BIS 687 Group 4 Final Report

Baijia Xu, Jinxuan Bian, Huangrui Chu

2024-05-02

Background

Alzheimer's disease (AD) is a major global health challenge, affecting millions of people worldwide. As the most common cause of dementia in older adults, AD is characterized by progressive neurodegeneration and cognitive decline. Despite extensive research, the pathophysiology of AD remains complex and multifactorial, involving genetic, molecular, and environmental factors that interact over time (Scheltens et al., 2021). The Alzheimer's Disease Neuroimaging Initiative (ADNI) has been instrumental in elucidating this complexity, providing an extensive database of clinical, cognitive, and biomarker information. However, the trajectory of AD progression varies greatly among individuals, and current predictive models often fail to capture the nuances of this variability. Incorporating time-varying factors into these models is an emerging area of research that holds promise for improving predictive accuracy. Innovative approaches that leverage dynamic data could enhance our understanding and prediction of AD progression, leading to improved patient outcomes and more effective allocation of healthcare resources.

Significance

The capability to enhance predictive models for Alzheimer's Disease (AD) by incorporating time-varying factors holds profound implications across multiple domains, ranging from early diagnosis and intervention to risk identification and resource allocation. This advancement in predictive modeling carries far-reaching consequences that extend beyond individual patient care, permeating into broader societal and healthcare spheres.

One of the most notable impacts lies in the realm of early diagnosis and early intervention. By augmenting the predictive power of these models, clinicians gain a heightened ability to detect AD at its nascent stages, enabling them to deliver timely diagnosis. Early diagnosis is of paramount importance because interventions implemented during the initial phases of AD can significantly decelerate the disease's progression, thereby markedly enhancing patients' quality of life and offering a glimmer of hope by delaying the detrimental effects on cognitive function.

Moreover, our refined predictive model can play an instrumental role in identifying individuals who are at an elevated risk of developing AD even before the manifestation of symptoms. This preemptive identification paves the way for the implementation of preventative measures and

meticulous monitoring, potentially postponing or even averting the onset of the disease altogether. For clinicians, this paradigm shift represents a transition towards a more proactive approach in the management of AD, placing a profound emphasis on prevention and early intervention rather than reactive treatment after diagnosis has been established.

The ramifications of more accurate AD predictions transcend the boundaries of individual patient care. By identifying risks and onset with greater precision and at earlier junctures, healthcare providers can allocate resources with enhanced efficiency, prioritizing preventative measures and early treatment for those who require it most urgently. This could culminate in a more judicious utilization of healthcare resources, alleviating the overall burden on healthcare systems and ensuring that patients receive the most appropriate care when they need it most.

Integrating time-varying factors into predictive models not only enhances prediction accuracy but also deepens our comprehension of AD's progression and its underlying mechanisms. This research endeavor could unveil novel insights into the biological, cognitive, and environmental factors that influence AD, potentially paving the way for the development of innovative therapeutic interventions and preventative strategies.

Furthermore, with more precise predictions at their disposal, families can better prepare for the future, gaining a more profound understanding of the disease's likely progression and thereby enabling them to plan accordingly. This can help mitigate some of the emotional and financial strains associated with caring for someone afflicted with AD, empowering families and caregivers to make informed decisions regarding care and support.

In essence, the significance of enhancing predictive models for AD with time-varying factors extends far beyond the confines of individual patient care, reverberating through broader societal and healthcare landscapes, offering a glimmer of hope for those affected by this debilitating condition and their loved ones.

Innovation

This study introduces an approach to Alzheimer's Disease (AD) progression modeling by integrating landmark analysis with random survival forests—a method not conventionally applied in this domain. This innovation lies in the dynamic incorporation of time-varying factors, such as biomarker changes or treatment responses, which are crucial for capturing the real-time progression of AD. By benchmarking these advanced machine learning techniques against traditional Cox proportional hazards models, the research promises to not only improve prediction accuracy but also to provide a more granular understanding of AD dynamics.

Research Plan

The research plan for evaluating and improving the predictive power of clinical and cognitive measures on the progression of Alzheimer's Disease (AD) by incorporating a time-varying factor includes several steps:

1) Data Collection and preprocessing

Utilize data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al., 2010), focusing on patients with comprehensive clinical and cognitive assessment records. Emphasize selecting individuals who have been monitored over extended periods, as their data is invaluable for a deeper analysis of time-varying factors that could influence the disease's progression. Extract relevant data such as demographic information, clinical test results, cognitive scores, neuroimaging data, and biomarkers.

2) Identification of Time-Varying Factors and Time-Static Factors

For Time-Varying Factors, undertake literature review to identify crucial factors that exhibit variability over time and have demonstrated significant associations with the progression of Alzheimer's Disease. Focus on factors that are consistently recorded at multiple time points within patient data.

For Time-Independent Factors, apply marginal screening to reduce the data to potentially relevant features, and then employ Principal Component Analysis (PCA) to further refine this selection by identifying the principal components that capture the most variance related to the disease's progression.

3) Development of Predictive Model

Implement a landmark analysis strategy to systematically identify and select cohorts of patients at specific milestones in the progression of Alzheimer's Disease (Bansal & Heagerty, 2019). For each identified landmark time, utilize random survival forests to construct predictive models that estimate the survival probabilities of patients (Zabor & Assel, 2023). This sophisticated ensemble method leverages decision trees designed for survival analysis, incorporating both time-dependent and time-independent variables to handle the complexity of Alzheimer's progression data.

4) Model Validation and Benchmarking

Utilize several approaches as competing methods to make comparisons. The first method is the traditional Cox Proportional Hazards model, which incorporates only time-static covariates. These are variables that do not change over time, such as genetic factors or baseline clinical measurements. The second approach involves an enhanced version of the Cox Proportional Hazards model (Zhang et al., 2018), which is capable of integrating both time-varying and time-static covariates. This model allows for the inclusion of exactly the same covariates as the

model we develop, offering a more comprehensive view of model comparisons. Another approach is the Random Survival Forest model (Wang et al., 2023), which also handles time-static covariates, but providing a more robust alternative to traditional models. The last approach is the Accelerated Failure Time Model, which assumes that the effect of covariates accelerates or decelerates the life-time of a patient.

Apply cross-validation techniques to evaluate the predictive performance of the models. This involves systematically partitioning the data into subsets, using each in turn to validate the model trained on the remaining data. This process helps to prevent overfitting and ensures that the models' performance is robust across different subsets of data.

Employ specific metrics such as the concordance index (C-index) to assess the predictive capabilities of each model. The C-index measures the accuracy of the model's ability to predict the event of interest, providing a clear, quantitative indication of model performance in terms of its prognostic accuracy.

5) Analysis and Interpretation
Translate the raw data and model outputs into meaningful insights about Alzheimer's Disease progression. Once the predictive model has been validated, conduct detailed statistical analysis to interpret the relationships between factors and the disease's progression, and impacts of incorporating time-varying factors. Examine the significance and impact of both time-varying and time-static factors identified during the modeling phase. Use visualizations to communicate findings clearly and effectively.

6) Limitations and Future Research
Acknowledge the constraints of the current study and outline areas for future investigation to enhance the understanding and prediction of Alzheimer’s Disease progression.

Details of each step are included as below.

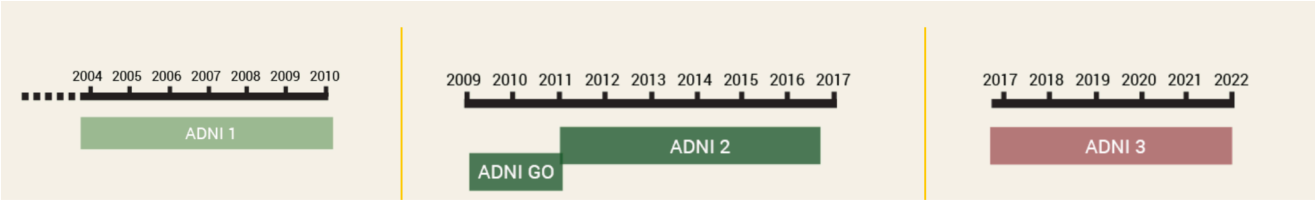
Research Strategy
Specific Aim
To Evaluate and Improve the Predictive Power of the Alzheimer's Disease Progression by Introducing One or More Additional Time-Varying Factors at Multiple Time Points

Hypothesis
We hypothesize that incorporating the time-varying factors (such as Mini Mental State Exam (MMSE), Brain Volume, Regional Thickness etc.) at multiple time points into predictive models with time-constant genetics measures and demographics information will significantly improve

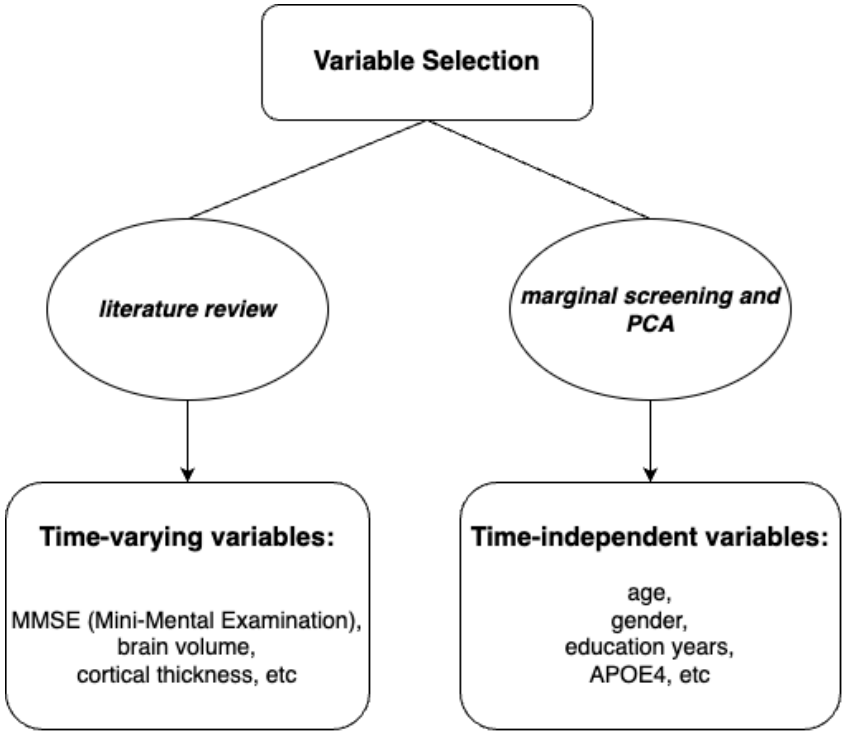
the predictions of Alzheimer's Disease (AD) progression, providing a more comprehensive assessment.

Experimental Approach

- **Data Collection and Preparation**: Data we utilized is collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI), focusing on patients with detailed records of clinical and cognitive assessments, alongside longitudinally collected factor data. The study can be divided into four stages – ADNI1, ADNI-GO, ADNI-2, and ADNI-3, of which the timeline is shown in the below figure.



- **Selection of Features:**
  Time-Varying Factors —- Identify key factors that vary over time and are likely to influence AD progression and that have been recorded at multiple time points. Begin with literature review to figure out which cognitive scores, biomarkers, and clinical assessments have shown significant correlations with disease progression. We refer to Zhang et al., 2021 and Su et al., 2021 to select several time-varying measures, such as MMSE, brain volume, and cortical thickness, which might be highly correlated with AD onset.
  Time-Independent Factors —- Conduct feature selection using marginal screening and Principal component analysis (PCA), identifying the most predictive variables from clinical scores, cognitive test outcomes, and time-varying factors such as medication use, lifestyle alterations, or biomarker levels.
  The figure below demonstrates the framework of feature selection.



- **Statistical Modeling and Analysis**: Employ landmark analysis to periodically select cohorts of patients who have reached certain points in the disease progression. Then, for each "landmark" time, we build random survival forests to predict survival probabilities based on the data, accounting for time-dependent nature of survival data and potentially more dynamic and accurate predictions over time.
- **Competing method**: Cox proportional hazards models that treat time-varying covariates as time-constant variables would serve as a competing benchmark. We intend to compare our landmark analysis model's performance against the Cox proportional hazards models, both with and without time-varying covariates, Random Survival Forest model, and Accelerated Failure Time model.
- **Model Validation**: Use cross-validation techniques to assess the predictive performance of the model. Performance metrics, such as concordance index (C-index) would be used to evaluate the model's predictive capabilities. Sensitivity analyses would also be conducted to further explore the robustness of the model, helping to understand how changes in time-varying factors or assumptions about the disease progression impact the predictions and providing a deeper insight into the dynamics of AD progression and enhancing the model's applicability in clinical settings.

Interpretation of Results

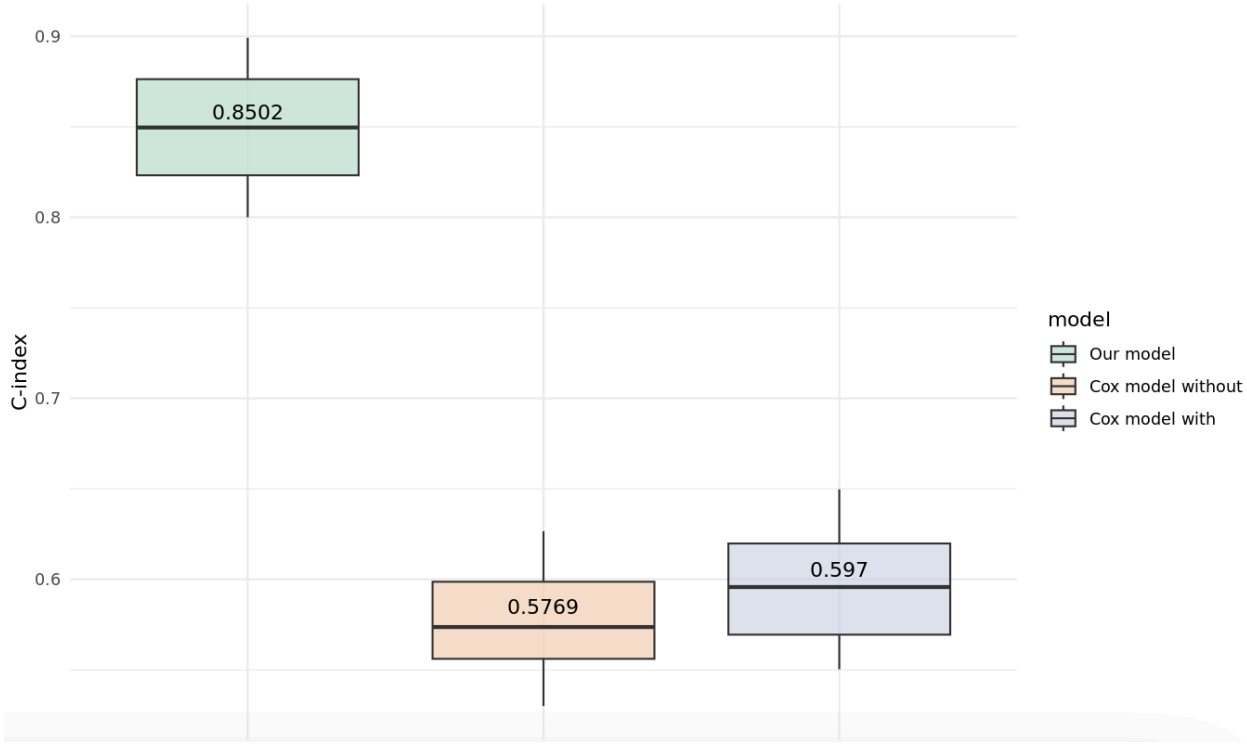
- **Feature selection**
  The final time-independent variables we select to fit the model are ***gender, age, education years*** and ***APOE4***, which is a genetic factor. We conducted the marginal screening in which we used only one of the variables to fit the model each time and then selected the four variables with highest C-index. The specific C-index values are shown in the following figure.

variable	C-index
gender	0.6987
age	0.7604
education years	0.6576
APOE4	0.8350

Table: Marginal Screening Results

  For the time-varying variables, we finally chose ***MMSE, standardized brain volume***, and ***cortical thickness*** by conducting literature review.

- **C-index and model evaluation**
  Our in-sample C-index came out to be approximately 0.90. The high C-index suggests that our model is somehow accurate in predicting the risk of MCI within the same dataset used to develop the model. To ensure the robustness of our model, we also conducted cross-validation and compared our model's performance against the traditional Cox proportional hazards model, both with and without time-varying covariates. The comparison results are shown in the below box plot. Our landmark analysis model, with a median C-index of around 0.85, significantly outperforms the Cox model both with and without time-varying covariates since they only have a median C-index of about 0.60 - 0.65. This demonstrates that incorporating time-varying variables and using our landmark analysis method greatly improves the model's predictive power.



The results we obtained provide insights into the progression of Alzheimer's Disease (AD). The hypothesis that the successful incorporation of time-varying factors—potentially biomarkers—into the predictive models would signify a notable improvement over the traditional method, is confirmed. This could lead to a more nuanced understanding of AD progression and potentially inform more personalized treatment approaches. The interpretation will focus on the magnitude and direction of the impact that the time-varying factor has on disease progression. A significant change in predictive power, as reflected by an increased C-index in the model including the time-varying factor, would support the hypothesis that incorporating such dynamic factors offers a more accurate and holistic assessment of AD risk over time. Additionally, the landmark analysis would provide snapshots of the disease progression at various stages, which could reveal critical windows for therapeutic intervention.

Conclusion

Based on the results of the models used in this study, the hypothesis that incorporating time-varying factors such as Mini Mental State Exam (MMSE), Brain Volume, and Cortical Thickness into predictive model alongside time-constant genetic measures and demographic information including age, gender, education level, and APOE4 gene would significantly improve the predictions of Alzheimer's Disease (AD) progression is strongly supported. The model we developed, which integrated these time-varying factors, achieved a concordance index

(C-index) of 0.9. This is a substantial improvement over the traditional Cox Proportional Hazards model with a C-index of 0.622, the Cox model with time-varying covariates at 0.66, and the Accelerated Failure Time model at 0.6279. The Random Survival Forest model had the lowest performance with a C-index of 0.4117712, indicating that the more sophisticated integration of time-varying and time-static variables in our developed model is particularly effective.

These findings suggest that the enhanced predictive accuracy of our model can be attributed to its ability to dynamically incorporate changes in key variables over time, reflecting more accurately the progression of Alzheimer's Disease. Therefore, the approach of integrating both time-static and time-varying factors in the analysis of AD progression offers a more comprehensive and accurate model, highlighting the importance of considering the dynamic nature of the disease in predictive analytics. This can potentially lead to more targeted and timely interventions based on a patient's changing condition over time.

Relating the conclusion back to the scientific question, by achieving a C-index of 0.9, our model demonstrates its potential to improve the accuracy of Alzheimer's diagnosis. This high level of diagnostic accuracy is crucial for timely intervention, allowing clinicians to initiate treatment plans at an early stage when they are likely to be more effective. The integration of time-varying factors such as MMSE, brain volume, and regional thickness at multiple time points allows our model to detect subtle changes that may indicate the onset of Alzheimer's before major symptoms manifest. This capability can transform the landscape of early detection, potentially leading to interventions that could delay or mitigate the progression of the disease. Also, the dynamic nature of our model, which adapts to changes in a patient's condition over time, supports the development of personalized monitoring plans. This approach enables clinicians to tailor monitoring and treatment to the individual's specific disease progression pattern, thus enhancing treatment responsiveness and potentially improving patient outcomes. With the model's robust predictive power, it provides clinicians with detailed insights into the likely progression of Alzheimer's in individual patients. This supports more informed decision-making in treatment planning, optimizing both the timing and nature of interventions. The ability to anticipate disease progression with greater accuracy ensures that care strategies are both proactive and precisely aligned with patient needs. Finally, the advanced capabilities of our model not only assist clinicians but also serve as a valuable educational tool for families of patients with Alzheimer's. By understanding the potential trajectory of the disease, families and caregivers can make more informed decisions regarding care management and are better prepared for future challenges. This knowledge empowers them to engage in more effective discussions about care options and long-term planning.

Limitations and future research

Despite the promising results of our study, several limitations must be acknowledged. First, while the model demonstrated high predictive accuracy, it is heavily reliant on the availability and quality of data concerning time-varying factors such as MMSE scores, brain volume, and cortical thickness. In practice, consistent and high-quality data collection may be challenging, potentially limiting the model's applicability in less controlled settings. Another limitation arises during data integration. When merging datasets, some time entries might be lost due to missing values in some variables, which can compromise the continuity and integrity of the data used in our models. Furthermore, our study primarily utilized data from a specific cohort, which may not fully represent the broader population of Alzheimer's patients, particularly those from diverse ethnic and socioeconomic backgrounds.

Future research should focus on several key areas to enhance the understanding and application of predictive models in Alzheimer’s disease. Firstly, expanding the dataset to include a more diverse patient population would help to validate and possibly refine the model’s predictive capabilities across different demographic groups. Additionally, exploring the integration of other potential predictors, such as lifestyle factors or more detailed genetic information, could further enhance the model's accuracy. Finally, longitudinal studies that track patient outcomes over longer periods would provide deeper insights into the long-term efficacy of using such predictive models in clinical practice. These efforts would help to address the limitations noted and broaden the model's applicability and impact in the field of Alzheimer’s research and care.

References

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., ... & Frisoni, G. B. (2021). Alzheimer's disease. The Lancet, 397(10284), 1577-1590.

Zhang, B., Lin, L., Wu, S., & Al-Masqari, Z. H. M. A. (2021). Multiple Subtypes of Alzheimer's Disease Based on Brain Atrophy Pattern. Brain Sciences, 11(2), 278. https://doi.org/10.3390/brainsci11020278

Su, Y., Dong, J., Sun, J., et al. (2021). Cognitive function assessed by Mini-mental state examination and risk of all-cause mortality: A community-based prospective cohort study. BMC Geriatrics, 21, 524. https://doi.org/10.1186/s12877-021-02471-9

Wang, Y., Deng, Y., Tan, Y. et al. A comparison of random survival forest and Cox regression for prediction of mortality in patients with hemorrhagic stroke. BMC Med Inform Decis Mak 23, 215 (2023). https://doi.org/10.1186/s12911-023-02293-2

Bansal, A., & Heagerty, P. J. (2019). A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. Diagnostic and prognostic research, 3, 14. https://doi.org/10.1186/s41512-019-0057-6

Dafni, U. (2011). Landmark analysis at the 25-year landmark point. Circulation: Cardiovascular Quality and Outcomes, 4, 363-371. https://doi.org/10.1161/CIRCOUTCOMES.110.957951

Zabor, E. C., & Assel, M. (2023). On the need for landmark analysis or time-dependent covariates. Journal of Urology. https://doi.org/10.1097/JU.0000000000003459

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. M. (2018). Time-varying covariates and coefficients in Cox regression models. Annals of translational medicine, 6(7), 121. https://doi.org/10.21037/atm.2018.02.12

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jr, Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology, 74(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Appendix

```
################################ Step 1. Prepare covariates ################################
ct <- read.csv("/gpfs/gibbs/project/UCSFFSX51_11_08_19_28Apr2024.csv")

data(adnimerge)
# data(ucberkeleyfbb)
# data(ucberkeleyav1451)
data(ucberkeleyav45)

df1 <- adnimerge %>%
  select(c("RID", "PTID", "VISCODE", "EXAMDATE", "AGE", "PTGENDER",
           "PTEDUCAT", "APOE4", "MMSE", "DX", "Years.bl", "Month.bl"))
# "WholeBrain"
# length(unique(df1$RID)) # 2404 subjects

df2 <- ucberkeleyav45 %>%
  select(c("RID", "EXAMDATE", "SUMMARYSUVR_WHOLECEREBNORM"))
# "WHOLECEREBELLUM_VOLUME"
# length(unique(df2$RID)) # 1341 subjects

df3 <- ct %>%
  select(c("RID", "EXAMDATE", "ST102TA"))

df3 <- df3 %>%
  group_by(RID, EXAMDATE) %>%
  summarise(CT = mean(ST102TA))

# length(unique(df3$RID)) # 1067 subjects
```

```r
# length(unique(df3$RID)) # 1067 subjects

overlapped_subj1 <- which(df1$RID %in% df2$RID)
df1 <- df1[which(df1$RID %in% df2$RID),] # 1340 subjects
df2 <- df2[which(df2$RID %in% df1$RID),] # 1340 subjects

df1 <- df1[which(df1$RID %in% df3$RID),] # 964 subjects
df2 <- df2[which(df2$RID %in% df3$RID),] # 964 subjects
df3 <- df3[which(df3$RID %in% df1$RID),] # 964 subjects

df1$RID <- as.character(df1$RID)
df2$RID <- as.character(df2$RID)
df3$RID <- as.character(df3$RID)

names(df1)[4] <- "EXAMDATE1"
names(df2)[2] <- "EXAMDATE2"
names(df3)[2] <- "EXAMDATE3"

df1$EXAMDATE1 <- as.Date(df1$EXAMDATE1)
df2$EXAMDATE2 <- as.Date(df2$EXAMDATE2)
df3$EXAMDATE3 <- as.Date(df3$EXAMDATE3)

d <- full_join(df1, df2, by = "RID") %>%
  mutate(date_diff = abs(as.numeric(difftime(EXAMDATE1, EXAMDATE2, units = "days")))) %>%
  filter(date_diff < 10) %>%
  select(-date_diff)

d <- d %>%
  full_join(df3, by="RID") %>%
  mutate(date_diff = abs(as.numeric(difftime(EXAMDATE1, EXAMDATE3, units = "days")))) %>%
  filter(date_diff < 10) %>%
  select(-date_diff)
```

```r
d <- d[complete.cases(d),]

# length(unique(d$RID)) # 486 subjects

# define event
d <- d %>%
  mutate(event = ifelse(DX=="CN", 0, 1))

# remove
d <- d %>%
  select(-c("Month.bl", "EXAMDATE2", "EXAMDATE3", "DX")) %>%
  mutate(age = AGE + Years.bl) %>%
  select(-c("AGE", "Years.bl"))

# write.csv(d, "/gpfs/gibbs/project/zhao_yize/bx69/datause.csv")

################################## Step 1. done ##################################
##################################################################
```

```r
################################## Step 2. Prepare time-varying covariates ##################################
# set time interval (last NACC - first NACC among sample)
################### first separate the variables into continous ones, and uncontinuous ones

# set time intervals
tvisit <- d %>%
  group_by(RID) %>%
  summarize(T1 = first(age),
            Tc = last(age))

seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)

# seq(min(id_status_change$age_event1[id_status_change$age_event1>0]),
#     max(id_status_change$age_event1),
#     length.out=7)

tk1 <- seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)[2]
tk2 <- seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)[3]
tk3 <- seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)[4]
tk4 <- seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)[5]
tk5 <- seq(min(tvisit$T1), max(tvisit$Tc), length.out=7)[6]

data.use <- d %>%
  select(c("RID", "MMSE", "SUMMARYSUVR_WHOLECEREBNORM", "CT", "age"))

dyn_con <- data.use[data.use$age<=tk1,] %>%
  group_by(RID)

dyn1_con <- data.frame("RID"=unique(dyn_con$RID))
```

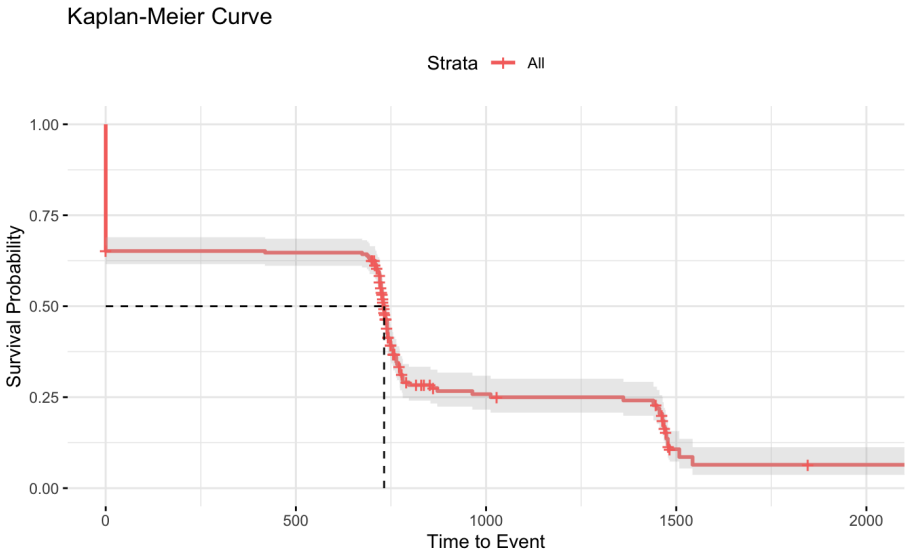
```r
for (j in 2:(ncol(dyn_con)-1)){
  name <- names(dyn_con)[j]
  dyn1 <-
    dyn_con %>%
    group_by(RID) %>%
    summarise_at(vars(name), list(mean), na.rm=TRUE)

  dyn1_con <- dyn1_con %>%
    left_join(dyn1, by="RID")
}

dyn_con <- data.use[data.use$age<=tk2&(data.use$age>tk1),] %>%
  group_by(RID)

dyn2_con <- data.frame("RID"=unique(dyn_con$RID))

for (j in 2:(ncol(dyn_con)-1)){
  name <- names(dyn_con)[j]
  dyn2 <-
    dyn_con %>%
    group_by(RID) %>%
    summarise_at(vars(name), list(mean), na.rm=TRUE)

  dyn2_con <- dyn2_con %>%
    left_join(dyn2, by="RID")
}

dyn_con <- data.use[data.use$age<=tk3&(data.use$age>tk2),] %>%
  group_by(RID)

dyn3_con <- data.frame("RID"=unique(dyn_con$RID))
```

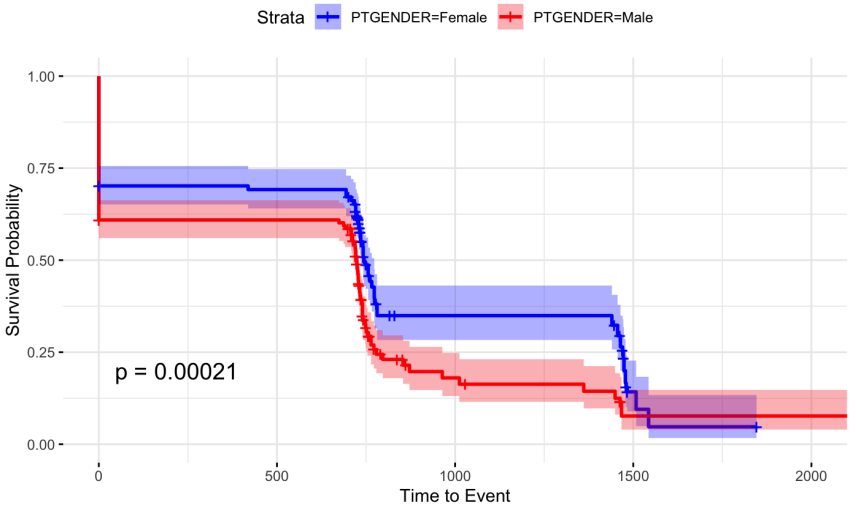
```r
for (j in 2:(ncol(dyn_con)-1)){
  name <- names(dyn_con)[j]
  dyn3 <-
    dyn_con %>%
    group_by(RID) %>%
    summarise_at(vars(name), list(mean), na.rm=TRUE)

  dyn3_con <- dyn3_con %>%
    left_join(dyn3, by="RID")
}

dyn_con <- data.use[data.use$age<=tk4&(data.use$age>tk3),] %>%
  group_by(RID)

dyn4_con <- data.frame("RID"=unique(dyn_con$RID))

for (j in 2:(ncol(dyn_con)-1)){
  name <- names(dyn_con)[j]
  dyn4 <-
    dyn_con %>%
    group_by(RID) %>%
    summarise_at(vars(name), list(mean), na.rm=TRUE)

  dyn4_con <- dyn4_con %>%
    left_join(dyn4, by="RID")
}

dyn_con <- data.use[data.use$age<=tk5&(data.use$age>tk4),] %>%
  group_by(RID)

dyn5_con <- data.frame("RID"=unique(dyn_con$RID))
```

```r
for (j in 2:(ncol(dyn_con)-1)){
  name <- names(dyn_con)[j]
  dyn5 <-
    dyn_con %>%
    group_by(RID) %>%
    summarise_at(vars(name), list(mean), na.rm=TRUE)

  dyn5_con <- dyn5_con %>%
    left_join(dyn5, by="RID")
}

w_dyn <- data.frame("RID"=unique(data.use$RID))
# add 1st
Nvar <- ncol(dyn1_con)-1
w_dyn <- w_dyn %>%
  left_join(dyn1_con, by="RID")
names(w_dyn)[2:ncol(w_dyn)] <- sprintf("c.%d.1.1", 1:(Nvar))

# ncol(w_dyn) #354

# add 2nd
w_dyn <- w_dyn %>%
  left_join(dyn2_con, by="RID")
names(w_dyn)[(2+Nvar):ncol(w_dyn)] <- sprintf("c.%d.2.1", 1:Nvar)

# add 3rd
w_dyn <- w_dyn %>%
  left_join(dyn3_con, by="RID")
names(w_dyn)[(2+Nvar*2):ncol(w_dyn)] <- sprintf("c.%d.3.1", 1:Nvar)
```

```r
# add 4th
w_dyn <- w_dyn %>%
  left_join(dyn4_con, by="RID")
names(w_dyn)[(2+Nvar*3):ncol(w_dyn)] <- sprintf("c.%d.4.1", 1:Nvar)

# add 5th
w_dyn <- w_dyn %>%
  left_join(dyn5_con, by="RID")
names(w_dyn)[(2+Nvar*4):ncol(w_dyn)] <- sprintf("c.%d.5.1", 1:Nvar)

### proposed preprocessing procedure
# left split
w_dyn[is.na(w_dyn)] <- -10^5
Nleft <- ncol(w_dyn)-1

# right split
for (i in 2:(Nleft+1)){
  new_col_name <- sub("1$", "2", names(w_dyn)[i])
  # Copy values from existing columns to new columns
  w_dyn[, new_col_name] <- w_dyn[, i]
  # Replace -10^5 with 10^5 in the new columns
  w_dyn[w_dyn[, new_col_name] == -10^5, new_col_name] <- 10^5
}

# ncol(w_dyn) # 31

################################## Step 2. done ##################################
################################## ################################
```

```
################################## Step 3. Combine survival data ###################################
w_dyn <- w_dyn %>%
  left_join(d[,c("RID", "PTGENDER", "PTEDUCAT", "APOE4", "event","age")], by = "RID")

################################## Step 4. Fit the model ###################################

#### use all measurements #########
train_id <- sample(nrow(w_dyn), nrow(w_dyn)*0.8)
train <- w_dyn[train_id,]
test <- w_dyn[-train_id,]

dyn_t1 <- ranger(Surv(age, event) ~ ., data = train,
                 keep.inbag = TRUE, min.node.size = 15)

Cindex <- 1-dyn_t1$prediction.error
```





Cox Proportional Hazards model

```
Call:
coxph(formula = Surv(Time_to_Event, event) ~ age + PTGENDER +
    PTEDUCAT + APOE4, data = merged)

  n= 643, number of events= 406

                 coef exp(coef)  se(coef)      z Pr(>|z|)
age           0.002625  1.002628  0.007263  0.361 0.717769
PTGENDERMale  0.380345  1.462789  0.103191  3.686 0.000228 ***
PTEDUCAT     -0.024843  0.975463  0.019900 -1.248 0.211881
APOE4         0.479627  1.615472  0.073528  6.523 6.89e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
age             1.0026     0.9974    0.9885     1.017
PTGENDERMale    1.4628     0.6836    1.1949     1.791
PTEDUCAT        0.9755     1.0252    0.9381     1.014
APOE4           1.6155     0.6190    1.3987     1.866

Concordance= 0.622  (se = 0.019 )
Likelihood ratio test= 55.56  on 4 df,   p=2e-11
Wald test            = 59.32  on 4 df,   p=4e-12
Score (logrank) test = 60.52  on 4 df,   p=2e-12

         chisq df     p
age      2.821  1 0.093
PTGENDER 0.381  1 0.537
PTEDUCAT 1.629  1 0.202
APOE4    1.483  1 0.223
GLOBAL   6.955  4 0.138
```

Cox Proportional Hazards model with time-varying covariates

```
Call:
coxph(formula = Surv(start_time, end_time, event) ~ age + PTGENDER +
    PTEDUCAT + APOE4 + MMSE + CT + SUMMARYSUVR_WHOLECEREBNORM,
    data = merged2)

  n= 299, number of events= 182

                               coef exp(coef) se(coef)      z Pr(>|z|)
age                         -0.02243   0.97782  0.01345 -1.668   0.0954 .
PTGENDERMale                -0.07348   0.92915  0.15558 -0.472   0.6367
PTEDUCAT                    -0.02080   0.97942  0.03234 -0.643   0.5202
APOE4                        0.17315   1.18904  0.13285  1.303   0.1924
MMSE                        -0.08583   0.91775  0.02102 -4.083 4.44e-05 ***
CT                          -0.11381   0.89243  0.42863 -0.266   0.7906
SUMMARYSUVR_WHOLECEREBNORM  0.47534   1.60857  0.42021  1.131   0.2580
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef) lower .95 upper .95
age                            0.9778     1.0227    0.9524    1.0039
PTGENDERMale                   0.9292     1.0763    0.6849    1.2604
PTEDUCAT                       0.9794     1.0210    0.9193    1.0435
APOE4                          1.1890     0.8410    0.9165    1.5427
MMSE                           0.9178     1.0896    0.8807    0.9563
CT                             0.8924     1.1205    0.3852    2.0674
SUMMARYSUVR_WHOLECEREBNORM     1.6086     0.6217    0.7059    3.6654

Concordance= 0.66  (se = 0.021 )
Likelihood ratio test= 36.15  on 7 df,   p=7e-06
Wald test            = 42.47  on 7 df,   p=4e-07
Score (logrank) test = 45.04  on 7 df,   p=1e-07
```

Random Survival Forest

```r
obj <- rfsrc(Surv(Time_to_Event, event) ~ age + PTGENDER + PTEDUCAT + APOE4,
             data = merged,
             ntree = 1000, importance = TRUE)
get.cindex(obj$yvar[,1], obj$yvar[,2], obj$predicted.oob)
```

```
[1] 0.4117712
```

Accelerated Failure Time

```
Call:
survreg(formula = surv_object ~ PTGENDER + PTEDUCAT + APOE4 +
    age, data = merged3, dist = "lognormal")
               Value Std. Error      z      p
(Intercept)   6.46660    0.30415  21.26 <2e-16
PTGENDERMale -0.15260    0.04678  -3.26 0.0011
PTEDUCAT      0.00219    0.00909   0.24 0.8097
APOE4        -0.11750    0.03652  -3.22 0.0013
age           0.00724    0.00340   2.13 0.0331
Log(scale)   -1.05079    0.05147 -20.42 <2e-16

Scale= 0.35

Log Normal distribution
Loglik(model)= -1350.8   Loglik(intercept only)= -1363.8
        Chisq= 26.06 on 4 degrees of freedom, p= 3.1e-05
Number of Newton-Raphson Iterations: 3
n= 299

Call:
concordance.formula(object = Surv(merged3$Time_to_Event, merged3$event) ~
    predicted_times, data = merged3)

n= 299
Concordance= 0.6279 se= 0.02646
concordant discordant     tied.x     tied.y    tied.xy
     16614       9844          0        218          0
```