

2021

Bayesian modeling of neuropsychological test scores

<https://hdl.handle.net/2144/43161>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**BAYESIAN MODELING OF NEUROPSYCHOLOGICAL TEST
SCORES**

by

MENGTIAN DU

BS in Applied Mathematics, BA in Psychology, University of
California, San Diego, 2013
MS in Biostatistics, Yale University, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2021

© 2021 by
Mengtian Du
All rights reserved

Approved by

First Reader

Paola Sebastiani, PhD
Adjunct Professor of Biostatistics

Second Reader

Stacy L. Andersen, PhD
Assistant Professor of Medicine

Third Reader

Yorghos Tripodis, PhD
Research Associate Professor of Biostatistics

Fourth Reader

Sara Lodi, PhD
Assistant Professor of Biostatistics

Fifth Reader

Stefano Monti, PhD
Associate Professor of Biostatistics

ACKNOWLEDGMENTS

I would first like to express my gratitude to my main advisor, Professor Paola Sebastiani, for all her support, guidance and encouragement throughout my time at Boston University. She not only guided me to find the direction of my thesis, but also led me to the path of being a good biostatistician. She had always encouraged me during times when I was not confident and taught me how to be persistent. I would like to thank the rest of my thesis committee members, Professor Stacy Andersen, Professor Yorghos Tripods, Professor Sara Lodi and committee chair Professor Stefano Monti, for their advice and comments on my dissertation thesis. I would also like to thank Dr. Thomas Perls, the lead of the Long Life Family Study group at Boston Medical Center, for including me as part of the study team and the opportunity to work on such interesting and exciting data.

I would also like to thank my friends and classmates here at the Biostatistics Department, with whom I share so many enjoyable memories with and I look forward to our continued friendships in the years to come. And last but not least, I would like to thank my parents for their continuous support and encouragement throughout my education.

BAYESIAN MODELING OF NEUROPSYCHOLOGICAL TEST SCORES

MENGTIAN DU

Boston University, Graduate School of Arts and Sciences, 2021

Major Professor: Paola Sebastiani, PhD, Professor of Biostatistics

ABSTRACT

In this dissertation we propose novel Bayesian methods of analysis of patterns of neuropsychological testing. We first focus attention to situations in which the goal of the analysis is to discover risk factors of cognitive decline using longitudinal assessment of tests scores. Variable selection in the Bayesian setting is still challenging, particularly for analysis of longitudinal data. We propose a novel approach to selection of the fixed effects in mixed effect models that combines a backward selection algorithm and a metrics based on the posterior credible intervals of the model parameters. The heuristic of this approach is based on searching for those parameters that are most likely to be different from zero based on their posterior credible intervals, without requiring ad hoc approximations of model parameters or informative prior distributions. We show via a simulation study that this approach produces more parsimonious models than other popular criteria such as the Bayesian deviance information criterion. We then apply this approach to test the hypothesis that genotypes of the *APOE* gene have different effects on the rate of cognitive decline of participants in the Long Life Family Study. In the second part of the dissertation we shift focus on analysis of neuropsychological tests administered using emerging digital technologies. The challenge of analyzing these data

is that for each study participant the test is a data stream that records time and spatial coordinates of the digitally executed test and the goal is to extract some useful and informative summary univariate variables that can be used for analysis. Toward this goal, we propose a novel application of Bayesian Hidden Markov Models to analyze digitally recorded Trail Making Tests. Applying the Hidden Markov Model enables us to perform automatic segmentation of the digital data stream and allows us to extract meaningful metrics that correlate the Trail Making Tests performance to other cognitive and physical function test scores. We show that the extracted metrics provide information in addition to the traditionally used scores.

CONTENTS

Acknowledgments	iv
Abstract	v
List of Tables	ix
List of Figures	xii
List of Symbols and Abbreviations	xiii
1 Motivation and Introduction	1
2 Bayesian Variable Selection Utilizing Posterior Probability Credible Intervals	4
2.1 Introduction	4
2.2 Method	6
2.3 Empirical Evaluation	10
2.3.1 Simulation setup	10
2.3.2 Simulation results	13
2.4 Conclusion	14
3 Application of the CI algorithm	18
3.1 Introduction	18
3.2 Method	19
3.2.1 Study Population	19
3.2.2 Cognitive Tests	20

3.2.3	Statistical Analysis	20
3.3	Results	24
3.4	Discussion	29
3.5	Supplementary Material	33
4	Analyzing Digitally Assessed Trail Making Test Using Hidden Markov Models	44
4.1	Introduction	44
4.2	Method	46
4.2.1	Study Population and Test Measures	46
4.2.2	Hidden Markov Model and Trail Making Tests	47
4.2.3	Extracted Metrics	50
4.2.4	Statistical Analysis	52
4.3	Results	53
4.4	Discussion	60
5	Discussion	76
	List of Journal Abbreviations	78
	Bibliography	80
	Curriculum Vitae	87

LIST OF TABLES

2.1	Simulation results comparing the CI algorithm and DIC.	16
3.3a	Parameter estimates of Animal Fluency and DSST by generation. . .	25
3.3b	Parameter estimates of the Digits Span tests by generation.	27
3.4	Parameter estimates of Logical Memory tests $\epsilon 4$ allele carriers vs. non- $\epsilon 4$ allele carriers.	28
3.1	Demographic characteristics and test scores of 1,785 older genera- tion (born in or before 1935) LLFS participants.	34
3.2	Demographic characteristics and test scores of 2,802 younger genera- tion (born after 1935) LLFS participants.	35
3.s1a	Demographic characteristics and test scores of 1,785 older genera- tion (born in or before 1935) LLFS participants, broken down by <i>APOE</i> . 36	36
3.s1b	Demographic characteristics and test scores of 1,785 older genera- tion (born in or before 1935) LLFS participants, broken down by <i>APOE</i> . 37	37
3.s2a	Demographic characteristics and test scores of 2,802 younger gen- eration (born after 1935) LLFS participants, broken down by <i>APOE</i> genotype.	38
3.s2b	Demographic characteristics and test scores of 2,802 younger gen- eration (born after 1935) LLFS participants, broken down by <i>APOE</i> genotype.	39
3.s3	Parameter estimates of <i>APOE2</i> analysis.	40
3.s4	Parameter estimates of <i>APOE4</i> analysis.	41

3.s5	Parameter estimates of Animal Fluency, DSST, Digits Span tests, without stratification of <i>APOE</i> genotype.	42
3.s6	Parameter estimates of Logical Memory tests.	43
3.s7	Parameter estimates of Logical Memory tests e4 allele carriers vs. non-e4 allele carriers, adjusting for CRP level.	43
4.1	Demographic characteristics and test scores of participants who completed the Trail Making Tests.	54
4.2	Matched differences of metrics between TMT-A and TMT-B.	54
4.3a	Parameter estimates of GEE models using completion time as predictor, TMT-A.	62
4.3b	Parameter estimates of GEE models using completion time as predictor, TMT-A.	62
4.4a	Parameter estimates of GEE models using extracted metrics as predictors, TMT-A.	63
4.4b	Parameter estimates of GEE models using extracted metrics as predictors, TMT-A.	64
4.5a	Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-A.	65
4.5b	Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-A.	66
4.6a	Parameter estimates of GEE models using completion time as predictor, TMT-B.	67
4.6b	Parameter estimates of GEE models using completion time as predictor, TMT-B.	67

4.7a	Parameter estimates of GEE models using extracted metrics as predictors, TMT-B.	68
4.7b	Parameter estimates of GEE models using extracted metrics as predictors, TMT-B.	68
4.8a	Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-B.	69
4.8b	Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-B.	70
4.9a	Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.	71
4.9b	Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.	71
4.10a	Parameter estimates of GEE models using extracted metrics as predictors, difference between TMT-B and TMT-A.	72
4.10b	Parameter estimates of GEE models using extracted metrics as predictors, difference between TMT-B and TMT-A.	73
4.11a	Parameter estimates of GEE models using completion time and extracted metrics as predictors, difference between TMT-B and TMT-A.	74
4.11b	Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.	75

LIST OF FIGURES

2.1	A three-parameter example of an iteration process of the CI algorithm.	9
2.2	Simulation results for both full model and main effect model.	17
3.s1	Forest plot for standardized mean difference comparing <i>APOE2</i> vs. <i>APOE3</i> and <i>APOE4</i> vs. <i>APOE3</i> at both visits.	33
4.1	An example of drawing in TMT-A.	49
4.2	An example of drawing in TMT-A after HMM segmentation.	49
4.3	An example of drawing in TMT-A with cluster points in red color. . .	51
4.4	Pairwise scatter plot matrix for metrics in TMT-A.	55
4.5	Pairwise scatter plot matrix for metrics in TMT-B.	56

LIST OF ABBREVIATIONS

AIC	Akaike's Information Criterion
APOE	Apolipoprotein E
BIC	Bayesian Information Criterion
BUGS	Bayesian inference Using Gibbs Sampling
CI	Credible Interval
DIC	Deviance Information Criterion
DSST	Digit Symbol Substitution Test
GEE	Generalized Estimating Equation
GVS	Gibbs Variable Selection
HMM	Hidden Markov Model
HVLT-R	Hopkins Verbal Learning Test - Revised
JAGS	Just Another Gibbs Sampler
LASSO	Least Absolute Shrinkage and Selection Operator
LLFS	Long Life Family Study
MCMC	Markov Chain Monte Carlo
MMSE	Mini-Mental State Examination
SSVS	Stochastic Search Variable Selection
TICS	Telephone Interview for Cognitive Status
TMT	Trail Making Test
VIF	Variance Inflation Factor
WAIS-R	Wechsler Adult Intelligence Test - Revised
WMS-R	Wechsler Memory Scale - Revised

CHAPTER 1

Motivation and Introduction

Intact cognition is a crucial aspect of healthy aging. Although cognitive decline is associated with the normal aging process (Wetherell et al. (2002)), cognitive impairment gives rise to an increased risk of dementia, disability and mortality (Dewey & Saz (2001); Plassman et al. (2008)). Studying the process of cognitive decline, especially in the early stages of mild cognitive impairment could help us better understand the underlying protective and risk factors that lead to dementia or intact cognitive function throughout life span. The Long Life Family Study (LLFS) is a multi-center longitudinal study of extreme human longevity and healthy aging. The study recruited over 5,000 participants including members of long-lived families as well as their spouses from three sites in the United States (Boston, New York, Pittsburgh) and one site in Denmark. The LLFS participants underwent two in-person visits that were approximately 8 years apart, both included a battery of neuropsychological testings and tests for physical functions. We aim to discover different patterns of cognitive decline and not only the risk factors associated with cognitive decline but also protective factors that help with maintaining good cognitive functions in elderly population. This dissertation is motivated by data collected in the LLFS and problems and challenges arose while analyzing these data.

One challenge we face when analyzing longitudinally collected data is variable selection. Several methods have been proposed and extensively used in the frequentist approach for linear mixed effects models, including adaptations of the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and shrinkage based methods such as the Least Absolute Shrinkage and Selection Operator (LASSO). However, variable selection methods that work well with mixed

effects models in the Bayesian setting have been lacking. There are currently two main approaches in Bayesian variable selection, the first approach includes methods based on computing model probabilities that are extensions of the spike and slab method introduced by Mitchell & Beauchamp (1988), and the second approach involves methods based on model choice criteria such as the Deviance Information Criterion (DIC) Spiegelhalter et al. (2002). These methods either rely on ad hoc approximations of model parameters or on informative prior distributions that may substantially influence the results. In chapter 2, we propose a novel approach to perform variable selection in Bayesian hierarchical models using a heuristic based algorithm. This algorithm utilizes the posterior distributions of parameter estimates and their credible intervals to select parameters that are most likely different from the null value in a backward order, without requiring to compute ad hoc approximations or need of informative priors. We show in a simulation study that this algorithm produces parsimonious results in comparison with the DIC. We will then show an application of this approach to the LLFS data to analyze the effect of the apolipoprotein E (*APOE*) gene on the longitudinal change of cognitive functions in chapter 3.

As a part of the neuropsychological testing battery during the second in-person visit in the LLFS, emerging digital tools were utilized in assessing the Trail Making Tests (TMTs). TMTs are commonly used and well-established neuropsychological tests for evaluations of organic brain damage and age related diseases. Traditionally, TMTs are administered using pen and paper, where a series of numbers and letters are displayed on a piece of paper, and participants are asked to draw lines to connect the numbers and letters in sequence as quickly as possible. The scoring of the TMTs consists of the total time to completion and number of errors made

during the drawing process. The LLFS has taken a step forward to record the drawing process using a digital pen, which records and timestamps drawing coordinates 75 times per second. In chapter 4 we propose to use the recorded digital data stream to decompose total time to completion into thinking time and drawing time, with the goal to provide deeper insights of which aspect of cognitive or physical functions contribute to the overall performance of the TMTs. We also propose a novel application of the Hidden Markov Models (HMMs) to perform automatic segmentation of the coordinate pairs to extract number of connections drawn and to use the summary statistics of the connections as new metrics of the TMTs. We will correlate these new metrics with other cognitive and physical function tests in the LLFS to test our hypothesis that the digitally recorded data stream could provide additional information in analyzing the underlying mechanism of the TMTs performance that is not described by the overall completion time.

CHAPTER 2

Bayesian Variable Selection Utilizing Posterior Probability Credible Intervals

2.1 INTRODUCTION

An important problem in statistics is choosing the model that best describes the data from a set of a priori plausible models. This problem is often reduced to variable selection from a set of explanatory variables assuming a general linear regression model, and there are many criteria and search procedures that are applicable to modeling data from cross-sectional studies Kadane & Lazar (2004). Longitudinal data and repeated measurements data are common data types collected in medical research. Many variable selection methods have been proposed and extensively used for linear mixed effects models, including adaptation of the information criteria such as AIC, BIC, and shrinkage based methods such as LASSO. A review of some of these methods is provided by Muller et al. (2013).

In recent years, there has also been a proliferation of Bayesian variable selection methods based on computing model probabilities or model choice criteria. Many methods based on computing model probabilities are extensions of the spike and slab method first introduced by Mitchell & Beauchamp (1988). The general idea of this approach is to assign a prior distribution that mixes a point mass distribution at the null model and a diffuse uniform distribution elsewhere for each candidate variable. For each model, the posterior probabilities for both the vector of inclusion and coefficients are calculated through Markov Chain Monte Carlo (MCMC) techniques. The final subset of predictors are then selected based on a pre-specified threshold on the posterior probability. The Stochastic Search Variable Selection (SSVS) method proposed by George & McCulloch (1993) uses a mixture of prior

distributions on possible models and uses the Gibbs sampling to identify the models with high posterior probability. The Gibbs Variable Selection (GVS) method suggested by Dellaportas et al. (2002) samples the parameter estimates from a mixture pseudo-prior that is concentrated around the posterior density of regression coefficients. Adaptive shrinkage methods such as Bayesian LASSO proposed by Park & Casella (2008) specify a prior distribution directly over the regression coefficients to induce sparseness of the model. Rather than placing prior probabilities directly on the regression coefficients of the individual covariates, one could view the model as a whole and place prior directly on the number of covariates and their coefficients. Methods using this approach include reversible jump MCMC first proposed by Green (1995), and composite model method introduced by Godsill (2001). Some of these approaches could achieve a relatively fast computation speed and good separation of variables as shown by O'Hara & Sillanpaa (2009). However, these methods rely on computing model posterior probabilities and are highly sensitive to the choice of priors so that the results may vary substantially with different prior distributions. The reviews by O'Hara & Sillanpaa (2009), Dellaportas et al. (2002), and George & McCulloch (1997) provide additional details and discussion.

There are limited options for Bayesian model selection criteria that work well with mixed effect models. A well-known criterion is the Deviance Information Criterion (DIC) that was proposed by Spiegelhalter et al. (2002). Similar to AIC, DIC penalizes for larger number of effective parameters in the model and models with smaller DIC are preferred. The criterion is implemented in OpenBUGS (BUGS, Bayesian inference Using Gibbs Sampling) and in JAGS (Just Another Gibbs Sampler). The latter implementation requires to run at least two parallel chains in the

model. Recently Gelman et al. (2019), introduced a Bayesian version of R^2 , but this criterion needs to be evaluated.

Here we propose a variable selection algorithm that utilizes the parameters posterior credible intervals to identify the variables to be retained in a model. The algorithm does not need "ad hoc" prior distributions of the regression parameters. It is computationally efficient and produces reasonable results. Vague conjugate priors can be assigned to the model coefficients without any need for tuning.

We will describe the algorithm in the next section. In Section 2.3, we describe the results of comprehensive simulations that show this algorithm on average produces more parsimonious models compared to DIC. In Section 4, we use the algorithm to test the hypothesis that genotypes of the *APOE* gene correlate with changes of cognitive function in a cohort of centenarians described in (Sebastiani & Perls (2012)). Conclusions and suggestions for future work are in Section 5.

2.2 METHOD

Consider a Bayesian model with outcome y_i for observation i ($i = 1, \dots, N$), a set of $h + p$ possible predictors consisting of h variables $Z = (z_1, \dots, z_h)$ to be kept in every model, and p candidate variables $X = (x_1, \dots, x_p)$, where only an unknown subset of the p candidate variables may be relevant. The Z and X variables could be, for example, main effects and interaction terms. We denote the set of possible parameter choices as $\gamma_m = (\gamma_1, \dots, \gamma_p)^T$, where γ_j takes on values:

$$\gamma_j = \begin{cases} 1, & \text{if } X_j \text{ is in model } m \\ 0, & \text{otherwise} \end{cases}$$

We assume that the variables $X = (x_1, \dots, x_p)$ can only have fixed effects, so that the regression model can then be expressed as

$$y_i = \xi_{0i} + \sum_{k=1}^h \theta_{i,k} z_{i,k} + \sum_{j=1}^p \gamma_j \beta_j x_{i,j} + e_i$$

where $e_i \sim N(0, \sigma^2)$ is the normally distributed error term with mean 0 and variance σ^2 with Gamma prior. The term $\xi_{0i} \sim N(\xi_0, \sigma_\xi^2)$ denotes a "random intercept" that we assume follows a normal distribution with mean ξ_0 and variance σ_ξ^2 , where ξ_0 is the "fixed effect intercept" with mean 0 and prior variance σ_ξ^2 . The parameter $\theta_{i,k} \sim N(\theta_k, \sigma_\theta^2)$ denotes the k th "random effect" for observation i , and θ_k denotes the k th "fixed effect" that we also assume follows a normal distribution with mean 0 and prior variance σ_θ^2 . The parameters β_j are fixed effects parameters that we assume are a priori independent and normally distributed with known mean and variance.

There are a total of $|\gamma| = 2^p$ plausible models based on the combinations of 0s and 1s in γ_m , and we want to select the model that best describes the data. Denote $C_\alpha = (C_{\alpha,1}, \dots, C_{\alpha,p})$ as the set of posterior credible intervals for the parameters $\beta = (\beta_1, \dots, \beta_p)$ of the p candidate variables, where $1-\alpha$ denotes the posterior coverage and each credible interval $C_{\alpha,j}$ consists of a lower bound $C_{\alpha,j,LB}$ and an upper bound $C_{\alpha,j,UB}$, $C_{\alpha,j} = (C_{\alpha,j,LB}, C_{\alpha,j,UB})$. The credible interval (CI) algorithm proposed in this paper utilizes the credible intervals and its lower and upper bounds to perform variable selection.

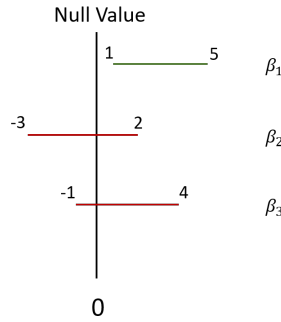
The main idea behind the CI algorithm is to use the backward elimination method, first introduced by Marill & Green (1963) in the early 1960s, and a Bayesian metric. In a traditional non-Bayesian multiple linear model, backward

elimination begins with all candidate variables in the model and removes the least significant (largest p-value) variable one at a time in an iterative way. In a Bayesian framework, we can utilize the posterior credible interval to help quantify the importance of each variable in the final model. We frame the variable selection problem as a hypothesis testing problem, in which rejecting the null hypothesis $H_0 : \beta_j = \beta_{0,null}$ leads to include the covariate X_j in the model. When the credible interval for the parameter β_j includes the null value, there are two alternative scenarios to consider. In the first scenario, the lower and upper bounds of the posterior credible interval are both far away from the null value, in other words, the null value falls well within the credible interval. In this case, we can say with some confidence that this variable is unlikely to be important. The second scenario is when either the lower or the upper bounds is close to the null value. In this case as more variables are dropped from the model, this variable will be more likely to be retained in the final model compared to the first scenario. In a non-Bayesian model, the second scenario could be interpreted as borderline significant. At each iteration step, we wish to remove the variable that is most likely to have a regression coefficient that satisfy the null hypothesis. The rationale for the CI algorithm is that, for any credible interval that contains the null value, the minimum of the absolute value of the difference of the two bounds from the null value represents the importance of this variable, consequently the CI algorithm identifies the variable with the minimum evidence against the null hypothesis to be removed at each iteration.

To perform variable selection, we first standardize all candidate variables to ensure that the regression coefficients are on the same scale. We start with the full model that includes all h fixed variables Z and all p candidate variables X . In

the first iteration, we obtain the posterior credible intervals $C_{\alpha,j}$ ($j = 1, \dots, p$) for the parameters β_j ($j = 1, \dots, p$). If the credible interval of β_j does not include the null value ($\beta_{j,null} \notin C_{\alpha,j}$), then x_j can remain in the model for the next iteration. Out of the β_j 's with $\beta_{j,null} \in C_{\alpha,j}$, we remove from the model the variable x_j with $\max_{j \in p'} \{\min(|C_{\alpha,j, LB} - \beta_{j,null}|, |C_{\alpha,j, UB} - \beta_{j,null}|)\}$, where p' is the subset of parameters with $\beta_{j,null} \in C_{\alpha,j}$, and we repeat this process until all remaining candidate variables have $\beta_{j,null} \notin C_{\alpha,j}$. A three-parameter example of this iteration process is illustrated in Figure 2.1.

Figure 2.1: A three-parameter example of an iteration process of the CI algorithm.



In this example, there are three candidate variables for variable selection, namely x_1 , x_2 , and x_3 , with corresponding posterior estimates β_1 , β_2 , and β_3 and 95% credible intervals $C_{0.95,1} = (1, 5)$, $C_{0.95,2} = (-3, 2)$, and $C_{0.95,3} = (-1, 4)$. The null value is 0 for all β s in this example. In the first iteration, the initial step is to identify the β s with 95% CI that include 0 ($0 \in C_{0.95,j}$), this case β_2 and β_3 , while β_1 has a 95% CI that does not include 0 and it can remain in the model for the next iteration. The second step is to identify which variable between x_2 and x_3 to remove by finding the maximum of the minimum of the absolute value of the lower and upper bounds of the 95% CI ($\max_{j \in p'} \{\min(|C_{\alpha,j, LB} - 0|, |C_{\alpha,j, UB} - 0|)\}$).

In this example, $\min(|C_{0.95,2,LB} - 0|, |C_{0.95,2,UB} - 0|) = \min(|-3|, |2|) = 2$ for β_2 , and $\min(|C_{0.95,3,LB} - 0|, |C_{0.95,3,UB} - 0|) = \min(|-1|, |4|) = 1$ for β_3 . Since 2 is greater than 1, we decide to remove x_2 in this iteration. We will refit the model with only x_1 and x_3 in the next iteration and repeat the previous steps until all remaining candidate variables have posterior estimates 95% CI not including 0.

A method that has been extensively used and well implemented in Bayesian model selection is the DIC proposed by Spiegelhalter, et al. in 2002 Spiegelhalter et al. (2002). DIC is calculated as:

$$DIC = \bar{D} + pD, \text{ where}$$

$$\bar{D} = E_{\theta|y}[D(\theta)], D(\theta) = -2\log P(y|\theta), \text{ and}$$

$$pD = E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta])$$

DIC is composed of two parts: the posterior mean of deviance (\bar{D}) and the effective number of parameters (pD). Similar to the AIC, DIC penalizes for larger number of effective parameters in the model. We will be comparing the proposed credible interval algorithm and the DIC in the next section.

2.3 EMPIRICAL EVALUATION

2.3.1 Simulation setup

We conducted a simulation study to evaluate the sensitivity and specificity of the proposed CI algorithm and to compare it with model selection based on DIC. Data used for simulation were generated based on the Digits Span Forward/Backward test from the Long Life Family Study (LLFS), which is a multi-center, longitudinal,

family-based study of healthy aging and longevity (Newman et al. (2011)). The Digits Span test is a neuropsychological test that assesses auditory attention and working memory with score ranging from 0 to 14. The test was administered to approximately 4,800 LLFS participants at enrollment, between 2006 and 2009. A second administration of the test occurred approximately 8 years later, in about 2500 participants, for a total N=7,289. We built two regression models of the test score, using as predictors age at enrollment, follow-up time, gender, years of education, and an indicator for familial longevity (whether they were a member of a long-lived family or a spouse control). One model included only main effects, and the second model included the pairwise interactions between age at enrollment, follow-up time, and familial longevity indicator, sex, and education. We then used the two models to simulate datasets and to evaluate the accuracy of the algorithm in identifying the generating model. To perform variable selection using the CI algorithm, we used standardized covariates and started from the full model and went through the iterative steps to select the final model. The final selected model was checked against the simulated model to obtain the level of concordance. The full model had the following form:

$$\begin{aligned}
y_{ij} = & \beta_0 \times (1 - rep.ind_i) + \beta_{0i} \times rep.ind_i + \\
& \beta_{age} \times age.b_i + \beta_{dage} \times dage_{ij} + \beta_{sex} \times sex_i + \beta_{educ} \times educ_i + \\
& \beta_{fam.ind} \times fam.ind_i + \\
& \beta_{sex \times age} \times sex_i \times age.b_i + \beta_{sex \times dage} \times sex_i \times dage_{ij} + \\
& \beta_{educ \times age} \times educ_i \times age.b_i + \beta_{educ \times dage} \times educ_i \times dage_{ij} + \\
& \beta_{fam.ind \times age} \times fam.ind_i \times age.b_i + \beta_{fam.ind \times dage} \times fam.ind_i \times dage_{ij} + \epsilon_{ij}
\end{aligned}$$

where y_{ij} represents the test score of the i^{th} individual at the j^{th} visit ($j=0$ for baseline and $j=1$ for follow-up) that we assumed to be normally distributed. To account for repeated measurements using a random intercept per study participant, we created an indicator variable *rep.ind* with value 1 if subject i had more than one measurements, and 0 otherwise. The covariates *age.b* and *dage* denoted age at enrollment and follow-up time in years. The variable *sex* was a binary variable with value 1 for males and 0 for females, and variable *educ* was an ordinal variable with values 0-17 that approximated years of education. The variable *fam.ind* was an indicator variable with value 1 if subject i was a member of a long-lived family and 0 if subject i was a spouse control. The random intercept term β_{0i} was assigned a normal prior distribution with mean β_0 and precision parameter τ , which had Gamma distribution with both shape and scale parameters equal to 1. All other covariates were assigned normal prior distributions with mean 0 and precision 0.1.

A mismatch occurred when the CI algorithm falsely detected an interaction term in the analysis of the data generated from the main effects model (a false positive) or failed to detect a real interaction in the analysis of the data generated from the full model (a false negative). We run the simulations with four different sample sizes: 500, 1,000, 5,000, and the largest sample size we had in our original dataset, 7,289, and generated 100 data set for each sample size. We also performed model selection in the same datasets using DIC. In each simulation setting (main effect model and full model) and with each sample size, we ran two parallel chains of the MCMC in the R package JAGS, and computed the DIC for each of the $2^p=64$ models and selected the final model with the smallest DIC. All Bayesian models in the CI algorithm were ran with 2,000 adaptations and 5,000 iterations, and all Bayesian models used in the DIC method were ran with 2,000 adaptations and 1,000

iterations to reduce the computing time.

All analyses were run in R3.5.1 using the rjags package version 4-6.

2.3.2 Simulation results

The results of the simulation are shown in Table 2.1 and Figure 2.2. When the data were generated from the full model both the CI algorithm and DIC detected the correct model with 100% accuracy for sample sizes > 500 . With the smallest sample size 500, the CI algorithm detected the correct model with 74% accuracy and 26% 1-mismatch rate (only five out of the six interaction terms were detected), and DIC detected the correct model with 87% accuracy and 13% 1 mismatch rate. This was the only scenario where the DIC performed better than the CI algorithm in the full model setting. When the data were generated from the model with main effects only, the CI algorithm detected the correct model with 100% accuracy with sample sizes 500, 1,000, and 7,289, and with 99% accuracy and 1% 2-mismatch rate (falsely detect two interaction terms) with sample size 5,000. However, DIC only detected the correct model with 37%, 25%, 9% and 3% accuracy in the sample sizes 500, 1,000, 5,000, and 7,289, respectively. As shown in Table 2.1 and Figure 2.2, DIC tended to falsely select interaction terms which are more concentrated with one, two, and three mismatches. The simulation results suggest that the CI algorithm favors more parsimonious models than DIC but this property may lead to miss important effects with small sample sizes. DIC appears to favor models with more redundant parameters and may be a better choice for predictive modeling. However, performing a full search using DIC is only computationally feasible with a relatively small number of parameters. In a model with p candidate variables, calculating the DIC for 2^p models could be extremely computationally intensive as

p gets large. Although one can also perform variable selection using a backward approach with DIC, it is still very computationally demanding because it requires running multiple chains for each model.

2.4 CONCLUSION

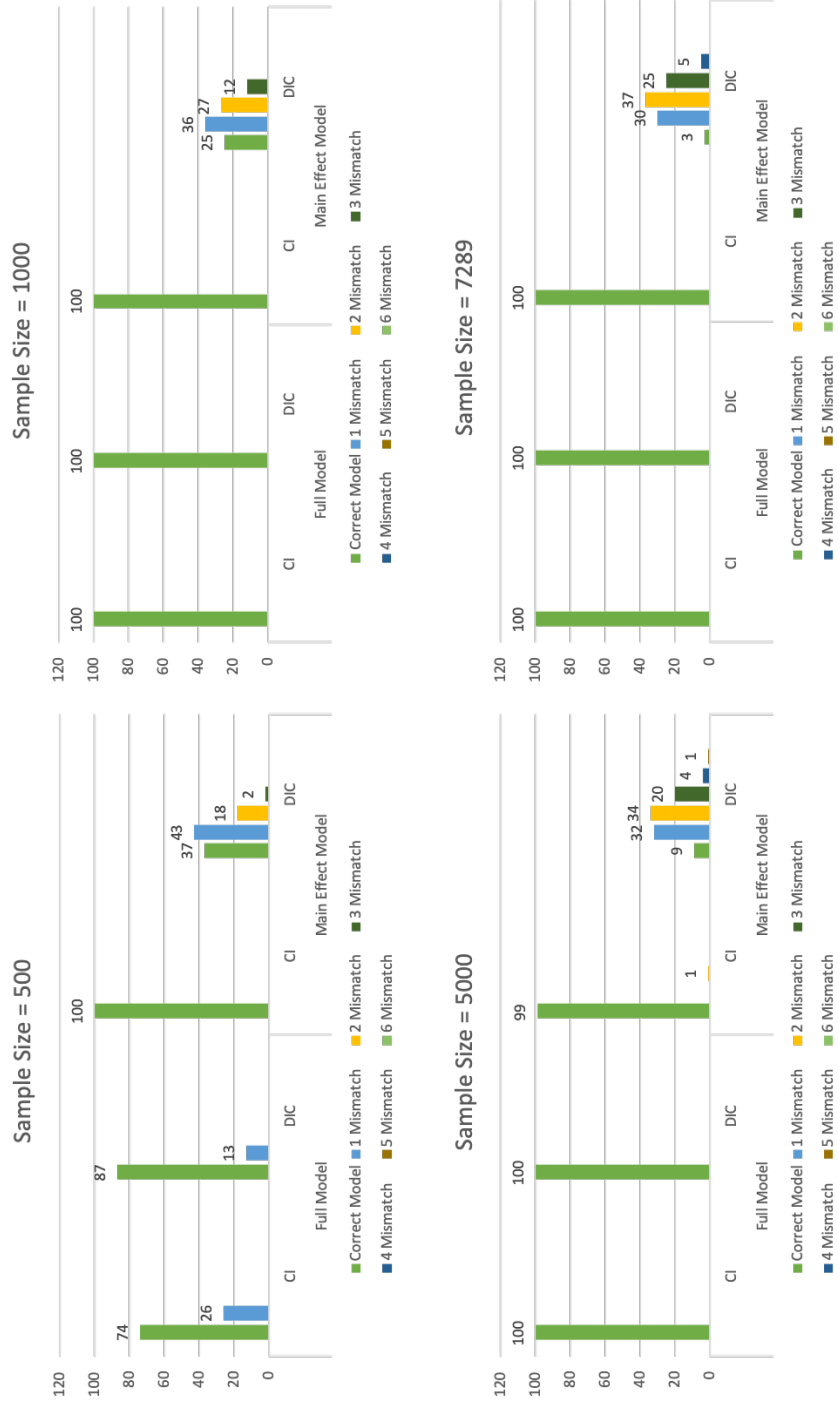
In this article, we proposed the CI algorithm: a novel approach to perform Bayesian variable selection utilizing the posterior credible intervals. Inspired by the backward elimination variable selection method in linear regression models, this algorithm removes candidate variables one at a time by quantifying and comparing the strength of the association of possible predictors with the outcome. We conducted a comprehensive simulation to assess the sensitivity and specificity of the algorithm and the simulation suggests that the algorithm is accurate with relatively large samples, and compared to DIC tends to produce more parsimonious models with smaller false positive rate.

There are a few advantages of our proposed CI algorithm relative to some variable selection methods based on computing model probabilities and methods based on computing model choice criteria such as DIC. First, the CI algorithm does not require to specify prior distributions on the model coefficients and the inclusion probabilities of each predictor, or a posterior threshold to decide whether or not to include a predictor. Secondly, our method produces parsimonious models. The final selected model all have 95% credible intervals not containing the null value. Thirdly, although in this paper we illustrated our proposed algorithm using a mixed effects model with random intercepts, it can be extended and applied to a more generalized form of linear models. For variables with random effects, we can still perform variable selection on the fixed part of the random effects. And

lastly, the CI algorithm is computationally efficient and the largest number iterations (number of Bayesian models to run) is equal to the number of candidate variables. DIC is computationally intensive since it requires to run at least two parallel MCMC chains of all possible 2^p models. Spike and slab based methods require to run long MCMC chains to search over model space, which also could be computationally demanding. However, computation load of our algorithm could still get immense in the case of high-dimensional data where $p \gg n$. Bondell & Reich (2012) proposed a Bayesian variable selection approach utilizing joint credible regions that can be suited for the high-dimensional case, though this approach yields higher false positive rate in the low-dimensional setting compared to our proposed algorithm based on their simulation study.

Overall, our proposed method performs well with reasonable number of parameters. It yields results with a parsimonious number of parameters and is computationally efficient.

Figure 2.2: Simulation results for both full model and main effect model.



CHAPTER 3

Application of the CI algorithm

Association between *APOE* alleles and Change of Neuropsychological Tests in the Long Life Family Study

3.1 INTRODUCTION

Cognitive decline, both normal and pathologic, is one of the most common complications of reaching older age. Preservation of good cognitive function or delaying the onset of cognitive decline is essential in maintaining quality of life in older adults and it is important to identify risk factors of onset and rate of cognitive decline that can suggest therapeutic interventions. Several known factors including cardiovascular risk factors, alcohol use, smoking, high systemic levels of inflammatory markers, as well as socioeconomic status contribute to the cognitive decline process (Gottesman et al. (2017); Krell-Roesch et al. (2017)). Cognitive decline patterns vary among older adults, and are genetically regulated (Fan et al. (2019)). The apolipoprotein E (*APOE*) gene is one of the most important genes related to cognition. The gene has 3 alleles, namely $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$ that result from the combination of the variations of two single nucleotide polymorphisms rs7412 and rs429358 (Liu et al. (2013)). Several studies have shown that carriers of the $\epsilon 4$ allele are at increased risk of dementia and Alzheimers disease, while the $\epsilon 2$ allele might have a protective effect against age-related neurodegenerative diseases (Henderson et al. (1995); Raber et al. (2004)), and is associated with extreme human longevity. (Sebastiani et al. (2019); Wolters et al. (2019)) A relatively small number of studies have investigated the effect of both alleles on the rate of cognitive decline using longitudinally collected data.(Blair et al. (2005); Caselli et al. (2004);

Kim et al. (2017)) The review by O'Donoghue et al. (2018) lists 40 studies of the association between *APOE* and cognition in longitudinal studies and only one study showed a protective effect of *APOE* ϵ 2 on verbal episodic memory, while other studies showed a negative effect of *APOE* ϵ 4 on various measures of cognition. Most of these studies were small (median sample size = 550), with largest sample size of 5,544 and length of follow up ranging between 2 and 30 years (median = 5.6). Several factors may contribute to inconsistent findings, including sample size, short follow-up time, neuropsychological tests used, neurobiological mechanisms, and population ancestry. The Long Life Family Study (LLFS) recruited over 5,000 individuals from longevous families. Participants underwent two in-person assessments, approximately 8 years apart, and attention, memory, and executive function were assessed through a battery of neuropsychological tests. The study included a relatively large sample of carriers of the *APOE* ϵ 2 allele and provides a unique opportunity to assess whether *APOE* is associated with cross-sectional or longitudinal cognitive decline in this healthy aging cohort. In line with previous findings, we hypothesize that carriers of the *APOE* ϵ 4 allele have increased risk for poorer cognitive function, while carriers of the ϵ 2 allele are protected against cognitive decline.

3.2 METHOD

3.2.1 Study Population

The LLFS is a multicenter longitudinal study for healthy aging and familial longevity that recruited 5,086 participants from three sites in the United States (Boston, New York, Pittsburgh) and one site in Denmark. The recruitment process and inclusion criteria for this study have been described in Newman et al.

(2011) and were based on a metric of familial longevity that was calculated from the aggregated survival probabilities of family members (Sebastiani et al. (2009)). The study recruited spouses of members of long-lived families as referents. The participants completed two in-person visits, where their physical and cognitive functions were assessed through questionnaires, performance measures, and neuropsychological tests. Approximately 4,700 participants provided blood samples for genotyping, and *APOE* alleles were determined from the SNPs rs7412 and rs429358 that were genotyped using real time PCR. *APOE* alleles were defined as $\epsilon 2$: rs7412=T; rs429358=T, $\epsilon 3$: rs7412=C; rs429358=T, $\epsilon 4$: rs7412=C; rs429358=C. All subjects provided informed consent and data are available via dbGaP (dbGaP Study Accession: phs000397.v1.p1).

3.2.2 Cognitive Tests

Six neuropsychological tests were the main outcomes in our analyses. These tests include Verbal Fluency (category fluency for animals) to assess semantic memory and generativity; Digit Symbol Substitution Test (DSST) from the Wechsler Adult Intelligence Test (WAIS-R, 5) for processing speed; Digit Span forward and backward to measure working memory and attention; and Logical Memory (immediate and delayed recall) from the Wechsler Memory Scale Revised (WMS-R, 4) to assess attention and episodic memory. The Mini-Mental State Examination (MMSE) was also administered but was not included in the analysis because of low variability.

3.2.3 Statistical Analysis

Participants were divided into three genotype groups defined as:

- *APOE2* group: carriers of the *APOE* genotypes $\epsilon 2\epsilon 2$ or $\epsilon 2\epsilon 3$;

- *APOE3* group: carriers of the genotype $\epsilon 3\epsilon 3$;
- *APOE4* group: carriers of the genotypes $\epsilon 3\epsilon 4$ or $\epsilon 4\epsilon 4$.

We used *APOE3* as reference group. We summarized participants characteristics using mean and standard deviation. We compared participants characteristics of the *APOE2* and *APOE4* groups to the *APOE3* group using t-tests or χ^2 tests. We analyzed the effect of $\epsilon 2$ and $\epsilon 4$ in two separate analyses using additive genetic models. To test for association between *APOE* alleles and each of the neuropsychological tests, we used Bayesian hierarchical modelling of the longitudinal values of each test score as a function of age at enrollment, follow-up time, gender, education, field center, birth cohort indicator (≤ 1935 , or > 1935), and the number of copies of $\epsilon 2$ or $\epsilon 4$ alleles. The full model for both *APOE2* and *APOE4* analyses had the following form:

$$\begin{aligned}
y_{ij} = & \beta_0 \times (1 - rep.ind_i) + \beta_{0i} \times rep.ind_i + \\
& \beta_{age} \times age.b_i + \beta_{dage} \times dage_{ij} + \beta_{sex} \times sex_i + \beta_{educ} \times educ_i + \\
& \beta_{APOE} \times APOE_i + \beta_{field.center} \times (fc_i - mu.fc_i) + \beta_{ind1935} \times ind1935_i + \\
& \beta_{sex \times age} \times sex_i \times age.b_i + \beta_{sex \times dage} \times sex_i \times dage_{ij} + \\
& \beta_{educ \times age} \times educ_i \times age.b_i + \beta_{educ \times dage} \times educ_i \times dage_{ij} + \\
& \beta_{APOE \times age} \times APOE_i \times age.b_i + \beta_{APOE \times dage} \times APOE_i \times dage_{ij} + \\
& \beta_{ind1935 \times age} \times ind1935_i \times age.b_i + \beta_{ind1935 \times dage} \times ind1935_i \times dage_{ij} + \epsilon_{ij}
\end{aligned}$$

where y_{ij} denotes the j^{th} ($j=1$ baseline, $j=2$ follow-up) test score of the i^{th} participant. The term ϵ_{ij} is the normally distributed random error with constant variance that was assigned an inverse Gamma prior distribution. The model intercept

$\beta_0 \times (1 - rep.ind_i) + \beta_{0i} \times rep.ind_i$ included the indicator $rep.ind_i$ that takes on value 1 if the i^{th} participant had repeated measurements, and 0 otherwise, the fixed effect β_0 , and the random effect β_{0i} with normal prior distribution with mean β_0 and precision parameter τ , which had Gamma prior distribution with both shape and scale parameters equal to 1. This parameterization used random effect only for participants with more than one assessments. Family structure and within family correlation were ignored in this model. The covariates $age.b$ and $dage$ coded ages at enrollment and follow-up time in years. The covariate sex took on value 1 for male and 0 for female, and the covariate $educ$ was an ordinal variable taking values 0-17 that approximates years of education. We did not use random slopes in the model to be able to include observations from every individual with at least one measurement of a test. Out of the four field centers, we used Boston as the referent site, and created three dummy variables for each of the other three field centers that take on values 1 and 0. The field center variables were centered to promote better convergence in the Markov Chain Monte Carlo (MCMC) chain. As reported in a previous analysis of the LLFS study (Sun et al. (2015)), the birth year cutoff 1935 was used to distinguish the older and younger generations, we created an indicator variable $ind1935$ to have value 1 if a participant was born after 1935 (the younger generation in LLFS), and 0 otherwise (the older generation in LLFS). We initially included the eight interaction terms to represent the varying effects of sex, education, generation indicator and *APOE* at age at enrollment cross-sectionally, and over different lengths of follow-up time. Besides the random intercept term β_{0i} , we modelled all main effects and interaction terms to follow Normal prior distributions with mean 0 and precision 0.1. We focused on testing the following hypotheses, adjusting for the other factors,

1. there is a significantly different effect of age at enrollment on neuropsychological test scores comparing the ϵ_2 and ϵ_4 alleles to the ϵ_3 allele ($\beta_{APOE \times age} \neq 0$);
2. there is a significantly different rate of change over time of neuropsychological test scores comparing the ϵ_2 and ϵ_4 alleles to the ϵ_3 allele ($\beta_{APOE \times dage} \neq 0$);
3. there is a significant difference in neuropsychological test scores comparing the ϵ_2 and ϵ_4 alleles to ϵ_3 allele ($\beta_{APOE} \neq 0$).

To test these hypotheses, we implemented an automated model selection algorithm that utilizes the credible intervals of the parameter estimates (Du et al., 2021, manuscript in preparation) to retain only significant interactions and main effects in the model. After the model selection process, the data of those tests that were not associated with either *APOE2* or *APOE4* groups,

Before running the model selection algorithm, we standardized the variables *age.b*, *dage*, *sex*, *educ*, *ind1935* and *APOE* (to mean=0 and SD=1) to keep the parameters on the same scale. The algorithm started from the full model described above, it approximated the posterior distributions of the parameters with MCMC using 2,000 adaptations and 10,000 iterations, and recorded the lower and upper bounds of the 95% credible interval (CI) of each of the 6 interaction terms (interactions between *sex*, *educ*, *APOE* and *age.b* and *dage*). We kept the interaction terms with a 95% CI that did not include 0 in the model for that iteration, while the interaction terms with 95% CI that did include 0 were candidates for removal. To choose the least important interaction to drop, we calculated the minimum of the absolute value of the two interval limits (2.5% and 97.5% quantiles) and we dropped the interaction term with the largest of these values, which would be the

least probable to be different from 0, and refitted the model. This procedure was repeated until all interaction terms remaining in the model had their 95% CI not including 0. This algorithm is computationally efficient since it will go through at most six iterations, and the selected model is guaranteed to have all interaction terms statistically significant. To obtain the parameter estimates on the original scale, main effect and interaction terms were scaled back by dividing by their standard deviations. The parameter estimates of age at baseline and follow-up time for the younger generation in Tables 3.3a, 3.3b & 3.4 were calculated by adding the main effects of age at baseline and follow-up time and their interactions with the birth cohort indicator, then scaled back by dividing by their standard deviations.

The LLFS data used in this analysis was frozen by June 2018.

3.3 RESULTS

Out of 5,086 LLFS participants we excluded 22 participants with missing sociodemographic data, 387 participants with missing *APOE* genotype, and 90 participants with *APOE* genotype $\epsilon 2\epsilon 4$. Tables 3.1 and 3.2 summarize demographic characteristics and cognitive test scores of the remaining 4,587 participants (1,785 in the older generation and 2,802 in the younger generation, Supplementary Tables 3.s1a, 3.s1b, 3.s2a & 3.s2b show similar information with breakdown of each *APOE* genotype). The *APOE3* group (n=3,038) was the most prevalent and was used as the referent group. Age at enrollment ranged from 71 to 110 years among the older generation (mean=88.3 years; SD:7.71), and from 25 to 73 years among the younger generation (mean=59.7 years; SD:7.10). In the older generation, the *APOE2* group was older (89.4 years vs. 88.3 years, $p=0.02$), while the *APOE4* group was younger (86.8 years vs. 88.3 years, $p<0.004$) than the *APOE3* group at enrollment. At visit

2, there were no other significant differences in age, sex, education, percent deceased among genotype groups. In the younger generation, there were no significant differences in the demographic characteristics among the genotype groups. At enrollment, the *APOE4* group had significantly lower DSST score (50.4 vs. 51.7, $p=0.03$) and lower Digit Span Backward score (6.6 vs. 6.9, $p=0.02$) than the *APOE3* group. At visit 2, the *APOE4* group had lower DSST score than the *APOE3* group (47.4 vs. 48.9, $p=0.02$). There were no significant differences in the cognitive test scores comparing the *APOE2* group to the *APOE3* in both generations at either visit. Tables 3.3a, 3.3b & 3.4 and Supplementary Tables 3.s3, 3.s4, 3.s5 & 3.s6 show parameter estimates generated using the model selected with the credible interval algorithm. The overall conclusion is that there was no significant effect of the $\epsilon 2$ allele on either the baseline assessment or the rate of change over follow-up time on any of the neuropsychological tests, while the $\epsilon 4$ allele had a negative effect on the two logical memory tests at baseline but had no effect on their rate of decline. We describe below the details of the analysis of each test. *Animal Fluency*. Neither

Table 3.3a: Parameter estimates of Animal Fluency and DSST by generation.

	Verbal Fluency		DSST	
	Older Generation	Younger Generation	Older Generation	Younger Generation
age	-0.17(-0.18,-0.15)	0.01(-0.03,0.06)	-0.67(-0.71,-0.64)	-0.4(-0.48,-0.31)
dage	-0.06(-0.09,-0.03)	0.18(0.11,0.25)	-0.48(-0.54,-0.43)	-0.06(-0.2,0.07)
sex, male	-0.06(-0.33,0.21)		-4.52(-5.05,-4)	
educ	0.45(0.4,0.49)		0.92(0.82,1.01)	
ind.1935	0.2(-0.1,0.53)		-0.41(-1.03,0.2)	
educ*age	-0.01(-0.01,-0.005)		-0.01(-0.01,-0.001)	
educ*dage				
sex*age	0.04(0.02,0.05)		0.12(0.08,0.15)	
sex*dage			0.11(0.01,0.22)	

$\epsilon 2$ nor the $\epsilon 4$ alleles of *APOE* were associated with performance on animal fluency

(Supplementary Table 3.s3 & 3.s4). Age at enrollment, follow-up time, and some of the interactions with generation, sex and education were significant (Supplementary Table 3.s3), suggesting that the cross-sectional and longitudinal effects of age were different in the younger and older generations and were modified by sex and education. 3.3a and 3.3b describes the estimated age and follow-up effects by generation. Older age at enrollment was associated with a lower score (age effect = -0.17, 95%CI: -0.18, -0.15) and, for every year of follow-up, the score decreased by -0.06 points (95%CI: -0.09, -0.03) in the older generation while, in the younger generation, the effect of age at enrollment was not significant (age effect = 0.01, 95%CI: -0.03, 0.06). The analyses also predicted a significant increase in score for each year of follow-up time in the younger generation (0.18, 95%CI: 0.11, 0.25) that could be caused by a practice effect among the younger participants. Higher education had a positive effect on the score but slightly diminished with older age at enrollment (educ*age interaction effect = -0.01, 95%CI: -0.01, -0.005). The age effect was smaller in males (sex*age interaction effect = 0.04, 95%CI: 0.02, 0.05).

DSST. Only age at enrollment, gender and education were significantly associated with DSST score, while the effects of $\epsilon 2$ and $\epsilon 4$ alleles of *APOE* were not significant (Supplementary Tables 3.s3 & 3.s4). In the older generation, an older year of age at enrollment was associated with a decrease of 0.67 points (95%CI: -0.71, -0.64) on the DSST (Table 3.3a). Follow-up time also had a negative effect on DSST (dage effect = -0.48, 95%CI: -0.54, -0.43). The negative effects of age at enrollment in the younger generation was smaller (-0.4, 95%CI: -0.48, -0.31), while the effect of follow-up time was not significant (-0.06, 95%CI: -0.20, 0.07). Higher education and female sex were associated with higher scores but the effect was reduced with older age at enrollment and longer follow up (educ*age interaction

effect = -0.01, 95%CI: -0.01, -0.001; sex*age interaction effect = 0.12, 95%CI: 0.08, 0.15; sex*dage interaction 0.11, 96%CI 0.01,0.22).

Table 3.3b: Parameter estimates of the Digits Span tests by generation.

	Digits Span - Forward		Digits Span - Backward	
	Older Generation	Younger Generation	Older Generation	Younger Generation
age	-0.03(-0.03,-0.02)	0.02(0.004,0.04)	-0.03(-0.04,-0.02)	0.01(-0.003,0.03)
dage	-0.12(-0.13,-0.11)	-0.07(-0.1,-0.05)	-0.04(-0.05,-0.03)	0.01(-0.02,0.04)
sex, male	0.13(0.04,0.23)		-0.08(-0.18,0.02)	
educ	0.12(0.1,0.14)		0.14(0.13,0.16)	
ind.1935	-0.04(-0.16,0.07)		0(-0.13,0.12)	
educ*age			-0.001(-0.002,-0.001)	
educ*dage	0.003(0.0003,0.01)			
sex*age				
sex*dage				

Digit Span — Forward. Neither $\epsilon 2$ nor $\epsilon 4$ alleles of *APOE* were associated with this test (Supplementary Tables 3.s3 & 3.s4). Age at enrollment, follow-up time, gender and education were associated with the digit span forward score, in both the older and younger generations (Table 3.3b). In the older generation, the score was expected to decrease by 0.03 points (95%CI: -0.03, -0.02) for every year of age at enrollment, and decrease by 0.12 points (95CI: -0.13, -0.11) for every year of follow-up time. In the younger generation, there was an estimated increase in forward span score as baseline age increased (0.02, 95%CI: 0.004, 0.04) and for each additional year in the follow-up time, the score decreased by -0.07 points (95%CI: -0.10, -0.05), thus suggesting a smaller rate of decline in the younger generation. Education was positively associated with the score (0.12 points, 95%CI: 0.10, 0.14) and the effect increased as follow-up time increased (educ*dage interaction effect 0.003, 95%CI: 0.0003, 0.01). Males tended to score higher by 0.13 points (95%CI: 0.04, 0.23) than females.

Digit Span — Backward. Similar to the Digit Span forward test, *APOE* was not associated with the backward span test score (Supplementary Tables 3.s3 & 3.s4). Older age at enrollment and longer follow-up time were negatively associated with the score only in the older generation (age effect = -0.03, 95%CI: -0.04, -0.02; dage effect = -0.04, 95%CI: -0.05, -0.03, Table 3.3b), and had no significant effect in the younger generation. Higher education was positively correlated with the score but the effect decreased with older age at enrollment (educ*age interaction effect = -0.001, 95%CI: -0.002, -0.001). We did not detect any gender difference in this test.

Table 3.4: Parameter estimates of Logical Memory tests $\epsilon 4$ allele carriers vs. non- $\epsilon 4$ allele carriers.

	Logical Memory - Immediate		Logical Memory - Delayed	
	Older Generation	Younger Generation	Older Generation	Younger Generation
age	-0.1(-0.12,-0.09)	0.09(0.06,0.12)	-0.12(-0.13,-0.1)	0.07(0.04,0.11)
dage	0.06(0.03,0.08)	0.2(0.14,0.25)	0.04(0.01,0.06)	0.19(0.13,0.25)
sex(male)	-0.87(-1.07,-0.67)		-1.08(-1.29,-0.87)	
educ	0.35(0.31,0.38)		0.35(0.31,0.39)	
<i>APOE4</i>	-0.31(-0.57,-0.05)		-0.37(-0.64,-0.1)	
ind.1935	-0.4(-0.63,-0.17)		-0.2(-0.45,0.05)	
educ*age	-0.003(-0.01,-0.001)		-0.003(-0.005,-0.001)	
educ*dage	-0.01(-0.02,-0.002)			
sex*age	0.02(0.002,0.03)		0.03(0.02,0.04)	
sex*dage				
<i>APOE4</i> *age				
<i>APOE4</i> *dage				

Logical Memory Recall Tests. As shown in Table 3.4, the $\epsilon 4$ allele had a negative effect on the logical memory tests (immediate recall $\epsilon 4$ allele effect = -0.31, 95%CI: -0.57, -0.05; delayed recall $\epsilon 4$ allele effect = -0.37, 95%CI: -0.64, -0.10) compared to carriers of $\epsilon 3$ or $\epsilon 2$. These effects were not modified by any of the other variables. In the older generation, older age at enrollment was associated with lower scores of both tests (immediate recall age effect = -0.10, 95%CI: -0.12, -0.09;

delayed recall age effect = -0.12, 95%CI: -0.13, -0.10). However, consistent with a possible practice effect, follow-up time had positive effects on both tests (immediate recall age effect = 0.06, 95%CI: 0.03, 0.08; delayed recall age effect = 0.04, 95%CI: 0.01, 0.06). The effects of age at baseline and follow-up time were different in the younger generation as indicated by the significant interactions of ind1935*age and ind1935*dage (Supplementary Table 3.s6). In the younger generation, both older baseline age and follow-up time were associated with higher scores of both tests. In the immediate recall test, the analysis estimated an increase of 0.09 points (95%CI: 0.06, 0.12, Table 3.4) with every additional year increase in baseline age, and an increase of 0.20 points (95%CI: 0.14, 0.25) with every year of follow-up time. Similarly, in the delayed recall test, for every one-year increase in baseline age the score increased by 0.07 points (95%CI: 0.04, 0.11), and by 0.19 points (95%CI: 0.13, 0.25) for every year of follow-up time. The age effects were modified by sex and education. Male sex reduced the effect of age at enrollment (immediate recall sex*age interaction effect = 0.02, 95%CI: 0.002, 0.03; delayed recall sex*age interaction effect = 0.03, 95%CI: 0.02, 0.04, Table 3.4). The advantage of higher education diminished slightly as baseline age increased (immediate recall educ*age interaction effect = -0.003, 95%CI: -0.01, -0.001; delayed recall educ*age interaction effect = -0.003, 95%CI: -0.005, -0.001), and also diminished as follow-up time increased in immediate recall (educ*dage interaction effect = -0.01, 95%CI: -0.02, -0.002).

3.4 DISCUSSION

We conducted a comprehensive analysis of the effect of *APOE* alleles on age-related change in various different cognitive domains. The analyses confirm the

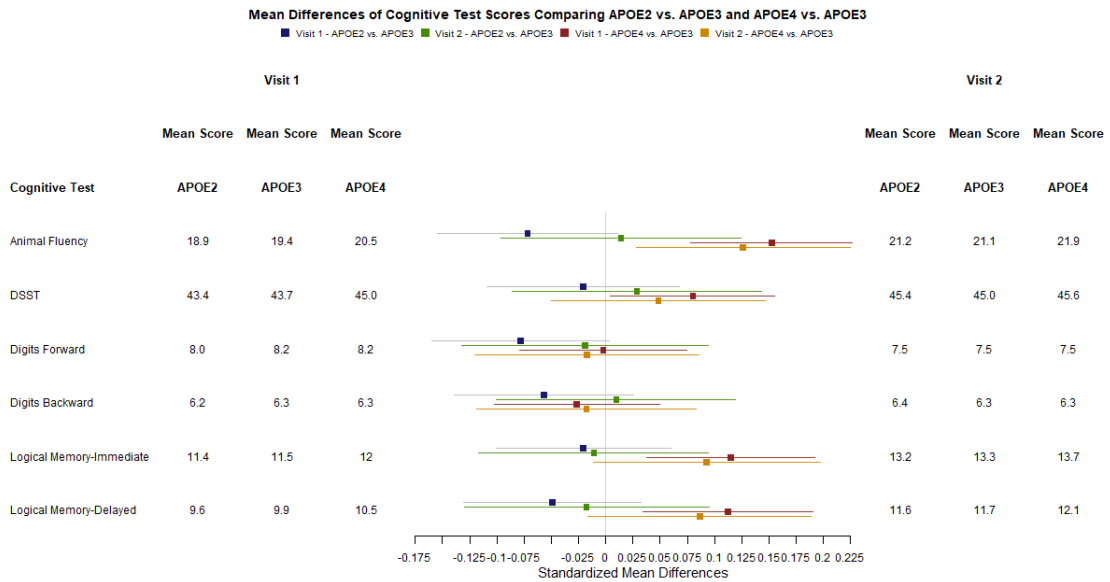
negative effect of $\epsilon 4$ allele on episodic memory assessed by immediate and delayed recall on logical memory: compared to the $\epsilon 2$ and $\epsilon 3$ alleles, carriers of one or more $\epsilon 4$ alleles scored lower in both tests although they did not exhibit a faster rate of decline. We did not detect any significantly protective effect of the $\epsilon 2$ allele compared to the $\epsilon 3$ allele. There is substantial literature linking the $\epsilon 4$ allele of *APOE* to poorer cognition at older age, faster rate of cognitive decline and higher risk for Alzheimers disease. The review by ODonoghue and colleagues[12] identified 12 cross-sectional and 15 longitudinal studies that reported a significant negative association between $\epsilon 4$ and episodic memory. Our findings show poorer memory in carriers of $\epsilon 4$ compared to other genotypes, but did not detect a significantly faster rate of decline. A study of centenarians (Xiang et al. 2020 preprint) also reported similar adverse effect of the $\epsilon 4$ allele on a memory cognitive test score using Beta regression modeling. Rawle et al. showed that $\epsilon 4$ homozygous carriers have a faster rate of cognitive decline compared to other genotype carriers in a study of comparable sample size (Rawle et al. (2018)). LLFS is a study of healthy aging and longevity with approximately 50% fewer $\epsilon 4$ carriers compared to the study in Rawle et al. and the smaller number of $\epsilon 4$ carriers may have reduced the power of our study. Alternatively, the lack of a difference in the rate of decline in the older generation may be due to a survivor bias. A review paper by Smith et al. (2019) had suggested that young $\epsilon 4$ carriers presented better mental performance compared to $\epsilon 3\epsilon 3$ carriers. A study by Caselli et al. (2009) aiming to address the transition from cognitive advantage to cognitive deficit in $\epsilon 4$ carriers showed that the longitudinal decline began before age 60 and had faster acceleration compared to $\epsilon 3\epsilon 3$ carriers. In LLFS with an average age of 87 years, $\epsilon 4$ carriers may be survivors with increased resilience to the risk conferred by the $\epsilon 4$ allele and therefore

are not showing the accelerated declines seen in other samples. In addition, Tao et al. (2018) suggested that the adverse effect of the allele might be activated by chronic low-grade inflammation. We conducted an additional analysis of the two Logical Memory Recall tests adjusting for baseline C-Reactive Protein (CRP) level and the results (Supplementary Table 3.s7) showed a reduction of the adverse effect of $\epsilon 4$ in the Logical Memory Immediate Recall test and no change of the $\epsilon 4$ effect on the Logical Memory Delayed Recall test.

Our findings on the effect of $\epsilon 4$ and cognition are consistent with other analyses conducted in the LLFS but expand the set of results to longitudinal assessments of cognitive function. For example, Kulminski et al. (2015) showed that the $\epsilon 4$ allele increases the lifetime risk of neurological disorders including dementia and Alzheimers disease by 98% in both LLFS men and women. Barral et al. (2017) defined exceptional cognitive performance using predominantly immediate and delayed memory, and showed that being in an exceptional cognitive performance family was significantly associated with being a non-carrier of the *APOE* $\epsilon 4$ allele. The fact that Logical Memory was the only cognitive test affected by $\epsilon 4$ is consistent with early episodic memory changes in Alzheimers disease. Reduced verbal fluency has also been posited as a marker of early Alzheimers disease that affects semantic memory (Gomez & White (2006)). In our sample there was no relationship between animal fluency and $\epsilon 4$. It is possible that the difference between phonemic and semantic fluency, that is poorer semantic fluency relative to phonemic fluency, is more indicative of early Alzheimers disease (Henry et al. (2004)) than semantic fluency in isolation. Our analyses did not detect any significant protective effect of $\epsilon 2$ on cognition although the data set included 314 carriers of one or more $\epsilon 2$ alleles. The results regarding the effect of the $\epsilon 2$ allele on cognition

have been mixed. The Religious Orders Study (Wilson et al. (2002)) found that $\epsilon 2$ carriers had an annual increase in episodic memory score while the $\epsilon 4$ subgroup decreased more rapidly compared to $\epsilon 3$ carriers. The study also found that $\epsilon 4$ carriers declined faster than $\epsilon 3$ in semantic memory and processing speed, but not in working memory. Four additional studies suggested the $\epsilon 2$ allele has a protective effect and is associated with reduced odds for developing cognitive impairment (Kim et al. (2017); Shinohara et al. (2016); Hyman et al. (1996); Helkala et al. (1996)). In contrast, the $\epsilon 2$ allele was not significantly associated with cognitive decline in Henderson et al. (1995). A study of 18,000 people by Marioni et al. (2016) did not detect any relationship between the $\epsilon 2$ allele and learning and episodic memory (Logical Memory), processing speed (DSST) and a fluency test. A clustering analysis of the LLFS cohort based on pattern of cognitive change by Sebastiani (2020) detected a cluster of slowest changers of DSST that was enriched for $\epsilon 2$ carriers, and a cluster of fastest changers that was enriched for $\epsilon 4$, suggesting that the effect of $\epsilon 2$ may be modified by other factors. Our study has some limitations. Only 55% of LLFS participants completed a second visit, though the dropout rates are comparable in each group ($E2=46\%$, $E3=45\%$, $E4=40\%$). Supplementary Figure 3.s1 shows a forest plot of the standardized mean differences of the neuropsychological tests among the three genotype groups at both visits and suggests that differences between participants who completed both visits and those that completed one visit should not effect the results. Secondly, having at most two time points restricts the analysis to a linear model rather than any nonlinear models. Lastly, our study sample is highly ethnically homogeneous that 99% of the study population are Caucasians. Therefore we cannot infer if *APOE* alleles have different effects in different ethnic groups. In conclusion, *APOE* $\epsilon 4$ allele was confirmed as a risk factor

Figure 3.s1: Forest plot for standardized mean difference comparing *APOE2* vs. *APOE3* and *APOE4* vs. *APOE3* at both visits.



for episodic memory in older adults, while *APOE* ϵ 2 allele was not significantly associated with any of the cognitive tests, and neither allele appear to modify the rate of cognitive decline.

3.5 SUPPLEMENTARY MATERIAL

Table 3.1: Demographic characteristics and test scores of 1,785 older generation (born in or before 1935) LLFS participants.

	APOE2($\epsilon 2\epsilon 2, \epsilon 2\epsilon 3$)	APOE3($\epsilon 3\epsilon 3$)	APOE4($\epsilon 3\epsilon 4, \epsilon 4\epsilon 4$)	p-value (t-test, comparing APOE2 and APOE3)	p-value (t-test, comparing APOE4 and APOE3)
N(%)	314(17.6%)	1239(69.4%)	232(13.0%)		
Age at Enrollment, mean(SD), years	89.4(7.9)	88.3(7.7)	86.8(7.4)	0.02	0.004
Age at visit 2, mean(SD), years	90.6(7)	91.1(6.8)	90.1(6.8)	0.54	0.25
Gender, male(%)	152(48.4%)	555(44.8%)	118(50.9%)	0.25	0.09
Education, college and above(%)	83(26.4%)	353(28.5%)	70(30.2%)	0.46	0.61
Deceased at follow up (%)	220(70.1%)	816(65.9%)	158(68.1%)	0.15	0.50
Test Scores at baseline (SD)					
MMSE	25.7(4.3)	25.8(4.2)	25.9(3.9)	0.58	0.75
Animal Fluency	14.4(5.5)	14.8(5.4)	15.3(5.7)	0.27	0.22
DSST	30.1(13.8)	30.8(12.6)	30.5(12.2)	0.47	0.70
Digits Forward	7.4(2.2)	7.6(2.2)	7.6(2.2)	0.18	0.98
Digits Backward	5.3(2.2)	5.5(2.1)	5.4(2.1)	0.15	0.30
Logical Memory-Immediate	8.7(4.3)	8.5(4.7)	8.8(4.8)	0.57	0.38
Logical Memory-Delayed	6.4(4.5)	6.5(4.8)	6.8(4.8)	0.66	0.46
Test Scores at follow-up (SD)					
MMSE	26(5.2)	26.2(4)	25.1(6)	0.66	0.15
Animal Fluency	15.2(6.2)	15(5.8)	15.6(6)	0.82	0.53
DSST	32.5(14.4)	30.3(12.3)	30.9(14.7)	0.20	0.79
Digits Forward	6.7(2.1)	6.8(2.1)	6.7(2.4)	0.65	0.73
Digits Backward	5.5(1.9)	5.3(2)	5.4(2.1)	0.49	0.76
Logical Memory-Immediate	10.1(4.4)	9.8(4.9)	10(5)	0.53	0.73
Logical Memory-Delayed	8(5.1)	7.6(5.3)	7.6(5.1)	0.52	1.00

Table 3.2: Demographic characteristics and test scores of 2,802 younger generation (born after 1935) LLFS participants.

	<i>APOE2</i> (e2e2, e2e3)	<i>APOE3</i> (e3e3)	<i>APOE4</i> (e3e4)	p-value (t-test, comparing <i>APOE2</i> and <i>APOE3</i>)	p-value (t-test, comparing <i>APOE4</i> and <i>APOE3</i>)
N(%)	419(15.0%)	1799(64.2%)	584(20.8%)	0.08	0.30
Age at Enrollment, mean(SD), years	59.1(7.3)	59.8(7.2)	60.1(6.7)	0.13	0.29
Age at visit 2, mean(SD), years	67.1(7)	67.8(6.9)	68.2(6.6)	0.10	0.29
Gender, male(%)	199(47.5%)	774(43%)	266(45.5%)	0.37	0.09
Education, college and above(%)	229(54.7%)	1027(57.1%)	310(53.1%)	0.38	0.98
Deceased at follow up (%)	18(4.3%)	95(5.3%)	31(5.3%)	0.52	0.18
Test Scores at baseline (SD)				0.43	0.70
MMSE	28.9(2.5)	29(1.8)	28.8(2.6)	0.59	0.03
Animal Fluency	22.3(5.9)	22.5(5.9)	22.4(5.4)	0.29	0.08
DSST	52(12.4)	51.7(12.2)	50.4(11.8)	0.81	0.02
Digits Forward	8.5(2.2)	8.6(2.2)	8.5(2.2)	0.57	0.29
Digits Backward	6.8(2.3)	6.9(2.4)	6.6(2.2)	0.49	0.24
Logical Memory-Immediate	13.3(3.9)	13.4(3.9)	13.2(4.1)	0.92	0.58
Logical Memory-Delayed	11.9(4.2)	12.1(4.2)	11.8(4.5)	0.74	0.94
Test Scores at follow-up (SD)				0.96	0.02
MMSE	29.1(2)	29.1(1.6)	29.2(1.4)	0.85	0.33
Animal Fluency	22.9(5.7)	22.8(6.1)	22.8(5.9)	0.88	0.08
DSST	48.9(11.7)	48.9(12)	47.4(11.4)	0.49	0.63
Digits Forward	7.7(2.7)	7.7(2.5)	7.6(2.5)	0.52	0.64
Digits Backward	6.6(2.1)	6.6(2.2)	6.4(2.1)		
Logical Memory-Immediate	14.1(3.6)	14.3(3.8)	14.1(4.2)		
Logical Memory-Delayed	12.7(4.2)	12.9(4.2)	12.7(4.3)		

Table 3.s1a: Demographic characteristics and test scores of 1,785 older generation (born in or before 1935) LLFS participants, broken down by APOE.

	$\epsilon 2\epsilon 2$	$\epsilon 2\epsilon 3$	$\epsilon 3\epsilon 3$	p-value (t-test, comparing $\epsilon 2\epsilon 2$ and $\epsilon 3\epsilon 3$)	p-value (t-test, comparing $\epsilon 2\epsilon 3$ and $\epsilon 3\epsilon 3$)
N(%)	13(0.7%)	301(16.9%)	1239(69.4%)		
Age at Enrollment, mean(SD), years	91.2(6)	89.4(7.9)	88.3(7.7)	0.10	0.04
Age at visit 2, mean(SD), years	92.8(4.6)	90.6(7.1)	91.1(6.8)	0.53	0.48
Gender, male(%)	6(46.2%)	146(48.5%)	555(44.8%)	0.93	0.25
Education, college and above(%)	4(30.8%)	79(26.2%)	353(28.5%)	0.87	0.43
Deceased at follow up (%)	10(76.9%)	210(69.8%)	816(65.9%)	0.38	0.19
Test Scores at baseline (SD)					
MMSE	25.7(3)	25.7(4.3)	25.8(4.2)	0.89	0.59
Animal Fluency	14.2(3.6)	14.4(5.6)	14.8(5.4)	0.58	0.30
DSST	28.3(13.4)	30.2(13.8)	30.8(12.6)	0.53	0.53
Digit Span - Forward	7.2(2.4)	7.4(2.2)	7.6(2.2)	0.60	0.21
Digit Span - Backward	6.6(2.6)	5.3(2.1)	5.5(2.1)	0.19	0.07
Logical Memory-Immediate	6.4(3.7)	8.8(4.3)	8.5(4.7)	0.07	0.37
Logical Memory-Delayed	4.2(3)	6.5(4.5)	6.5(4.8)	0.02	0.91
Test Scores at follow-up (SD)					
MMSE	26.8(1)	25.9(5.3)	26.2(4)	0.36	0.63
Animal Fluency	12.2(4.6)	15.3(6.3)	15(5.8)	0.31	0.68
DSST	31.2(15.3)	32.6(14.5)	30.3(12.3)	0.91	0.20
Digit Span - Forward	6.5(3.3)	6.7(2)	6.8(2.1)	0.86	0.68
Digit Span - Backward	7.7(2.1)	5.4(1.9)	5.3(2)	0.19	0.72
Logical Memory-Immediate	9.2(4.9)	10.2(4.4)	9.8(4.9)	0.84	0.49
Logical Memory-Delayed	5.5(5.4)	8.1(5.1)	7.6(5.3)	0.50	0.41

Table 3.s1b: Demographic characteristics and test scores of 1,785 older generation (born in or before 1935) LLFS participants, broken down by APOE.

	$\epsilon 3\epsilon 4$	$\epsilon 4\epsilon 4$	$\epsilon 3\epsilon 3$	p-value (t-test, comparing $\epsilon 3\epsilon 4$ and $\epsilon 3\epsilon 3$)	p-value (t-test, comparing $\epsilon 4\epsilon 4$ and $\epsilon 3\epsilon 3$)
N(%)	222(12.4%)	10(0.6%)	1239(69.4%)		
Age at Enrollment, mean(SD), years	86.9(7.4)	83.4(6.5)	88.3(7.7)	0.01	0.04
Age at visit 2, mean(SD), years	90.3(6.8)	85(1)	91.1(6.8)	0.39	0.001
Gender, male(%)	114(51.4%)	4(40%)	555(44.8%)	0.07	0.78
Education, college and above(%)	69(31.1%)	1(10%)	353(28.5%)	0.44	0.10
Deceased at follow up (%)	152(68.5%)	6(60%)	816(65.9%)	0.44	0.73
Test Scores at baseline (SD)					
MMSE	25.9(3.9)	26.4(2.1)	25.8(4.2)	0.80	0.47
Animal Fluency	15.3(5.7)	14.7(5.6)	14.8(5.4)	0.20	0.95
DSST	30.4(12.2)	32.1(12.3)	30.8(12.6)	0.65	0.76
Digit Span - Forward	7.6(2.2)	7.1(2.1)	7.6(2.2)	0.91	0.47
Digit Span - Backward	5.3(2)	6.2(2.5)	5.5(2.1)	0.21	0.44
Logical Memory-Immediate	8.9(4.9)	8.6(2.8)	8.5(4.7)	0.38	0.93
Logical Memory-Delayed	6.8(4.9)	5.6(3.7)	6.5(4.8)	0.39	0.52
Test Scores at follow-up (SD)					
MMSE	25.1(6.1)	24.5(4.9)	26.2(4)	0.17	0.71
Animal Fluency	15.7(6.1)	12.5(0.7)	15(5.8)	0.46	0.06
DSST	30.8(14.9)	34(NA)	30.3(12.3)	0.82	-
Digit Span - Forward	6.7(2.5)	6(1.4)	6.8(2.1)0.79	0.56	
Digit Span - Backward	5.5(2.1)	3.5(0.7)	5.3(2)	0.60	0.16
Logical Memory-Immediate	10.2(4.8)	NA	9.8(4.9)	0.54	-
Logical Memory-Delayed	7.8(5)	NA	7.6(5.3)	0.84	-

Table 3.s2a: Demographic characteristics and test scores of 2,802 younger generation (born after 1935) LLFS participants, broken down by APOE genotype.

	$\epsilon 2\epsilon 2$	$\epsilon 2\epsilon 3$	$\epsilon 3\epsilon 3$	p-value (t-test, comparing $\epsilon 2\epsilon 2$ and $\epsilon 3\epsilon 3$)	p-value (t-test, comparing $\epsilon 2\epsilon 3$ and $\epsilon 3\epsilon 3$)
N(%)	20(0.7%)	399(14.2%)	1799(64.2%)		
Age at Enrollment, mean(SD), years	58.2(8.9)	59.1(7.2)	59.8(7.2)	0.44	0.10
Age at visit 2, mean(SD), years	68.1(9)	67(7)	67.8(6.9)	0.90	0.11
Gender, male(%)	25%	194(48.6%)	774(43%)	0.09	0.04
Education, college and above(%)	12(60%)	217(54.4%)	1027(57.1%)	0.80	0.33
Deceased at follow up (%)	0(0%)	18(4.5%)	95(5.3%)	< 0.001	0.51
Test Scores at baseline (SD)					
MMSE	29.1(0.9)	28.9(2.6)	29(1.8)	0.59	0.49
Animal Fluency	22.5(5.9)	22.3(5.9)	22.5(5.9)	0.98	0.42
DSST	49.2(13.1)	52.2(12.4)	51.7(12.2)	0.42	0.46
Digit Span - Forward	8.7(2.4)	8.5(2.2)	8.6(2.2)	0.90	0.27
Digit Span - Backward	7.2(2.5)	6.8(2.3)	6.9(2.4)	0.63	0.72
Logical Memory-Immediate	14(4.2)	13.3(3.9)	13.4(3.9)	0.54	0.47
Logical Memory-Delayed	12.3(4.7)	11.9(4.2)	12.1(4.2)	0.85	0.45
Test Scores at follow-up (SD)					
MMSE	28.9(1.2)	29.1(2)	29.1(1.6)	0.59	0.87
Animal Fluency	21.4(3.3)	23(5.7)	22.8(6.1)	0.15	0.60
DSST	45.9(11.4)	49(11.7)	48.9(12)	0.35	0.89
Digits Forward	7.1(2.3)	7.7(2.7)	7.7(2.5)	0.35	0.98
Digits Backward	6.4(2)	6.6(2.1)	6.6(2.2)	0.61	0.96
Logical Memory-Immediate	14.1(3.9)	14.1(3.6)	14.3(3.8)	0.91	0.49
Logical Memory-Delayed	13.4(4)	12.6(4.3)	12.9(4.2)	0.60	0.45

Table 3.s2b: Demographic characteristics and test scores of 2,802 younger generation (born after 1935) LLFS participants, broken down by APOE genotype.

	$\epsilon 3\epsilon 4$	$\epsilon 4\epsilon 4$	$\epsilon 3\epsilon 3$	p-value (t-test, comparing $\epsilon 3\epsilon 4$ and $\epsilon 3\epsilon 3$)	p-value (t-test, comparing $\epsilon 4\epsilon 4$ and $\epsilon 3\epsilon 3$)
N(%)	545(19.5%)	39(1.4%)	1799(64.2%)		
Age at Enrollment, mean(SD), years	60.1(6.7)	59.5(6.7)	59.8(7.2)	0.26	0.83
Age at visit 2, mean(SD), years	68.3(6.6)	66.7(6.4)	67.8(6.9)	0.18	0.35
Gender, male(%)	250(45.9%)	16(41%)	774(43%)	0.24	0.81
Education, college and above(%)	288(52.8%)	22(56.4%)	1027(57.1%)	0.08	0.93
Deceased at follow up (%)	27(5%)	4(10.3%)	95(5.3%)	0.76	0.32
Test Scores at baseline (SD)					
MMSE	28.8(2.7)	28.9(1.5)	29(1.8)	0.19	0.65
Animal Fluency	22.4(5.4)	22.1(5.9)	22.5(5.9)	0.76	0.68
DSST	50.5(11.6)	49.7(13.4)	51.7(12.2)	0.04	0.37
Digit Span - Forward	8.5(2.2)	7.6(2.3)	8.6(2.2)	0.24	0.01
Digit Span - Backward	6.4(2.5)	6.9(2.4)	0.03	0.25	
Logical Memory-Immediate	13.2(4)	13.2(4.7)	13.4(3.9)	0.29	0.82
Logical Memory-Delayed	11.8(4.4)	11.8(5.1)	12.1(4.2)	0.24	0.80
Test Scores at follow-up (SD)					
MMSE	29.2(1.4)	29.2(1.3)	29.1(1.6)	0.62	0.74
Animal Fluency	22.8(5.9)	22.3(5.9)	22.8(6.1)	0.97	0.64
DSST	47.6(11.5)	45.1(11.1)	48.9(12)	0.04	0.07
Digits Forward	7.7(2.5)	6.7(2.4)	7.7(2.5)	0.64	0.02
Digits Backward	6.4(2.1)	6.1(2.2)	6.6(2.2)	0.13	0.22
Logical Memory-Immediate	14.2(4.1)	14.1(4.7)	14.3(3.8)	0.64	0.88
Logical Memory-Delayed	12.7(4.4)	13(4.1)	12.9(4.2)	0.60	0.89

Table 3.s3: Parameter estimates of APOE2 analysis.

	Animal Fluency	DSST	Digit Span - Forward	Digit Span - Backward	Logical Memory- Immediate	Logical Memory- Delayed
age	-0.17(-0.19,-0.15)	-0.68(-0.72,-0.64)	-0.03(-0.03,-0.02)	-0.03(-0.04,-0.02)	-0.11(-0.12,-0.09)	-0.13(-0.14,-0.11)
dage	-0.06(-0.1,-0.03)	-0.49(-0.55,-0.43)	-0.13(-0.14,-0.11)	-0.04(-0.06,-0.03)	0.05(0.03,0.08)	0.04(0.01,0.06)
sex(male)	-0.06(-0.36,0.24)	-4.59(-5.16,-4.01)	0.18(0.07,0.29)	-0.05(-0.16,0.07)	-0.91(-1.13,-0.69)	-1.08(-1.32,-0.84)
educ	0.46(0.41,0.52)	0.92(0.81,1.01)	0.12(0.1,0.14)	0.14(0.12,0.17)	0.34(0.3,0.38)	0.35(0.31,0.39)
APOE2	-0.17(-0.52,0.18)	0.12(-0.12,-0.54)	-0.04(-0.16,0.09)	0.02(-0.12,0.15)	0.08(-0.17,0.33)	0.03(-0.25,0.31)
fc.DK	2.86(2.42,3.29)	-5.3(-6.14,-4.46)	-1.67(-1.83,-1.51)	-1.09(-1.25,-0.92)	1.38(1.06,1.69)	1.35(1.01,1.69)
fc.NY	0.12(-0.3,0.54)	0.71(-0.13,1.51)	0.63(0.48,0.79)	0.31(0.15,0.47)	-0.21(-0.51,0.1)	-0.42(-0.75,-0.09)
fc.PT	-0.17(-0.58,0.23)	0.23(-0.54,0.96)	0.77(0.63,0.92)	0.36(0.21,0.51)	-0.14(-0.43,0.15)	-0.35(-0.66,-0.04)
ind1935	0.2(-0.13,0.55)	-0.53(-1.17,0.14)	-0.04(-0.16,0.09)	0.01(-0.12,0.14)	-0.36(-0.61,-0.1)	-0.21(-0.46,0.05)
educ*age	-0.01(-0.01,-0.01)	-0.01(-0.01,-0.001)	-0.001(-0.003,-0.0004)	-0.001(-0.003,-0.0004)	-0.004(-0.01,-0.002)	-0.003(-0.01,-0.001)
educ*dage		0.01(0.001,0.01)			-0.01(-0.02,-0.002)	
sex*age	0.04(0.03,0.06)	0.12(0.08,0.15)			0.02(0.01,0.03)	0.03(0.02,0.05)
sex*dage		0.12(0.003,0.24)				
APOE2*age						
APOE2*dage						
ind1935*age	0.17(0.13,0.21)	0.3(0.22,0.39)	0.05(0.04,0.07)	0.05(0.03,0.06)	0.19(0.15,0.22)	0.18(0.14,0.21)
ind1935*dage	0.24(0.16,0.32)	0.45(0.3,0.61)	0.04(0.01,0.07)	0.05(0.02,0.08)	0.14(0.07,0.2)	0.15(0.08,0.21)

Table 3.s4: Parameter estimates of APOE4 analysis.

	Animal Fluency	DSST	Digit Span - Forward	Digit Span - Backward	Logical Memory- Immediate	Logical Memory- Delayed
age	-0.17(-0.19,- 0.14)	-0.67(-0.71,- 0.63)	-0.03(-0.03,- 0.02)	-0.03(-0.04,- 0.02)	-0.1(-0.12,- 0.09)	-0.12(-0.13,- 0.1)
dage	-0.07(-0.1,- 0.03)	-0.49(-0.55,- 0.43)	-0.12(-0.14,- 0.11)	-0.04(-0.06,- 0.03)	0.05(0.03,0.08)	0.03(0,0.06)
sex(male)	-0.15(- 0.45,0.14)	-4.74(-5.31,- 4.16)	0.15(0.04,0.26)	-0.1(-0.21,0.01)	-0.84(-1.06,- 0.62)	-1.06(-1.3,- 0.83)
educ	0.44(0.39,0.5)	0.88(0.78,0.98)	0.12(0.1,0.14)	0.14(0.12,0.16)	0.35(0.31,0.38)	0.34(0.3,0.38)
APOE4	-0.25(- 0.57,0.07)	-0.45(- 1.07,0.18)	0(-0.12,0.12)	-0.1(-0.22,0.03)	-0.3(-0.54,- 0.06)	-0.36(-0.62,- 0.11)
fc.DK	2.84(2.4,3.28)	-5.6(-6.42,- 4.74)	-1.69(-1.84,- 1.53)	-1.08(-1.24,- 0.91)	1.5(1.17,1.82)	1.5(1.17,1.85)
fc.NY	0.21(- 0.23,0.64)	0.12(- 0.72,0.96)	0.6(0.45,0.76)	0.26(0.09,0.42)	-0.13(- 0.44,0.19)	-0.4(-0.73,- 0.06)
fc.PT	-0.19(-0.6,0.21)	-0.28(- 1.03,0.47)	0.72(0.58,0.86)	0.33(0.19,0.48)	-0.1(-0.39,0.19)	-0.33(-0.64,- 0.02)
ind1935	0.29(- 0.05,0.65)	-0.2(-0.88,0.48)	-0.06(- 0.18,0.07)	0.03(-0.1,0.17)	-0.4(-0.65,- 0.14)	-0.2(-0.47,0.06)
educ*age	-0.01(-0.01,- 0.01)	-0.01(-0.01,- 0.001)	-0.001(-0.002,- 0.0003)	-0.002(-0.003,- 0.001)	-0.003(-0.01,- 0.001)	-0.003(-0.01,- 0.001)
educ*dage					-0.01(-0.02,- 0.0004)	
sex*age	0.05(0.03,0.07)	0.1(0.06,0.14)			0.02(0.01,0.03)	0.03(0.02,0.05)
sex*dage						
APOE4*age						
APOE4*dage						
ind1935*age	0.16(0.12,0.2)	0.25(0.17,0.34)	0.05(0.04,0.07)	0.04(0.03,0.06)	0.19(0.16,0.22)	0.19(0.15,0.22)
ind1935*dage	0.23(0.14,0.31)	0.48(0.31,0.64)	0.05(0.02,0.08)	0.05(0.02,0.09)	0.15(0.08,0.22)	0.18(0.11,0.24)

Table 3.s5: Parameter estimates of Animal Fluency, DSST, Digits Span tests, without stratification of APOE genotype.

	Animal Fluency	DSST	Digit Span - Forward	Digit Span - Backward
age	-0.17(-0.18,-0.15)	-0.67(-0.71,-0.64)	-0.03(-0.03,-0.02)	-0.03(-0.04,-0.02)
dage	-0.06(-0.09,-0.03)	-0.48(-0.54,-0.43)	-0.12(-0.13,-0.11)	-0.04(-0.05,-0.03)
sex(male)	-0.06(-0.33,0.21)	-4.52(-5.05,-4)	0.13(0.04,0.23)	-0.08(-0.18,0.02)
educ	0.45(0.4,0.49)	0.92(0.82,1.01)	0.12(0.1,0.14)	0.14(0.13,0.16)
fc.DK	2.68(2.3,3.07)	-5.47(-6.23,-4.71)	-1.67(-1.82,-1.53)	-1.07(-1.22,-0.93)
fc.NY	0.12(-0.26,0.51)	0.41(-0.31,1.14)	0.57(0.43,0.71)	0.27(0.13,0.42)
fc.PT	-0.2(-0.55,0.16)	0.02(-0.68,0.7)	0.71(0.58,0.85)	0.33(0.19,0.47)
ind1935	0.2(-0.1,0.53)	-0.41(-1.03,0.2)	-0.04(-0.16,0.07)	-0.002(-0.13,0.12)
educ*age	-0.01(-0.01,-0.005)	-0.01(-0.01,-0.001)		-0.001(-0.002,-0.001)
educ*dage			0.003(0.00029,0.01)	
sex*age	0.04(0.02,0.05)	0.12(0.08,0.15)		
sex*dage		0.11(0.01,0.22)		
ind1935*age	0.18(0.15,0.22)	0.28(0.2,0.35)	0.05(0.03,0.06)	0.04(0.03,0.06)
ind1935*dage	0.24(0.16,0.31)	0.42(0.28,0.57)	0.05(0.02,0.08)	0.05(0.02,0.08)

Table 3.s6: Parameter estimates of Logical Memory tests.

	Logical Memory- Immediate	Logical Memory-Delayed
age	-0.1(-0.12,-0.09)	-0.12(-0.13,-0.1)
dage	0.06(0.03,0.08)	0.04(0.01,0.06)
sex(male)	-0.87(-1.07,-0.67)	-1.08(-1.29,-0.87)
educ	0.35(0.31,0.38)	0.35(0.31,0.39)
<i>APOE4</i>	-0.31(-0.57,-0.05)	-0.37(-0.64,-0.1)
ind1935	-0.4(-0.63,-0.17)	-0.2(-0.45,0.05)
educ*age	-0.003(-0.01,-0.001)	-0.003(-0.005,-0.001)
educ*dage	-0.01(-0.02,-0.002)	
sex*age	0.02(0.002,0.03)	0.03(0.02,0.04)
sex*dage		
<i>APOE4</i> *age		
<i>APOE4</i> *dage		
ind1935*age	0.19(0.17,0.22)	0.19(0.16,0.22)
ind1935*dage	0.14(0.08,0.2)	0.16(0.09,0.22)

Table 3.s7: Parameter estimates of Logical Memory tests e4 allele carriers vs. non-e4 allele carriers, adjusting for CRP level.

	Logical Memory - Immediate		Logical Memory - Delayed	
	Older Generation	Younger Generation	Older Generation	Younger Generation
age	-0.1(-0.12,-0.09)	0.09(0.05,0.12)	-0.12(-0.13,-0.1)	0.07(0.03,0.1)
dage	0.06(0.04,0.08)	0.2(0.14,0.25)	0.04(0.01,0.06)	0.2(0.14,0.26)
sex(male)	-0.85(-1.05,-0.64)		-1.08(-1.29,-0.87)	
educ	0.35(0.31,0.39)		0.35(0.31,0.39)	
<i>APOE4</i>	-0.25(-0.51,0.01)		-0.35(-0.63,-0.06)	
CRP	0.12(0.03,0.21)		0.08(-0.02,0.17)	
ind.1935	-0.42(-0.67,-0.18)		-0.21(-0.47,0.03)	
educ*age	-0.003(-0.01,-0.001)		-0.003(-0.005,-0.001)	
educ*dage	-0.01(-0.02,-0.002)			
sex*age	0.01(0.001,0.03)		0.03(0.01,0.04)	
sex*dage				
<i>APOE4</i> *age				
<i>APOE4</i> *dage				

CHAPTER 4

Analyzing Digitally Assessed Trail Making Test Using Hidden Markov Models

4.1 INTRODUCTION

Trail Making Tests (TMT) is one of the most commonly used and well-established neuropsychological tests for clinical evaluation of brain damage and diagnosis of age related diseases such as Alzheimers Disease. The original test was first introduced in the Army Individual Test Battery (War Department (1944)) as well as the Halstead-Reitan Neuropsychological Test Battery (Mazur-Mosiewicz & Dean (2011)) in the 1940s. The most widely used paper-based version of the TMT consists of two parts, namely Part A and Part B. In Trail Making Test Part A (TMT-A), a series of numbers are displayed on a piece of paper, and participants are instructed to use a pen to draw lines to connect the numbers in sequence. In Trail Making Test Part B (TMT-B), a series of numbers and letters are displayed and participants are instructed to connect the numbers and letters in alternate sequence. Traditionally, the TMT scoring consists of total time to completion and number of errors made recorded by examiners. The TMT was first utilized in the hospitals as an indicator for certain effects of brain damage (Reitan (1955); Reitan (1958); Spreen & Benton (1965)). Recent studies have shown that the performance of the TMT-A is related to cognitive domains such as visual attention and processing speed, while the performance on TMT-B is associated with more complex cognitive abilities including set shifting and mental flexibility (Arbuthnott & Frank (2000); Crowe (1998); Lezak et al. (2004); Oosterman et al. (2010); Sánchez-Cubillo et al. (2009)).

Although the TMT has proven to be a highly sensitive test for diagnosing brain impairment, specific mechanisms underlying the performances of TMT are not

captured by the overall time to completion. For example, a poorly scored TMT test might be a result of prolonged thinking, hesitation, or difficulties with holding pens and drawing lines, or both. The former implies impairment in cognitive abilities such as visual searching and scanning, and the latter reveals dysfunction in physical abilities such as grip strength. Thus, decomposing total time to completion into thinking time and drawing time might provide insights of cognitive or physical abilities that contribute to the overall performance of the TMT. Another goal of this chapter is to utilize TMT to not only detect individuals with apparent impairments, but also to pick up subtle differences or patterns among cognitively and physically healthy individuals. We are interested in whether these patterns or characteristics associate with other cognitive and physical tests that were administered in the LLFS and can be more informative in predicting cognitive decline.

During the second in-person visit of the LLFS, participants completed both parts of the TMT using a digital pen that records and timestamps coordinates 75 times per second, or approximately every 13 milliseconds. In this digitally recorded version of the TMT, 25 numbers are displayed in TMT-A and 13 numbers and 12 letters are displayed in TMT-B. The recorded data streams is sectioned by each pen stroke, defined by a continuous drawing without lifting up the pen. During the data extraction process, the digital pen automatically deletes coordinate pairs from the same spot and intermediate coordinate pairs in a straight line. Other information that are less relevant to the current analysis such as force and color of the pen stroke is also recorded.

In this chapter we propose to use the digitally recorded data stream to decompose total time to completion, which would provide deeper insights of the cognitive function or physical function underlying the overall performance of the TMT.

We also propose a novel application of Bayesian Hidden Markov Models (HMMs) to perform automatic segmentation of the recorded drawings, and use the results to estimate the number of connections drawn on papers and summary statistics of these connections to be used as new metrics of the two tests. Our hypothesis is that digitally recorded data stream could provide additional information of cognitive state even among cognitively intact individuals that are not described by the overall time used to complete the test. In the next section we will briefly describe Markov chains and HMMs and introduce an application of HMMs to the digitally recorded TMT-A and TMT-B tests.

4.2 METHOD

4.2.1 Study Population and Test Measures

The study population from the LLFS study has been described in section 3.2.1. In addition to the six neuropsychological tests described in section 3.2.2, we included a modified version of the Telephone Interview for Cognitive Status (TICS) score in this analysis. This modified TICS sums the scores to questions Counting Backward, Word List Learning and Subtractions. We also included the Hopkins Verbal Learning Test - Revised (HVLT-R) test to test assess learning and memory. Two physical function scores including gait speed and grip strength were also included to measure motor functions and strength. TICS score was measured longitudinally with multiple measurement per participant and all other measures were assessed once at the second in-person visit.

4.2.2 Hidden Markov Model and Trail Making Tests

To analyze information provided by the digital stream of coordinates, we implemented a novel way to perform automatic segmentation of the trail drawing data using HMMs. HMMs, first published by Baum et al. (1970), are a type of stochastic model that assumes that the underlying process is a Markov chain, which embodies the Markov assumption. A Markov chain model describes a sequence of discrete states in which a probability is associated with transitioning to one another. Denote a set of N states $S = s_1, s_2, \dots, s_N$, a transition probability matrix $P = p_{11}, p_{12}, \dots, p_{n1}, \dots, p_{nn}$, where p_{ij} represents the transition probability from state i to state j such that $\sum_{j=1}^n p_{jn} = 1, \forall i$, and a set of initial probabilities $\Pi = \pi_1, \pi_2, \dots, \pi_n$, where π_i represents the initial probability of being in state i . The Markov assumption states that the conditional probability of state t only depends on the previous state $t - 1$, and not any other states before that. The Markov assumption can be expressed as follows:

$$P(s_t | s_1, s_2, \dots, s_{t-1}) = P(s_t | s_{t-1})$$

The Markov chain model allows us to compute event probabilities in the case when events are directly observable. However, in many applications, the underlying states are not directly observable, or hidden. HMMs enable us to incorporate the observations into the model. In addition to the Markov chain model components, we denote a set of T observations $Y = y_1, y_2, \dots, y_T$, and a set of emission probabilities $K = k_i(y_t), i = 1, \dots, n, t = 1, \dots, T$, where $k_i(y_t)$ represents the probability of generating observation y_t from state i . And in addition to the Markov assumption, the HMMs have three more assumptions. The first two assumptions state that both

the observations Y and the hidden states N come from known, finite sets. HMMs also makes the conditional independence assumption that the probability of observing y_t only depends on the immediate hidden state s_t . This assumption can be expressed as:

$$P(y_t|y_1, y_2, \dots, y_{t-1}, s_1, s_2, \dots, s_{t-1}) = P(y_t|s_t)$$

In the digitally recorded TMT data, we directly observe the sequence of coordinates and timestamps of drawings on the paper, but we do not know which *connection* (a connection is a line drawn to connect two symbols in sequence) a particular coordinate pair is in. If a participant were able to draw a perfectly straight line for each connection, the correct number of segmentation would be 24 for both TMT-A and TMT-B. In practice, a connection could be made up of multiple segmentations when there is a detectable turn of direction, or curvature, in the drawing. Figure 4.1 and Figure 4.2 illustrate an example of the HMM segmentation from TMT-A. Figure 4.1 recreates the drawing by connecting the recorded coordinate pairs, while Figure 4.2 shows the segmentation from the HMM with each color representing a unique segment assignment.

In these TMT data, the coordinates can be viewed as the observable sequence of observations, while their corresponding underlying connections can be viewed as unobservable, or hidden states. The nature of these data satisfies the four assumptions of HMMs such that the observed coordinates are finite in number and the underlying segmentations are assumed to be finite with a specified measuring sensitivity; the conditional probability for a particular pair of coordinates to be in segment i only depends on the segment of the previous pair of coordinates; and lastly the probability of observing a particular segment only depends on the immediate segment. We assume that the connections between numbers or letters

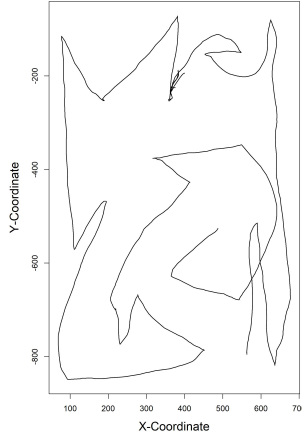


Figure 4.1: An example of drawing in TMT-A.

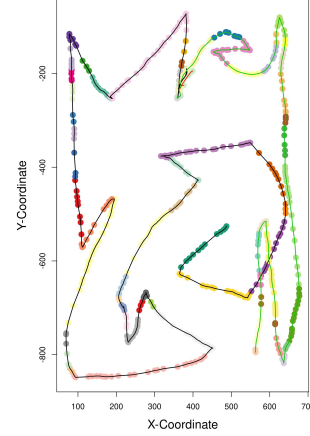


Figure 4.2: An example of drawing in TMT-A after HMM segmentation.

in sequence can be approximated by straight lines. We used the following model specification for the HMM:

$$\varepsilon[t] \sim \text{Cat}(p.\varepsilon[\varepsilon[t-1],])$$

$$y[t] \sim \text{Norm}(\mu.\varepsilon[t], \tau)$$

$$\mu.\varepsilon[t] = b_0[\varepsilon[t]] + b_1[\varepsilon[t]] \times x[t]$$

The random variable $\varepsilon[t]$ denotes the hidden state (segment) at time t that we assumed follows a categorical distribution with probability of transition that only depends on the state at time $t-1$, say $p.\varepsilon[\varepsilon[t-1],]$. We assumed a maximum of 40 hidden states, and assumed that the vector of transition probabilities from state $t-1$ follows a Dirichlet prior distribution with $\alpha = 1, \forall i = 1, \dots, 40$. The pairs of $(x[t], y[t])$ coordinates were modelled using a linear regression of $y[t]$, conditional on $x[t]$, and we assumed that $y[t]$ followed a normal distribution, with precision τ , that had Gamma prior distribution with both shape and scale parameters equal

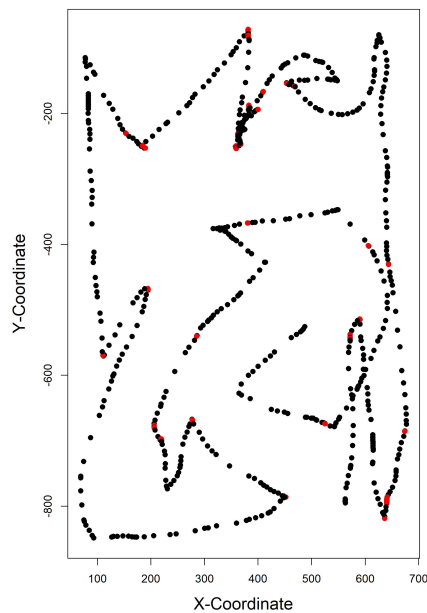
to 1. We modelled the expected value $\mu_{\cdot\varepsilon}[t]$ of $y[t]|x[t]$ using a linear relationship with intercept term b_0 and slope b_1 for each hidden state ε . We assumed that b_0 had a normal prior distribution with mean 0 and precision $\tau_0 = 10^{-6}$, and b_1 had normal prior distribution with mean 0 and precision $\tau_1 = 10^{-4}$. We estimated the values of the hidden states and the parameters of each segmentation using Markov Chain Monte Carlo methods, with 2,000 adaptations and 10,000 iterations and monitored the hidden state for each recorded coordinate pair. To smooth some of the estimated segmentation, we re-assigned any segment with only one point that was different from both the segment before and after it, to the previous assigned segment. For example, if coordinate pairs 1 through 5 were assigned to segment 1, coordinate pair 6 to segment 8, and coordinate pairs 7 through 10 to segment 15, in this case coordinate pair 6 with segment 8 would be re-assigned to segment 1. We also re-assigned any series of non-consecutive segment to a new segment number. For example, if coordinate pairs 1 through 5 were assigned to segment 1 and later coordinate pairs 75 through 85 were also assigned segment 1, then the segment number of coordinate pairs 75 through 85 were re-assigned to a new number that is different from any other existing segments.

4.2.3 Extracted Metrics

Using data directly available from the digital pen, we extracted and derived several time metrics for each test of TMT-A and TMT-B. We first defined a set of intuitive raw time variables, namely raw drawing time and raw thinking time. Raw drawing time was defined as the time spent while the digital pen was on the paper, suggesting a drawing motion was in place. Raw thinking time was defined as the time spent while the digital pen was lifted away from the paper, suggesting

the participant was likely thinking or looking for the next number or letter in sequence. A cluster of points was defined as a group of coordinates with pairwise distance less than $\sqrt{2}$ coordinate units. Examples of clusters of points are illustrated in Figure 4.3 using the same TMT-A drawing as in Figure 4.1 and Figure 4.2, with clusters of points marked in red color. A cluster of points indicates that

Figure 4.3: An example of drawing in TMT-A with cluster points in red color.



the digital pen had moved less than one coordinate unit in both the vertical and horizontal directions, where one coordinate unit is equivalent to 0.3 millimeters. Given that the time spent in these cluster coordinate pairs also suggested thinking or hesitation, we then defined a set of derived time variables, namely derived drawing time and derived thinking time. Derived drawing time was calculated by subtracting the time spent in cluster points from the raw drawing time, and derived thinking time was calculated by adding time spent in cluster points to the

raw thinking time. The derived time variables would provide us a more accurate understanding of the decomposition of the total completion time. The ratio between derived thinking time and derived drawing time was also calculated to be used as a metric in the subsequent analysis. Using the results of the HMM segmentation, we extracted several metrics including number of segmentations and summary statistics of the segments such as the minimum, maximum, median and mean length (in coordinate units) of the segments, for both TMT-A and TMT-B.

4.2.4 Statistical Analysis

We examined the association between these new derived metrics and more traditional metrics of cognitive and physical function using Generalized Estimating Equations (GEE) with exchangeable correlation structure to account for family clustering. The analyses were limited to the subset of participants who had successfully completed the tests, since only these participants would get examiner-timed scores on their performances. The analyses were conducted in three settings. In all settings, the GEE models used the traditional cognitive and physical scores as the outcome variables, and were adjusted for age at test, sex, education level, and familial longevity if significant. In the first setting, the completion time in seconds, which is the commonly used score of the TMT tests, was used as the only additional predictor in the model. In the second setting, the new TMT metrics including derived drawing time, ratio between derived thinking time and drawing time, number of HMM segments and mean segment length were used as additional predictors. Derived thinking time was not included in the analyses due to high correlation with other variables and high variance inflation factor (VIF) in an exploratory analysis, to avoid multicollinearity. In the third setting, both comple-

tion time and extracted TMT metrics were included in the model as additional predictors. In all three settings, we first performed stepwise variable selection always keeping age, sex and education level in the model. Using the selected variables, GEE models with exchangeable correlation structures were fitted accounting for within subject and within family correlations for TICS, and accounting for within family correlations for all other measures. We conducted this set of analyses separately in TMT-A and TMT-B, as well as the difference between the metrics of TMT-A and TMT-B. We retained variables with significance level less than 0.05, but will only discuss results with significance levels that pass the Bonferroni correction of multiple testing, which is $0.5/10=0.005$. The HMM analysis was conducted in R using the rjags package and the annotation analysis was implemented in SAS 9.4 using PROC HPGENSELECT for variable selection and PROC GENMOD for the final parameter estimates of GEE models.

4.3 RESULTS

Out of 3,349 LLFS participants who were alive at the time of the second in-person visit, 2,364 participants received the TMT-A and 2,181 successfully completed the test. Only 2,330 participants received the TMT-B and 2,083 successfully completed the test. Table 4.1 summarizes the demographic characteristics and test scores of the participants who were included in this analysis. At the second in-person visit, participants who completed either test had age ranged from 43 to 106 years old, and those who completed TMT-B were slightly younger than participants who completed TMT-A (70.9 years vs. 71.8 years, $p=0.005$). Approximately 44.9% participants were male and 52.8% had college degrees or above. Participants who completed TMT-B had significantly higher test scores in animal fluency (21.8 vs.

Table 4.1: Demographic characteristics and test scores of participants who completed the Trail Making Tests.

	TMT-A	TMT-B	p-value(t-test)
N	2181	2083	
Age at visit 2 mean(SD), years	71.8	70.9	0.005
Gender, male(%)	982(45%)	931(44.7%)	0.83
Education, college and above(%)	1134(52%)	1118(53.7%)	0.27
Test Scores at Visit 2 (SD)			
TICS	15.7(4.1)	15.9(3.8)	0.13
Animal Fluency	21.3(6.4)	21.8(6.2)	0.02
DSST	45.1(13.8)	46.3(12.9)	0.01
Digits Forward	7.4(2.3)	7.4(2.3)	0.59
Digits Backward	6.3(2)	6.4(2)	0.19
Logical Memory-Immediate	13.5(4.3)	13.8(4.1)	0.03
Logical Memory-Delayed	12(4.8)	12.3(4.6)	0.03
HVLT	23.6(6.2)	24.1(5.8)	0.01
Gait Speed	1.002(0.3)	1.022(0.2)	0.01
Grip Strength	28(11)	28.6(10.9)	0.05

21.3, $p=0.02$), DSST (46.3 vs. 45.1, $p=0.01$), Logical Memory Recall tests (13.8 vs. 13.5, $p=0.03$ for Immediate Recall; 12.3 vs. 12.0, $p=0.03$ for Delayed Recall) and Gait speed (1.022 vs. 1.002, $p=0.01$). Paired comparisons of extracted metrics between TMT-A and TMT-B are shown in Table 4.2 . On average TMT-B took longer

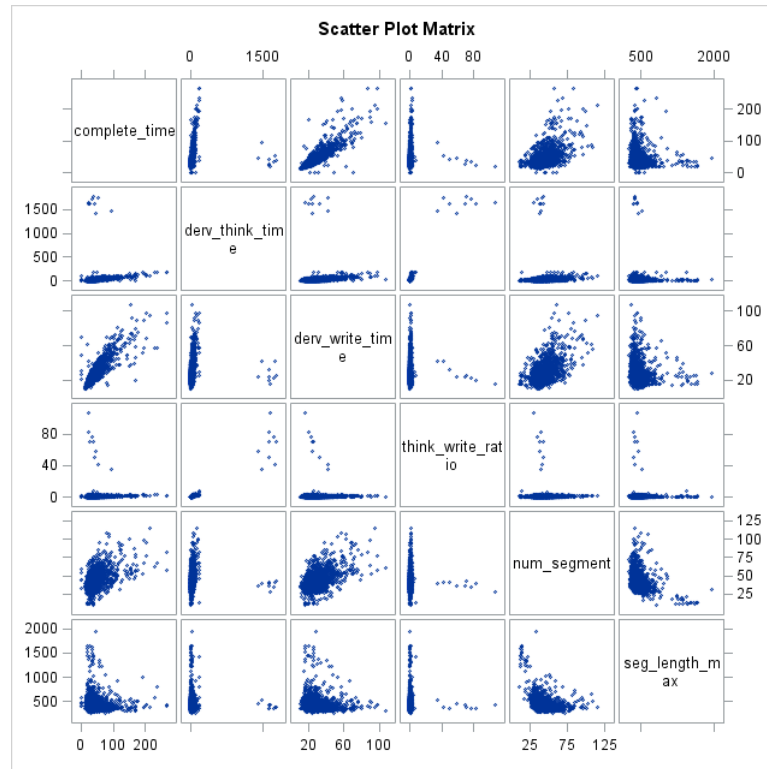
Table 4.2: Matched differences of metrics between TMT-A and TMT-B.

	TMT-A	TMT-B	p-value(t-test)
Completion time	42.3(25.2)	99.5(55.8)	<0.001
Derived thinking time	27.9(105.4)	67.7(123)	<0.001
Derived drawing time	27.7(10.9)	47.6(17.2)	<0.001
Think/draw ratio	1(4.4)	1.5(6.7)	0.003
Number of HMM segments	43(11.3)	59(18.8)	<0.001
Maximum length of segments	443.3(145.8)	694.3(198.6)	<0.001

to complete and required both longer derived thinking time and longer derived drawing time comparing to TMT-A. The ratio between derived thinking time and drawing time was also higher in TMT-B compared to TMT-A (1.5 vs. 1, $p=0.003$).

On average TMT-B resulted in more HMM segments (59 vs. 43, $p < 0.001$) with longer maximum length (694.3 vs. 443.3, $p < 0.001$) compared to TMT-A. Figure 4.4 and Figure 4.5 show the pairwise scatter plots of the new metrics and completion

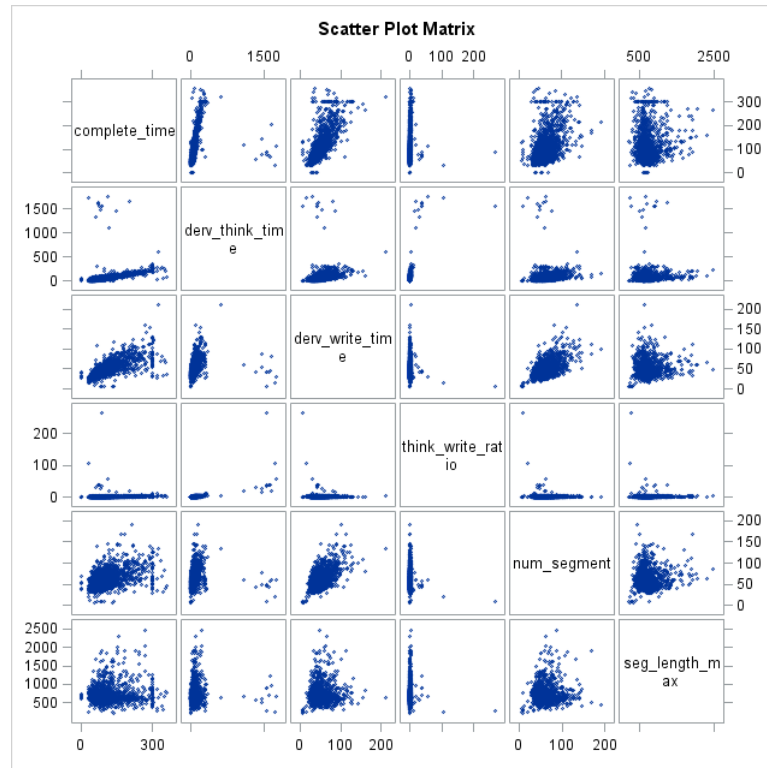
Figure 4.4: Pairwise scatter plot matrix for metrics in TMT-A.



time in TMT-A and TMT-B, respectively. Completion time and derived drawing time exhibit a strong correlation with $\rho = 0.85$ in TMT-A and $\rho = 0.73$ in TMT-B. A weak to moderate correlation is shown between completion time and number of HMM segments, with $\rho = 0.51$ in TMT-A and $\rho = 0.48$ in TMT-B.

Tables 4.3a, 4.3b, 4.4a, 4.4b, 4.5a & 4.5b show the GEE parameter estimates of the analysis of the new metrics derived from the digital administration of the TMT-A, while Tables 4.6a, 4.6b, 4.7a, 4.7b, 4.8a & 4.8b show the GEE parameter estimates of the analysis of the new metrics derived from the digital administration of the TMT-

Figure 4.5: Pairwise scatter plot matrix for metrics in TMT-B.



B. Tables 4.9a, 4.9b, 4.10a, 4.10b, 4.11a & 4.11b show the GEE parameter estimates of the analysis of difference between these two sets of new metrics. Overall, completion time was significantly associated with all cognitive and physical test measures when it was the only additional predictor in the GEE models. Derived drawing time was significantly associated with all test scores while the other extracted metrics including ratio between derived thinking time and derived drawing time, number of HMM segments and maximum length of HMM segments were significantly associated with selected test scores. When adjusted for completion time, the extracted metrics remained significant in selected test scores. Parameter estimates of the GEE models with significance levels pass the Bonferroni correction for multiple testing are summarized below.

TMT-A. As shown in Table 4.3a and 4.3b, in the first setting where completion time was used as the only predictor in addition to age, gender and education level, completion time had significant negative associations with all the traditional cognitive and physical test scores. This suggested participants with longer completion time were expected to have lower test scores. In the second setting shown in Table 4.4a and 4.4b, where the GEE models included the extracted metrics but not the completion time, the derived drawing time had a significant negative associations with all traditional measures of cognitive and physical function, suggesting that longer time spent drawing connections correlates with worse cognitive and physical functions, adjusting for age, gender and education level. The number of HMM segments was negatively associated with the DSST and gait speed. For each additional segment, the DSST score was expected to decrease by 0.11 points (SD: 0.02, $p < 0.0001$), and gait speed to decrease by 0.002m/s (SD: 0.0006, $p = 0.01$). Lastly, in the third setting (Table 4.5a and 4.5b), where we included both completion time and extracted metrics in the GEE models, the completion time remained significantly associated with all test scores with the exception of Logical Memory delayed recall and gait speed. Derived drawing time was negatively associated with these two test scores suggesting that the time spent in making connections between letters had more significant associations with the two test scores compared to the traditionally used completion time (parameter estimate = -0.05, SD:0.01, $p = 0.0002$ in Logical Memory delayed recall; parameter estimate = -0.001, SD:0.0005, $p = 0.01$ in gait speed). Derived drawing time was also significantly associated with the DSST score in addition to completion time (parameter estimate = -0.33, SD: 0.04, $p < 0.0001$). This suggested the derived drawing time explained additional variance of the DSST score that completion time did not explain. The ratio between derived

thinking time and derived drawing time was significantly associated with TICS, and Logical Memory immediate recall. For every one unit increase in the derived ratio, the TICS score was expected to increase by 0.4 points (SD: 0.12, $p=0.0009$), and Logical Memory immediate recall to increase by 0.05 points (SD:0.02, $p=0.003$). Higher number of HMM segment was associated with lower test scores in DSST and gait speed, with gait speed having the same model selected as in the second setting. For each additional segment, the DSST score was expected to decrease by 0.10 points (SD: 0.02, $p<0.0001$).

TMT-B. Similar to the analyses of TMT-A, completion time was significantly associated with all test measures when it was the only additional predictor in the model (Table 4.6a and 4.6b). In the second setting where we included only the extracted metrics in the GEE models as shown in Table 4.7a and 4.7b, derived drawing time was negatively associated with all test scores. The ratio between derived thinking time and derived writing time was associated with lower test scores in DSST, backward digit span and HVLT. For each unit higher in the derived ratio, DSST score was expected to decrease by 0.11 points (SD: 0.03, $p<0.0001$), backward digit span score to decrease by 0.02 points (SD: 0.01, $p=0.003$) and HVLT score to decrease by 0.03 points (SD: 0.01, $p=0.004$). This suggested longer thinking time and shorter drawing time implied poorer cognitive functions. As shown in Table 4.8a and 4.8b, in the third setting where the GEE analyses included both completion time and extracted metrics, the completion time was negatively associated with all test scores except for grip strength, for which derived drawing time had a significant negative association in place of completion time. Derived drawing time was also significant with outcome measures DSST and HVLT, in addition to the presence of completion time in the model. Each additional unit increase in

the derived thinking time and drawing time ratio was associated with lower test scores in DSST by 0.06 points (SD: 0.02, $p=0.0003$) and backward digit span by 0.01 points (SD: 0.004, $p=0.0003$). Number of HMM segments was associated with lower gait speed by 0.001m/s (SD: 0.0003, $p=0.001$). Lastly, each unit increase in maximum length of HMM segments was associated with higher score in DSST by 0.004 points (SD: 0.001, $p<0.0001$).

Difference between TMT-A and TMT-B. Similar to the analyses in both TMT-A and TMT-B, the difference in completion time had significant negative associations with all test measures except for grip strength as shown in Table 4.9a and 4.9b. This suggested as the additional time required to complete TMT-B increases, the test scores were expected to decrease. In Table 4.10a and 4.10b, when only the extracted metrics were included as predictors in the GEE models, the difference in derived drawing time was also negatively associated with all test measures, suggesting the longer it took the participants to draw the lines in TMT-B compared to TMT-A, the lower the test scores were expected. The difference in thinking and drawing time ratio was negative associated with DSST and backward digit span. For each unit higher in the ratio difference, the DSST score was expected to decrease by 0.08 points (SD: 0.02, $p<0.0001$) and the backward digit span score to decrease by 0.01 points (SD: 0.003, $p<0.0001$). Larger difference in the ratio between two tests suggested in terms of time allocation, participants spent more time to think than to draw in TMT-B compared to TMT-A. Lastly Table 4.11a and 4.11b show the GEE models adjusting for both completion time and extracted metrics. Derived drawing time was associated with grip strength in place of completion time, however, the association did not pass for the Bonferroni correction for multiple testing. The differences of metrics between TMT-A and TMT-B were not significantly associa-

tion with the cognitive and physical test measures.

4.4 DISCUSSION

In this chapter we extracted metrics from digitally recorded TMT by deriving time variables and using HMM to perform automatic segmentation of the recorded coordinates. We then analyzed the associations between these TMT metrics and other cognitive and physical function test scores. The overall results suggested that the extracted metrics could provide additional information of the underlying mechanism among cognitively and physically healthy individuals in addition to the time used to complete the tests.

The analyses suggest that the overall time employed to complete each test is predictive of cognitive and physical functions. However, when the completion time is analyzed together with the new derived metrics, the effects of completion time is explained by drawing time in logical memory delayed recall in TMT-A and grip strength in TMT-B, and a combination of drawing time and number of segments in gait speed in TMT-A. This indicates that the new metrics could serve as better predictors for outcome measures involve episodic memory, functional mobility and muscular strength. The results also show that both metrics from the HMM provide additional information in predicting cognitive and physical functions when analyzed together with total completion time. Higher number of segments is associated with lower score in DSST and gait speed, posing higher risk of decline in processing speed, attention, working memory and functional mobility. Longer maximum length of segments serves as an predictor of higher score in DSST, and could be an indicator of better processing speed, attention and working memory. The LLFS has also implemented several other tests using digital pen,

including the DSST and the Clock Drawing Test. In the study by Andersen et al. (2019), they categorized participants into different performance trajectories across the DSST that significantly correlate to performance on verbal fluency and episodic memory. In an analysis of the digital Clock Drawing Test, results also suggest there are significant associations between digital metrics of the Clock Drawing Test and APOE genotype (Du et al. (2021a), manuscript in preparation). This reaffirms that digital metrics help disentangle the underlying effect of cognitive and physical functions that is not captured in the traditional testing formats.

Several studies have employed fully-digital versions of the TMT, using iPads or Android based applications on tablets or personal computers (Fellows et al. (2017); Lunardini et al. (2019); Makizako et al. (2013)). Dahmen et al. (2017) and Fellows et al. (2017) implemented a digital version of the TMT and extracted information such as pauses, lifts of the pen, time spent inside circles and time between circles, with more sophisticated metrics extracted for TMT-B including average time before numbers or letters. Our version of the studies captures similar information such as lifts, pauses (in form of cluster coordinate pairs) and pressure. While in our approach we do not classify the coordinate pairs by inside or outside of circles, or before or after letters, the HMM segmentations provide information about detectable turn of direction in the drawing. Our innovative application of the HMM in this digital version of the TMT allows us to mathematically quantify and classify the recorded drawings in a holistic perspective, while not introducing potential confounders caused by low familiarity with technology in elderly participants that may influence the test results.

Table 4.3a: Parameter estimates of GEE models using completion time as predictor, TMT-A.

	TICS	DSST	Verbal Fluency	Digit Span - Forward	Digit Span - Backward	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	22.99	<.0001	79.86	<.0001	34.09	<.0001
age	-0.12	<.0001	-0.48	<.0001	7.70	<.0001
sex(male)	-1.07	<.0001	-0.19	<.0001	0.01	-0.02
education	0.26	<.0001	-3.99	<.0001	-0.01	0.94
spouse			0.85	<.0001	0.11	<.0001
completion time	-0.04	<.0001	-0.89	0.04	-0.01	<.0001

Table 4.3b: Parameter estimates of GEE models using completion time as predictor, TMT-A.

	Logical Memory - Immediate	Logical Memory - Delayed	HVLT	Gait Speed	Grip Strength	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	18.60	<.0001	18.74	<.0001	35.21	<.0001
age	-0.09	<.0001	-0.12	<.0001	-0.17	<.0001
sex(male)	-0.85	<.0001	-1.11	<.0001	-2.68	<.0001
education	0.24	<.0001	0.28	<.0001	0.38	<.0001
spouse					0.00	0.01
completion time	-0.03	<.0001	-0.03	<.0001	-0.06	<.0001
					-0.02	<.0001
					-0.05	<.0001

Table 4.4a: Parameter estimates of GEE models using extracted metrics as predictors, TMT-A.

	TICS		DSST		Verbal Fluency		Digit Span - Forward		Digit Span - Backward	
	Estimate	$p(> Z)$	Estimate	$p(> Z)$	Estimate	$p(> Z)$	Estimate	$p(> Z)$	Estimate	$p(> Z)$
(Intercept)	23.41	<.0001	85.31	<.0001	35.57	<.0001	7.90	<.0001	6.69	<.0001
age	-0.12	<.0001	-0.42	<.0001	-0.19	<.0001	-0.02	0.00	-0.02	<.0001
sex(male)	-1.12	<.0001	-3.95	<.0001	-0.10	0.67	-0.02	0.83	-0.06	0.44
education	0.28	<.0001	0.82	<.0001	0.26	<.0001	0.12	<.0001	0.15	<.0001
spouse			-0.98	0.02						
derived drawing time	-0.09	<.0001	-0.46	<.0001	-0.11	<.0001	-0.02	0.0001	-0.03	<.0001
think/draw ratio										
number of segments			-0.11	<.0001	-0.02	0.04				
maximum length of segments										

Table 4.4b: Parameter estimates of GEE models using extracted metrics as predictors, TMT-A.

	Logical Memory - Immediate		Logical Memory - Delayed		HVLFT		Gait Speed		Grip Strength	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	19.55	<.0001	19.72	<.0001	37.07	<.0001	1.81	<.0001	52.94	<.0001
age	-0.11	<.0001	-0.13	<.0001	-0.18	<.0001	-0.01	<.0001	-0.37	<.0001
sex(male)	-0.90	<.0001	-1.15	<.0001	-2.68	<.0001	0.03	<.0001	14.33	<.0001
education	0.25	<.0001	0.29	<.0001	0.38	<.0001	0.004	0.01	-0.16	0.00
spouse										
derived drawing time	-0.04	0.00	-0.05	0.00	-0.11	<.0001	-0.004	<.0001	-0.12	<.0001
think/draw ratio										
number of segments					-0.03	0.03	-0.002	0.001		
maximum length of segments							-0.0001	0.02		

Table 4.5a: Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-A.

	TICS		DSST		Verbal Fluency		Digit Span - Forward		Digit Span - Backward	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	22.82	<.0001	83.23	<.0001	34.62	<.0001	7.70	<.0001	6.39	<.0001
age	-0.12	<.0001	-0.40	<.0001	-0.19	<.0001	-0.01	0.01	-0.02	0.00
sex(male)	-1.08	<.0001	-3.86	<.0001	-0.03	0.89	-0.01	0.94	-0.05	0.55
education	0.27	<.0001	0.80	<.0001	0.25	<.0001	0.11	<.0001	0.15	<.0001
spouse			-0.99	0.02						
completion time	-0.04	<.0001	-0.08	0.0001	-0.05	<.0001	-0.01	<.0001	-0.01	<.0001
derived drawing time			-0.33	<.0001						
think/draw ratio	0.40	0.00								
number of segments			-0.10	<.0001	-0.03	0.02				
maximum length of segments										

Table 4.5b: Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-A.

	Logical Memory - Immediate	Logical Memory - Delayed	HVLT	Gait Speed	Grip Strength
	Estimate	Estimate	Estimate	Estimate	Estimate
	$p(> Z)$	$p(> Z)$	$p(> Z)$	$p(> Z)$	$p(> Z)$
(Intercept)	18.54	19.72	35.21	1.81	52.87
age	<.0001	<.0001	<.0001	<.0001	<.0001
sex(male)	-0.09	-0.13	-0.17	-0.01	-0.37
education	<.0001	<.0001	<.0001	<.0001	<.0001
spouse	0.24	0.29	0.38	0.004	-0.16
completion time	<.0001		<.0001		<.0001
derived drawing time	-0.03	-0.05	-0.06	-0.004	-0.04
		0.00	0.00	<.0001	
think/draw ratio	0.05				0.07
number of segments	0.00			-0.002	-0.04
maximum length of segments				-0.0001	0.04
				0.001	
				0.02	

Table 4.6a: Parameter estimates of GEE models using completion time as predictor, TMT-B.

	TICS		DSST		Verbal Fluency		Digit Span - Forward		Digit Span - Backward	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	22.16	<.0001	76.59	<.0001	32.00	<.0001	7.33	<.0001	5.52	<.0001
age	-0.09	<.0001	-0.38	<.0001	-0.14	<.0001	0.002	0.77	0.01	0.23
sex(male)	-1.13	<.0001	-4.00	<.0001	-0.15	0.55	0.03	0.75	-0.04	0.59
education	0.19	<.0001	0.68	<.0001	0.21	<.0001	0.09	<.0001	0.13	<.0001
completion time	-0.02	<.0001	-0.09	<.0001	-0.03	<.0001	-0.01	<.0001	-0.01	<.0001

Table 4.6b: Parameter estimates of GEE models using completion time as predictor, TMT-B.

	Logical Memory - Immediate		Logical Memory - Delayed		HVLIT		Gait Speed		Grip Strength	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	17.58	<.0001	17.87	<.0001	33.48	<.0001	1.62	<.0001	52.41	<.0001
age	-0.06	<.0001	-0.08	<.0001	-0.14	<.0001	-0.01	<.0001	-0.38	<.0001
sex(male)	-0.85	<.0001	-1.09	<.0001	-2.78	<.0001	0.03	0.00	14.78	<.0001
education	0.19	<.0001	0.22	<.0001	0.33	<.0001	0.00	0.01	-0.20	<.0001
completion time	-0.02	<.0001	-0.02	<.0001	-0.03	<.0001	-0.001	<.0001	-0.01	<.0001

Table 4.8b: Parameter estimates of GEE models using completion time and extracted metrics as predictors, TMT-B.

	Logical Memory - Immediate	Logical Memory - Delayed	HVLT	Gait Speed	Grip Strength
	Estimate	Estimate	Estimate	Estimate	Estimate
	p(> Z)	p(> Z)	p(> Z)	p(> Z)	p(> Z)
(Intercept)	17.58	17.87	32.34	1.65	53.14
age	-0.06	-0.08	-0.12	-0.01	-0.38
sex(male)	-0.85	-1.09	-2.79	0.03	14.81
education	0.19	0.22	0.32	-0.001	-0.19
completion time	-0.02	-0.02	-0.02	-0.001	-0.05
derived drawing time			-0.03	0.001	<.0001
think/draw ratio			0.02	-0.001	0.001
number of segments			0.001	0.02	
maximum length of segments			0.02	0.02	

Table 4.9a: Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.

	TICS		DSST		Verbal Fluency		Digit Span - Forward		Digit Span - Backward	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	22.24	<.0001	80.20	<.0001	33.22	<.0001	7.39	<.0001	5.72	<.0001
age	-0.10	<.0001	-0.51	<.0001	-0.18	<.0001	-0.01	0.30	-0.004	0.39
sex(male)	-1.18	<.0001	-4.24	<.0001	-0.12	0.62	0.01	0.94	-0.07	0.42
education	0.21	<.0001	0.78	<.0001	0.23	<.0001	0.10	<.0001	0.14	<.0001
completion time	-0.02	<.0001	-0.08	<.0001	-0.03	<.0001	-0.01	<.0001	-0.01	<.0001

Table 4.9b: Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.

	Logical Memory - Immediate		Logical Memory - Delayed		HVLTL		Gait Speed		Grip Strength	
	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)	Estimate	p(> Z)
(Intercept)	17.68	<.0001	17.90	<.0001	34.16	<.0001	1.67	<.0001	52.89	<.0001
age	-0.08	<.0001	-0.10	<.0001	-0.17	<.0001	-0.01	<.0001	-0.41	<.0001
sex(male)	-0.88	<.0001	-1.13	<.0001	-2.78	<.0001	0.03	0.001	14.75	<.0001
education	0.22	<.0001	0.26	<.0001	0.36	<.0001	0.004	0.04	-0.17	0.0004
completion time	-0.02	<.0001	-0.02	<.0001	-0.02	<.0001	-0.001	<.0001	-0.01	0.01

Table 4.10a: Parameter estimates of GEE models using extracted metrics as predictors, difference between TMT-B and TMT-A.

	TICS	DSST	Verbal Fluency	Digit Span - Forward	Digit Span - Backward
	Estimate	Estimate	Estimate	Estimate	Estimate
	$p(> Z)$	$p(> Z)$	$p(> Z)$	$p(> Z)$	$p(> Z)$
(Intercept)	22.86	84.53	34.81	7.82	6.44
age	-0.12	-0.63	-0.22	-0.02	-0.02
sex(male)	-1.16	-4.25	-0.11	0.01	-0.06
education	0.24	0.92	0.27	0.11	0.16
derived drawing time	-0.05	-0.13	-0.05	-0.01	-0.02
think/draw ratio	-0.37	-0.08			-0.01
number of segments					
maximum length of segments					-0.0004
					0.03

Table 4.10b: Parameter estimates of GEE models using extracted metrics as predictors, difference between TMT-B and TMT-A.

	Logical Memory - Immediate	Logical Memory - Delayed	HVLT	Gait Speed	Grip Strength
	Estimate p(> Z)	Estimate p(> Z)	Estimate p(> Z)	Estimate p(> Z)	Estimate p(> Z)
(Intercept)	18.72 <.0001	19.06 <.0001	35.55 <.0001	1.71 <.0001	53.32 <.0001
age	-0.10 <.0001	-0.13 <.0001	-0.20 <.0001	-0.01 <.0001	-0.41 <.0001
sex(male)	-0.86 <.0001	-1.10 <.0001	-2.75 <.0001	0.03 0.002	14.78 <.0001
education	0.25 <.0001	0.29 <.0001	0.39 <.0001	0.01 0.01	-0.16 0.001
derived drawing time	-0.03 0.003	-0.03 0.001	-0.05 <.0001	-0.002 0.001	-0.03 0.01
think/draw ratio					
number of segments			0.02		
maximum length of segments			0.04		

Table 4.11b: Parameter estimates of GEE models using completion time as predictor, difference between TMT-B and TMT-A.

	Logical Memory - Immediate	Logical Memory - Delayed	HVLT	Gait Speed	Grip Strength
	Estimate	Estimate	Estimate	Estimate	Estimate
	p(> Z)	p(> Z)	p(> Z)	p(> Z)	p(> Z)
(Intercept)	17.68	17.90	34.16	1.67	52.89
age	-0.08	-0.10	-0.17	-0.01	-0.41
sex(male)	-0.88	-1.13	-2.78	0.03	14.75
education	0.22	0.26	0.36	0.004	-0.17
completion time	-0.02	-0.02	-0.02	-0.001	-0.03
derived drawing time					0.01
think/draw ratio					0.05
number of segments					
maximum length of segments					0.01

CHAPTER 5

Discussion

In this dissertation we present novel methods and applications to analyze neuropsychological test scores using Bayesian models. The dissertation work described in chapter 2 and chapter 3 proposed a novel approach to perform variable selection in Bayesian hierarchical models with a simulation study showing this proposed algorithm produces more parsimonious results compared to DIC, as well as a real world example to show an application of this algorithm. The work described in chapter 4 provided a novel application of the HMM to analyze data from neuropsychological tests administered using emerging digital technologies. These works can be extended and future directions are discussed below.

We present our proposed variable selection algorithm that works in analogue to a backward selection, though this it can be extended to both forward and stepwise selection methods with careful modification of the algorithm logic. To apply this idea in a forward selection, the logic behind the algorithm needs to be reversed. In stead of identifying the variable that is least probable to be different from 0, we will need to identify and add the variable that is most probable to be different from 0. Similarly in a stepwise selection, after each variable is added to the model, the posterior credible intervals of all variables currently in the model should be checked, and the variable that is least probably to be different from 0 should be removed before refitting the model.

Future steps for applying HMM to analyze TMT data include extracting metrics focus on local rather than global characteristics. One way to accomplish this is to outline coordinate perimeters for each node (a number or a letter). By creating a non-overlapping region around each node, we can extract metrics such as

number of segments between two nodes, drawing speed leading up to a particular node, and arriving or departure drawing speed of numbers compared to letters. With these extracted local metrics, we can then consider performing hierarchical clustering of these characteristics to explore the associations between patterns of drawing and cognitive or physical function states, as well as other factors such as socioeconomic characteristics and genetic factors including the *APOE* gene.

LIST OF JOURNAL ABBREVIATIONS

Am J Epidemiol	American Journal of Epidemiology
Ann Neurol	Annals of Neurology
Arch Clin Neuropsychol	Archives of Clinical Neuropsychology
Bayesian Anal	Bayesian Analysis
Clin Neuropsychol	The Clinical Neuropsychologist
Eur J Hum Genet	European Journal of Human Genetics
Front Genet	Frontiers in Genetics
Geriatr Gerontol Int	Geriatrics & Gerontology International
IEEE Transactions on Informa- tion Theory	Institute of Electrical and Electronics Engi- neers Transactions on Information Theory
Int J Mol Sci	International Journal of Molecular Sciences
J Gerontol ABiol Sci Med Sci	The journals of gerontology. Series A, Biologi- cal sciences and medical sciences
J Head Trauma Rehabil	The Journal of Head Trauma Rehabilitation
JAMA Netw Open	Journal of the American Medical Association Network Open
JAMA Neurol	Journal of the American Medical Association Neurology
JConsult Psychol	Journal of Consulting and Clinical Psychology
N Engl J Med	The New England Journal of Medicine
Nat Rev Neurol	Nature Reviews Neurology
Neurobiol Aging	Neurobiology of Aging

Neurosci Lett

PLoS One

Rejuvena-tion Res

Sci Rep

Statist Sci

TechnolHealth Care

Transl Psychiatry

Neuroscience Lettersă

Public Library of Science One

Rejuvenation Research

Scientific Reports

Statistical Science

Technology and Health Care

Translational Psychiatry

BIBLIOGRAPHY

- Alosco, M. L., Aslan, M., Du, M., Ko, J., Grande, L., Proctor, S. P., Concato, J., & Vasterling, J. J. (2016). Consistency of recall for deployment-related traumatic brain injury. *J Head Trauma Rehabil*, 31(5), 360–8.
- Andersen, S. L., Sweigart, B., Cosentino, S., Wojczynski, M. K., Glynn, N. W., Thyagarajan, B., Mengel-From, J., Thielke, S., Perls, T. T., & Sebastiani, P. (2019). Digital technology identifies distinct performance patterns on the digit symbol substitution test among cognitively healthy adults. *Alzheimer's Dementia*, 15, P1555–P1555.
- Arbuthnott, K., & Frank, J. (2000). Trail making test, part b as a measure of executive control: Validation using a set-switching paradigm. *Journal of Clinical and Experimental Neuropsychology*, 22(4), 518–528.
- Barral, S., Singh, J., Fagan, E., Cosentino, S., Andersen-Toomey, S. L., Wojczynski, M. K., Feitosa, M., Kammerer, C. M., Schupf, N., & Long Life Family, S. (2017). Age-related biomarkers in llfs families with exceptional cognitive abilities. *J Gerontol A Biol Sci Med Sci*, 72(12), 1683–1688.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1), 164–171.
- Blair, C. K., Folsom, A. R., Knopman, D. S., Bray, M. S., Mosley, T. H., Boerwinkle, E., & Atherosclerosis Risk in Communities Study, I. (2005). Apoe genotype and cognitive decline in a middle-aged cohort. *Neurology*, 64(2), 268–76.
- Bondell, H. D., & Reich, B. J. (2012). Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(23482517), 1610–1624.
- Caselli, R. J., Dueck, A. C., Osborne, D., Sabbagh, M. N., Connor, D. J., Ahern, G. L., Baxter, L. C., Rapcsak, S. Z., Shi, J., Woodruff, B. K., Locke, D. E., Snyder, C. H., Alexander, G. E., Rademakers, R., & Reiman, E. M. (2009). Longitudinal modeling of age-related memory decline and the apoe epsilon4 effect. *N Engl J Med*, 361(3), 255–63.
- Caselli, R. J., Reiman, E. M., Osborne, D., Hentz, J. G., Baxter, L. C., Hernandez, J. L., & Alexander, G. G. (2004). Longitudinal changes in cognition and behavior in asymptomatic carriers of the apoe e4 allele. *Neurology*, 62(11), 1990–5.

- Crowe, S. F. (1998). The differential contribution of mental tracking, cognitive flexibility, visual search, and motor speed to performance on parts a and b of the trail making test. *Journal of clinical psychology, 54*(5), 585–591.
- Dahmen, J., Cook, D., Fellows, R., & Schmitter-Edgecombe, M. (2017). An analysis of a digital variant of the trail making test using machine learning techniques. *Technol Health Care, 25*(2), 251–264.
- Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing, 12*(1), 27–36.
- Dewey, M. E., & Saz, P. (2001). Dementia, cognitive impairment and mortality in persons aged 65 and over living in the community: a systematic review of the literature. *International Journal of Geriatric Psychiatry, 16*(8), 751–761.
- Du, M., Andersen, S. L., Schupf, N., Feitosa, M. F., Barker, M. S., Perls, T., & Sebastiani, P. (2021a). Association between apoe alleles and change of neuropsychological tests in the long life family study. *Journal of Alzheimer's Disease, 79*, 117–125.
- Du, M., Andersen, S. L., Schupf, N., Feitosa, M. F., Barker, M. S., Perls, T. T., & Sebastiani, P. (2021b). Association between apoe alleles and change of neuropsychological tests in the long life family study. *Journal of Alzheimer's Disease, 79*, 117–125.
- Du, M., Van Ness, S., Gordeuk, V., Nouraie, S. M., Nekhai, S., Gladwin, M., Steinberg, M. H., & Sebastiani, P. (2018). Biomarker signatures of sickle cell disease severity. *Blood cells, molecules diseases, 72*, 1–9.
- Fan, J., Tao, W., Li, X., Li, H., Zhang, J., Wei, D., Chen, Y., & Zhang, Z. (2019). The contribution of genetic factors to cognitive impairment and dementia: Apolipoprotein e gene, gene interactions, and polygenic risk. *Int J Mol Sci, 20*(5).
- Fellows, R. P., Dahmen, J., Cook, D., & Schmitter-Edgecombe, M. (2017). Multicomponent analysis of a digital trail making test. *Clin Neuropsychol, 31*(1), 154–167.
- Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for bayesian regression models. *The American Statistician, 73*(3), 307–309.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association, 88*(423), 881–889.
- George, E. I., & McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica, 7*(2), 339–373.

- Godsill, S. J. (2001). On the relationship between markov chain monte carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2), 230–248.
- Gomez, R. G., & White, D. A. (2006). Using verbal fluency to detect very mild dementia of the alzheimer type. *Arch Clin Neuropsychol*, 21(8), 771–5.
- Gottesman, R. F., Albert, M. S., Alonso, A., Coker, L. H., Coresh, J., Davis, S. M., Deal, J. A., McKhann, G. M., Mosley, T. H., Sharrett, A. R., Schneider, A. L. C., Windham, B. G., Wruck, L. M., & Knopman, D. S. (2017). Associations between midlife vascular risk factors and 25-year incident dementia in the atherosclerosis risk in communities (aric) cohort. *JAMA Neurol*, 74(10), 1246–1254.
- Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4), 711–732.
- Harvey, P. D., Aslan, M., Du, M., Zhao, H., Siever, L. J., Pulver, A., Gaziano, J. M., & Concato, J. (2016). Factor structure of cognition and functional capacity in two studies of schizophrenia and bipolar disorder: Implications for genomic studies. *Neuropsychology*, 30(1), 28–39.
- Helkala, E. L., Koivisto, K., Hanninen, T., Vanhanen, M., Kervinen, K., Kuusisto, J., Mykkanen, L., Kesaniemi, Y. A., Laakso, M., & Riekkinen, S., P. (1996). Memory functions in human subjects with different apolipoprotein e phenotypes during a 3-year population-based follow-up study. *Neurosci Lett*, 204(3), 177–80.
- Henderson, A. S., Eastel, S., Jorm, A. F., Mackinnon, A. J., Korten, A. E., Christensen, H., Croft, L., & Jacomb, P. A. (1995). Apolipoprotein e allele epsilon 4, dementia, and cognitive decline in a population sample. *Lancet*, 346(8987), 1387–90.
- Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9), 1212–22.
- Hyman, B. T., Gomez-Isla, T., Briggs, M., Chung, H., Nichols, S., Kohout, F., & Wallace, R. (1996). Apolipoprotein e and cognitive change in an elderly population. *Ann Neurol*, 40(1), 55–66.
- Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465), 279–290.
- Kim, Y. J., Seo, S. W., Park, S. B., Yang, J. J., Lee, J. S., Lee, J., Jang, Y. K., Kim, S. T., Lee, K. H., Lee, J. M., Lee, J. H., Kim, J. S., Na, D. L., & Kim, H. J. (2017). Protective effects of apoe e2 against disease progression in subcortical vascular

- mild cognitive impairment patients: A three-year longitudinal study. *Sci Rep*, 7(1), 1910.
- Krell-Roesch, J., Vemuri, P., Pink, A., Roberts, R. O., Stokin, G. B., Mielke, M. M., Christianson, T. J., Knopman, D. S., Petersen, R. C., Kremers, W. K., & Geda, Y. E. (2017). Association between mentally stimulating activities in late life and the outcome of incident mild cognitive impairment, with an analysis of the apoe epsilon4 genotype. *JAMA Neurol*, 74(3), 332–338.
- Kulminski, A. M., Arbeev, K. G., Culminskaya, I., Ukraintseva, S. V., Stallard, E., Province, M. A., & Yashin, A. I. (2015). Trade-offs in the effects of the apolipoprotein e polymorphism on risks of diseases of the heart, cancer, and neurodegenerative disorders: insights on mechanisms from the long life family study. *Rejuvenation Res*, 18(2), 128–35.
- Lezak, M. D., Howieson, D. B., Loring, D. W., & Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press, USA.
- Liu, C. C., Liu, C. C., Kanekiyo, T., Xu, H., & Bu, G. (2013). Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol*, 9(2), 106–18.
- Lunardini, F., Luperto, M., Daniele, K., Basilico, N., Damanti, S., Abbate, C., Mari, D., Cesari, M., Ferrante, S., & Borghese, N. A. (2019). Validity of digital trail making test and bells test in elderlies. In *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, (pp. 1–4).
- Makizako, H., Shimada, H., Park, H., Doi, T., Yoshida, D., Uemura, K., Tsutsumimoto, K., & Suzuki, T. (2013). Evaluation of multidimensional neurocognitive function using a tablet personal computer: test-retest reliability and validity in community-dwelling older adults. *Geriatr Gerontol Int*, 13(4), 860–6.
- Marill, T., & Green, D. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9, 11–17.
- Marioni, R. E., Campbell, A., Scotland, G., Hayward, C., Porteous, D. J., & Deary, I. J. (2016). Differential effects of the apoe e4 allele on different domains of cognitive ability across the life-course. *Eur J Hum Genet*, 24(6), 919–23.
- Mazur-Mosiewicz, A., & Dean, R. S. (2011). *Halstead-Reitan Neuropsychological Test Battery*, (pp. 727–731). Boston, MA: Springer US.
- Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404), 1023–1032.
- Muller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statist. Sci.*, 28(2), 135–167.

- Newman, A. B., Glynn, N. W., Taylor, C. A., Sebastiani, P., Perls, T. T., Mayeux, R., Christensen, K., Zmuda, J. M., Barral, S., Lee, J. H., Simonsick, E. M., Walston, J. D., Yashin, A. I., & Hadley, E. (2011). Health and function of participants in the long life family study: A comparison with other cohorts. *Aging (Albany NY)*, 3(1), 63–76.
- O'Donoghue, M. C., Murphy, S. E., Zamboni, G., Nobre, A. C., & Mackay, C. E. (2018). Apoe genotype and cognition in healthy individuals at risk of alzheimer's disease: A review. *Cortex*, 104, 103–123.
- O'Hara, R. B., & Sillanpaa, M. J. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian Anal.*, 4(1), 85–117.
- Oosterman, J. M., Vogels, R. L. C., van Harten, B., Gouw, A. A., Poggesi, A., Scheltens, P., Kessels, R. P. C., & Scherder, E. J. A. (2010). Assessing mental flexibility: neuroanatomical and neuropsychological correlates of the trail making test in elderly people. *The Clinical Neuropsychologist*, 24(2), 203–219.
- Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686.
- Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, M. B., Burke, J. R., Hurd, M. D., Potter, G. G., Rodgers, W. L., et al. (2008). Prevalence of cognitive impairment without dementia in the united states. *Annals of internal medicine*, 148(6), 427–434.
- Raber, J., Huang, Y., & Ashford, J. W. (2004). Apoe genotype accounts for the vast majority of ad risk and ad pathology. *Neurobiol Aging*, 25(5), 641–50.
- Rawle, M. J., Davis, D., Bendayan, R., Wong, A., Kuh, D., & Richards, M. (2018). Apolipoprotein-e (apoe) epsilon4 and cognitive decline over the adult life course. *Transl Psychiatry*, 8(1), 18.
- Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *J Consult Psychol*, 19(5), 393–4.
- Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8(3), 271–276.
- Sebastiani, P., Andersen, S. L., Sweigart, B., Du, M., Cosentino, S., Thyagarajan, B., Christensen, K., Schupf, N., & Perls, T. T. (2020). Patterns of multi-domain cognitive aging in participants of the long life family study. *GeroScience*, 42(5), 1335–1350.

- Sebastiani, P., Gurinovich, A., Nygaard, M., Sasaki, T., Sweigart, B., Bae, H., Andersen, S. L., Villa, F., Atzmon, G., Christensen, K., Arai, Y., Barzilai, N., Puca, A., Christiansen, L., Hirose, N., & Perls, T. T. (2019). Apoe alleles and extreme human longevity. *J Gerontol A Biol Sci Med Sci*, 74(1), 44–51.
- Sebastiani, P., Hadley, E. C., Province, M., Christensen, K., Rossi, W., Perls, T. T., & Ash, A. S. (2009). A family longevity selection score: ranking sibships by their longevity, size, and availability for study. *Am J Epidemiol*, 170(12), 1555–62.
- Sebastiani, P., & Perls, T. T. (2012). The genetics of extreme longevity: lessons from the new england centenarian study. *Front Genet*, 3, 277.
- Sebastiani, S. B. D. M. C. S. . P. T. e. a., Anderson SL (2020). Patterns of multi-domain cognitive aging in participants of the long life family study. *GeroScience*, 79(5), 758–774.
- Shinohara, M., Kanekiyo, T., Yang, L., Linthicum, D., Shinohara, M., Fu, Y., Price, L., Frisch-Daiello, J. L., Han, X., Fryer, J. D., & Bu, G. (2016). Apoe2 eases cognitive decline during aging: Clinical and preclinical evaluations. *Ann Neurol*, 79(5), 758–774.
- Smith, C. J., Ashford, J. W., & Perfetti, T. A. (2019). Putative survival advantages in young apolipoprotein 4 carriers are associated with increased neural stress. *Journal of Alzheimer's disease : JAD*, 68(3), 885–923.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Spreen, O., & Benton, A. L. (1965). Comparative studies of some psychological tests for cerebral damage. *Journal of Nervous and Mental Disease*.
- Sun, F., Sebastiani, P., Schupf, N., Bae, H., Andersen, S. L., McIntosh, A., Abel, H., Elo, I. T., & Perls, T. T. (2015). Extended maternal age at birth of last child and women's longevity in the long life family study. *Menopause*, 22(1), 26–31.
- Sánchez-Cubillo, I. ., Periañez, J., Adrover-Roig, D., Rodríguez-Sánchez, J., Rios-Lago, M., Tirapu, J., & Barcelo, F. (2009). Construct validity of the trail making test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society: JINS*, 15(3), 438.
- Tao, Q., Ang, T. F. A., DeCarli, C., Auerbach, S. H., Devine, S., Stein, T. D., Zhang, X., Massaro, J., Au, R., & Qiu, W. Q. (2018). Association of chronic low-grade inflammation with risk of alzheimer disease in apoe4 carriers. *JAMA Netw Open*, 1(6), e183597.

- War Department, W. D., Adjunct Generals Office (1944). Army individual test battery manual of directions and scoring.
- Wetherell, J. L., Reynolds, C. A., Gatz, M., & Pedersen, N. L. (2002). Anxiety, Cognitive Performance, and Cognitive Decline in Normal Aging. *The Journals of Gerontology: Series B*, 57(3), P246–P255.
- Wilson, R. S., Bienias, J. L., Berry-Kravis, E., Evans, D. A., & Bennett, D. A. (2002). The apolipoprotein e epsilon 2 allele and decline in episodic memory. *J Neurol Neurosurg Psychiatry*, 73(6), 672–7.
- Wolters, F. J., Yang, Q., Biggs, M. L., Jakobsdottir, J., Li, S., Evans, D. S., Bis, J. C., Harris, T. B., Vasan, R. S., Zilhao, N. R., Ghanbari, M., Ikram, M. A., Launer, L., Psaty, B. M., Tranah, G. J., Kulminski, A. M., Gudnason, V., Seshadri, S., & investigators, E. C. (2019). The impact of apoe genotype on survival: Results of 38,537 participants from six population-based cohorts (e2-charge). *PLoS One*, 14(7), e0219668.

CURRICULUM VITAE

