A Comparison of Linear Mixed Effects Models and RE-EM Trees for Prediction of Cognitive Decline

Kristina Vatcheva[1]

[1] School of Mathematical & Statistical Sciences, University of Texas Rio Grande Valley, Brownsville, TX 78520, USA

Abstract

Cognitive decline is common with ageing, but other risk factors may influence this process. Cognitive decline can have profound implications for individuals' well-being and its prediction and early detection can prevent and improve lives and decrease hospitalization cost. We compared the performance of linear mixed effects model and RE-EM tree on predicting cognitive decline. Data from five waves of the English Longitudinal Study of Aging (ELSA) were analyzed. RE-EM trees using 1 and 6 iterations and three linear mixed effects models, with predictors selected by RE-EM trees and with all predictors, with random intercept and a slope for time variable, were fitted on training data. Models' prediction abilities were evaluated on test data using root mean squared error (RMSE). Data were unbalanced and comprised of 12, 212 participants with a total of 42, 560 records. All liner mixed effects models resulted with better prediction performance compared to the fitted RE-EM trees (RMSE=3.57, RMSE=3.60, RMSE=3.63 vs. RMSE=3.67 and RMSE=3.68, respectively).

Key Words: Linear mixed effects model, RE-EM tree, prediction, cognitive decline.

1. Introduction

Cognitive decline is common with ageing, however other risk factors including family history, education level, brain injury, medications, physical inactivity, Parkinson's disease, Alzheimer's disease, heart disease and stroke, diabetes, and depression may influence this process as well (CDC & Alzheimer's Disease and Healthy Aging Program Home). Cognitive decline, ranging from mild cognitive impairment to dementia, can have profound implications for an individual's overall health and well-being (CDC & Alzheimer's Disease and Healthy Aging Program Home). The prediction and identification of people who are showing early signs of cognitive decline can prevent and improve the lives of many adults as well as decrease hospitalization cost (CDC & Alzheimer's Disease and Healthy Aging Program Home).

Longitudinal studies are particularly useful when studying development and lifespan issues (Caruana et al., 2015; Singer & Willett, 2003). Longitudinal studies collect measures for subjects at multiple follow-ups over a period (Caruana et al., 2015; Singer & Willett, 2003). In the recent years, longitudinal studies of ageing in multiple countries have emerged (Stanziano et al., 2010). Longitudinal studies of aging allow researchers to collect broad range of factors, to characterize the change in individuals at different points in life, to provide some of the reasons of these developmental shifts, as well as identification of periods in the lifespan when interventions will potentially have their greatest impact (Caruana et al., 2015; Zaninotto et al., 2018).

Special methods of statistical longitudinal data analysis are needed to take into account the within-subject measurements correlation over time and the change in variance of longitudinal data with time. Ignoring within-subject correlation leads to underestimation of the variance of time-independent predictor variables and overestimation of variance for time-dependent predictor variables (Dunlop, 1994). Potential correlation and variation may be combined in a covariance structure that must be considered in the analysis to draw valid statistical inferences.

Popular methods for the analysis of longitudinal data are the general linear mixed models for continuous response variable and generalized linear mixed models for non-normal response variable (Verbeke et al., 2010). These mixed models (known variously as individual growth models, random coefficient models, multilevel models, and hierarchical linear models) are extensions of the general linear models and generalized linear models, respectively, by the addition of random effect parameters and by allowing a more flexible specification of the covariance matrix of the random errors (Singer & Willett, 2003; Verbeke et al., 2010). In mixed model the random effects are subject-specific regression parameters, reflecting how the response evolves over time for each subject, while the fixed effects are population specific regression parameters, assumed to be the same for all subjects. The mixed effects model handles unbalanced data with unequally spaced time points and subjects observed at different time points (Singer & Willett, 2003).

Data mining methodology, such as decision trees, artificial neural networks, and support vector machines has increasingly been applied for analysis and prediction to health sciences and clinical research (Bellazzi & Zupan, 2008; Islam et al., 2018). A new data mining tool for analysis of longitudinal data, called the RE-EM tree, was proposed by Sela and Simonoff (2012) (Sela & Simonoff, 2012). RE-EM tree combines the structure of mixed effects models for longitudinal data with the flexibility of tree-based estimation methods (Sela & Simonoff., 2012). More specifically, the focus of the RE-EM tree is using regression trees, implemented with R package rpart and incorporating the object specific random effects to take into account the longitudinal nature of the data (Breiman et al., 1984; Sela & Simonoff, 2012; Therneau & Atkinson, 2010). If the random effects in the model are known the tree can be fit to estimate the fixed effects. If the population-level effects are known then the random effects, can be estimated using traditional mixed effect regression model. Using expectation–maximization (EM) algorithm, there is an alternation between estimating regression tree, assuming that the estimated random effects are correct, and estimating the random effects assuming regression tree is correct (Sela & Simonoff, 2012).

The objective of this study was to compare the predictive ability of linear mixed effect model with that of RE-EM tree method on cognitive decline using data from the English Longitudinal Study of Aging (ELSA).

2. Methods

2.1 Data source

In this study we used data from wave 1 to wave 5 (2002-2011) of the English Longitudinal Study of Ageing (ELSA) (Banks et al., 2019). The ELSA is a prospective and nationally representative cohort of men and women aged 50 years and over living in England. A detailed description of the goals, design and methods of the ELSA has been published elsewhere (Banks et al., 2019; Marmot et al, 2017; Steptoe et al., 2013; Zaninotto et al., 2018). A total of 12,212 participants with 2 or more waves resulting in total of 42,160 records were included in this study. The individuals had unequal number and variably spaced waves resulting in unbalanced and time-unstructured data.

The outcome variable was index of memory function with a range 0-24 (memtotb). This variable combines results from 3 memory tests: respondents today's date, carry out an instruction given to the respondent earlier in the interview and remembering a word list both immediately and after a delay (Steel et al., 2003). The predictor variables considered in this analysis were type 2 diabetes history, age, sex, marital status, employment status, level of education, having difficulties in preforming activities, mobility, current smoking status, drinking status, hypertension, depression score, stroke, cancer history, lung disfunction, dementia, Altzhimer, health history (Jacqmin-Gadda et al., 1997; Matthews et al., 2012; Wu et al., 2003; Zaninotto et al., 2018; Zheng et al., 2018).

2.2 Data Analysis

2.2.1 Linear mixed effects model

The general form of the linear mixed effects model, written in matrix notation, is:

$$Y = X\beta + Z\gamma + \varepsilon$$

where Y is a column vector of observed responses, X is the design matrix of predictor variables, β is the vector of unknown fixed effects regression parameters, Z is the design matrix of random variables, reflecting how the response evolves over time for each subject, γ is the vector of unknown random effects parameters, and ε is the vector of random errors, representing the within-subject variation.

The model assumptions are that: 1) the relationship between the response variable and (set of) predictor variables is linear, 2) the random effects are independent and identically distributed with means 0 and a positive definite covariance matrix $G, \gamma \sim N(0, G)$, 3) the random errors are normally distributed with means of 0 and a positive definite covariance matrix $R, \varepsilon \sim N(0, R)$, and no longer required to be independent and homogeneous as in the general linear model, and 4) the random effects and random errors are independent of each other. Based on these we obtain the variance of Y is $V = ZGZ' + R$.

To fit the linear mixed effects model, we used SAS PROC MIXED (SAS institute inc. 2015). We model V by setting up the random effects design matrix Z and by specifying covariance structures for G and R. The covariance structure of the random effects G models the natural between-subjects variability using RANDOM statement in PROC MIXED. It is recommended that the unstructured covariance structure be specified in the RANDOM statement, indicating that we do not want to impose any structure on the variances for intercept and variances for slopes, and on the covariance between the intercept and slopes (SAS Institute Inc, 2002). REPEATED statement directly models the correlation within subject, which is directly related to the spacing of measurements, by specifying a covariance structure for the R matrix (SAS institute Inc, 2015).

We determined the covariance structures appropriate for our model. First, using RANDOM statement with unstructured covariance structure we fitted two simpler models: the unconditional means model and the unconditional growth model (Singer & Willett, 2003). These unconditional models partition and quantify the outcome variation across people without regard to time, and across both people and time, respectively, and allow to determine if there is systematic variation in the outcome and where that variation resides (within or between individuals) (Singer & Willett, 2003). Based on the unconditional means model, we computed the intraclass correlation coefficient (ICC), computed as a ratio ICC = (between-person variance) / (between-person variance + within-person variance), to describe the proportion of the total outcome variation that lies between individuals. Next, we fitted models only with REPEATED statement and restricted maximum likelihood (REML) method of estimation and explored models with different covariance structures for *R:* independent or variance components (VC) (0 within-subject correlation), compound symmetry (CS) (constant within-subject correlation regardless of the distance between time points), unstructured (UN) (the observations for each pair of times have their own unique correlations), first-order autoregressive (AR(1)) (the correlation between observations is a function of the number of time points apart, ρ^{d} for d units apart; Toeplitz (TOEPH) (similar to AR(1) but instead of ρ^{d}, observations d units apart have correlation ρ_{d}; and special power (SP(POW)) structure (allows for unequal spacing and is generalization of AR(1)). The best covariance structure was selected based on the smallest Akaike's Information Criterion (AIC) and finite sample version of the AIC the (AICC) (Akaike, 1974). Finally, the linear mixed effects models with RANDOM and REPEATED statements and selected covariance structure for R were refitted.

The linear mixed effects models were fitted using the set of all predictor variables of interest. In addition, the liner mixed effect models were refitted using the set of the predictor variables based on the resultant RE-EM tree models.

Linear mixed effects models were also estimated using the lmer function of the lme4 package in R (Bates et al., 2015).

2.2.2 RE-EM tree

Sela and Simonoff (2012) constructed the *RE-EM* tree algorithm for longitudinal and clustered data as an iterative two-step process (Sela & Simonoff, 2012) described as follows. After the first initialization of the estimated random effects $\hat{\gamma}$ to 0, in the first step the regression tree approximating β based on the target value $y - Z\hat{\gamma}$ and the specified x ignoring the longitudinal structure is estimated using CART algorithm. In this step a set of indicator variables $I(x \in g_p)$ over all terminal nodes g_p in the regression tree is obtained. These indicator variables are then used in the second step, to fit a linear mixed effects model $Y = Z\gamma + I(x_{it} \in g_p)\mu_p + \varepsilon$ and extract $\widehat{\gamma}$. They replace the predicted response at each terminal node with the estimated population level predicted response $\widehat{\mu_p}$. These two steps are repeated until the estimated random effects $\hat{\gamma}$ converge using ML or REML method of estimation.

RE-EM tree algorithm in the first step is based on binary tree growing and pruning rules. A binary recursive splitting is a greedy top-down approach that select the best split in selected predictor variable x based on maximizing the reduction in sum of squares for the nodes (Breiman, et al., 1984; Gareth et al., 2013). Splitting continues as long as the increase in the proportion of variability account for by the tree is greater than 0.001 (the complexity parameter, cp>0.001) and the number of observations in the node being considered for splitting is greater than 20. The grown tree is then pruned based on 10-fold cross validation. The best tree is selected as the smallest tree with the largest cp value and corresponding estimated 10-fold cross validated error rate that is within one standard error of the minimum (also known as 1-SE rule) (Breiman et al., 1984; Gareth et al., 2013). In the case of missing data, the binary tree growing algorithm uses surrogate splits to automatically handle missing values in the variables.

In our analysis we fitted RE-EM trees with 1-iteration (i.e., reporting a final tree estimated with one time iteration through the two aforementioned steps) and RE-EM tress with 6-iterations using all candidate predictor variables.

RE-EM trees were fitted using REEMtree function in the R package REEMtree (Sela & Simonoff, 2012).

2.2.3 Comparison of predictive models

The study data were randomly divided into training and validating components based on the number of individuals in wave 1. Seventy five percent (75%) of the data (n=31,849) were used to derive the models and 25% of the data (n=10,311) were used to evaluate the prediction performance of the models (Picard, 1990). The prediction performance of the fitted models was evaluated using the root mean square error (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n_2} \frac{(y_i - \hat{y}_i)^2}{n_2}}$$

where y_i and $\hat{y}_i$ represent the actual and predicted value for an observation i, respectively; n_2 is the total number of observations in the testing data.

3. Results

Table 1 presents the estimated random effects covariance parameters based on fitted unconditional means and unconditional growth models for the memory function score with unstructured covariance structure, and random intercept and random intercept and slope for Time variable (measured in number of years), respectively. Based on the unconditional means model, the ICC=11.0813/(11.0813+7.15)=60.8% indicated that 60.8% of the total variation in memory function score was attributable to differences among subjects (Table 1) indicating that random effects model is appropriate to analyze these longitudinal data (Musca et al., 2011). Comparing the residual covariance parameters estimates between the two unconditional models in Table 1, there was 5.6% decrease in within-subject variation associated with Time. The unconditional growth model suggested that very small of the within-subjects variation was attributable to linear Time and that there were significant between-subjects variations in both true initial status and true rate of change.

Table 1. Random effects covariance parameter estimates based on unconditional means and unconditional growth models for the memory function score

Unconditional Means Model					Unconditional Growth Model				
Covariance Parameter	**Estimate**	**Standard Error**	**Z Value**	**P-value**	**Covariance Parameter**	**Estimate**	**Standard Error**	**Z Value**	**P-value**
UN(1,1) (In initial status)	11.0813	0.1734	63.91	<.0001	UN(1,1) (In initial status)	10.0276	0.1924	52.13	<.0001
Residual (Within-subject)	7.15	0.05818	122.9	<.0001	UN(2,1) (Covariance)	0.1392	0.02327	5.98	<.0001
					UN(2,2) (In rate of change)	0.04232	0.00417	10.15	<.0001
					Residual (Within-subject)	6.7477	0.064	105.43	<.0001
Fit Statistics					**Fit Statistics**				
-2 Res Log Likelihood	226604.2				-2 Res Log Likelihood	226249.1			
AIC (Smaller is Better)	226608				AIC (Smaller is Better)	226257			
AICC (Smaller is Better)	226608				AICC (Smaller is Better)	226257			
BIC (Smaller is Better)	226623				BIC (Smaller is Better)	226287			

Table 2 presents the fit statistics for linear mixed effects models with various covariance structures for R matrix (i.e., models fit with RANDOM statement only in SAS PROC MIXED). The results show that the model with the smallest fit statistics is the model fitted with unstructured covariance structures for R matrix.

Table 2. Fit statistics for linear mixed effects models with various covariance structures for R matrix

Fit Statistics	**VC* (Independent)**	**UN**	**CS**	**AR (1)**	**SP(POW)**	**TOEP**
-2 Res Log Likelihood	243171.7	226187.6	226613	228816	226439.7	226432
AIC (Smaller is Better)	243174	226218	226617	228820	226446	226442
AICC (Smaller is Better)	243174	226218	226617	228820	226446	226442

BIC (Smaller is Better)	243181	226329	226632	228835	226468	226480

*Default if REPEATED statement is not specified.

Figure 1 and Figure 2 show the fitted RE-EM trees with 1 and 6 iterations, respectively. The selected predictors based on 1-iteration RE-EM tree were age, education level, employment status, drinking status and sex. The selected predictors based on 6-iterations RE-EM tree were education level, age, and sex.

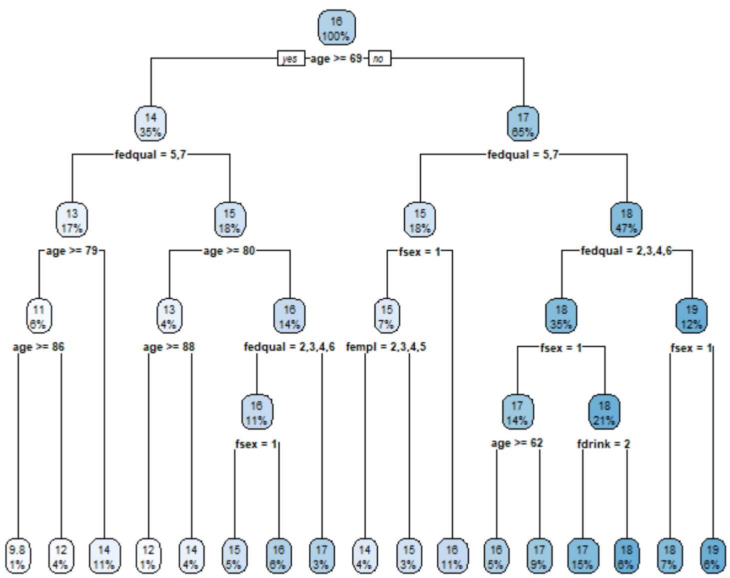

Figure 1. Estimated 1-iteration RE-EM tree for memory function score with random intercept and slope for Time variable using REML method (n=31,849)

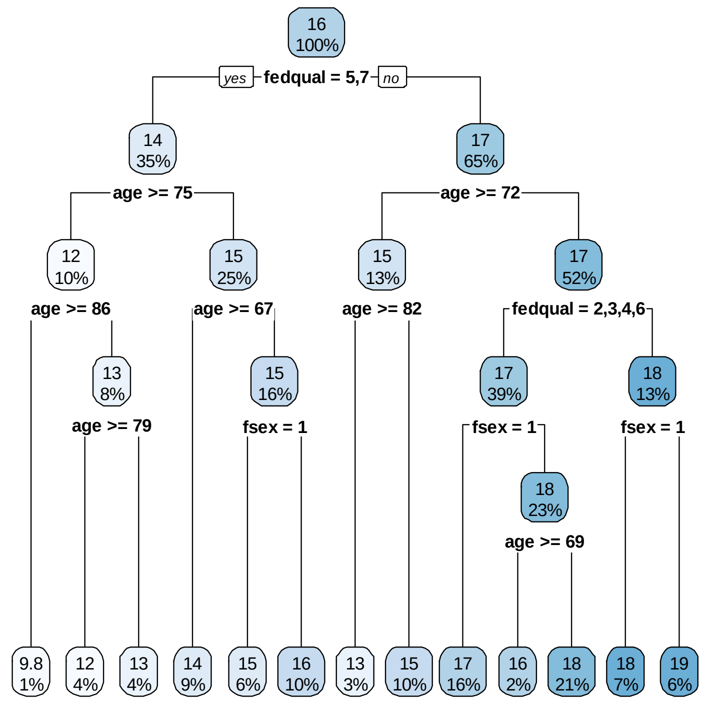

Figure 2. Estimated 6-iterations RE-EM tree for memory function score with random intercept and slope for Time variable using REML method (n=31,849)

Table 3 summarizes the computed RMSEs for various fitted linear mixed effects models and RE-EM tress. The fitted 1-iteration RE-EM tree performs better in predicting the outcome variable memory function score (RMSE=3.668023) than the fitted 6-iterations RE-EM tree (RMSE= 3.676194). The linear mixed effects model fitted using all covariates of interest as well as the reduced linear mixed effects models fitted using covariates selected based on the fitted RE-EM trees resulted with smaller RMSEs indicating that they are more effective than the RE-EM trees in predicting the memory function score (Table 3).

Table 3. Predictive performance of various fitted linear mixed effects models and RE-EM tress

Models	Root Mean Square Error (RMSE)
Linear mixed effects model with all covariates	3.565972

1-iteration RE-EM tree	3.668023
Linear mixed effects model reduced based on 1-iteration RE-EM tree	3.602161
6-iterations RE-EM tree	3.676194
Linear mixed effects model reduced based on 6-iterations RE-EM tree	3.624616

4. Discussion and future work

In this study we used empirical longitudinal data to compare the predictive performance of fitted linear mixed effects models with random intercept and random slope for Time variable and RE-EM trees estimated with 1 and 6 iterations. The results showed that the linear mixed effects models had the best predictive ability compared to the RE-EM trees.

We did not fit linear mixed effects models with both within-subject correlation and random effects covariance structures due to computational errors. Fitting a model with random effects and serial correlation might sometimes overparameterize the covariance structure, because the random effects are often able to represent the serial correlations among the measurements (SAS institute Inc, 2015; Chi et al., 1989; Diggle et al., 1994). When we specify the intercept and Time variable in a RANDOM statement, the random coefficient model indirectly models the serial correlation within subject and enable the correlations within subject to change over time (SAS institute Inc, 2015). In addition, by specifying the unstructured covariance structure we do not impose any structure on the variances for intercepts and variances for slopes, and on the covariance between the intercepts and slopes. By specifying the random intercept and slope for Time variable, the unequal time intervals are considered as well because the Z matrix is used in the computation of the V matrix (SAS institute Inc, 2015).

In our study the memory function score variable was approximately normally distributed as well as all linear mixed effects model's assumptions were satisfied. In addition, the data used in the analyses had only 0.2% missing values. Future studies need to be conducted to compare predictive performance of the mixed effects models and RE-EM trees using simulated data with different scenarios for sample size, distribution of the outcome variable, and different percentage of missing values.

References

SAS institute Inc. 2015. SAS/STAT® 14.1 user's guide. Cary, NC: SAS institute Inc.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Banks, J., Batty, GD., Coughlin, K., Dangerfield, P., Marmot, M., Nazroo, J., Oldfield, Z., Steel, N., Steptoe, Wood, M., Zaninotto, P. (2019). English Longitudinal Study of Ageing: Waves 0-9, 1998-2019. [data collection]. 33rd Edition. UK Data Service. SN: 5050, http://doi.org/10.5255/UKDA-SN-5050-20

Bates D, Mächler M, Bolker B, and Walker S. (2015). Fitting linear mixed-effects models using lme4 . *Journal of Statistical Software, 67*(1), 1-48.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics, 77*(2), 81-97. doi:S1386-5056(06)00274-7.]

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees* Wadsworth.

Caruana, E. J., Roman, M., Hernandez-Sanchez, J., & Solli, P. (2015). Longitudinal studies. *Journal of Thoracic Disease, 7*(11), 537. doi:10.3978/j.issn.2072-1439.2015.10.63.

CDC, & Alzheimer's Disease and Healthy Aging Program Home.Cognitive impairment: A call for action, now! https://www.cdc.gov/aging/pdf/cognitive_impairment/cogimp_poilicy_final.pdf. Accessed 20 May 2021.

Chi, E.M., Reinsel, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association,* (84), 452-459.

Diggle, P.J., Liang, K., Zeger, S.L. (1994). Analysis of longitudinal data New York: Oxford University Press.

Dunlop, D. D. (1994). Regression for longitudinal data: A bridge from least squares regression. *The American Statistician,* (48), 299-303.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning: With applications in R Springer, New York, NY.

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare (Basel, Switzerland), 6*(2), 10.3390/healthcare6020054. doi:E54.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., & Dartigues, J. F. (1997). A 5-year longitudinal study of the mini-mental state examination in normal aging. *American Journal of Epidemiology, 145*(6), 498-506. doi:10.1093/oxfordjournals.aje.a009137.

Marmot M, Oldfield Z, Clemens S, et al. (2017). English longitudinal study of ageing: Waves 0-7, 1998-2015. [data collection].*UK Data Service. Sn: 5050, 10.5255/Ukda-Sn-5050-14,* (27th edition).

Matthews, F., Marioni, R., Brayne, C., & Medical Research Council Cognitive Function and Ageing Study. (2012). Examining the influence of gender, education, social class and birth cohort on MMSE tracking over time: A population-based prospective cohort study. *BMC Geriatrics, 12*, 45-45. doi:10.1186/1471-2318-12-45.

Musca, S. C., Kamiejski, R., Nugier, A., Meot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology, 2*, 74. doi:10.3389/fpsyg.2011.00074.

Picard RR, B. K. (1990). Data splitting. *The American Statistician, 44*, 140–147.

SAS Institute Inc. (2002). Longitudinal data analysis with discrete and continuous responses course notes Cary, NC: SAS Institute Inc.

Sela, R.J. , Simonoff, J.S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning,* (86), 169-207.

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence (1st ed.) Oxford University Press.

Stanziano, D. C., Whitehurst, M., Graham, P., & Roos, B. A. (2010). A review of selected longitudinal studies on aging: Past findings and future directions. *Journal of the American Geriatrics Society, 58 Suppl 2*, 292. doi:10.1111/j.1532-5415.2010.02936.x [doi]

Steel, N., F. Huppert, B. McWilliams and D. Melzer. (2003). Physical and cognitive function. In marmot et al (eds), health, wealth and lifestyles of the older population in england: The 2002 english longitudinal study of ageing, london: IFS.)

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: The english longitudinal study of ageing. *International Journal of Epidemiology, 42*(6), 1640-1648. doi:10.1093/ije/dys168.

Therneau, T. M. and Atkinson B. (2010). Rpart: Recursive partitioning. R port by brian ripley. *R package version 3.*

Verbeke G., Molenberghs G., Rizopoulos D. (2010). Random effects models for longitudinal data. In van Montfort K., Oud J., Satorra A. (Ed.), Longitudinal research with latent variables. Berlin, Heidelberg.: Springer.

Wu, J. H., Haan, M. N., Liang, J., Ghosh, D., Gonzalez, H. M., & Herman, W. H. (2003). Impact of diabetes on cognitive function among older latinos: A population-based cohort study. *Journal of Clinical Epidemiology, 56*(7), 686-693. doi:S0895435603000775.

Zaninotto, P., Batty, G. D., Allerhand, M., & Deary, I. J. (2018). Cognitive function trajectories and their determinants in older people: 8 years of follow-up in the english longitudinal study of ageing. *Journal of Epidemiology and Community Health, 72*(8), 685-694. doi:10.1136/jech-2017-210116.

Zheng, F., Yan, L., Yang, Z., Zhong, B., & Xie, W. (2018). HbA1c, diabetes and cognitive decline: The english longitudinal study of ageing. *Diabetologia, 61*(4), 839-848. doi:10.1007/s00125-017-4541-7.