APPLICATION OF ADVANCED STATISTICAL METHODS IN AN AGING DATASET

by

Karina Nelly Alvarez

BS, Carnegie Mellon University, 2009

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Karina Nelly Alvarez

It was defended on

April 13, 2012

and approved by

Thesis Advisor:
Lisa Weissfeld, PhD, MA
Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Sachin Yende, MD, MS
Associate Professor
Critical Care Medicine
University of Pittsburgh Medical School
University of Pittsburgh

Committee Member:
Caterina Rosano, MD, MPH
Associate Professor
Epidemiology
Graduate School of Public Health
University of Pittsburgh

APPLICATION OF ADVANCE STATISTICAL METHODS IN AN AGING DATASET

Karina Nelly Alvarez, M.S.

University of Pittsburgh, 2012

The focus of this thesis was to explore the application of advanced statistical methods in the **Ginkgo Evaluation of Memory (GEM) Study.** GEMS enrolled 3,069 participants age 75 or older with normal cognition or mild cognitive impairment. Those with dementia were excluded from participation. After extensive medical and neuropsychological screening, participants were randomly assigned to receive twice-daily doses of either 120 milligrams of ginkgo extract or an identical-appearing placebo. The 240 milligrams daily dose of ginkgo was selected based on current dosage recommendations and prior clinical studies indicating possible effectiveness at this dosage. The products used in the study were supplied by Schwabe Pharmaceuticals, a German company. We focused on two methods, a flexible Cox model (Gray's model) and a trajectory procedure based on a mixture model that is implemented in the SAS procedure PROC TRAJ. The spline-based extension of the Cox model was applied to biomarker data; specifically: Cystatin-C, Beta Amyloid 40, Beta Amyloid 42, and a ratio of Beta Amyloid 42 over Beta Amyloid 40. We wanted to determine if the estimate of the log-hazard ratio changed over time for each of the biological measures. The trajectory analysis was used to determine if a patient's illness trajectory continued on the same path towards demented or non-demented before experiencing a pneumonia event. The trajectory analysis was applied to the longitudinal trajectories of activities of daily living (ADL), independent activities of daily living (IADL) and modified mini-mental status exam (3MSE). The Cox Spline analysis resulted in no statistically significant information added to the models using the spline analysis. Trajectory analysis concluded that patients on a downward trajectory at baseline only escalated before the pneumonia event. As the average life expectancy continues in increase in humans, it is important to evaluate statistical methods in the elderly population to identify subpopulations that need more medical attention than the population at large. Thus, the public health significance of this thesis is that by identifying these subgroups that are distinctly different from the overall population, we can provide preventative care where needed more efficiently.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Ginkgo biloba is among the most extensively studied herbs in use today. There have been claims that the herb helps to treat blood disorders and enhance memory. Some scientific studies have found evidence that supports these claims, others have found no association. Regardless of the conflicting findings, it is still believed by many that ginkgo may be help treat dementia, including Alzheimer's disease, and intermittent poor circulation in the legs and shows promise for enhancing memory in older adults (University of Maryland Medical Center).

In an attempt to formally answer questions about Ginkgo biloba's effectiveness as a treatment for a variety of illnesses, a randomized, double-blind clinical trial was created. The trial is known as the Ginkgo Evaluation of Memory Study (GEMS). The primary focus of the study was to determine the effect of 240mg/day Ginkgo biloba in decreasing the incidence of dementia and specifically Alzheimer's disease (AD). Secondary outcomes included: number of participants with the indicated cardiovascular disease or mortality and the progression of cognitive decline in standardized z-score scale (Steven T. DeKosky). More information about the trial can be found at clinicaltrials.gov, using identifier NCT00010803.

From this trial, came a rich dataset. The dataset was filled with years and years of measures from the same 3,069 participants that included everything from age, gender and BMI to information on cardiovascular illnesses. Our analysis focused on the cognitive measures collected from the subjects. We wanted to utilize the advanced statistical methods of trajectory analysis and the flexible Cox Model, also known as Gray's model, to determine the relationship between dementia, pneumonia, biomarkers and time.

In order to analyze the relationship between dementia and pneumonia, understanding cognitive function and functional status prior to hospitalization for pneumonia is important. Pneumonia could merely be a marker for those with rapidly declining cognitive function, and the relationship between pneumonia and dementia may not be causal. Therefore, longitudinal measures of cognition and functional status prior to pneumonia could be an important confounding factor. Using trajectory analysis, we estimated the probability of group membership for different trajectories for GEMS participants.

Similarly, although the exact cause if AD is still unknown, the consensus is that the accumulation of beta amyloid (Abeta) peptides in the senile plaques is one of the hallmarks of the progression of the disease (Rajendran L). Using the flexible Cox model, we were able to calculate the varying log-hazard ratios of biomarkers, like Abeta, over time to determine if there was an increased hazard of developing dementia associated with time.

1.1 RCT DESIGN

The Ginkgo biloba trial was designed to be a randomized, double-blind, placebo controlled clinical trial. By randomizing the subjects, we eliminated the potential bias due to differences in subject characteristics, both known and unknown. Double blinding ensures that neither the subjects nor the administrators know who is receiving the active treatment (G. Biloba) or the placebo treatment. The study was conducted in five academic medical centers in the United States between 2000 and 2008 (Steven T. DeKosky and al.)

1.2 OBJECTIVES

1.2.1 Pneumonia

Determine if pneumonia hospitalization increases the risk of dementia in older adults. Using trajectory analysis, we examined: functional status, activities of daily (ADL) and independent ADL (IADL); for cognitive function, we used the Teng's modified mini-mental status examination (3MSE).

1.2.2 Dementia

Using flexible Cox regression models, we wish to determine if the amount of biomarker fluid (Cystatin-C, beta amyloid 40 (Beta-40), beta amyloid 42 (Beta-42)) found in the body are stable from the start of the clinical trial to when subject developed dementia.

2.0 METHODOLOGY

2.1 PARTICIPANTS

The Ginkgo Evaluation of Memory (GEM) Study enrolled 3,069 community volunteers aged 75 or older with normal cognition or mild cognitive impairment. Starting in September 2000 to June 2002, subjects were recruited using voter registration and other purchased mailing lists from 4 US communities with academic medical centers: Hagerstown, Maryland (Johns Hopkins); Pittsburgh, Pennsylvania (University of Pittsburgh); Sacramento, California (University of California–Davis); and Winston-Salem and Greensboro, North Carolina (Wake Forest University). Participants were required to identify a proxy willing to be interviewed every 6 months at the time of each study visit. Signed informed consent was obtained from participants (Steven T. DeKosky and al.). Those with dementia were excluded from participation. An exhaustive exclusionary list and criteria for dementia and mild cognitive impairment can be found elsewhere (Steven T. DeKosky and al.)

2.2 MEASURES

Diagnosis of dementia was a primary endpoint of GEM study and was determined using DSM-IV criteria and included full neuropsychological evaluation, neurological exam, and magnetic resonance imaging when participants when they showed decline on cognitive testing battery, had memory problems, or were prescribed medications to improve memory. Pneumonia was identified from hospitalization records of adverse events. ADL, IADL, and 3MSE measurements were obtained every 6 months. Cystatin-C, Beta-40 and Beta-42 were taken only at baseline and in the 9th year of the study.

2.3 EXPERIMENTAL DESIGN

Participants were randomly assigned to receive twice-daily doses of either ginkgo extract or an identical-appearing placebo. Assignment to treatments G biloba or placebo was determined by permuted-block design by site to ensure balanced allocation between groups. All clinical, coordinating personnel and participants were blinded to treatment assignment. The only exceptions were site personnel responsible for monitoring serious adverse events and reporting to the study's data and safety monitoring board and the study pharmacist, who allocated the medication into batches, knew which medication was active. All of these personnel were unaware of participant information and had no contact with participant (Beth E. Snitz and al.).

2.4 MEDICATION ADMINISTRATION

For participants receiving the active treatment, they received twice-daily doses of either 120 milligrams of ginkgo extract. Controls were given an identical-appearing placebo. The 240 milligrams daily dose of ginkgo was selected based on current dosage recommendations and prior clinical studies indicating possible effectiveness at this dose. The products used in the study were supplied by Schwabe Pharmaceuticals, a German company. Subjects were given 6-month supplies of treatment in blister packs upon every visit. They were asked to return the blister packs during the next 6-month visit.

2.5 DATA ANALYSIS

2.5.1 Trajectory Models

The first was a trajectory modeling technique that created clusters of similar trajectories. The trajectory model is based on a discrete mixture model. This model allows for data grouping using different parameter values for each group distribution. This allows us to identify distinct subpopulations in the data that would not be seen if the data were to be analyzed assuming the same parameter values. Three types of distributions are offered within the trajectory method: censored (or regular) normal (CNORM), zero inflated (or regular) Poisson (ZIP), and Bernoulli distributions (logistic model). The method can also handle data with average values changing smoothly as a function of a dependent variable such as time; some sharp changes can be handled through the inclusion of time dependent covariates. We adapted the ZIP and CNORM model in particular to our dataset.

Zero-inflated Poisson regression is used to model count data that has an excess of zero counts. It assumes that with probability p, the only possible observation is 0 and with probability 1 – p, a Poisson (λ) random variable is observed. From a regression model, p is calculated from the coefficients as follows:

$$\mathrm{p(x)} = \begin{cases} 0 & \mathit{with\ probability}\ \rho \\ Poisson\ (\lambda) & \mathit{with\ probability}\ 1 - \ \rho \end{cases}$$

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \ \beta_2 X_2 + \ \ldots\ldots + \beta_x x_x$$

$$\rho = \frac{e^{\beta_0 + \ \beta_1 X_1 + \ \beta_2 X_2 + \cdots + \ \beta_x x_x}}{1 + e^{\beta_0 + \ \beta_1 X_1 + \ \beta_2 X_2 + \cdots + \ \beta_x x_x}}$$

Here, lambda (λ) is the mean of the Poisson distribution. The regression coefficients can also be combined to calculate the lambda of the distribution. Due to the exponential feature of the passion distribution, one must take the log of the sum of the coefficients to calculate lambda as shown above.

The second model used in the dataset was the CNORM. The CNORM distribution follows all the requirement of the normal distribution along with user defined minimum (a) and maximum (b) values. The function is definite as $\mathrm{P}(\mathrm{Y}_\mathrm{i} = y_i | \mathrm{C}_\mathrm{i} = k,\ \mathrm{W}_\mathrm{i} = w_i) =$:

$$\prod_{y_{ij=Min}} \Phi\left(\frac{\mathrm{Min} - \ \mu_{\mathrm{ijk}}}{\sigma}\right) \prod_{Min<y_{ij}<Max} \frac{1}{\sigma}\varphi\left(\frac{y_{ij} - \ \mu_{ij}}{\sigma}\right) \prod_{y_{ij}=Max} \left(1 - \ \Phi\left(\frac{\mathrm{Max} - \ \mu_{\mathrm{ijk}}}{\sigma}\right)\right)$$

$$\mu_{ijk} = \ \beta_{0k} + \mathrm{X}_{1ij}\beta_{1,k} + \mathrm{X}_{2ij}\beta_{2k} + \ \ldots + \ w_{ij}\delta_k$$

where Φ denotes the cumulative distribution function of the standard normal distribution and φ denotes the probability density function of the standard normal. The mean of the CNORM distribution is the sum of the beta coefficients from the regression model.

We assume the risk factors for each subject, Z' and the data trajectory for each subject consisting of the repeated measurements over T measurement periods, Y', are independent given the grouping, Ci.

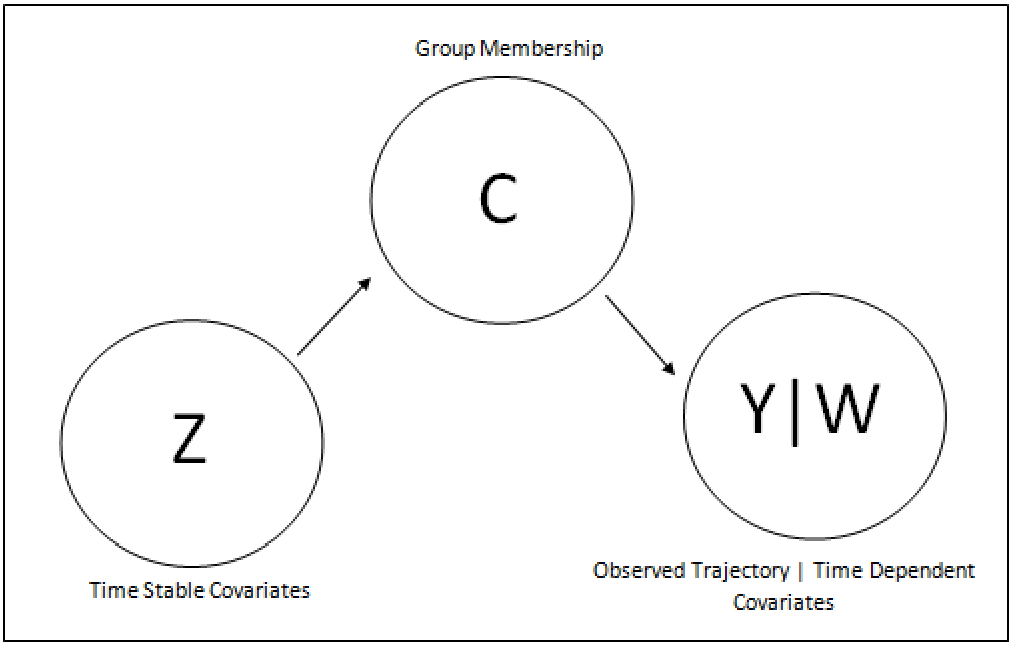

Figure 1. Adapted from Jones's Graphical Representation of the Independence Assumption in Trajectories (Jones, Nagin and Roeder)

The conditional distribution of the observable data for subject i, given risk factors and a time-dependent covariate, W', given K groups equals:

$$f(\boldsymbol{y}_i|\mathbf{z}_i, \mathbf{w}_i) = \sum_{k=1}^{K} P(C_i = k|\boldsymbol{Z}_i = \boldsymbol{z}_i)P(\boldsymbol{Y}_i = \boldsymbol{y}_i|C_i = k, \boldsymbol{W}_{\boldsymbol{i}} = \boldsymbol{w}_i)$$

The time-stable covariate effect on group membership is modeled, with a generalized logit function:

$$P(C_i = k \mid Z_i = z_i) = \frac{exp(\theta_k + \boldsymbol{\lambda_k z_i})}{\sum_{l=1}^{k} \exp(\theta_l + \boldsymbol{\lambda_l z_i})}$$

Here Pr(Yi = yi | Ci = k,Wi = wi) is the section of the modeling that the user chooses; it is here where Proc Traj allows us the option of modeling three different distributions (Jones, Nagin and Roeder).

2.5.2 Trajectory Criteria

The best model will be selected using two important factors: convergence and Bayesian Information Criterion (BIC). A sequence of numbers, Y_n, converges in law to **Y** if and only if $E[f(Y_n)] \rightarrow E[f(\mathbf{Y})]$ for every bounded continuous real-valued function *f* (E.L. Lehmann). Within the trajectory macro, convergence is assessed and its status is given along with the output. Because the function, *f*, and order of degree are determined by the user and not the program, it is important to check the convergence for every model run. From here, we will determine the best model via significance levels of coefficients. Model adjustment will be done by increasing or decreasing the degree of the models and/or by removing or adding a group.

2.5.3 Cox Spline Models

The flexible Cox spline model, more commonly known as Gray's model, consists of a flexible approach to using fixed knot splines with a small number of knots to model aspects of the data. The penalized partial likelihood was used to estimate the parameters of the model and determine significance. The models allow both linear terms and flexible spline functions of covariates. Let x be the linear terms and z be the covariates for the spline terms, for subject i. The covariates can be fixed or time-varying. The hazards model for the hazard for the ith subject then becomes:

$$\lambda(\mathrm{t}|x_i, z_i) = \lambda_0(\mathrm{t})\exp\left\{\mathrm{x'}_{\mathrm{i}}\beta + \sum_{\mathrm{j}=1}^{\mathrm{s}} \mathrm{f}_{\mathrm{j}}\left(\mathrm{z}_{\mathrm{ij}}\right)\right\}$$

Where $\lambda_0(\mathrm{t})$ the unspecified underlying hazard function and β is is a (column) vector of unknown parameters. M is the number of interior knots in each of the splines. Knot locations are roughly equal numbers of data points apart. The spline parameterization is as follow:

$$f_j(z) = \theta_{j0}z + \sum_{k=1}^{M+2} \theta_{jk}B_{jk}(z)$$

Here, Bjk is equal to the standard cubic B-spline basis functions (Gray). The fully penalized log-likelihood function is calculated with:

$$L_p(\eta) = L(\eta) - \frac{1}{2}\sum_{j=1}^{s} \lambda_j \theta'_j P_j \theta_j$$

where λj controls the amount of smoothing applied, with $\lambda j = 0$ corresponding to no penalty and $\lambda j \to \infty$ forcing $fj(z) = \theta joz$. This creates the vector θj. $L(\eta)$ is equal to Cox's log partial-likelihood and Pj is a nonnegative definite matrix . Explanation of the creation of the fully penalized log-likelihood function goes outside the scope of this paper.

2.5.4 Cox Spline Assumptions & Diagnostics

One of the most important assumptions of the proportional hazard model is that the effects of the covariates on the hazard do not change with time (Gray). The flexible spline analysis allows for this type of modeling. The final model will be selected through a trial and error approach that will determine the minimum number of knots needed to adequately represent the changes in the log-hazard ratio. Diagnostics will take the form of comparing log-likelihoods as outputted from the R package.

2.5.5 Missing Data

Data will be assumed missing at random (MAR). Some of the missingness is monotonic, meaning that once a subject does not have their information recorded for visit(i), all visits proceeding it, visit(i+1) are likely to also be missing. This is not always the case. There is no missingness in dementia outcome or pneumonia outcome. The two methods were conducted on two different subsets of the data. The Cox Spline analysis is based on a subset containing n = 2491. This subset was created by dropping subjects whose endpoint was anything other than end of study or dementia. The amount of missingness in this subset is ~1%. For the pneumonia dataset, we were able to use data from all subjects. Missingness in the total dataset for our variables is ~0.2%.

3.0 RESULTS

3.1 DEMOGRAPHIC CHARACTERISTICS

The models created in this analysis were unadjusted; no demographic information was used nor multivariable analysis conducted. We summarized some of the characteristics of the participants for both datasets in Appendix A.

3.2 TRAJECTORY ANALYSIS

Pneumonia hospitalizations information was collected from the subjects from record of adverse events. Of the 3,069 subjects, 221 cases of pneumonia were identified. The trajectory analysis was used to determine if functional status and cognitive abilities were affected after a pneumonia event. Cognitive abilities were measured using the 3MSE. Functional status was determined by self-reported difficulty with at least one activity of daily living (ADL) or one independent activity of daily living (IADL).

To create the pre-pneumonia and post-pneumonia event values, two different censoring variables were created from the patient data. A censoring variable was created for each variable. The pre-pneumonia variables (pre-3MSE, pre-adl, pre-iadl) were created by censoring out values of the respective variables by replacing any measure taken after the pneumonia event with a missing value (“.” in SAS). The post-pneumonia variables (post-3MSE, post-adl, post-iadl) were created in a similar but opposite fashion; all values taken before the pneumonia event were replaced by a missing value. For subjects that did not experience a pneumonia event, their values were all contained in the pre-pneumonia variables.

3.2.1 ADL/IADL

Subjects’ ADL/IADL values are considered to be worse as the values increase over time. This indicates an increasing number of daily functions a subject has trouble with.

ADL

The final model for ADL consists of three groups; the three groups have different estimates representing each one. For members in group one, it required a linear model to represent their illness trajectory. Of the cohort, 60.95% of the subjects were selected to be in this group. Group 1 members represent a subset of the population that experiences very little change in their ADL over time. In addition, Group 1 members were also the “most functional” subset in that their ADL scores were all around zero. All coefficients were statistically different from zero, p-value < 0.05.

Group 2 members have a very similar model. The main difference between the two groups is that the intercept for group 2 was higher, $\beta_{0(\text{ADL, G1})} = -3.34$ vs. $\beta_{0(\text{ADL, G2})} = -0.85$, indicating that group 2 members started worse in daily functions than group 1 members. Group 2 members also experienced little change over time; 32.39% of the cohort was selected to be in group 2. All coefficients were statistically different from zero, p-value < 0.05.

In contrast to the previous ADL group, group 3 members required a cubic model to represent their change in ADL over time. First off, their intercept was $\beta_{0(\text{ADL, G3})} = 0.16$, the only positive intercept. Group 3 members started with the highest value of ADL and only continued to increase over time. All coefficients were statistically different from zero, p-value < 0.05. Group 3 represented 6.66% of the cohort.

Table 1. Group Membership for ADL

Group Membership	Parameter Type	Estimate	Standard Error	Prob > |t|	Percent of Cohort
1	Intercept	-3.34	0.11	0.0000	60.95 %
	TIME	0.0004	0.00007	0.0000	
2	Intercept	-0.85	0.45	0.0000	32.39 %
	TIME	0.0002	0.00003	0.0000	
3	Intercept	0.16	0.07	0.0162	6.66 %
	TIME	0.0006	0.0001	0.0000	
	TIME^2	-0.000*	0.000*	0.0022	

*Number too small, not reported by macro; BIC = -14310.32

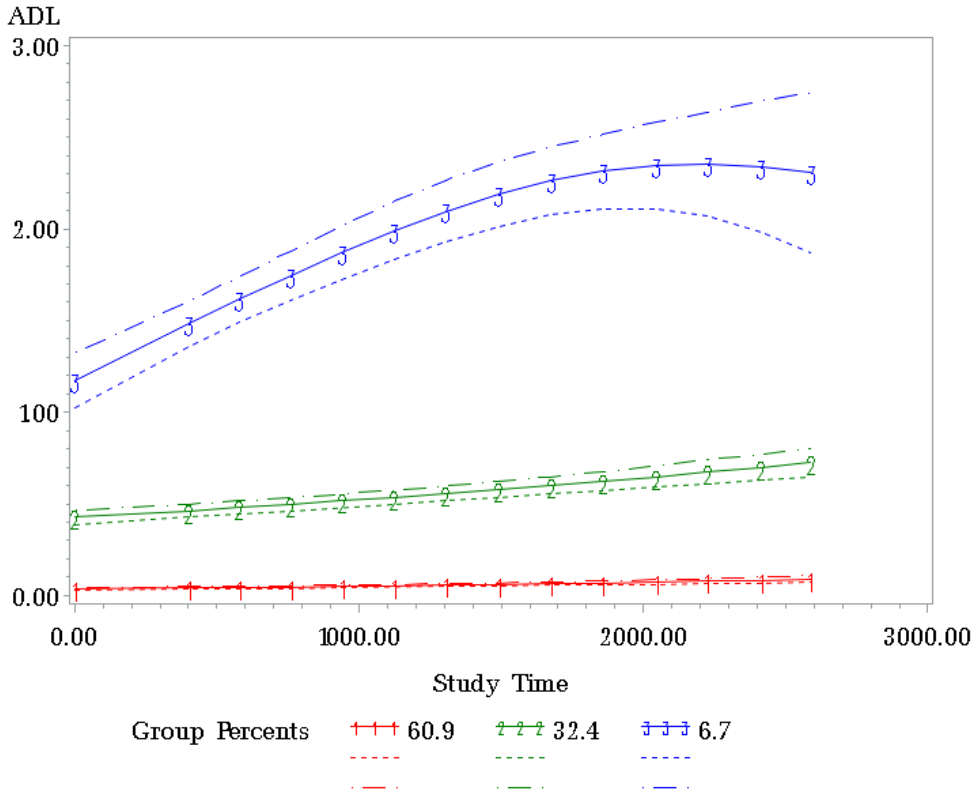

Figure 2. ADL Trajectory

IADL

The final model for IADL also consists of three groups. For members in group one, it required a quadratic model to represent their illness trajectory. Of the cohort, 46.99% of the subjects were selected to be in this group. Group 1 members represent a subset of the population that experiences very little change in their IADL over time. Similar to Group 1 in ADL, Group 1 members were also the "most functional" subset. The coefficient for squared variable was statistically different from zero, p-value < 0.05. The linear variable was marginally statistically different from zero, p-value < 0.10. This was enough significance for us to allow the addition of a higher order variable.

Group 2 members have a similar quadratic model. Again, the main difference between the two groups is that the intercept for group 2 was higher, $\beta_{0(\mathrm{IADL,\ G1})} = -2.64$ vs. $\beta_{0(\mathrm{IADL,\ G2})} = -$

0.67. Group 2 members also experienced little change over time; this represented 43.21% of the cohort. Only the coefficient for the linear variable was statistically different from zero, p-value < 0.05. Group 2 was given a second order IADL variable because convergence could not be reached without it.

Group 3 members required a cubic model to represent their change in IADL over time. Their intercept was $\beta_{0(\mathrm{IADL},\ \mathrm{G3})} = 0.18$, also the only positive intercept. Group 3 members started with the highest value of ADL and only continued to increase over time. All coefficients were statistically different from zero, p-value < 0.05. Group 3 represented 9.80% of the cohort.

Table 2. Group Membership for IADL

Group Membership	Parameter Type	Estimate	Standard Error	Prob > |t|	Percent of Cohort
1	Intercept	-2.64	0.11	0.0000	46.99 %
	TIME	-0.0003	0.0002	0.0819	
	TIME^2	0.000*	0.000*	0.0004	
2	Intercept	-0.67	0.04	0.0000	43.21 %
	TIME	0.0002	0.00007	0.0005	
	TIME^2	0.000*	0.000*	0.6906	
3	Intercept	0.18	0.05	0.0004	9.80 %
	TIME	0.0008	0.0001	0.0000	
	TIME^2	-0.000*	0.000*	0.0000	
	TIME^3	0.000*	0.000*	0.0000	

*Number too small, not reported by macro; BIC = -18879.88

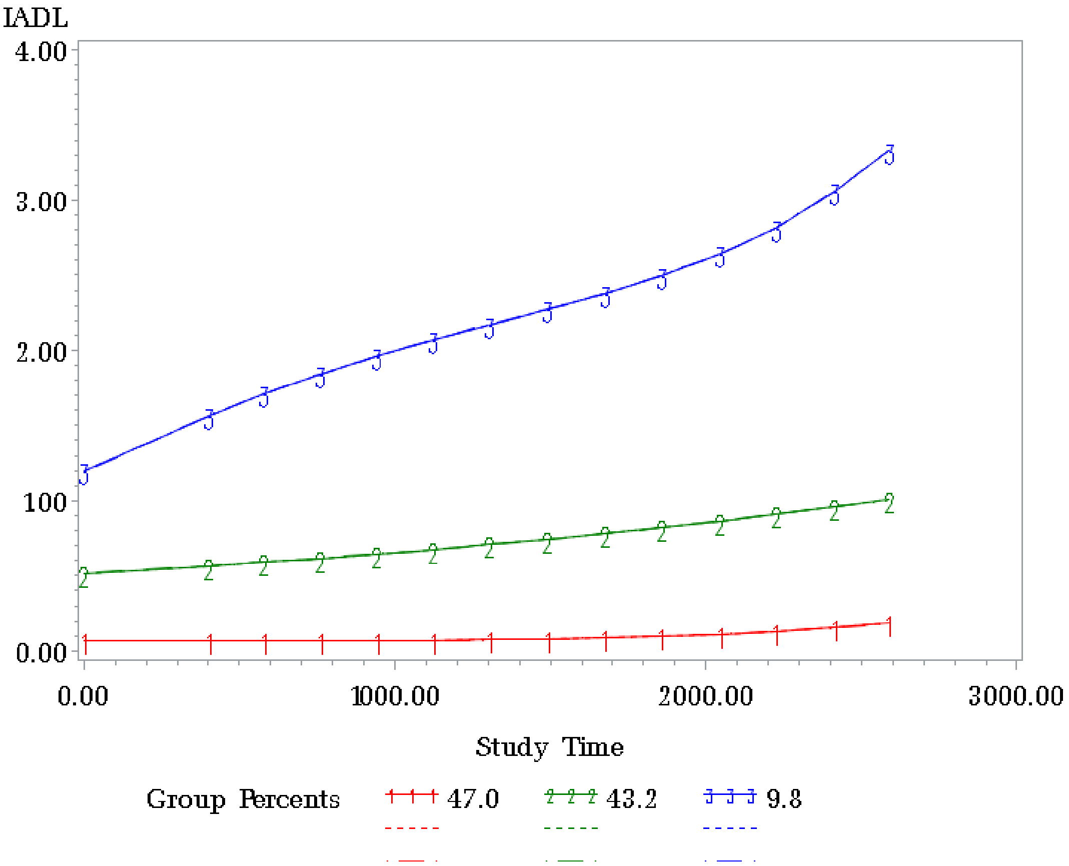

Figure 3. IADL Trajectory

3.2.2 3MSE

In contrast to the ADL/IADL scoring system, the higher the 3MSE value is, the better the subjects is doing cognitively. A score of 70 or below is considered demented. Three groups were selected to represent the subjects. All groups required cubic models. All coefficients in all groups were statistically different from zero, p-value < 0.05. The main different between the groups is their intercepts: $\beta_{0(3MSE,\ G1)} = 87.57$ vs. $\beta_{0(3MSE,\ G2)} = 92.22$ vs. $\beta_{0(3MSE,\ G3)} = 96.47$. In this framework, the most cognitively functioning group is group 3. In contrast, group 1 approaches lower measures, but nothing that would imply a demented status. The breakdown of the groups is as follows: Group 1 – 13.64%, Group 2 – 34.97%, and Group 3 – 51.40%.

Table 3. Group Membership for 3MSE

Group Membership	**Parameter Type**	**Estimate**	**Standard Error**	**Prob > |t|**	**Percent of Cohort**
1	Intercept	87.57	0.160	0.0000	13.64 %
	TIME	-0.003	0.0004	0.0000	
	$TIME^2$	0.000*	0.0000	0.0004	
2	Intercept	92.22	0.105	0.0000	34.97 %
	TIME	0.003	0.0002	0.0000	
	$TIME^2$	-0.000*	0.0000	0.0000	
3	Intercept	96.47	0.079	0.0000	51.40 %
	TIME	0.002	0.0001	0.0000	
	$TIME^2$	-0.000*	0.0000	0.0000	

*Number too small, not reported by macro; BIC = -87595.18

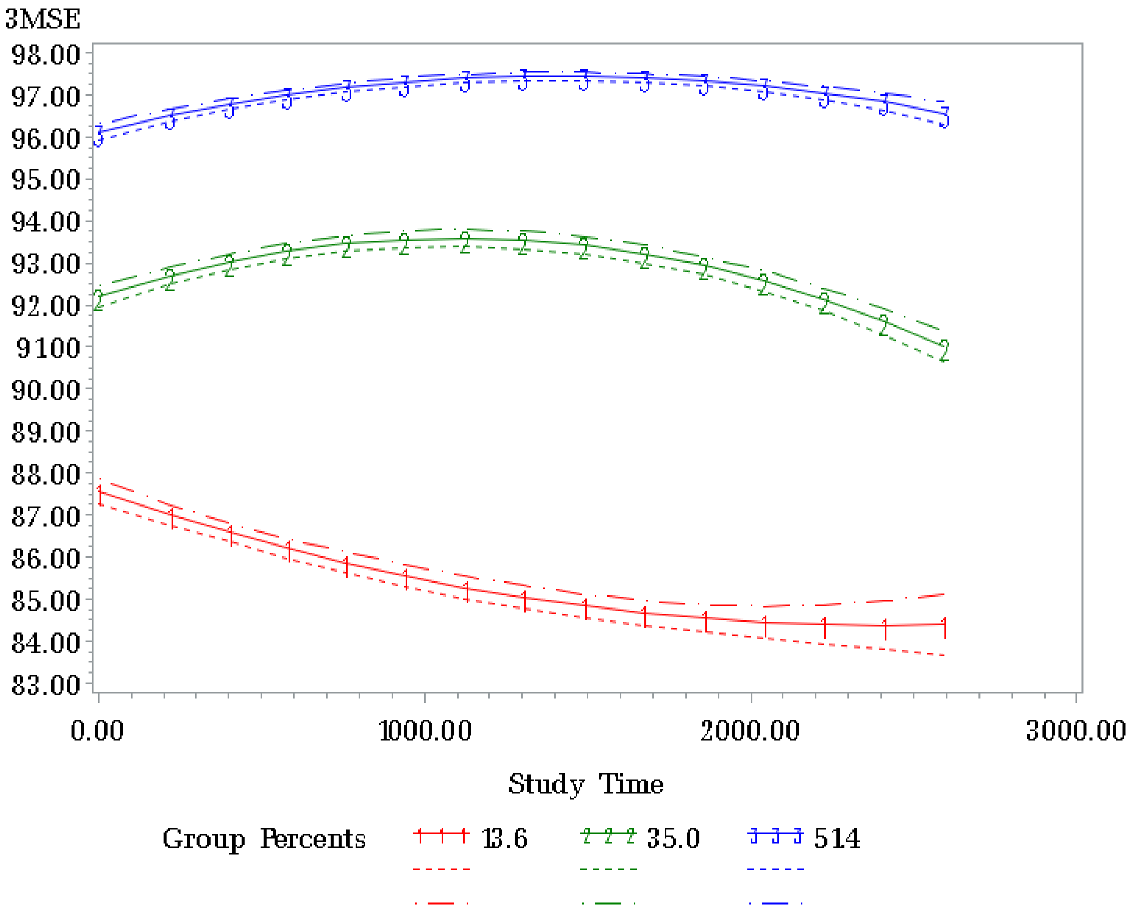

Figure 4. 3MSE Trajectory

3.3 COX SPLINE ANALYSIS

There were 523 cases of dementia developed over the 7 year period. Application of Gray's model was used to measure changes in beta amyloid fluid and Cystatin-C over time relative to when a patient developed dementia. It is believed that Cystatin-C, Aβ1-40 amyloid, and Aβ1-42 amyloid are most closely associated with the risk of developing dementia and therefore the focus of this part of the analysis. Cystatin-C, Aβ1-40 amyloid, and Aβ1-42 amyloid were measure at baseline and in the 9th year of the study. Subjects were measured at different times; giving us a range of beta-amyloid and Cystatin-C measures with corresponding times. We also explored the composite variable ratio (Aβ1-42/Aβ1-40) to determine if it better explained the behavior and relationship of beta amyloid and dementia. After several model runs to determine the number of knots to use, we decided that each of the variables could not be measured accurately with the same number of knots.

Because of the difference in knot numbers, the variables were plotted separately. The log-hazard ratio is the primary output of Gray's model. This is interpreted the same as a standard hazard ratio where an increasing value indicates an increasing risk of hazard.

3.3.1 Beta Amyloid 40

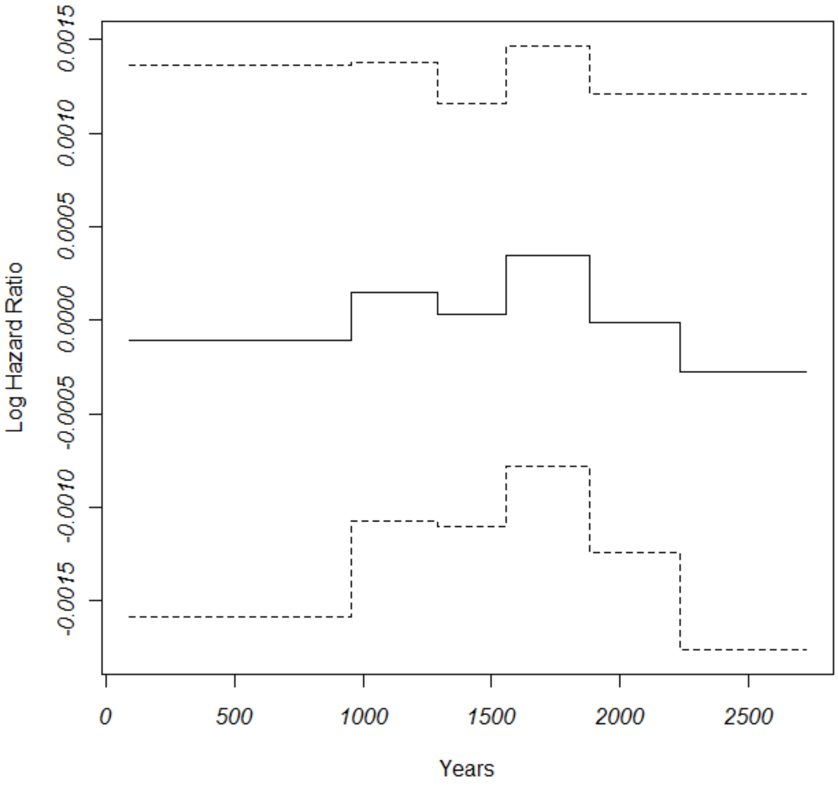

Figure 5. Beta-40 Log -- Hazard Ratio over time with 95% CI

The estimates for the log hazard ratio for Beta-40 ranged from -0.0003 to 0.0003. This indicates that during the study, Beta-40 was both protective and increased the risk of harm of dementia. The final model required 10 knots to completely describe the movement of the log-hazard estimated. The final model had a fully penalized log-likelihood of -3869.37.

3.3.2 Beta Amyloid 42

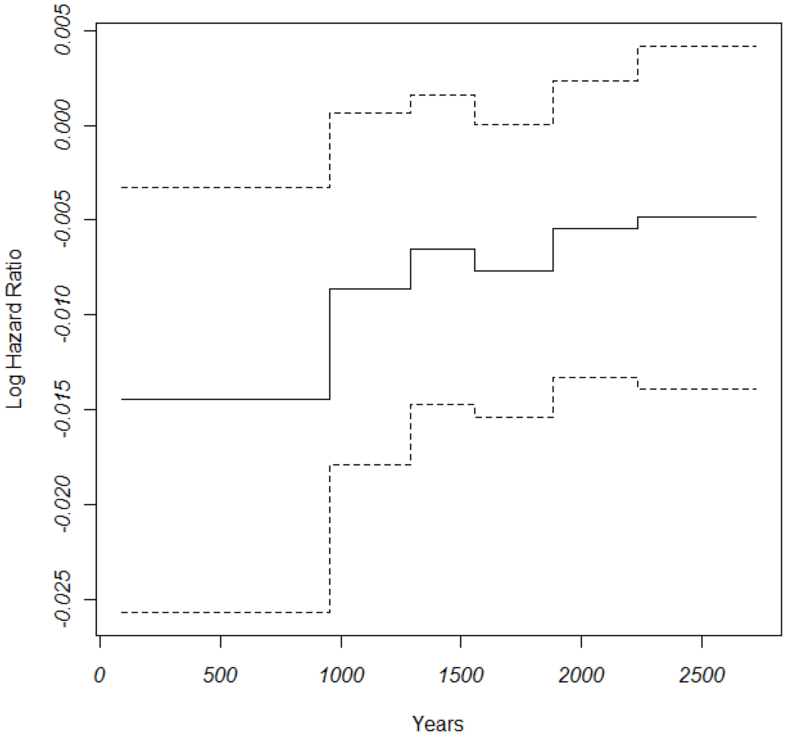

Figure 6. Beta-42 Log – Hazard Ratio over time with 95% CI

The estimates for the log hazard ratio for Beta-42 ranged from -0.01329 to -0.00396. This indicates that during the study, Beta-42 decreased the risk of harm of dementia. The final model required 7 knots to completely describe the movement of the log-hazard estimated. The final model had a fully penalized log-likelihood of -3807.47.

3.3.3 Beta Amyloid (42/40)

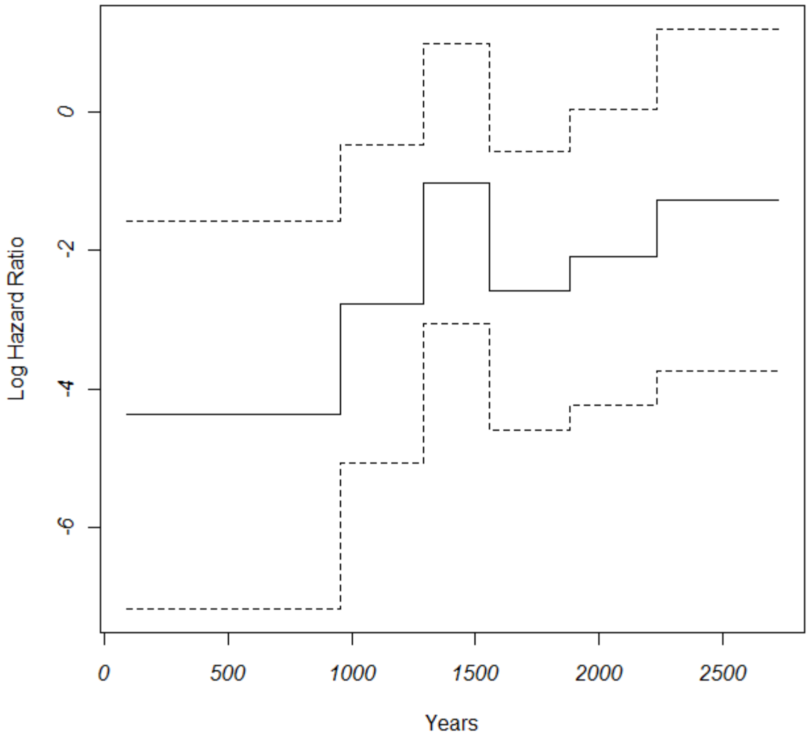

Figure 7. Beta ratio (42/40) Log – Hazard Ratio over time with 95% CI

The estimates for the log hazard ratio for the ratio Beta (42/40) ranged from -3.727 to -1.108. This indicates that during the study, Beta (42/40) decreased the risk of harm of dementia. The final model required 8 knots to completely describe the movement of the log-hazard estimated. The final model had a fully penalized log-likelihood of -3805.24.

3.3.4 Cystatin-C

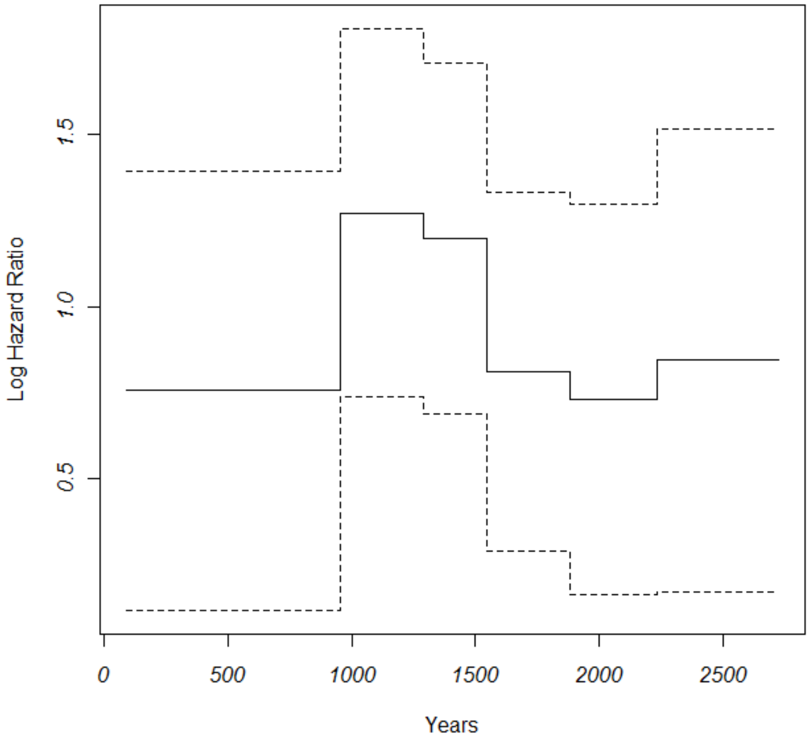

Figure 8. Cystatin-C Log – Hazard Ratio over time with 95% CI

The estimates for the log hazard ratio for Cystatin-C ranged from 0.732 to 1.273. This indicates that during the study, Cystatin-C increased the risk of harm of dementia. The final model required 5 knots to completely describe the movement of the log-hazard estimated. The final model had a fully penalized log-likelihood of -3850.79.

3.3.5 Biomarker Comparison

The range of values for each of the biological measures is shown in the table below. The table also includes the log-likelihood when parameters were set to zero, fully penalized log-likelihood and the difference between the minimum and maximum values (Δ):

Table 4. Biomarker Comparison

Variable	**Minimum**	**Maximum**	**Δ (Max – Min)**	**L(η)**	**L_p (η)**
Cystatin-C	0.7316	1.2730	0.5414	-3862.232	-3850.790
Aβ1-40	-0.0002901	0.0002785	0.000569	-3869.370	-3869.102
Aβ1-42	-0.013290	-0.003959	0.009331	-3812.614	-3807.471
Ratio (Aβ1-42/ Aβ1-40)	-3.727	-1.108	2.619	-3812.393	-3805.237

The largest change in value occurred in the ratio between the beta amyloid values.

4.0 DISCUSSION

4.1 CONCLUSIONS

4.1.1 Trajectory Analysis

The trajectory analyses resulted in each of the variables, ADL, IADL and 3MSE, breaking into three subgroups. For all variables, the three groups provided unique models and coefficients for each group. All coefficients were statistically significant, with the exception of the quadratic variable in IADL in group 2. That variable was kept in the model in order to reach convergence. The graphs portray three types of subjects: the healthy group, the intermediate group, and the group with the highest risk.

4.1.2 Flexible Cox Model

The flexible Cox model provided us with a range of log-hazard estimates for each of the variables of interest but when comparing the final, fully penalized log-likelihood for each of the models, there was little evidence to suggest that the additional information was statistically significant. In the beta ratio (42/40), which had the largest change in values, the initial log-likelihood was -3812.39 and the final model produced a log-likelihood of -3805.24. Even though there is no formal way to test the two log-likelihoods due the initial model not being nested in the final model, by looking at the two values, it can be seen that there was not a significant decrease in the final log-likelihood. Therefore, there was a minimal amount of information gained when allowing the hazard to vary over time when looking at biomarker data for dementia.

4.2 PUBLIC HEALTH ASPECT

As the population in the US continues to live longer, it is important to know how to analyze this subset of the population correctly. This thesis showcases two methods that can be used to address the issue of analyzing elderly Americans.

Trajectory analysis can be used to locate subgroups in the elderly population. Assuming that everyone ages the same way and run the same age-risks is naïve. By identifying the subgroups in the elderly population, we can target important medical treatments to people that really need them, improve longer-term health through preventative measures using the same identification process, and help create an overall, healthier older population.

Flexible spline analysis can be used to determine if the hazard of a disease changes over time. By identifying the times where hazard is at its highest, we can create time sensitive interventions that can help prevent or delay the disease from occurring. This again helps in creating a healthier elderly population.

APPENDIX A

VARIABLE SUMMARIES

Table 5. Comparison of dementia cases to non-cases at baseline from 2000 to 2008

	Remained non-demented	Incident dementia	χ^2/t-test	p-value
Number of subjects	N = 1966	N = 523		
Age	78.07 (2.97)	79.94 (3.63)	-10.89	**<.0001**
Education level	14.44 (2.88)	14.19 (3.23)	1.63	0.1041
Race, Whites	1886 (95.93%)	491 (93.88%)	4.35	**0.0445**
Gender, Female	897 (45.63%)	256 (48.95%)	1.83	0.1756
3MSE scores	94.12 (4.25)	90.67 (5.34)	13.69	**<.0001**
CES-D	3.25 (3.22)	4.58 (4.02)	-7.03	**<.0001**
APOE-4 allele	335 (20.44%)	145 (36.62%)	46.31	**<.0001**
Cerebrovascular risk factors				
Hypertension	818 (41.99%)	222 (42.94%)	0.1535	0.9261
Heart disease	519 (25.40%)	181 (34.61%)	13.77	**0.0002**
Diabetes mellitus	157 (8.09%)	49 (9.57%)	1.156	0.2822
Laboratory tests				
Cystatin-C	0.80 (0.19)	0.85 (0.22)	-4.15	**<.0001**
Creatinine	0.98 (0.23)	1.00 (0.26)	-1.15	0.2486
Vitamin B12	503.77 (251.35)	503.59 (243.55)	0.02	0.9880
TSH	2.26 (1.80)	2.16 (1.61)	1.23	0.2187
Plasma amyloid				
Aβ1-40	188.19 (97.79)	190.29 (67.04)	-0.97	**0.5681**
Aβ1-42	15.79 (28.40)	12.87 (15.84)	3.07	**0.0022**
Aβ1-42/Aβ1-40	0.08 (0.11)	0.07 (0.06)	2.74	**<.0001**

Table 6. Log – Hazard Ratio for Beta-40 over time

Time	Log – Hazard Ratio	Variance	Time	Log – Hazard Ratio	Variance
90	-0.0003	5.29e-07	1673	0.0002	2.80e-07
661	-0.0003	5.29e-07	1673	0.0003	2.87e-07
661	-0.00004	4.40e-07	1857	0.0003	2.87e-07
983	-0.00004	4.40e-07	1857	0.0001	3.09e-07
983	0.0002	3.63e-07	2021	0.0001	3.09e-07
1156	0.0002	3.630e-07	2021	-0.00002	3.55e-07
1156	0.0001	3.14e-07	2212	-0.00002	3.55e-07
1309	0.0001	3.14e-07	2212	-0.0002	4.34e-07
1309	0.00002	2.87e-07	2416	-0.0002	4.34e-07
1490	0.00002	2.87e-07	2416	-0.0002	5.24e-07
1490	0.0002	2.80e-07	2728	-0.0002	5.24e-07

Table 7. Log – Hazard Ratio for Beta-42 over time

Time	Log – Hazard Ratio	Variance	Time	Log – Hazard Ratio	Variance
90	-0.013	2.98e-05	1548	-0.010	1.43e-05
791	-0.013	2.98e-05	1837	-0.010	1.43e-05
791	-0.009	2.34e-05	1837	-0.007	1.38e-05
1132	-0.009	2.34e-05	2057	-0.007	1.38e-05
1132	-0.008	1.86e-05	2057	-0.004	1.47e-05
1314	-0.008	1.86e-05	2366	-0.004	1.47e-05
1314	-0.008	1.60e-05	2366	-0.004	1.93e-05
1548	-0.008	1.60e-05	2728	-0.004	1.93e-05

Table 8. Log – Hazard Ratio for Beta ratio (42/40) over time

Time	Log – Hazard Ratio	Variance	Time	Log – Hazard Ratio	Variance
90	-3.726734	1.8266167	1679	-2.318040	0.9265520
759	-3.726734	1.8266167	1679	-2.862105	0.9338653
759	-3.044987	1.4369813	1882	-2.862105	0.9338653
1112	-3.044987	1.4369813	1882	-2.532740	0.9974097
1112	-2.453121	1.1300149	2124	-2.532740	0.9974097
1289	-2.453121	1.1300149	2124	-1.173355	1.1894872
1289	-1.108119	0.9967553	2386	-1.173355	1.1894872
1481	-1.108119	0.9967553	2386	-1.643314	1.5120233
1481	-2.318040	0.9265520	2728	-1.643314	1.5120233

Table 9. Log – Hazard Ratio for Cystatin-C over time

Time	Log Hazard Ratio	Variance	Time	Log Hazard Ratio	Variance
90	0.7558062	0.10229205	1548	0.8115762	0.06836820
952	0.7558062	0.10229205	1881	0.8115762	0.06836820
952	1.2729696	0.07190890	1881	0.7315651	0.08059763
1291	1.2729696	0.07190890	2233	0.7315651	0.08059763
1291	1.1979169	0.06542322	2333	0.8443062	0.11362867
1548	1.1979169	0.06542322	2728	0.8443062	0.11362867

APPENDIX B

GLOSSARY OF ACRONYMS

3MSE – Modified Mini–Mental State Examination

ADL – Activities of Daily Living

IADL – Independent Activities of Daily Living

Beta-40 – plasma Amyloid Aβ1-40

Beta-42 - plasma Aβ1-42

BIC - Bayesian Information Criterion

CI – Confidence Interval

GEMS - Ginkgo Evaluation of Memory Study

MAR – Missing at Random

PROC TRAJ – Procedure Trajectory

APPENDIX C

SAS SOURCE CODE

```
data adl2;
set gems_p.adl_1;
keep IDNO visitno dldt dlsiadl dlsadl;
run;

data end2;
set gems_p.gemsend;
keep IDNO SSTATUSC VDT;
run;

proc sort data = adl2; by IDNO; run;
proc sort data = end2; by IDNO; run;

data measures;
merge adl2 end2;
by IDNO;
run;

data measures2;
set measures;
time_adl = dldt - VDT;
time_iadl = dldt - VDT;
run;

data Gems_Pneu;
set gems_p.Gems_pneu_time;
keep IDNO PNEUMONIA DAYSTOPNEUMONIA PNEUMONIATOENDPOINT;
run;

proc freq data = gems_pneu;
tables PNEUMONIA;
run;

proc sort data = gems_pneu; by idno; run;
proc sort data = gems_p.measures2; by idno; run;

data measures3;
```

```
merge gems_pneu gems_p.measures2;
by idno;
run;

data ADL_long_p2; *** Saved ***;
set  measures3;
mark = 2;
if PNEUMONIA = 1 and time_adl < DAYSTOPNEUMONIA then mark = 1;
if mark = 1 then adl_cen = DLSADL;
if mark = 2 then adl_cen = . ;
if mark = 1 then iadl_cen = DLSIADL;
if mark = 2 then iadl_cen = . ;
if time_adl < 0 and time_adl ne . then  time_adl = 0;
if time_iadl < 0 and time_iadl ne . then  time_iadl = 0;
if PNEUMONIA = 1 and mark = 2 then adl_cen_1 = DLSADL;
if PNEUMONIA = 1 and mark = 2 then iadl_cen_1 = DLSiADL;
if PNEUMONIA = 0 then adl_cen = DLSADL;
if PNEUMONIA = 0 then iadl_cen = DLSIADL;
run;

proc sort data = ADL_LONG_P2; by pneumonia; run;

proc means data= gems_p.Adl_long_p2 N nmiss mean std min max;
var dlsadl DLSIADL time_adl time_iadl;
run;


*** Wide form is ADL_wide_p2 ***;

** IADL **;

PROC TRAJ DATA=gems_p.Adl_wide_p2 OUT=OF OUTPLOT=OP OUTSTAT=OS CI95M;
VAR iadl_cen1 iadl_cen4 iadl_cen5 iadl_cen6 iadl_cen7 iadl_cen8 iadl_cen9
iadl_cen10 iadl_cen11 iadl_cen12
      iadl_cen13 iadl_cen14 iadl_cen15 iadl_cen16;
INDEP time_iadl1 time_iadl4 time_iadl5 time_iadl6 time_iadl7 time_iadl8
time_iadl9 time_iadl10 time_iadl11
      time_iadl12 time_iadl13 time_iadl14 time_iadl15 time_iadl16;
MODEL ZIP;
ORDER 2 2 3;
RUN;

%TRAJPLOTNEW (OP, OS,,,"IADL","Study Time");

** ADL **;

PROC TRAJ DATA=gems_p.Adl_wide_p2 OUT=OF OUTPLOT=OP OUTSTAT=OS CI95M;
VAR adl_cen1 adl_cen4 adl_cen5 adl_cen6 adl_cen7 adl_cen8 adl_cen9 adl_cen10
adl_cen11 adl_cen12
      adl_cen13 adl_cen14 adl_cen15 adl_cen16;
INDEP time_iadl1 time_iadl4 time_iadl5 time_iadl6 time_iadl7 time_iadl8
time_iadl9 time_iadl10 time_iadl11
      time_iadl12 time_iadl13 time_iadl14 time_iadl15 time_iadl16;
MODEL ZIP;
ORDER 1 1 2;
RUN;
```

```
%TRAJPLOTNEW (OP, OS,,,"ADL","Study Time");

proc means data = gems_p.adl_long_p2 N nmiss mean std min max;
var adl_cen iadl_cen time_iadl time_adl;
run;

*** MSE, starting with MSE3 ***;

proc means data = gems_p.mse3 N nmiss mean std min max;
var cfscore cfdt;
run;

data Gems_Pneu;
set gems_p.Gems_pneu_time;
keep IDNO PNEUMONIA DAYSTOPNEUMONIA PNEUMONIATOENDPOINT;
run;

proc sort data = gems_p.mse3; by idno; run;
proc sort data = Gems_Pneu; by idno; run;

*** Saved ***;
data mse4;
merge gems_pneu gems_p.mse3;
by idno;
run;

proc means data = gems_p.mse4 N nmiss mean std min max;
var cfscore cfdt DAYSTOPNEUMONIA;
run;

proc freq data = gems_p.mse4;
tables PNEUMONIA;
run;

*** Censored at Pneumonia ***;
data gems_p.mse4;
set  gems_p.mse4;
mark = 2;
if PNEUMONIA = 1 and time_mse < DAYSTOPNEUMONIA then mark = 1;
if mark = 1 then mse_cen = cfscore;
if mark = 2 then mse_cen = . ;
if PNEUMONIA = 1 and mark = 2 then mse_cen_1 = cfscore;
if PNEUMONIA = 0 then mse_cen = cfscore;
run;

proc sort data = gems_p.mse4; by PNEUMONIA; run;

*** Brought back in wide form as mse5_wide ***;

proc means data = gems_p.mse5_wide n nmiss mean std min max;
var time_mse1 time_mse3-time_mse10 time_mse11 - time_mse16 mse_cen1 mse_cen3
- mse_cen16 mse_cen_11 mse_cen_13 mse_cen_14 mse_cen_15
      mse_cen_16 mse_cen_17 mse_cen_18 mse_cen_19 mse_cen_110 mse_cen_111
mse_cen_112 mse_cen_113 mse_cen_114 mse_cen_115 mse_cen_116;
run;

*** Minimum of Time will be set to zero ***;
```

```
data gems_p.mse5_wide;
set gems_p.mse5_wide;
if time_mse1 < 0 and time_mse1 ne . then time_mse1 = 0;
run;

proc means data = gems_p.mse5_wide n nmiss mean std min max;
var time_mse1;
run;

PROC TRAJ DATA=gems_p.mse5_wide OUT=OF OUTPLOT=OP OUTSTAT=OS CI95M;
VAR mse_cen1 mse_cen3 - mse_cen16;
INDEP time_mse1 time_mse3-time_mse10 time_mse11 - time_mse16;
MODEL CNORM; min 0; max 100;
ORDER 2 2 2;
RUN;

%TRAJPLOTNEW (OP, OS,,,"3MSE","Study Time");
```

APPENDIX D

R SOURCE CODE

```
Code:
oscar1 <- read.csv(file="oscar6.csv",head=TRUE,sep=",")
sub1_oscar1 <- oscar1[, c(1, 10, 15, 29, 154, 157, 153)]
## ID, censor time, dementia status, Cystatin-C, Beta-40, Beta-42, ratio ##

u2 <- cox.spline("t",sub1_oscar1$CENSOR_T,sub1_oscar1$DEMENTI2 ,sub1_oscar1
[,4],nknot=5)

u2$est
summary(u2$est[,2])
plot(u2$est[,1], u2$est[,2], type="b", xlim = (c(0, 3000)), ylim = (c(0, 1.5)),
xlab="Time", ylab = "Spline Estimates", main="Spline Estimates Over Time for C-
cystatin")

### Summary for Cystatin-C ###
> summary(u2$est[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7316  0.7558  0.8279  0.9357  1.1980  1.2730

u2 <- cox.spline("t",sub1_oscar1$CENSOR_T,sub1_oscar1$DEMENTI2 ,sub1_oscar1
[,5],nknot=10)
summary(u2$est[,2])
plot(u2$est[,1], u2$est[,2], type="b", xlim = (c(0, 3000)), xlab="Time", ylab =

"Spline Estimates", main="Spline Estimates Over Time for Beta-40")
abline(h=0)

### Summary for Beta-40 ###
> summary(u2$est[,2])
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.901e-04 -1.300e-04  2.471e-05  2.247e-05  1.566e-04  2.785e-04

u2 <- cox.spline("t",sub1_oscar1$CENSOR_T,sub1_oscar1$DEMENTI2 ,sub1_oscar1

[,6],nknot=7)
summary(u2$est[,2])
plot(u2$est[,1],  u2$est[,2],  type="b",  xlim  =  (c(0,  3000)),  ylim  =  (c(-0.02,  0)),
xlab="Time", ylab = "Spline Estimates", main="Spline Estimates Over Time for Beta-42")

### Summary for Beta-42 ###
```

```
> summary(u2$est[,2])
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.013290 -0.009335 -0.008207 -0.007953 -0.006272 -0.003959

u2 <- cox.spline("t",sub1_oscar1$CENSOR_T,sub1_oscar1$DEMENTI2 ,sub1_oscar1

[,7],nknot=8)
summary(u2$est[,2])
plot(u2$est[,1], u2$est[,2], type="b", xlim = (c(0, 3000)), xlab="Time", ylab =

"Spline Estimates", main="Spline Estimates Over Time for Beta Ratio (42/40)")

### Summary for Ratio ###

> summary(u2$est[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -3.727  -2.862  -2.453  -2.318  -1.643  -1.108

> u2$loglik
[1] -3862.232 -3850.790 -3849.823
> u3$loglik
[1] -3869.370 -3869.102 -3868.933
> u4$loglik
[1] -3812.614 -3807.471 -3807.059
> u5$loglik
[1] -3812.393 -3805.237 -3804.016
```

BIBLIOGRAPHY

Beth E. Snitz, PhD and et al. "Ginkgo biloba for Preventing Cognitive Decline in Older Adults, A Randomized Trial." JAMA (2009): 2663 - 2270.

E.L. Lehmann, George Casella. Theory of Point Estimation. New York City: Springer-Verlag New York Inc., 1998.

Gray, Robert J. "Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis." Journal of the American Statistical Association (1992): 942-951.

Jones, Bobby L., et al. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories." SOCIOLOGICAL METHODS & RESEARCH (2001): 374-393.

Rajendran L, Honsho M, Zahn TR, Keller P, Geiger KD, Verkade P, Simons K. "Alzheimer's disease beta-amyloid peptides are released in association with exosomes." Proc Natl Acad Sci USA (2006): 11172-11177.

Steven T. DeKosky, M.D. Ginkgo Biloba Prevention Trial in Older Individuals. 1 November 2010. 17 April 2012 <http://clinicaltrials.gov/ct2/show/NCT00010803?term=NCT00010803&rank=1>.

Steven T. DeKosky, MD and et al. "Ginkgo biloba for Prevention of Dementia, A Randomized Controlled Trial." JAMA (2008): 2253 - 2262.

University of Maryland Medical Center. Ginkgo biloba. 13 December 2010. 17 April 2012 <http://www.umm.edu/altmed/articles/ginkgo-biloba-000247.htm>.