Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Jiayue Qiu April 12, 2020

An Exploration of the Association between Mild Cognitive Impairment Subgroups and Dementia Progression Time

by

Jiayue Qiu

Dr. John Hanfelt

Adviser

Department of Biology

Dr. John Hanfelt

Adviser

Dr. Felicia Goldstein

Committee Member

Dr. Nicole Vega

Committee Member

2020

An Exploration of the Association between Mild Cognitive Impairment Subgroups and Dementia Progression Time

By

Jiayue Qiu

Dr. John Hanfelt

Adviser

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Biology

2020

Abstract

An Exploration of the Association between Mild Cognitive Impairment Subgroups and Dementia Progression Time

By Jiayue Qiu

MCI patients have a wide range of clinical presentations. To understand the heterogeneity of MCI disorders and related diseases, researchers have used latent class analysis with non-cognitive features such as neuropsychiatric symptoms as indicators. Using the National Alzheimer's Coordinating Center's Uniform Data Set (UDS) with 6,034 participants, this research aims to establish more informative MCI subtypes with neuropsychiatric features and to show their distinct associations with progression time to dementia which might suggest distinct disease etiologies. Latent class analysis and a subsequent proportional hazards regression model were used to analyze this association. We used a weight-adjusted three-step approach to better account for the uncertainty issues of the latent class membership assignment. As a result, we found 4 latent classes with varied neuropsychiatric characteristics, including two classes characterized by either uniformly mild or more severe neuropsychiatric features, a cluster characterized by a combination of high depression, anxiety, and apathy and another cluster characterized by both high agitation and high irritability. The subsequent statistical results from the proportional hazards model provide estimates of different relationships between MCI subtypes and subsequent times to conversion to dementia. We found different hazard levels that associate certain neuropsychiatric features, such as irritability and agitation, with earlier risk of dementia compared to the others. We believe the statistical results from this research may aid in the early recognition of dementia in a clinical setting.

An Exploration of the Association between Mild Cognitive Impairment Subgroups and Dementia Progression Time

By

Jiayue Qiu

Dr. John Hanfelt

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2020

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. John Hanfelt for his mentorship and continuous support and guidance throughout this research. His constant technical and editorial advice was essential for the completion of this thesis.

I would like to thank Dr. Felicia Goldstein who provided many valuable comments from clinical perspectives that helped improve this thesis. I am grateful for her willingness to share her clinical expertise and to discuss the possible clinical implications of the project.

I also thank Dr. Nicole Vega for her support as my thesis committee member. I am grateful for her thoughtful comments and questions which aided me in the process of writing and editing this thesis.

I would also like to thank the National Alzheimer's Coordinating Center for providing the data used in this research. I thank the ADC participants for their willingness to devote their time to research. I very much appreciate the effort of the ADC staff members who worked tirelessly to make the research viable.

Table of Contents

List of Figures and Tables

Figures

Tables

Introduction

1.1 Opportunities and Challenges in Early Recognition of Symptomatic Dementia

Alzheimer's disease and related diseases have raised significant social awareness and concerns in modern American society. Alzheimer's disease accounts for 60-80% of dementia cases. The current dementia diagnosis depends upon a physician's judgements based on a patient's medical history and a series of physical tests and examinations. Dementia is usually diagnosed with uncertainty and especially uncertainty to its underlying etiology. Being able to categorize mild cognitive impairment into subtypes and associate them with dementia provides insights into early recognition and diagnosis of dementia.

Early recognition of symptomatic dementia is considered vital to the effective treatment and care of dementia patients. Research has been conducted in various medical-related fields to help aid in its early diagnosis. For example, researchers are seeking practical imaging markers for Alzheimer's Disease, and companies have been developing screening instruments for dementia patients (Thyrian, 2018).

This research aims to explore a possible statistical association between latent classes of MCI with neuropsychiatric features and progression time to dementia. The statistic model developed would not only expand the classification of MCI but also suggest potential distinct disease etiologies of dementia. This statistical association may also help shed new light on the relationship between the presence of specific features with various progression path to dementia. And thus, the model may aid in future diagnosis tool developed to estimate the progression time to dementia based on patient symptoms and which MCI subclass they fall in.

1.2 Mild Cognitive Impairment Subtypes

Mild cognitive impairment (MCI) is defined as the intermediate stage between normal aging and dementia (Gauthier, 2006), which is a series of progressive conditions owing to brain degeneration, including dementia due to Alzheimer's disease, frontotemporal dementia, and vascular dementia. MCI patients have a wide range of clinical presentations. Researchers have made general agreement to divide MCI into 4 subtypes in 2004: amnestic (memory-impairment only), multidomain-amnestic (multiple domains including memory), multidomain-nonamnestic (multiple nonmemory domains), and single nonamnestic (Winblad, 2004). This classification was mostly based on cognitive assessments. It is increasingly recognized, however, that non-cognitive characteristics are also insightful for understanding the heterogeneity of MCI and the underlying neurodegenerative etiologies.

In previous research by Dr. John Hanfelt et al. (2011), the subclasses were further expanded and explored incorporating neuropsychiatric features such as depression, irritability, and apathy, as well as functional status in a sample of 1,655 MCI patients. Subsequent research has shown that MCI subgroups with neuropsychiatric and functional features are more likely to develop not only a "pure" form of Alzheimer's disease pathology but also a "mixed" pathology consisting of both Alzheimer's disease and cerebrovascular diseases (Hanfelt et al., 2018). A critical limitation of these previous studies is that they only partially incorporated information about neuropsychiatric features in the data analysis. In this research, we will conduct a more in-depth study using a larger data set of 6,034 patients to establish more informative MCI subtypes with neuropsychiatric features that might suggest distinct disease etiologies.

1.3 Subgrouping and Latent Class Analysis

Subgrouping of MCI patients in this research is accomplished via a statistical method named Latent Class Analysis (LCA). LCA is essentially the classification of similar objects into groups, in which the number of groups, as well as their forms, are unknown, where forms are defined as the parameters specific to each cluster/class such as means, variances, and such (Vermunt, Magidson). In this research, we use LCA to analyze the neuropsychiatric features of MCI patients. Patients belonging to the same latent class are similar in terms of these observed variables, and their observed scores are assumed to be from the same probability distributions.

$$f(y_i|\theta) = \Sigma_{k=1}^{K} \pi_k f_k(y_i|\theta)$$

Above shows the most basic LC cluster model, where $y_i's$ are indicators, K is the number of clusters, θ is the model parameter, and π_k is the probability of belonging to latent class k. Yet this basic model only incorporates a single indicator variable. In our case, we use a more generalized model for nominal variables:

$$f(y_i|\theta) = \Sigma_{k=1}^{K} \pi_k \Pi_{j=1}^{J} f_k\left(y_{ij}\middle|\theta_{jk}\right)$$

In this model, J is the total number of indicators and j, a particular indicator. This model allows each indicator to take on its proper univariate distribution function instead of being forced into one multivariate distribution function (Vermunt, Magidson).

1.4 Survival Analysis and Estimating Time of Progression to Dementia

Survival analysis is a statistical methodology used to assess the relationships between baseline covariates and the expected time duration to an event such as the death of an organism. In this case, we are trying to estimate a survival model that relates the MCI latent classes with the time duration

until patients progress into dementia. This methodology accommodates right-censored observations, where not every patient progresses into dementia by the end of the study. If the patient is undemented, we record the length of their follow-up time.

It is often helpful to first understand the distribution of the outcome variable which records survival times (*Time*). A probability distribution function (pdf or *f(t)*) of *Time* is defined as the probability of observing *Time* at a specific time *t* among all the other possible survival times. A cumulative distribution function of *Time*, (cdf or *F(t)*) describes the probability of observing *Time* at and before a specific time *t*. It is obtained by integrating the pdf over the range from 0 to *t*:

$$F(t) = \int_0^t f(t)dt$$

However, the exact distribution of survival times is usually not known prior to the analysis. Yet we can estimate the pdf using various methods including nonparametric methods in SAS with proc univariate (UCLA: Statistical Consulting Group, 2016).

While cumulative density function *F(t)* describes the probability of observing survival time at or before a specific time *t*, the cumulative survival function *S(t)* describes the probability of observing a survival time past time *t*, given by

$$S(t) = 1 - F(t)$$

The primary focus of survival analysis is to estimate the hazard rate (*h(t)*), which is calculating using a relationship between pdf and survival function:

$$h(t) = \frac{f(t)}{S(t)}$$

Hazard function or hazard rate (*h(t)*) thus calculates the instant probability of the outcome occurring at time *t* (*f(t)*) given that the individual has survived up to time *t* (*S(t)*). It is the instant probability of failure at a certain time *t.*

It is also useful to describe cumulative hazards. The cumulative hazard function (*H(t)*) is calculated by integrating hazard functions over time. It describes how much risk has accumulated up to time *t.*

$$H(t) = \int_0^t h(t)dt$$

We use the Cox proportional hazards regression model to estimate the relationship between latent classes and dementia progression time. In a proportional hazard model, it is assumed that the covariates, which are the predictor or independent variables, relate multiplicatively with the hazards. In this case, we examine how being in one MCI latent class may have a multiplicative effect on the hazards compared to the baseline covariate which is the subgroup that we set as the baseline level. An example could be that being in a more severe MCI latent class may double the hazard of being demented compared to the baseline normal group.

The hazard function of a Cox proportional hazards model has the form:

$$h(t) = h_0(t)\exp(x\beta_x)$$

The model consists of a baseline hazard function ($h_0(t)$) which describes how the risks change over the course of time at a baseline level. It also estimates effect parameters (β_x) that describe each covariate's (x) effect on the hazard rate. The exponential function allows the hazard rate to equal the baseline hazard rate when all covariates are equal to zeros.

The ratio of the hazard rates between the two groups are assumed to stay constant over time when the covariates are fixed. This constant ratio is called the hazard ratio (HR). For a covariate x:

$$HR = \frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)\exp\,(x_2\beta_x)}{h_0(t)\exp\,(x_1\beta_x)} = \exp\,(\beta_x(x_2 - x_1))$$

The covariate effect of x is thus a constant independent of time. The proportional hazards model, therefore, give us the advantage to estimate the predictor effect without specifying the baseline hazard rate $(h_0(t))$ (since the covariate effect is independent of the baseline hazard rate (UCLA: Statistical Consulting Group, 2016).

1.5 Three-Step Approach to Relate External Variable with Latent Classes

We consider two possible ways to relate the latent classes to an external variable (e.g. distal outcome), which are the traditional three-step approach and an adjusted three-step approach that accounts for the uncertainty in latent class assignments. The three-step approach, as its name suggests, relates an external variable with latent classes in three steps illustrated below in Fig. 1. First, the latent classes (Y) are established based on input variable information (X); second, latent class membership (W) is assigned to each subject based on his/her variable scores; third, the predicted class membership variable (W) is used for analyzing the relationship between latent classes and external variable Z (Bakk, 2018). Researchers have pointed out possible bias arising from this approach, owing to uncertainties in the class membership assignments (W), potentially causing the relationship between latent classes and the external variable to be underestimated (Block et. al., 2004). Thus, in this research, we will also design and implement improvements of the three-step approach to estimate the association between MCI latent classes and dementia progression time.

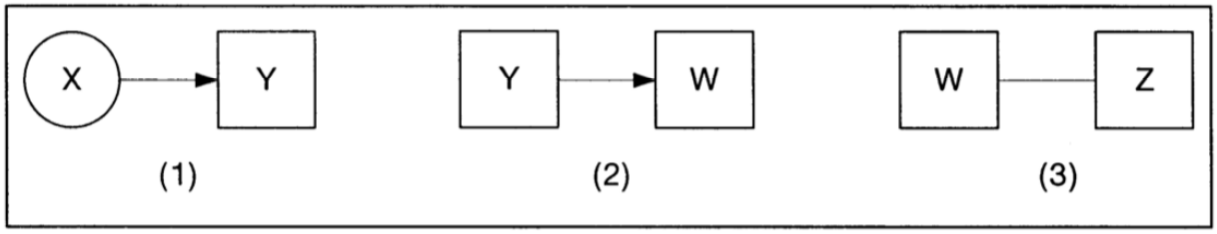

Fig. 1. Graphical depiction of the three-step approach (Bakk, 2018)

Methods

2.1 Participants

The data set used in this research is from the Uniform Data Set (UDS), a standardized assessment and data protocol maintained by the National Alzheimer's Coordinating Center with 29 NIH-supported Alzheimer's Disease Center nationwide (e.g., Hanfelt et.al., 2011). The data set contains a common set of clinical observations collected longitudinally on the ADC participants until December 2019. Data from 6,034 participants with MCI were available for this analysis.

2.2 Measures

Neuropsychiatric features were evaluated using the Neuropsychiatric Inventory Questionnaire, which assesses 12 behaviors including delusions, hallucinations, agitation/aggression, depression, apathy, elation, anxiety, disinhibition, irritability, aberrant motor behavior, nighttime behaviors, and appetite/eating. Each neuropsychiatric feature is scored , where 1 = present and 0 = absent.

Each participant's Mini-Mental State Exam (MMSE) score was obtained. The MMSE is one of the most commonly used mental status tests. During an MMSE, participants are asked a series of questions that assess a range of everyday mental skills. The maximum MMSE score is 30 points.

A score of 20 to 24 suggests mild dementia, 13 to 20 suggests moderate dementia, and less than 13 suggests severe dementia (Pangman, 2000). In our data set, the MMSE baseline scores range from 22 to 30. Time to progression to dementia was assessed by clinical experts at each center.

2.3 Analysis

2.3.1 Latent Class Analysis

The twelve neuropsychiatric variables (each scored 1=present, 0=absent) were entered into the LCA model for analysis. As is standard in LCA, variables were assumed to be independent given the latent classes. Missing-valued cases were retained on the assumption of missing at random, i.e. the missingness mechanism was assumed to be independent of the patient's missing variable scores. Model parameters were estimated by maximum likelihood.

The data set was fitted with various number of latent classes. An objective model selection criteria, Bayesian Information Criterion (BIC), was used to select the best fit model. The model with a minimized BIC value was selected as the best-fitting model. LCA generates a modal/latent class assignment for each individual in the data set as well as the probabilities of the individual belonging to each possible latent class. The software package used in this research for cluster analysis was LatentGold 5.0.

2.3.2 Adjustment to Variables

Before running the proportional hazards model, we formatted the age variable into the age in decades at baseline centered at age 75, i.e. BASEAGE75 = (Age-75)/10. Centering age at 75 and

compacting its range into units of decades help generate more meaningful intercept and parameter interpretations in the following regression model.

2.3.3 The Cox Proportional Hazards Regression Model

In the Cox proportional hazards regression model, the dependent variable was time (in years) from baseline until the patient progressed to dementia (data set label: clin_event_time), or if right-censored the total follow-up time. This dependent variable was regressed on 5 baseline predictor variables, including the cluster assignment of the individual (clu_), gender (female), education level in years (baseeduc), age (BASEAGE75), and MMSE score (BASEMMSE). We used PROC PHREG in SAS to fit the Cox model. Below is the SAS code:

```
proc phreg data=newdata plots(overlay)=(survival);
class clu_(REF="Mild");
model clin_event_time*clin_event_type(0) = clu_ female baseage75
basemmse baseeduc;
baseline covariates=covs out=base /rowid=clu_;
hazardratio clu_ / DIFF = REF ALPHA = 0.05 CL = WALD;
run;
```

In the MODEL statement, the response variable (clin_event_time) is crossed with censoring variable clin_event_type which recorded “0” if the patient was undemented as of December 2019. The values of clin_event_time are considered censored if the value of clin_event_type is “0”, in which case the survival time is essentially follow up time up until the patient’s most recent visit to the clinic. Otherwise, the survival time values are considered event times or progression time to dementia.

The CLASS statement specifies the reference group. We chose cluster 1 or the “Mild” group as the baseline for the model since they show overall mild symptoms and are considered the least dementia-inclined normal group. The BASELINE statement that specifies covariate values are used to generate survival plots by the cluster group.

The HAZARDRATIO statement is used to obtain hazard ratios of different cluster groups against the baseline reference “Mild” group. We chose to report a 95% confidence interval for the hazard ratio and thus specified ALPHA at 0.05. We also specified that WALD confidence limits are desired in the CL option.

In this study, p-values less than 0.05 in the Cox proportional hazards analysis were regarded as statistically significant.

2.3.4 Adjustment to the traditional three-step approach

As mentioned in the introduction, the estimates characterizing the relationship between the external variable and the latent class membership from the traditional three-step approach will always be smaller than those characterizing relationships between the input variables and the final external variable (Bolck et.al., 2004). Moreover, there might be uncertainty associated with using a clean-cut modal assignment for the regression analysis since we are not sure that a particular individual 100% belongs to one cluster than the others. To reduce the effect of this uncertainty, we adjusted the relationship between dementia progression time and cluster membership for the class weights or the probability of being assigned to one cluster but not the others.

The correction method is proved viable by showing the relationship between the X-Z distribution (the relationship between input variables and external variable) and the W-Z distribution (the relationship between latent class membership and external variable) in Figure 1. Figure 2 is another way of showing the relationship between variables W, X, Y and Z in Figure 1.

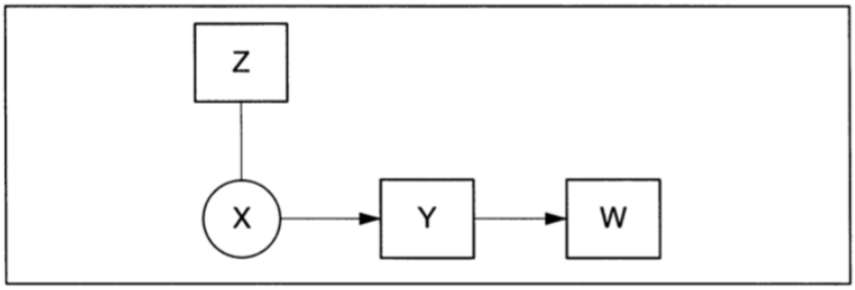

Fig. 2. Relationship between variables W, X, Y, and Z in the three-step approach (Bakk, 2018)

According to Bakk, the marginal distribution of W and Z can be finally derived to the following equation:

$$P(W = s, Z = z) = \sum_{t} P(X = t, Z = z)P(W = s|X = t)$$

Therefore, the W-Z distribution, $P(W = s, Z = z)$, is the weighted sums of the X-Z distribution, $P(X = t, Z = z)$. And the weights are the misclassification probabilities $P(W = s|X = t)$ that quantifies the probability of a certain class assignment conditional to the true class. In our case, we assumed that the modal assignment from Latent Class Analysis to be the true class (t) so s=t in this case. The misclassification probabilities thus equal the probabilities of one individual belonging to the assigned latent class from the LCA. And by using this probability as a weight, we adjusted the relationship between latent classes and the dementia progression time. This method may, therefore, better account the classification errors. Adjustment is accomplished by adding a WEIGHT statement in the PROC PHREG procedure in SAS.

Results

3.1 Results from Latent Class Analysis

Twelve Neuropsychiatric features were used as indicators for latent classes. Models with up to 5 maximum classes have been used to fit the data. The 4-cluster model is selected as the best-fitting model due to the lowest BIC(LL) value (showed in Table 1).

Table 1. 12 Features Model BIC Value Overview

Cluster Models	**BIC (LL)**
1 - Class	53313.2354
2 - Class	48282.4518
3 - Class	47887.4852
4 - Class	47736.7968
5 - Class	47766.5676

Table 2. Proportions of presence of the 12 neuropsychiatric features in the 4 latent classes

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster Size	0.5039	0.2633	0.1417	0.0911
Indicators				
DEL				
0	0.9955	0.9764	0.9321	0.7969
1	0.0045	0.0236	0.0679	0.2031
HALL				
0	0.9970	0.9797	0.9902	0.9130
1	0.0030	0.0203	0.0098	0.0870
AGIT				
0	0.9873	0.9257	0.3392	0.3097
1	0.0127	0.0743	0.6608	0.6903
DEPD				
0	0.9351	0.3920	0.6580	0.2014
1	0.0649	0.6080	0.3420	0.7986
ANX				
0	0.9482	0.5747	0.7356	0.3048
1	0.0518	0.4253	0.2644	0.6952
ELAT				
0	0.9958	0.9930	0.9660	0.8569
1	0.0042	0.0070	0.0340	0.1431
APA				
0	0.9762	0.6322	0.7459	0.3502
1	0.0238	0.3678	0.2541	0.6498
DISN				
0	0.9892	0.9326	0.8016	0.4955
1	0.0108	0.0674	0.1984	0.5045
IRR				
0	0.9295	0.7126	0.2780	0.1535
1	0.0705	0.2874	0.7220	0.8465
MOT				
0	0.9948	0.9428	0.9291	0.7063
1	0.0052	0.0572	0.0709	0.2937
NITE				
0	0.9231	0.6209	0.7318	0.3797
1	0.0769	0.3791	0.2682	0.6203
APP				
0	0.9680	0.7699	0.8467	0.5826
1	0.0320	0.2301	0.1533	0.4174

Table 2 shows more details about the 4 clusters. As seen in the table, the clusters exhibited varied characteristics in terms of specific neuropsychiatric features.

For each neuropsychiatric feature, 0 denotes the absence of symptoms and 1 denotes the presence of symptoms. The first row of Table 2 shows cluster sizes (i.e., proportions of the overall sample of 6034 participants), with cluster 1 being the largest group and cluster 4 being the smallest. Within cluster 1, members are mostly absent any of the neuropsychiatric features, i.e. probability of 0s is more than 90% for all features. However, for cluster 4, the probability of having 0s or absences of symptoms is the lowest among the 4 clusters for each neuropsychiatric feature; hence, cluster 4 represents a class with the uniformly highest rates of neuropsychiatric features.

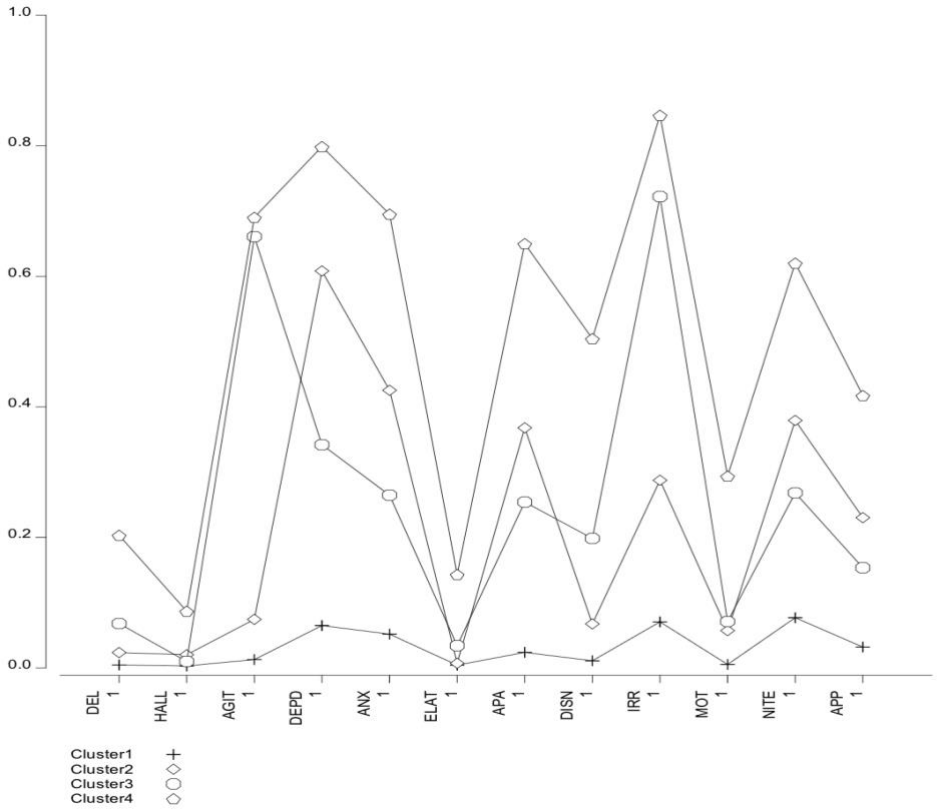

Fig. 3 Plot summary of the 4-cluster model

Figure 3 is a visual plot summary of the 4 clusters model. It shows the proportion of symptom presence(y-axis) for each of the 12 neuropsychiatric features (x-axis) across the 4 latent classes. The graph aligns with results showed in Table 2 where cluster 1 (showed in the cross) lies at the bottom of the graph and cluster 4 (showed in pentagons) tops all the three other clusters. It is also

worth noticing that cluster 2 is characterized by a relatively high proportion of depression, anxiety, and apathy symptoms and cluster 3 is characterized by relatively high agitation and irritability. Therefore, we renamed the clusters with their respective characteristics to better label them: we name cluster 1 “Mild” group, cluster 2 “DEPR+ANX+APA” group, cluster 3 “AGIT+IRR” group, and cluster 4 “Severe” group respectively.

3.2 Summary statistics for the predictor variables

Table 3. Summary of the frequencies of sex and dementia from PROC FREQ procedure

The FREQ Procedure

female	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3007	49.83	3007	49.83
1	3027	50.17	6034	100.00

clin_event_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3932	65.17	3932	65.17
1	2101	34.83	6033	100.00

The above is the frequency tables of gender and dementia diagnosis. Among the participants, 3007 of them are male and 3027 are female. Approximately 65% of the participants are undemented (clin_event_type =0) and are subject to censoring while 2101 participants (35%) were diagnosed with dementia.

Table 4. Summary of the mean values of predictor and response variables from PROC MEANS procedure

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
basemmse	6034	27.2999669	2.0578836	22.0000000	30.0000000
baseage75	6034	-0.1025269	0.9365487	-5.3149213	2.9476386
followup_time	6033	3.4243605	3.2632324	0	13.9166667
clin_event_time	6033	2.6537792	2.8478801	0	13.9166667
baseeduc	6034	15.1249586	3.2615115	0	25.0000000

The above table summarizes the means and ranges of predictor variables MMSE score (basemmse), age (baseage75), and education level (baseeduc). It can be inferred that the participants in this study have a relatively high education level with a mean of 15 years of education. The table also gives the mean of overall follow up time (followup_time) as 3.4 years, taking into account both demented and undemented participants, and the mean of the censored time until dementia (clin_event_time) as 2.7 years. As expected, the censored time until dementia is shorter than the overall follow up time since participants may continue to be followed in the clinics after being diagnosed with dementia.

3.3 Distribution of event time

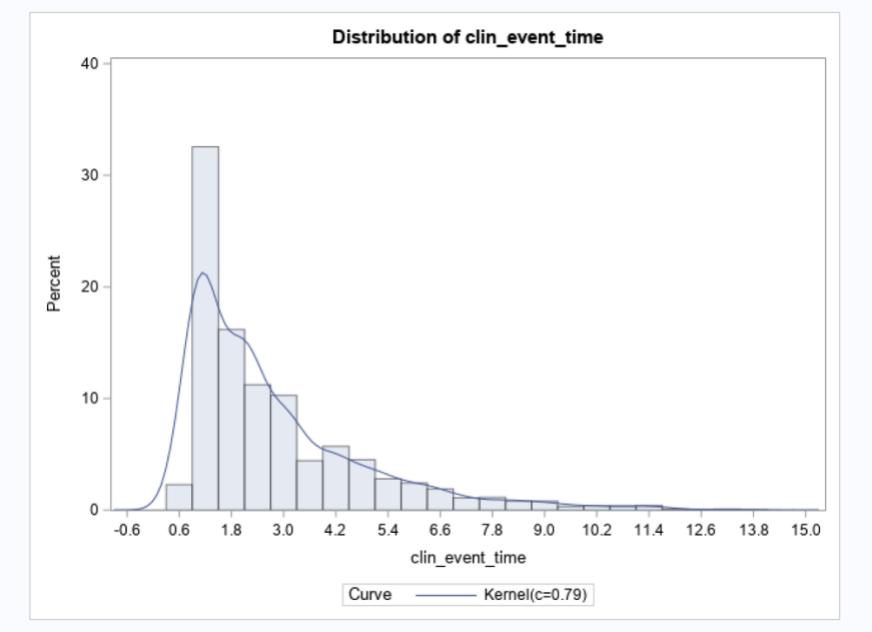

Fig. 4 Conditional Pdf of dementia progression time given progression to dementia by December 2019

Figure 4 plots the conditional distribution of dementia progression time (clin_event_time when clin_event_type = 1) among those patients who were not censored. We can see that the dementia progression time is slightly left-skewed, centering around the time of 2 years. Most of the participants were diagnosed with dementia after 1 year of visiting the clinics. Since this plot omits participants who did not covert to dementia by December 2019, it is important to note that the distribution shown in the plot is not representative of all the MCI patients, but only the faster converting ones.

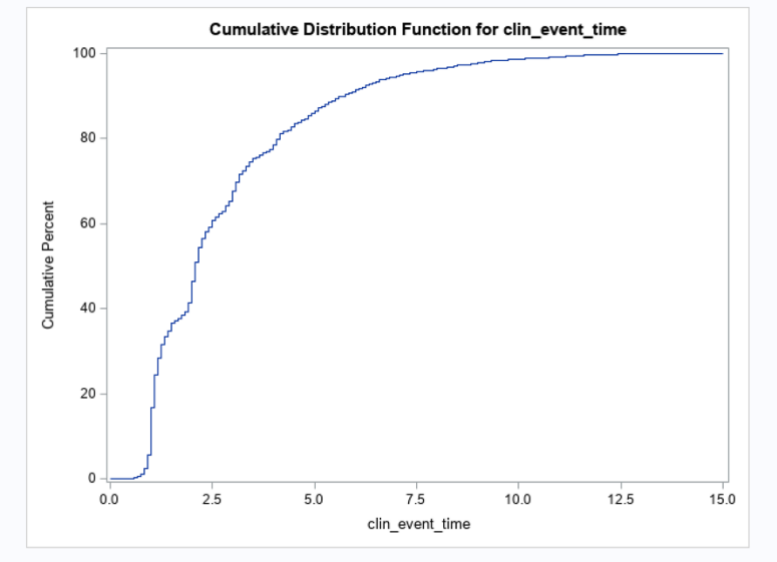

Fig. 5 Conditional Cdf of dementia progression time given progression to dementia by December 2019.

Figure 5 shows the conditional cumulative distribution function of dementia progression time among those patients who were not censored. We see that by around 2.5 years, the participants have already accumulated half of the risks of becoming demented.

3.4 Latent classes, age, MMSE score, and education are statistically significant in predicting dementia progression

Table 5. Summary of maximum likelihood estimates analysis from PROC PHREG using our adjusted 3-step approach and the first class ("Mild") as the reference class.

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
clu_	Agit+Irr	1	0.58423	0.07971	53.7236	<.0001	1.794	Cluster modal Agit+Irr
clu_	Depr+Anx+Apa	1	0.51001	0.05904	74.6158	<.0001	1.665	Cluster modal Depr+Anx+Apa
clu_	Severe	1	0.79800	0.08912	80.1801	<.0001	2.221	Cluster modal Severe
female		1	0.06067	0.05043	1.4473	0.2290	1.063	
baseage75		1	0.23396	0.02879	66.0578	<.0001	1.264	
basemmse		1	-0.22423	0.01190	355.0304	<.0001	0.799	
baseeduc		1	0.05011	0.00812	38.0442	<.0001	1.051	

Table 5 above summarizes the results of the maximum likelihood estimate analysis for the Cox proportional hazards regression model. All the predictors, except for gender (labeled as "female" in data set) were statistically significant to the prediction of dementia progression time with a p-value of less than 0.0001.

According to Table 5, both age (labeled as "baseage75") and education level ("baseeduc") increase the risk of dementing. Each unit increase in "baseage75" (i.e., each decade increase in age) led to a 1.264 fold increase in the hazards of dementing comparing to baseline. Each additional year of education also led to a 1.051 fold increase in the hazards of dementing comparing to baseline.

Surprisingly, the MMSE scores were inversely related to dementia progression. Since the parameter estimate is a negative value of -0.22423, yielding a hazard ratio less than one (0.799), an increase in MMSE scores decreases the hazards of dementing compared to the baseline.

As noted, gender was not a statistically significant predictor for this model, since its p-value, 0.2290, is rather large. Therefore, the results show that females and males don't seem to have separated progression paths for dementia and gender is not a factor of whether dementia progresses for an individual, once one takes into account age, education, MMSE and neuropsychiatric classification.

3.5 Dementia progression patterns for the 4 latent classes

Table 6. 95% hazard ratio confidence intervals using our adjusted 3-step approach and the first class ("Mild") as the reference class. Results are adjusted for age, education, gender, and MMSE score at baseline.

Hazard Ratios for Cluster modal			
Description	Point Estimate	95% Wald Confidence Limits	
clu_ Agit+Irr vs Mild	1.794	1.534	2.097
clu_ Depr+Anx+Apa vs Mild	1.665	1.483	1.870
clu_ Severe vs Mild	2.221	1.865	2.645

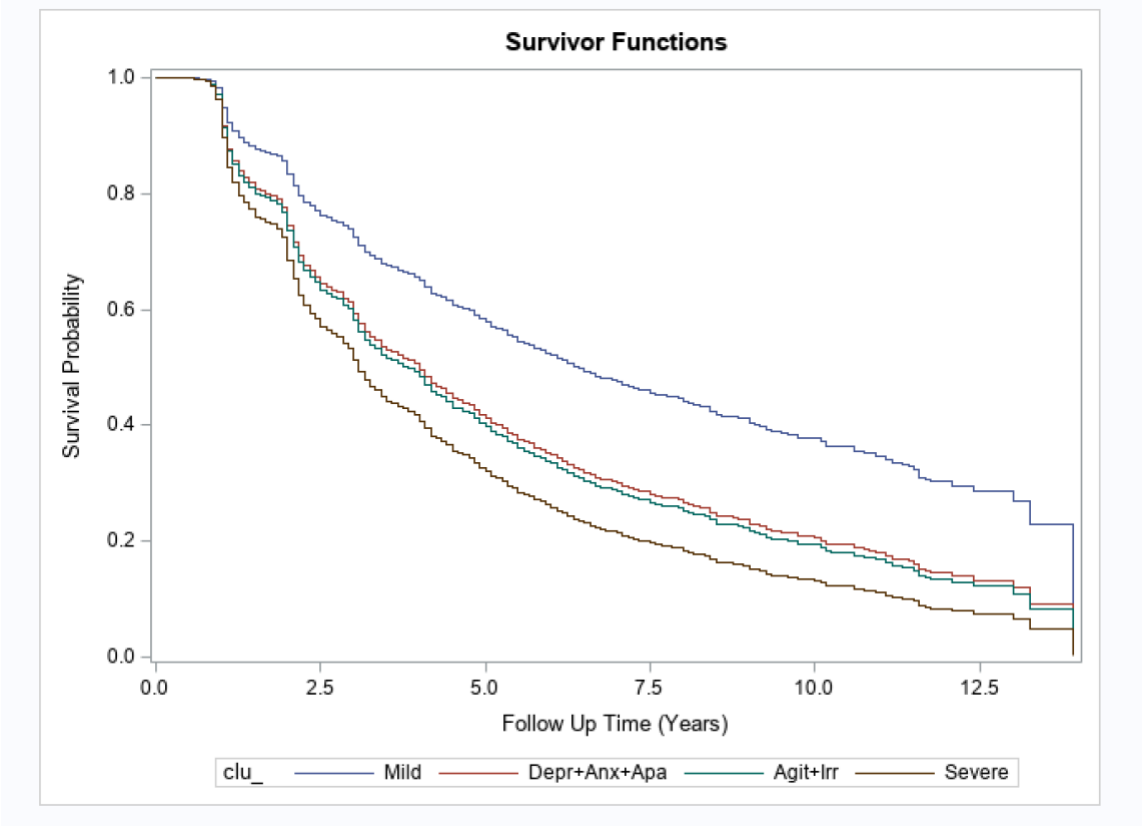

Fig. 6 Plot of the class-specific probability of being dementia-free versus follow-up time by December 2019

Figure 6 shows the relationship between the probability of being undemented and the 4 latent classes across the follow-up times. The "Mild" group has the highest probability of being undemented and healthy and is chosen as the baseline reference group. The rest of the three latent classes have a varied multiplicative effect that increases the hazards and thus lowers the probability of being undemented proportionally.

According to Table 6, the "Severe" group had the highest hazard ratio of 2.221 compared to the "Mild" baseline group. Therefore, in Figure 4, the "Severe" group, shown as the brown line, had the lowest probability of being undemented over time. The "Agit+Irr" group increases the hazards by 1.794 fold and the "Depr+Anx+Apa" group increases the hazards by 1.665 fold, thus they also

have a lower probability of being undemented compared to the "Mild" group but not as low as the "Severe" group.

Comparing the "Agit+Irr" group and the "Depr+Anx+Apa" group, we can conclude from the results that there is at least weak evidence based on this proportional hazards model that agitation and irritability are slightly more indicative neuropsychiatric features for risks of dementing compared to depression, anxiety, and apathy. This evidence is not very strong, however, since the 95% confidence intervals for the hazard ratios of these two groups overlapped considerably (Agit+Err group, 1.53 to 2.10; Depr+Anx+Apa group, 1.48 to 1.87) (Table 6).

3.6 Comparing the results between the traditional and the adjusted three-step approaches

Table 7. 95% hazard ratio confidence intervals from the traditional three-step approach

Hazard Ratios for Cluster modal			
Description	Point Estimate	95% Wald Confidence Limits	
clu_ Agit+Irr vs Mild	1.755	1.531	2.011
clu_ Depr+Anx+Apa vs Mild	1.577	1.424	1.746
clu_ Severe vs Mild	2.102	1.798	2.456

Comparing the traditional three-step approach and our adjusted three-step approach, our results on hazard ratios support the claim, by Bolck et al., that the traditional approach underestimates the relationship between latent classes and distal outcomes. As shown in Table 7, the hazard ratios for each cluster against the baseline "Mild" group from the traditional three-step approach are always smaller compared to the results in Table 6 from the adjusted three-step approach. This shows that the effect of latent classes on dementia progression hazards may be underestimated using the traditional three-step approach. Our adjusted three-step approach may be better at accounting for uncertainties of the class assignment and thus yields less biased estimates.

The final two tables below further support our adjustment to the three-step approach. We can see that all the model fit statistics of the adjusted three-step model are lower than those of the traditional three-step model. This indicates that the adjusted three-step approach generates a better fit model for the data compared to the traditional three-step approach.

Table 8. Model fit statistics for the adjusted three-step approach

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	25135.619	24567.481
AIC	25135.619	24581.481
SBC	25135.619	24621.032

Table 9. Model fit statistics for the traditional three-step approach

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	32366.756	31685.822
AIC	32366.756	31699.822
SBC	32366.756	31739.373

Discussion

We performed a more in-depth latent class analysis to understand the heterogeneity of MCI. The previous study explored MCI latent classes with neuropsychiatric features based on solely the total number, not the type, of neuropsychiatric features (e.g. Hanfelt et.al., 2011). This thesis presents finer-grained characteristics of each latent class, especially a cluster characterized by a combination of high depression, anxiety, and apathy and another cluster characterized by both high

agitation and high irritability. These distinct characteristics between MCI latent classes might suggest distinct disease etiologies that affect the brain cognitive functions in different areas. Since this thesis mainly considers the effect of latent classes on whether the patient progress to dementia, we did not explore the various types of dementia etiologies such as frontotemporal dementia, vascular dementia, and Alzheimer's disease, etc. Further competing outcome analysis can be conducted to better understand latent classes' effect on distinct dementia etiologies.

The results of this thesis show a statistical relationship between neuropsychiatric features and dementia. It is increasingly recognized that neuropsychiatric symptoms of dementia have a heterogeneous clinical presentation and thus shall not be treated as a collective syndrome (Phan et. al., 2019). The latent class analysis thus supports this heterogeneous presentation. Also, statistical results from the proportional hazards model in the thesis provide estimate of this different relationships between neuropsychiatric features and dementia. We found different hazard levels that associate certain neuropsychiatric features, such as irritability and agitation, more with dementia compared to the others.

We believe that a possible clinical implication of our findings concerns early recognition of dementia. Neuropsychiatric features are core features that are currently known to manifest in early and prodromal phases of Alzheimer's disease and related diseases (Lyketsos et.al., 2011). Therefore, being able to recognize MCI subtypes with neuropsychiatric features may aid clinicians in diagnosing dementia, modifying therapies for Alzheimer's disease and related diseases, as well as possibly estimating the progression time to dementia based on the patient's specific clinical presentation of neuropsychiatric symptoms.

In this thesis, we used a larger data set with 6,034 participants with longer follow up time up to December 2019, compared to the previous study with a smaller size of 1,655 participants. The UDS we used, however, is not a community-based sample. The participants are encouraged to participate in research at each clinic based on concerns with a family history of dementia and they do not represent the community. Not all participants in the data set progress to dementia and we considered using the right-censoring method. Also, whether the patient progresses to dementia is based on clinical judgment by healthcare professionals at clinical visits which might blur the exact timeline of dementia progression. The latent classes are statistical results and therefore require further clinical validation. The thesis uses a simple three-step approach adjustment to account for the latent class classification errors. The bias due to misclassification of latent classes on dementia which leads to underestimated hazard is confirmed. Given additional time in the future, a Monte Carlo simulation can be conducted to further adjust the three-step approach to account for this bias.

Reference:

1. Bakk Z, Tekle F, Vermunt J, “Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches”. American Sociological Association 2018
2. Bhat, H. S.; Kumar, N "On the derivation of the Bayesian Information Criterion" 2010
3. Bolck A, et. al., “Estimating latent structure models with categorial variables: one-step versus three-step estimators”. Political Analysis 2004
4. Breslow, N.E. “Analysis of Survival Data under the Proportional Hazards Model”. International Statistical Review / Revue Internationale de Statistique 43(1):45-57 doi: 10.2307/1402659. JSTOR 1402659
5. Gauthier S, Reisberg B, Zaudig M, et al, “Mild cognitive impairment”. Lancet 2006
6. Hanfelt J, Wuu J, Sollinger A, et al, “An exploration of subgroups of mild cognitive impairment based on cognitive, neuropsychiatric and functional features: analysis of data from the National Alzheimer’s Coordinating Center”. Am J Geriatr Psychiatry 2011
7. Hanfelt et. al., “Latent classes of mild cognitive impairment are associated with clinical outcomes and neuropathology: Analysis of data from the National Alzheimer's Coordinating Center”. Neurobiology of Disease 2018
8. Introduction to SAS. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/sas/modules/sas-learning-moduleintroduction-to-the-features-of-sas/
9. Lyketsos et.al., “Neuropsychiatric symptoms in Alzheimer's disease” PubMed 2011 doi: 10.1016/j.jalz.2011.05.2410

10. Morris JC et.al., “The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers” PubMed 2016
11. Pangman et. al., “An Examination of Psychometric Properties of the Mini-Mental State Examination and the Standardized Mini-Mental State Examination: Implications for Clinical Practice” PubMed 2000 PMID: 11078787 DOI: 10.1053/apnr.2000.9231
12. Phan et.al., “Neuropsychiatric Symptoms in Dementia: Considerations for Pharmacotherapy in the USA” PubMed 2019doi: 10.1007/s40268-019-0272-1
13. Thyrian J, Hoffman W, Eichler T, “Early recognition of dementia in primary care – current issues and concepts” Current Alzheimer Research 2018 Vol. 15 No. 1
14. Therneau, TM, Grambsch, PM. . Modeling Survival Data: Extending the Cox Model. Springer: New York. 2000
15. Vermunt J, Magidson J, “Latent Class Cluster Analysis”
16. Winblad B, Palmer K, Kivipelto M, et al, “Mild cognitive impairment—beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment”. J Internal Med 2004

Appendix

SAS code used for the proportional hazard analysis:

```
data newdata;
    merge distaloutcomes lg_pred;
    by NACCID;
      if clu_ = 1 then classweight = clu_1;
      if clu_ = 2 then classweight = clu_2;
      if clu_ = 3 then classweight = clu_3;
      if clu_ = 4 then classweight = clu_4;
      format clu_ LC.;
run;

proc format;
VALUE LC
1 = 'Mild'
2 = 'Depr+Anx+Apa'
3 = 'Agit+Irr'
4 = 'Severe';
Run;

/*Getting summary statistics for the variables*/
proc freq data=newdata;
      tables female clin_event_type clu_;
run;

proc means data=newdata;
      var basemmse baseage75 followup_time clin_event_time baseeduc;
run;

/*To understand the distribution of clin_event_time*/
proc univariate data = newdata(where=(clin_event_type=1));
var clin_event_time;
histogram clin_event_time/kernel;
run;

proc univariate data = newdata(where=(clin_event_type=1));
var clin_event_time;
cdfplot clin_event_time;
run;
```

```
/*obtaining hazard functions*/
proc lifetest data = newdata plots=hazard;
time clin_event_time*clin_event_type(0);
run;

proc phreg data= newdata plots=cumhaz;
model clin_event_time*clin_event_type(0) = female baseage75 basemmse
baseeduc;
run;

/*Covariate data set for PHREG*/
data covs;
input clu_ female baseage75 basemmse baseeduc;
format clu_ LC.;
datalines;
1 0.532705 0.003062 27.330289 15.126105
2 0.522468 -0.221683 27.240107 15.194500
3 0.387786 -0.171780 27.415267 15.227481
4 0.377282 -0.374625 27.123732 14.770791
;

run;

/*Adjusted three-step approach, the traditional three-step simply leaves
out the weight statement*/
proc phreg data=newdata plots(overlay)=(survivial);
class clu_(REF="Mild");
model clin_event_time*clin_event_type(0) = clu_ female baseage75 basemmse
baseeduc;
baseline covariates=covs out=base /rowid=clu_;
weight classweight;
hazardratio clu_ / DIFF = REF ALPHA = 0.05 CL = WALD;
label clin_event_time = "Follow Up Time (Years)";
/*title "Progression to Dementia By Latent Class";*/
FORMAT clu_ LC.;
run;
```