposed methods. A second major issue results from the numerical 'fragility' of the models, at least as fit in Winbugs. For example, attempts to include nonlinear effects for mean trends in the prediction of dementia onset parallel to those used for variance lead to numerical overflow ('trap') errors, as did use of flatter hyperprior parameters in S_0 for $p(\Sigma^{-1})$. Direct computation of the relevant conditional distributions for the Gibbs algorithm, although more time consuming, may allow for direct control of overflow errors; alternatively, model approaches such as use of growth mixture models to classify mean trends into categorical predictors may provide more robust numercial results.

Acknowledgements

The authors wish to thank two reviewers and the editor for their comments, which greatly improved the manuscript. This work was supported by Grant Number R03AG031980 from the National Institute of Aging.

References

1. Barnard J, McCulloch R, Meng X-L. Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statisticia Sinica* 2000; **10**:1281–1311.
2. Carrol RJ. Variances are not always nuisance parameters. *Biometrics* 2003; **59**:211–220.
3. Harlow SD, Lin X. Ho MJ. Analysis of menstrual diary data across the reproductive life span: applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology* 2000; **53**:722–733. DOI: 10.1016/S0895-4356(99)00202-4.
4. Sammel MD, Wang Y, Ratcliffe SJ, Freeman E, Propert KJ. Models for within-subject heterogeneity as predictors for disease. *Proceedings of The American Statistical Association*, Biometrics Section, Atlanta GA, 2001.
5. Elliott MR. Identifying latent clusters of variability in longitudinal data. *Biostatistics* 2007; **8**:756–771. DOI: 10.1093/biostatistics/kxm003.
6. Kikuya M, Ohkubo T, Metoki H, Asayama K, Hara A, Obara T, Inoue R, Hoshi H, Hashimoto J, Totsune K, Satoh H, Imai Y. Day-by-day variability of blood pressure and heart rate at home as a novel predictor of prognosis: the Ohasama study. *Hypertension* 2008; **52**:1045–1050. DOI: 10.1161/HYPERTENSIONAHA.107.104620.
7. Fouque JP, Papanicolaou G, Sircar KR. *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press: Cambridge UK, 2000.
8. Muthén B, Brown CH, Masyn BJ, Khoo S-T, Wang C-P, Kellman SG, Carlin J, Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics* 2002; **3**:459–475. DOI: 10.1093/biostatistics/3.4.459.
9. Ye W, Lin X, Taylor JMG. Semiparametric modeling of longitudinal measurements and time-to-event data – a two-stage regression calibration approach. *Biometrics* 2008; **64**:1238–1246. DOI: 10.1111/j.1541-0420.2007.00983.x.
10. Anstey KJ, Smith GA. Interrelationships among biological markers of aging, health, activity, acculturation, and cognitive performance in late adulthood. *Psychology and Aging* 1999; **14**:605–618.
11. Christensen H, Mackinnon AJ, Korten AE, Jorm AF, Henderson AS, Jacomb P, Rodgers B. An analysis of diversity in the cognitive performance of elderly community dwellers: individual differences in change scores as a function of age. *Psychology and Aging* 1999; **14**:365–379.
12. Hultsch DF, Hertzog C, Small BJ, McDonald-Miszczak L, Dixon RA. Short-term longitudinal change in cognitive performance in later life. *Psychology and Aging* 1992; **7**:571–584.
13. Nesselroade JR, Intraindividual variability and short-term change. Commentary. *Gerontology* 2004; **50**:44–47. DOI: 10.1159/000074389.
14. Martin M, Hofer SM. Intraindividual variability, change, and aging: conceptual and analytical issues. *Gerontology* 2004; **50**:7–11. DOI: 10.1159/000074382.
15. Schaie KW. The impact of longitudinal studies on understanding development from young adulthood to old age. *International Journal of Behavioral Development* 2000; **24**:257–266. DOI: 10.1159/000074382.
16. Darby D, Maruff P, Collie A, McStephen M. Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology* 2002; **59**:1042–1046.
17. Kliegel M, Sliwinski M. MMSE cross-domain variability predicts cognitive decline in centenarians. *Gerontology* 2004; **50**:39–43. DOI: 10.1159/000074388.
18. Collie A, Maruff P, Currie J. Behavioral characterization of mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology* 2002; **24**:720–733. DOI: 10.1076/jcen.24.6.720.8397.
19. Juster FT, Suzman R. An overview of the Health and Retirement Study. *Journal of Human Resources* 1995; **30**:S7-S56.
20. Taylor Jr. D H, Ostbye T, Langa KM, Weir D, Plassman BL. The accuracy of Medicare claims as an epidemiological tool: the case of dementia revisited. *Journal of Alzheimer's Disease* 2009; **17**:807–15. DOI: 10.3233/JAD-2009-1099.
21. Gelfand AE, Smith AMF. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**:389–409.
22. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC: Boca Raton, FL, 2004.
23. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 1992; **41**:337–348.
24. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989; **8**:551–561.
25. Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* 1996; **6**:733–807.
26. Lin H, Turnbull BW, Mc Culloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* 2002; **97**:5365.
27. Proust-Lima C, Letenneur L, Jacqmin-Gadda H. A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statistics in Medicine* 2007; **26**:2229–2245.
28. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*, 2nd edition. Wiley: New York, NY, 2002.