An empirical evaluation of methods for the prediction of survival with many longitudinally-measured predictors

Feng, Y.

Universiteit
Leiden
The Netherlands

An empirical evaluation of methods for the prediction of survival with many longitudinally-measured predictors

Yibin Feng

Thesis advisor: Dr. Mirko Signorelli
Mathematical Institute, Leiden University

Defended on 16 August, 2022

MASTER THESIS
STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Contents

Abstract

Survival analysis deals with the study of the time until an event of interest occurs. The Cox Proportional Hazards model (Cox model) is commonly used to model the relationship between a survival outcome and a set of cross-sectional covariates, but it cannot handle longitudinal covariates, i.e. covariates that are repeatedly measured over time. Traditional ways to deal with longitudinal covariates include joint modelling, landmarking and the time-dependent Cox model, but to date their applicability has mostly been restricted to problems with a small number of longitudinal covariates.

Recently, the increasing availability of repeated measurements in biomedical studies has motivated the development of statistical methods specifically designed to predict survival from a large (potentially high-dimensional) number of longitudinal covariates. Due to the fact that such methods are still quite new, little is known about how these methods may perform in practice.

The aim of this thesis is to compare the performance of various statistical methods to predict survival on a real dataset where many longitudinal covariates are available as predictors. Four methods were chosen for comparison, including two novel methods employing different techniques to harness the longitudinal information, Penalized Regression Calibration (PRC) and Multivariate Functional Principal Component Cox (MFPCCox) model, and penalized Cox models using landmarking (last observation carried forward method) and baseline measurements respectively.

These methods were applied to the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study in the context of dynamic prediction of time to develop dementia. The ADNI study monitored the development of dementia in cohort of elderly individuals, and collected an extensive, heterogeneous set of markers over multiple years of follow-ups. Predictions were computed using a total of 26 covariates, of which 21 were longitudinal. The predictive performance of the models was evaluated considering three performance measures (time-dependent AUC, C index, and Brier score).

The results showed that the best performing method depended on the choice of performance measure, landmark time, and prediction time. Landmarking was the best performing method when looking at the time-dependent AUC and C index, whereas PRC was the best performing method in terms of Brier score. Landmarking, PRC, and MFPCCox outperformed the baseline model that ignored the follow-up information, suggesting that the longitudinal information in the ADNI data can be used to improve predictions for dementia. Overall, our results seem to indicate that for the ADNI data a simple approach such as landmarking may be enough to deliver accurate predictions, when compared to more sophisticated approaches (PRC and MFPCCox) that model the trajectories of longitudinal covariates.

Foreword

Acknowledgements

First, I would like to express my deep gratitude to my supervisor, Dr. Mirko Signorelli, for his thorough guidance and generous support over the course of this thesis. Mirko's thoughtful feedback has been vital to this project. Moreover, it was an enjoyable and precious experience to learn statistical thinking and statistical computing from such a great teacher.

I would also like to thank the Mathematical Institute for inviting me to the statistical group meetings which broadened my horizons, and the Academic Leiden Interdisciplinary Cluster Environment (ALICE) team for allowing me to use their high performance cluster for running the computations for this thesis.

A special thanks to the authors of (i) *Thinking, Fast and Slow* (Daniel Kahneman), (ii) *Factfulness: Ten Reasons We're Wrong About the World — and Why Things Are Better Than You Think* (Hans Rosling, Anna Rosling Rönnlund, Ola Rosling), and (iii) *Soccermatics: Mathematical Adventures in the Beautiful Game* (David Sumpter) whose ideas led me to the rabbit role of statistics and data science.

Last but not least, I would like to sincerely thank my family and friends who encouraged and supported me from the very beginning of this journey, especially in the turmoil times these years.

Data availability statement

The data used in this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu/`). The investigators within the ADNI provided us with the ADNI data, but they did not participate in the project described in this thesis. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.`

Chapter 1

Introduction

Survival analysis deals with the study of the time until an event of interest occurs, which we refer to as *time to event* or *survival time*. Examples of events that can be studied through survival analysis are death, developing a certain disease, marriage, failure of a mechanical component, etc. A common trait of survival data is *censoring*, which occurs when the actual survival time cannot be observed. Censoring may be due to different reasons, such as the fact that a subject does not experience an event before the end of the study, or he/she is lost to follow-up during study, or he/she is withdrawn from study. A fundamental task in survival analysis is to estimate the survival or hazard function from data. Typically, statistical modelling of survival relates the survival outcome to cross-sectional covariates that are independent of time. The Cox Proportional Hazards model (Cox model) (Cox, 1972) is one of the most popular methods for this task.

However, this is not the case in longitudinal studies where interest lies in the relation between the survival outcome and repeated measurements of certain variable over the course of the study, which we refer to as *longitudinal covariate* or *time-dependent covariate*. For example, Tsiatis et al. (1995) evaluated the potential of CD4-lymphocyte counts as a marker for human immune virus (HIV) trials by analyzing its trajectory with the clinical progression. The problem that the Cox model cannot naturally handle the longitudinal covariate has led to the development of various methods to deal with such information in survival analysis:

- the time-dependent Cox model (Fisher and Lin, 1999) is an extension to the Cox model where the hazard function is coupled with a longitudinal covariate whose value is assumed to be constant in the time interval that occurs between two subsequent repeated measurements. It has limitations that the longitudinal covariate should be exogenous, and using step function to model the longitudinal covariate could become highly unrealistic in some circumstances. Along with landmarking introduced below, time-dependent Cox model are two most commonly used approaches for survival analysis involving longitudinal covariates (Putter and van Houwelingen, 2016);
- landmarking (or landmark approach) (Anderson et al., 1983) is a simpler approach than the time-dependent Cox model. It involves first setting a landmark time t_l, then using the last observation (or temporal aggregation such as the average measurements within an observation window) prior to the landmark time of the longitudinal covariate as predictor in a Cox model. The landmarking model is estimated based on the landmark dataset which only considers individuals still at risk at the landmark time. The performance of landmarking depends on the landmark choice: a landmark time too early is more likely to have more imbalanced data, leading to misclassification at longer follow-up, whereas a

landmark time too late will omit a high proportion of events, hence reducing the power (Dafni, 2011). Landmarking may lead to bias when last observations are carried forward for a longitudinal covariate of which the last observed value is remote from the true underlying value. Comprehensive comparisons between landmarking and time-dependent Cox model can be found in (Dafni, 2011; Putter and van Houwelingen, 2016; Bull et al., 2020);

- joint modelling (Rizopoulos, 2011) is a modelling approach that jointly estimates two submodels: one for the longitudinal covariates, and one for the survival outcome. The linear mixed model is a popular choice for the first submodel of the trajectories of longitudinal covariates. The second submodel is a Cox model depending on the same longitudinal covariates. Despite the advantages of being more data efficient than landmarking and of modelling the joint distribution of longitudinal data and survival data, estimation of joint model is computationally intensive; to date, its estimation remains computationally prohibitive when there are more than a few (e.g., 5) longitudinal covariates (Hickey et al., 2016; Mauff et al., 2020). Moreover, it can be quite sensitive to misspecification of the longitudinal trajectory (Putter and Houwelingen, 2022).

The aforementioned models were developed having in mind prediction problems that would involve a limited number of longitudinal covariates. Nowadays, technological and methodological advances in biomedical studies have made it more common for longitudinal studies to measure many longitudinal covariates. Recently, three novel methods have been proposed to address this challenge:

- Multivariate Functional Principal Component Cox model (MFPCCox) (Li and Luo, 2019) carries out dimension reduction and feature extraction on longitudinal covariates by multivariate functional principal component analysis (MFPCA) and uses the resultant MFPC scores as predictors in a Cox model;

- Penalized Regression Calibration (PRC) (Signorelli et al., 2021) involves a three-step approach: first, the longitudinal covariates are modelled using mixed effect models; then, subject-specific summaries of the longitudinal trajectories are derived from the fitted mixed models; lastly, the summaries of the trajectories are used to predict survival using a penalized Cox model;

- Functional Ensemble Survival Tree (Jiang et al., 2021) is similar to MFPCCox but the MFPC scores are used as predictors in an ensemble survival tree (random survival forest) (Ishwaran et al., 2008) instead of a Cox model for estimating the survival outcome.

These novel methods employed different techniques to handle the complexity arising from having numerous longitudinal covariates. Since these methods were developed very recently, little is known about their performance on real data. This thesis represents a first attempt to compare these methods with each other, and to simpler prediction approaches that may not properly handle many longitudinal covariates.

1.1 Research question

The aim of this study is to compare the predictive performance of various statistical methods for the dynamic prediction of survival in the presence of many longitudinal covariates using real data. We will compare a total of four different methods, two methods (PRC and MFPCCox) that use sophisticated statistical methods to summarize the longitudinal covariates; a simpler approach (landmarking) that uses the last observation available from each longitudinal covariate for prediction; and a penalized Cox model that ignores the repeated measurements, using only baseline measurements for prediction. A concise summary of these methods is presented in Table 1.1. Some of the aforementioned methods were excluded from comparison for various reasons:

- the time-dependent Cox model was excluded as its assumption would be violated by the endogeneity of the longitudinal covariates;
- joint modelling was excluded as its estimation was computationally prohibitive for this thesis (using > 20 longitudinal covariates);
- functional ensemble survival tree was excluded as the package `funest` did not work with the ADNI data successfully due to data compatibility issue.

The predictive performance of the methods will be evaluated in the context of dynamic prediction of developing dementia, using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Weiner et al., 2010). Dementia is a general term of syndrome corresponding to the deterioration in cognitive function which is different from normal biological aging process (WHO, 2021). It is one of the most common neurological disorders, and the seventh leading cause of death as of 2019 according to estimates from the World Health Organization. As dementia usually displays chronic or progressive nature over several years, dementia risk models are often developed with the goal of quantifying the probability that an individual may develop dementia to facilitate the decision on personalized intervention for individuals from either general population or specific subpopulation (Tang et al., 2015, 2017; Hou et al., 2018; Licher et al., 2019). In this regard, dynamic prediction enables these predictions to be updated when more biomarker data have been collected from new follow-up visits. The ADNI study provides an extensive set of heterogeneous and longitudinal dementia-related markers repeatedly measured over a long follow-up period.

Table 1.1: Applicability of prediction methods to high-dimensional data and longitudinal data for survival prediction

Method	**Suitable for**		**Reference**
	high-dimensional covariates	**longitudinal covariates**	
Penalized Cox model	Yes	No	Hastie and Tibshirani (2004)
Landmarking	Yes	Yes	Anderson et al. (1983); Putter and van Houwelingen (2016)
Penalized regression calibration	Yes	Yes	Signorelli et al. (2021)
Multivariate functional principal component and Cox model	Yes	Yes	Li and Luo (2019)
Time-dependent Cox model	No	Yes	Fisher and Lin (1999)
Joint modelling	No	Yes	Rizopoulos (2011); Hickey et al. (2016); Mauff et al. (2020)
Functional ensemble survival tree	Yes	Yes	Jiang et al. (2021)

1.2 Structure

The remainder of this thesis is organized as follow:

- in Chapter 2 we describe the statistical methods that we consider in our comparison;
- in Chapter 3 we introduce the ADNI study and its study characteristics. We also provide an overview of the ADNI data, and describe the procedure of data preparation and data transformation;
- in Chapter 4 we describe the experimental setup used to compare the different methods, including model specification, candidate covariates, model development, validation and performance measures;
- in Chapter 5 we present the results of this study, comparing the predictive performance of the different methods;
- in Chapter 6 we interpret the results, conclude the findings and discuss the limitations of our study.

Chapter 2

Methods for the prediction of survival outcomes with longitudinal covariates

This Chapter contains an overview of various statistical methods to predict survival using longitudinal covariates that will be compared in this thesis. We start by introducing the notation and the problem framework in Section 2.1. Then we describe the penalized Cox model in Section 2.2, followed by landmarking in Section 2.3. For novel methods, we describe Penalized Regression Calibration in Section 2.4 and Multivariate Functional Principal Component Cox model in Section 2.5 respectively.

2.1 Notation, data structure and problem definition

We consider a longitudinal study involving n subjects. For subject $i \in \{1, \dots, n\}$, an event of interest can either be observed or not observed throughout the period of observation. The latter case is called censoring which may occur when the subject :

- does not experience an event before the study ends;
- is lost to follow-up;
- is withdrawn from study.

When the subject is censored, the true survival time cannot be determined. Depending on the way the true survival time is being cut off by the observation interval, the censoring can be classified as:

- right-censoring: true survival time is equal to or greater than the observed time;
- interval-censoring: true survival time is within a known time interval;
- left-censoring: true survival time is less than or equal to the observed time.

In this chapter we will focus on situations with right-censoring. Let T_i^* denote the true survival time of subject i, and T_i^C the censored time in the case of right-censoring. The observed survival time is equal to $T_i = \min(T_i^*, T_i^C)$. We denote the status indicator by $\delta_i = 1$ when an event is observed at T_i and $\delta_i = 0$ in case of censoring.

Next, we introduce two basic quantities essential to the survival analysis, the survival function and hazard function.

2.1.1 Survival function and hazard function

The survival function $S(t)$ measures the probability that a subject will survive beyond some specified time t:

$$S(t) = P(T > t). \tag{2.1}$$

The survival function is a decreasing function over time. It can be estimated using the Kaplan-Meier estimator (Kaplan and Meier, 1958), which is a nonparametric method using the product limit formula.

The hazard function $h(t)$ measures the instantaneous chance that an individual will experience the event of interest, given that the event has not occurred until time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2}$$

The hazard function is non-negative and does not have an upper bound.

The relation between the survival function and the hazard function can be derived by substituting:

$$\begin{aligned} P(t \leq T < t + \Delta t | T \geq t) &= 1 - P(T \geq t + \Delta t | T \geq t) \\ &= 1 - \frac{P(T \geq t + \Delta t)}{P(T \geq t)} \\ &= 1 - \frac{S(t + \Delta t)}{S(t)} \end{aligned}$$

into (2.2):

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1 - \frac{S(t+\Delta t)}{S(t)}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t \; S(t)} \\ &= \frac{-1}{S(t)} \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t} \\ &= \frac{-dS(t)/dt}{S(t)} \\ &= -\frac{d}{dt} \log S(t) \end{aligned}$$

By considering the cumulative hazard function:

$$H(t) = \int_0^t h(u)du, \tag{2.3}$$

the relation between $S(t)$ and $H(t)$ can then be expressed as:

$$S(t) = \exp(-H(t)), \tag{2.4}$$

and

$$H(t) = -\log S(t). \tag{2.5}$$

The cumulative hazard function can be estimated using the Nelson-Aalen estimator (Nelson, 1969, 1972; Aalen, 1978).

2.1.2 Predictors of survival

Next, we consider two groups of variables that will be used to predict survival: a vector of r baseline covariates $\mathbf{a}_i$ and a matrix of p longitudinal covariates $\mathbf{y}_i$ (indexed by $s \in \{1, \dots, p\}$) that are observed repeatedly after the baseline. A baseline covariate is time-independent variable within the analysis context, for example, subject background characteristics such as gender, ethnicity, education, etc. A longitudinal covariate could be a marker measured repeatedly that is directly or indirectly associated with the event of interest.

In a longitudinal study, the longitudinal covariates are measured repeatedly from baseline until an observed survival outcome, following either a balanced or an unbalanced design, such that for each subject $i \in \{1, \dots, n\}$, $m_i \geq 1$ repeated measurements are taken at random times t_{ij} for $j \in \{1, \dots, m_i\}$ where $t_{ij} \leq T_i$. Correspondingly, we denote y_{sij} as the value of the s the longitudinal covariate that is measured on subject i at the j-th repeated measurement, where $j \in \{1, \dots, m_i\}$. From above, $\mathbf{y}_i = \{y_{1i1}, \dots, y_{sim_i}\}$ is the matrix containing all m_i repeated measurements of all items observed for subject i from baseline to survival outcome.

The goal of the survival prediction problem is to predict the (conditional) survival probability for subject i given a set of covariates $\mathbf{a}_i$ and $\mathbf{y}_i$.

2.2 Penalized Cox Proportional Hazards model

The Cox Proportional Hazards model (Cox model or Cox PH model) (Cox, 1972) is a common method to model the effect of baseline covariates $\mathbf{a}_i$ on the survival time T. The Cox model models the hazard for subject i as:

$$h(t|\mathbf{a}_i) = h_0(t) \exp\{\mathbf{a}_i^T \boldsymbol{\beta}\}, \tag{2.6}$$

where $h_0(t)$ is the baseline hazard rate, and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\}$ is a vector of regression coefficients.

The Cox model can be estimated by maximizing the partial likelihood in low-dimensional setting ($p < n$), but this is not possible in high-dimensional setting ($p > n$). Verweij and Houwelingen (1994) proposed the approach to add a penalty function of the regression coefficients to the partial likelihood to overcome this limitation. Below we present three commonly used penalty functions.

The lasso penalty (Tibshirani, 1996), also known as l_1 penalty, is defined as:

$$p(\boldsymbol{\beta};\lambda) = \lambda \sum_{s=1}^{p} |\beta_s|, \tag{2.7}$$

where λ is a tuning parameter. The lasso penalty leads to both shrinkage and variable selection when fitting a regression model, which may result in a sparse solution i.e. some or many zero's in the estimated regression parameters. The effect of variable selection is a desirable one when it comes to inference problems because a more parsimonious model is usually preferred to model the relationship between the dependent variable and the predictor(s). However, its benefit does not extend to prediction problems because the variable selection unnecessarily disregard covariates that are potentially predictive. Moreover, the non-uniqueness of lasso estimator will lead to unstable solutions in the case of collinear covariates. Since in this thesis we are interested in prediction models, we will not consider the lasso penalty.

The ridge penalty (Hoerl and Kennard, 1970), also known as l_2 penalty, is defined as:

$$p(\boldsymbol{\beta};\lambda) = \lambda \sum_{s=1}^{p} \beta_s^2, \tag{2.8}$$

where λ is a tuning parameter. The ridge penalty leads to shrinkage: ridge estimators shrink regression parameters towards zero (but not exactly zero i.e. no variable selection effect). This property makes ridge penalty usually preferable over lasso penalty for prediction tasks, especially in the presence of many correlated covariates.

The elasticnet penalty (Park and Hastie, 2007; Simon et al., 2011) is a linear combination of the lasso penalty and ridge penalty. It is defined as:

$$p(\boldsymbol{\beta};\lambda,\alpha) = \lambda \left(\alpha \sum_{s=1}^{p} |\beta_s| + (1-\alpha) \sum_{s=1}^{p} \beta_s^2 \right), \tag{2.9}$$

where $\alpha \in [0,1]$ is a tuning parameter determining the relative weights of the lasso penalty and ridge penalty. The elasticnet penalty is equivalent to the lasso penalty when $\alpha = 1$, and the ridge penalty when $\alpha = 0$. The elasticnet penalty combines the characteristics of lasso penalty for favoring a sparse solution and characteristics of ridge penalty for scaling all regression coefficients towards zero but not exactly zero which handles correlated covariates better. As prediction is the focus in this thesis, elasticnet penalty will not be considered as the inclusion of lasso penalty term does not seem to offer any conceivable benefit over the ridge penalty solely.

The penalized Cox model can be estimated by penalized maximum likelihood using the `R` package `glmnet`. The optimal value of tuning parameter λ for the lasso or ridge penalty can be determined by cross validation, whereas the (λ, α) for the elasticnet penalty can be determined by nested cross validation.

2.3 Landmarking

The idea of landmarking in prediction context is to reduce the complexity in longitudinal covariate by fixing its values based on a given landmark time t_l. The trajectory of such covariate is represented by the last observation before the landmark time (also known as last observation carried forward, LOCF) or some summary derived from repeated observations before the landmark time (for example, the average of all observations gathered until the landmark time). The summary of the longitudinal covariates can then be used in a Cox model in time-independent fashion to estimate the hazard function conditioned on the landmark dataset which only considers considering subjects at risk at the landmark time. Landmarking does not require the modelling of the longitudinal covariate, so it has the advantage of simple implementation. The estimation of landmarking model is identical to that in Section 2.2 as the survival outcome is modelled using a Cox model or penalized Cox model.

2.4 Penalized Regression Calibration

The penalized calibration regression (PRC) method proposed by Signorelli et al. (2021) models the survival function through a penalized Cox model, using as predictors subject-specific summaries of the longitudinal covariates. It offers two variants called PRC-LMM and PRC-MLPMM that differ in the modelling techniques (univariate linear mixed models and multivariate latent process mixed models respectively) used to model the longitudinal covariates. As we only applied the PRC-LMM in this study, below we describe the three modelling steps of PRC based on the PRC-LMM formulation.

First, each longitudinal covariate $\mathbf{y}_{si}$ is modelled using a linear mixed model (LMM) with correlated random intercept and random slope specified as:

$$y_{sij} = \beta_{s0} + b_{s0i} + (\beta_{s1} + b_{s1i})a_{ij} + \epsilon_{sij}, \tag{2.10}$$

where a_{ij} denotes the age of i-th subject at the j-th visit, $b_{si} = (b_{s0i}, b_{s1i}) \backsim N_2(0, D_s)$ comprises the random intercept b_{s0i} and random slope b_{s1i} respectively, and $\epsilon_{si} \backsim N_{m_i}(0, \sigma^2_{\epsilon_s} I_{m_i})$.

Equation (2.10) can also be expressed in matrix notation as follows:

$$\mathbf{y}_{si} = \mathbf{X}_i\boldsymbol{\beta}_s + \mathbf{Z}_i\mathbf{b}_{si} + \boldsymbol{\epsilon}_{si}, \tag{2.11}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices that correspond to the fixed effects coefficients $\boldsymbol{\beta}_s$ and the random effects $\mathbf{b}_{si} \backsim N(0, D_s)$, and $\boldsymbol{\epsilon}_{si} \backsim N(0, \sigma^2_s I_{m_i})$ is a Gaussian error term.

Second, the predicted random effects $\hat{\mathbf{b}}_{si}$ are computed as summary measures of $\mathbf{y}_{si}$:

$$\hat{\mathbf{b}}_{si} = E(\mathbf{b}_{si}|\mathbf{Y}_{si} = \mathbf{y}_{si}) = \hat{\mathbf{D}}_s\mathbf{z}_i^T\hat{\mathbf{V}}_{si}^{-1}(\mathbf{y}_{si} - \mathbf{X}_i\hat{\boldsymbol{\beta}}_s), \tag{2.12}$$

where $\mathbf{V}_{si} = \mathbf{Z}_i\mathbf{D}_s\mathbf{Z}_i^T + \sigma^2_{\epsilon_s}\mathbf{I}_{m_i}$ is the marginal covariance matrix of subject i. The random intercept represents the subject-specific variability around β_{s0i} and the random slope represents the subject-specific variability in terms of rate of progression of $\mathbf{y}_{si}$.

Finally, the predicted random effects are included as predictors in a penalized Cox model alongside with relevant baseline covariates:

$$h(t_i|\mathbf{a}_i, \hat{\mathbf{b}}_{0i}, \hat{\mathbf{b}}_{1i}) = h_0(t_i) \exp\left(\sum_{k=1}^{r} \tau_k a_{ki} + \sum_{s=1}^{p} \gamma_s \hat{b}_{s0i} + \sum_{s=1}^{p} \delta_s \hat{b}_{s1i}\right), \tag{2.13}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\tau}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ are vectors of regression coefficients corresponding to baseline covariates $\mathbf{a}_i$, predicted random effects $\hat{\mathbf{b}}_{0i}$ and $\hat{\mathbf{b}}_{1i}$ respectively.

If there are r baseline covariates and p longitudinal covariates, model (2.13) will comprise $r + 2p$ covariates, which is generally a high number and potentially leads to high-dimensionality. Similar to the penalized Cox regression in Section 2.2, PRC employs penalized likelihood estimation to overcome this problem. Signorelli et al. (2021) compared PRC models using lasso penalty, ridge penalty and elasticnet penalty (see Section 2.2 for details of each penalty type), reporting that the ridge penalty and the elasticnet penalty led to similar predictive performance and outperformed the lasso penalty. The ridge penalty yielded more stable solutions over the elasticnet penalty.

The elasticnet penalty in PRC is defined as:

$$p(\gamma, \delta; \lambda, \alpha) = \lambda \left[\alpha \left(\sum_{s=1}^{p} |\gamma_s| + \sum_{s=1}^{p} |\delta_s|\right) + (1-\alpha)\left(\sum_{s=1}^{p} \gamma_s^2 + \sum_{s=1}^{p} \delta_s^2\right)\right]. \tag{2.14}$$

PRC can be estimated using the `R` package `pencal`.

2.5 Multivariate Functional Principal Component Cox model

The Multivariate Functional Principal Component Cox (MFPCCox) model is a two-stage approach proposed by Li and Luo (2019) that predicts survival with a Cox model using as predictors features extracted from longitudinal covariates through Multivariate Functional Principal Component Analysis (MFPCA).

The method assumes that the observed value y_{sij} is a noisy measurement of a latent outcome process $X_{si}(t)$, for $t \in [0, t_l)$ where t_l is the landmark time [1]. It can be expressed as follows:

$$y_{sij} = X_{si}(t_{ij}) + \epsilon_{sij}, \tag{2.15}$$

where t_{ij} is the time at j-th visit of i-th subject, ϵ_{sij} are independent measurement errors with mean zero and variances $\sigma^2_{\epsilon_q}$.

In step 1 of MFPCCox, univariate FPCA is used to extract the changing patterns in each longitudinal covariate. First, we assume the s-th latent outcome process composes of a unknown smoothed mean function $\mu_q(t)$ and covariance function $\Sigma_s(t, t') = cov\{X_{si}(t), X_{si}(t')\}$ to model the correlation between observations at any two time points. The spectral decomposition of the covariance function from Mercer's theorem (Mercer, 1909) yields:

[1] The original formulation in Li and Luo (2019) used a different definition: $t \in [0, t_{max}]$ where t_{max} is the latest survival time observed, implying that it would use all repeated measurements for model fitting regardless of the landmark time. In this thesis, we only consider the repeated measurements before the landmark time to ensure a fair and consistent comparison in the context of dynamic prediction (see Section 4.2).

$$\Sigma_s(t, t') = \sum_{l=1}^{\infty} \lambda_{sl}\phi_{sl}(t)\phi_{sl}(t'), \tag{2.16}$$

where λ_{sl} are nonincreasing eigenvalues, and $\phi_{sl}(t)$ are the corresponding orthonormal eigenfunctions.

By Karhunen-Loéve expansion, the latent outcome process introduced in (2.15) can be expressed as follows:

$$X_{si}(t) = \mu_q(t) + \sum_{l=1}^{\infty} \xi_{sil}\phi_{sl}(t), \tag{2.17}$$

where the FPC scores $\xi_{sil} \backsim N(0, \lambda_{sl})$ are uncorrelated random variables. The eigenfunctions $\phi_{sl}(t)$ represent the l-th changing pattern within the s-th longitudinal covariate; the FPC scores ξ_{sil} measure the subject-specific association with the corresponding changing pattern.

Given a fixed proportion of variance explained (PVE) $\pi \in (0, 1)$, the latent outcome process can be approximated using a finite integer l_s of components as follows:

$$X_{si}(t) \approx \mu_q(t) + \sum_{l=1}^{l_s} \xi_{sil}\phi_{sl}(t), \tag{2.18}$$

where l_s is a minimum positive integer such that $\sum_{l=1}^{l_s} \lambda_{sil} / \sum_{l=1}^{\infty} \lambda_{sil} \geq \pi$.

The FPCA above is estimated using the principal analysis by conditional estimation (PACE) algorithm, which produces the estimated mean function $\hat{\mu}_q(t)$, error variances $\hat{\sigma}_{\epsilon_q}$, covariance function $\hat{\Sigma}_q(t, t')$, eigenvalues $\hat{\lambda}_{sl}$, and eigenfunctions $\hat{\phi}_{sl}(t)$ from a set observations for the s-th longitudinal covariate.

The subject-specific FPC scores are obtained as:

$$\hat{\xi}_{sil} = \hat{\lambda}(\hat{\boldsymbol{\phi}}_{sil}^{T} \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_{si}}^{-1} (\mathbf{Y}_{si} - \hat{\mu}_{si})), \tag{2.19}$$

where $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_{si}}$ is a $J_i \times J_i$ matrix with the (j, j') entry $(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_{si}})_{j,j'} = \hat{\boldsymbol{\Sigma}}_q(t_i, t_{ij'}) + \hat{\sigma}^2_{\epsilon_q}\delta_{j,j'}$ and $\delta_{j,j'} = 1$ if $j = j'$ and $\delta_{j,j'} = 0$ if $j \neq j'$.

Under multivariate setting, the PACE algorithm is applied to each longitudinal covariate and obtain the estimated eigenfunctions and estimated FPC scores respectively, at a chosen l_s determined by the PVE. We denote a vector $\hat{\boldsymbol{\xi}}_i$ of length $l_+ = \sum_{s=1}^{p} l_s$ to contain all FPC scores over all p longitudinal covariates for the i-th subject.

In step 2 of the MFPCA, in order to account for the potential correlations between the longitudinal covariates, we consider correlations among the FPC scores estimated in previous step as a proxy to indirectly approximate the actual correlations and apply multivariate FPCA.

First, we denote $\boldsymbol{\Theta}_{n \times l_+} = \{\hat{\boldsymbol{\xi}}_1^T, \ldots, \hat{\boldsymbol{\xi}}_n^T\}$. Then, the matrix eigenanalysis of the $l_+ \times l_+$ matrix $H = (n-1)^{-1}\boldsymbol{\Theta}^T\boldsymbol{\Theta}$ gives the estimated eigenvalues $\mathbf{v}_k$ and orthonormal eigenvectors $\mathbf{c}_k$ for $k = 1, \ldots, l_+$. The estimates for the multivariate eigenfunctions for the s-th longitudinal covariate are given by:

$$\hat{\psi}_{sk}(t) = \sum_{l=1}^{l_s} [\mathbf{c}_k]_l^{(s)} \hat{\phi}_{sl}(t), \tag{2.20}$$

where $[\mathbf{c}_k]_l^{(s)}$ denote the s-th block of the orthonormal eigenvector $\mathbf{c}_k$. The multivariate eigenfunctions represents the k-th changing pattern in the s-th longitudinal covariate.

The subject-specific MFPC scores can be estimated by:

$$\hat{\rho}_{ik} = \sum_{s=1}^{p} \sum_{l=1}^{l_s} [\mathbf{c}_k]_l^{(s)} \hat{\xi}_{sil}, \tag{2.21}$$

Given a fixed proportion of variance explained (PVE) denoted by $\pi \in (0, 1)$, the p longitudinal covariates can be approximated using the first $d \leq l_+$ MFPC scores $\hat{\boldsymbol{\rho}}_i = \{\hat{\rho}_{i1}, \dots, \hat{\rho}_{id}\}$. The approximate trajectory of the s-th longitudinal covariate can be computed using (2.20) and (2.21):

$$E(Y_{si}(t)) = \hat{X}_{iq}(t) \approx \hat{\mu}_q(t) + \sum_{k=1}^{d} \hat{\rho}_{ik} \hat{\psi}_{sk}(t). \tag{2.22}$$

Finally, the survival outcome is modelled through a Cox model with the baseline covariates and the MFPC scores $\hat{\boldsymbol{\rho}}_i$ included as predictors:

$$h(t_i | \mathbf{a}_i, \hat{\boldsymbol{\rho}}_i) = h_0(t_i) \exp\{\mathbf{a}_i^T \boldsymbol{\tau} + \hat{\boldsymbol{\rho}}_i^T \boldsymbol{\beta}\}, \tag{2.23}$$

where $h_0(t)$ is the baseline hazard function, and $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ are two parameter vectors. The Cox model can be estimated using maximum likelihood.

The MFPCCox model can be estimated using the scripts available at `https://github.com/kan-li/MFPCCox` (Li and Luo, 2019).

Chapter 3

Data description and manipulation

This Chapter comprises an overview of the Alzheimer's Disease Neuroimaging Initiative (ADNI) data and a description of the data manipulation steps taken before modelling. We start by introducing the ADNI study, especially its data collection and study characteristics, in Section 3.1. Data screening to select a subset of subjects suitable for modelling is elaborated in Section 3.2. Data preparation to arrange the ADNI data into formats suitable for modelling is described in Section 3.3. Then, we will explore the survival data and the longitudinal data in Section 3.4 and discuss the missingness in Section 3.5. Lastly, we will describe the data transformation in Section 3.6.

3.1 ADNI Data

Dementia is a syndrome that leads to a degeneration in cognitive function that is more severe than the one due to mere biological ageing. Various diseases and injuries that primarily or secondarily affect the brain may lead to dementia, usually of a chronic and progressive nature (WHO, 2021). Alzheimer's disease (AD) is a specific brain disease that contribute to most dementia cases (between 60 and 70% of global dementia cases according to estimates of the World Health Organization [1]).

The ADNI study (Weiner et al., 2010) is an ongoing multi-phase prospective longitudinal study commenced in 2004 that was designed to identify / validate biomarkers related to progression of AD. The general goals of the ADNI study are: (i) to improve AD detection at the earliest possible stage and identify biomarkers to track the progression of AD; (ii) to support advances in the intervention, prevention and treatment of AD; and (iii) to continually maintain the ADNI's data-accessibility to facilitate scientific research.

The ADNI study consists of 4 subsequent phases of data collection, respectively called ADNI1, ADNIGO, ADNI2 and ADNI3, which were tasked with distinct research goals regarding biomarkers associated with progression / prediction of AD. The first and the last phase began in 2004 and 2016 respectively; each phase typically lasted for 5 years. During the follow-up visits within each phase, the participants were assessed on dementia along with a variety of cognitive assessments, biospecimen sampling and/or brain imaging analysis according to the assigned data collection protocol. The visit schedule generally adhered to the following n-month intervals: 0, 3, 6, 12, 18,

[1] Source: `https://www.who.int/news-room/fact-sheets/detail/dementia`

24, 36, 48 and onward annually. When possible, participants who who entered the study in an earlier phase (e.g. ADNI1) were also carried forward to later phases for continual monitoring. Note that the exact data collection protocols differed among phases as certain month or certain type of measurement were sometimes skipped, resulting in an unbalanced dataset with irregular observation times when data from all phases are combined in single analysis.

Across the four phases, a total of 2,379 participants aged between 55 and 90 were recruited from 57 research centers in the USA and the Canada. They were either diagnosed as (i) cognitive normal (CN), (ii) with mild cognitive impairment (MCI) (as early MCI or late MCI), or (iii) with dementia in the initial assessment, representing different stages of cognitive decline.

The data collection protocol with respect to each phase is shown in Figure 3.1. Overall, the ADNI study provides an extensive set of heterogeneous, longitudinal data on clinical, cognitive, imaging, genetic and biochemical markers.

The ADNI repository (`http://adni.loni.usc.edu/`) provides various types of dataset including case reports, biomarker lab summaries, imaging data, sequencing data etc. For this thesis, we use the `adnimerge` dataframe in the `R` package `ADNIMERGE` obtained from the repository that is merged from multiple datasets to include several key variables across all phases and all data types. The data used in this thesis are based on the version retrieved on 30 January 2022.

In our analysis, we focus on the progression to dementia for participants initially diagnosed as CN or MCI. The survival outcome is the time until a dementia diagnosis. The baseline visit ($t = 0$) is set as the time of enrollment. In reality, the true time to dementia is likely to take place between two consecutive visits, thus it cannot be measured exactly and results in interval censoring. But for simplicity we treat the survival times as right censored when applying the methods in Chapter 2. The censoring may be due to not developing dementia up to last visit, loss to follow-up, or competing risk such as death. The modelling of competing risks is beyond the scope of this thesis.

The variables with potential predictive value can be categorized into two groups, namely baseline covariates and longitudinal covariates, depending on whether they are measured repeatedly (values are time-dependent). The selection of candidate covariates for modelling will be reported in Section 4.1.

3.2 Data screening

Before modelling, we applied data screening to exclude participants that are not of modelling interest or lack sufficient information for modelling. Figure 3.2 summarizes the screening process based on three exclusion criteria:

- participants already diagnosed with dementia at baseline are excluded because they are not in the at-risk group;
- participants with missing values in baseline covariates are excluded for complete case analysis, as the methods to be compared here don't automatically account for such missingness, and imputation of missing values is beyond the goal of this thesis;
- participants without any event status (CN/MCI/dementia) available in follow-up are excluded due to lack of response value.

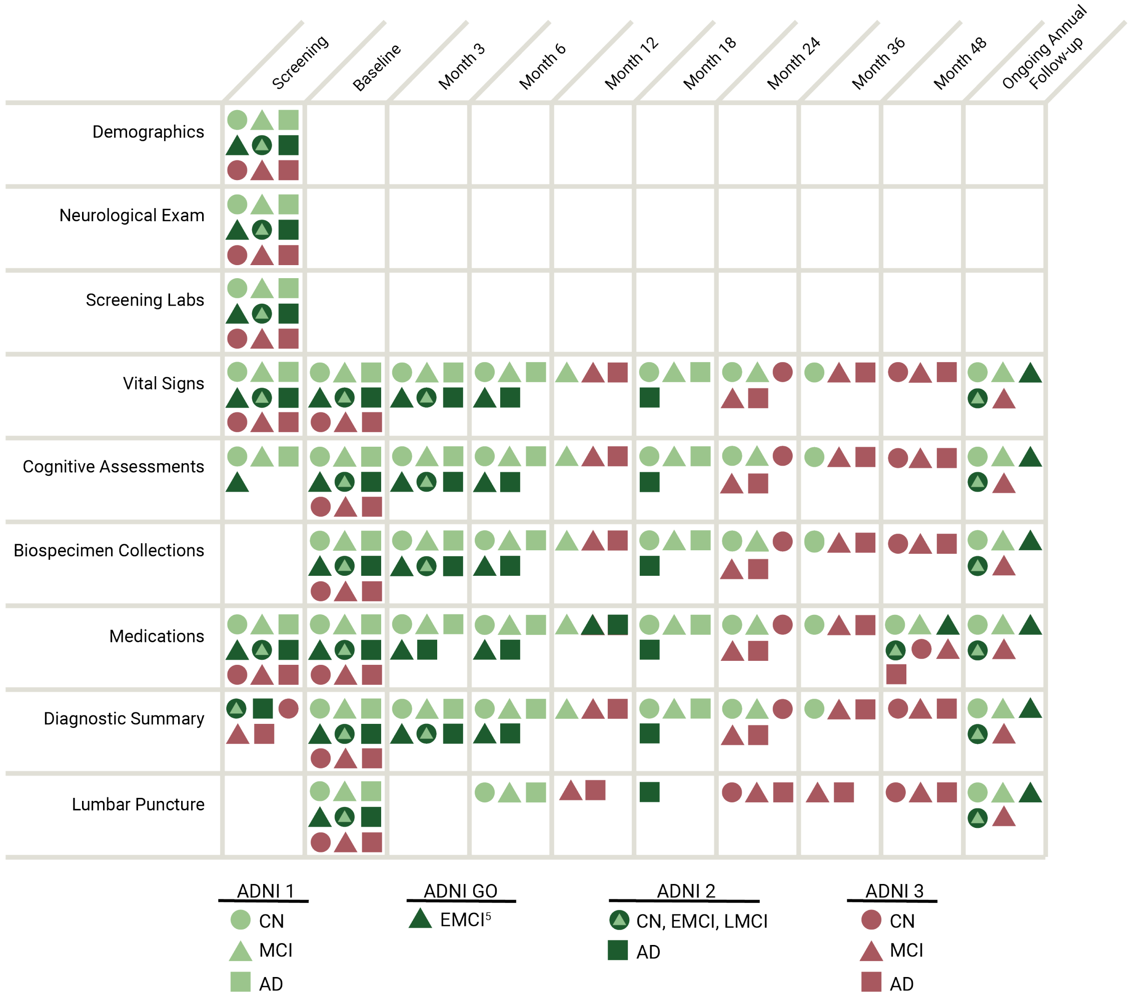


Figure 3.1: ADNI data collection protocol (source: `https://adni.loni.usc.edu/study-design/`).

After data screening, the data eligible for modelling and analysis consists of 1,615 subjects and 9,758 unique visits. In the final data, the average follow-up period is 4.1 years, the average number of visits is 6.0 and the average event rate is 25%. The study characteristics are summarized in Table 3.1.

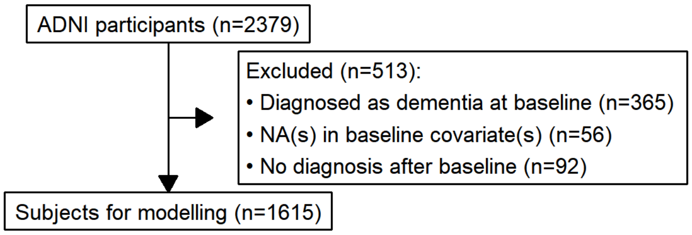

ADNI participants (n=2379)
Excluded (n=513):
• Diagnosed as dementia at baseline (n=365)
• NA(s) in baseline covariate(s) (n=56)
• No diagnosis after baseline (n=92)
Subjects for modelling (n=1615)

Figure 3.2: Eligibility for analysis. The following criteria have been applied to filter ADNI participants before they are included in this study: (i) made more than one visit; (ii) baseline diagnosis is available; (iii) baseline diagnosis is cognitive normal (CN) / mildly cognitively impaired (MCI), but not dementia; (iv) diagnosis after baseline is available; (v) baseline covariates (age, gender, education, number of APOE4 alleles) are available.

Table 3.1: Study characteristics in final data. Subjects are grouped by their original data collection protocol i.e. phase. Statistics are computed based on data after cleaning and screening, and only consider subjects diagnosed as CN or MCI at baseline.

	ADNI1	**ADNIGO**	**ADNI2**	**ADNI3**	**Combined**
	n = 600	*n = 114*	*n = 603*	*n = 298*	*n = 1615*
Start date	Oct 2004	Sep 2009	Sep 2011	Sep 2016	-
Duration (years)	5	2	5	5	-
Follow-up period (years)					
Mean	4.6	5.4	4.4	2.1	4.1
SD	3.9	2.9	2.8	0.8	3.2
Median	3.0	5.0	4.0	2.0	3.0
Range	0.5-15.7	0.5-11.2	0.4-10.5	0.9-4.3	0.4-15.7
Number of visits					
Mean	7.5	8.5	6.2	1.9	6.0
SD	5.5	3.2	2.8	0.7	4.4
Median	6	9	6	2	5
Range	1-22	2-14	1-13	1-4	1-22
Event rate	0.40	0.18	0.20	0.06	0.25

3.3 Data preparation

Before proceeding to modelling the data from the ADNI study, we need to arrange the data from the `ADNIMERGE` package in a format that is suitable for our analyses. This requires the creation of two datasets from the ADNI data:

- `surv` dataset: a dataset in wide format that contains the diagnosis (event status) and survival time of each subject in each row, accompanied with a set of baseline/time-independent covariates;

- `long` dataset: a dataset in long format that contains the measurements of longitudinal/time-dependent covariates from a single visit of a subject in each row, up to the last visit before a survival outcome (dementia/censoring) is determined.

The preparation of the raw `adnimerge` data into formatted `surv` and `long` datasets are illustrated in the examples for two example subjects with different outcomes (identified by the subject identifiers `RID=5` and `RID=41`) presented in Tables 3.2, 3.3 and 3.4 respectively. Note that only repeated measurements before the diagnosis of dementia are retained in the longitudinal data for the participant with `RID=41`. In the `adnimerge` data, each row represents a single visit that is uniquely identified by the date of visit (`EXAMDATE`) and contains the corresponding baseline covariates e.g. age (`AGE`) and gender (`PTGENDER`) and longitudinal covariates e.g. cognitive assessment (`ADAS13`). Besides the exact date, the time of visit is also represented in visit code (`VISCODE`) with respect to the study schedule and in a continuous time variable years since baseline (`Years.bl`). The `DX.bl` and `DX` records the diagnosis result at baseline and subsequent visits. To prepare the `surv` dataset, the survival time `time` and survival outcome `event` are obtained by finding the earliest time to dementia (encoded as `event=1`) or the latest time without dementia (i.e. censored, encoded as `event=0`) in a subject. For the `long` dataset, it is essential to keep the subject identifier `RID`, observation time `Years.bl`, baseline covariates and longitudinal covariates e.g. `ADAS13`. The observations made at or after the survival time is excluded. Note that the `long` dataset will be further formatted into a 3-dimensional array $\{$*number of subjects* $\times$ *observation times* $\times$ *number of longitudinal covariates*$\}$ for the application of MFPCCox described in Section 2.5.

Table 3.2: Example of raw data from the ADNI study. The first participant (RID = 5) entered the study as cognitively normal (DX=CN) and became censored at the 7th visits. The second participant (RID = 41) entered the study as mildly cognitively impaired (DX=MCI) and was diagnosed with dementia at the 4th visit.

id	RID	VISCODE	EXAMDATE	Years.bl	DX.bl	DX	AGE	PTGENDER	ADAS13	MMSE
1	5	bl	09/07/2005	0.000	CN	CN	73.7	Male	14.7	29
1	5	m06	03/09/2006	0.501	CN	CN	73.7	Male	15.0	29
1	5	m12	09/05/2006	0.994	CN	CN	73.7	Male	*NA*	30
1	5	m18	03/09/2007	1.500	CN	NA	73.7	Male	*NA*	*NA*
1	5	m24	09/07/2007	1.999	CN	CN	73.7	Male	11.0	29
1	5	m30	05/02/2008	2.650	CN	NA	73.7	Male	*NA*	*NA*
1	5	m36	09/10/2008	3.009	CN	**CN**	73.7	Male	11.7	30
2	41	bl	11/14/2005	0.000	LMCI	MCI	70.9	Female	28.3	25
2	41	m06	05/15/2006	0.498	LMCI	MCI	70.9	Female	25.7	25
2	41	m12	11/13/2006	0.997	LMCI	MCI	70.9	Female	27.0	24
2	41	m18	05/14/2007	1.495	LMCI	**Dementia**	70.9	Female	32.3	24
2	41	m24	11/07/2007	1.979	LMCI	Dementia	70.9	Female	30.0	24
2	41	m30	05/12/2008	2.491	LMCI	NA	70.9	Female	*NA*	*NA*
2	41	m36	11/12/2008	2.995	LMCI	Dementia	70.9	Female	35.3	23
2	41	m48	01/14/2010	4.167	LMCI	Dementia	70.9	Female	41.7	17

Table 3.3: `surv` dataset derived from Table 3.2

id	RID	time	event	status.bl	AGE	PTGENDER	ADAS13	MMSE
1	5	3.01	0	CN	73.7	Male	14.7	29
2	41	1.49	1	MCI	70.9	Female	28.3	25

Table 3.4: `long` dataset derived from Table 3.2. Note that some measurements of ADAS13 and MMSE are missing for the first participant.

id	RID	VISCODE	Years.bl	status.bl	AGE	PTGENDER	ADAS13	MMSE
1	5	bl	0.000	CN	73.7	Male	14.7	29
1	5	m06	0.501	CN	73.7	Male	15.0	29
1	5	m12	0.994	CN	73.7	Male	*NA*	30
1	5	m18	1.500	CN	73.7	Male	*NA*	*NA*
1	5	m24	1.999	CN	73.7	Male	11.0	29
1	5	m30	2.650	CN	73.7	Male	*NA*	*NA*
2	41	bl	0.000	MCI	70.9	Female	28.3	25
2	41	m06	0.498	MCI	70.9	Female	25.7	25
2	41	m12	0.997	MCI	70.9	Female	27.0	24

3.4 Data exploration

In this section, we provide some descriptive summaries to explore the final data after data screening.

As described in Section 3.2, our analysis focuses on 1,615 individuals who were dementia free upon enrollment (participants diagnosed with dementia at enrollment were excluded during the data screening and are not described here). There were 398 participants diagnosed with dementia in follow-up visits and 1,217 participants censored. The overall incidence rate of dementia was 25%. Figure 3.3 shows the the cumulative number of events and censored observations, and the number of subjects still at risk at every year after baseline. The number of subjects still at risk drops from 1,615 at baseline to less than 500 in the fifth year from baseline.

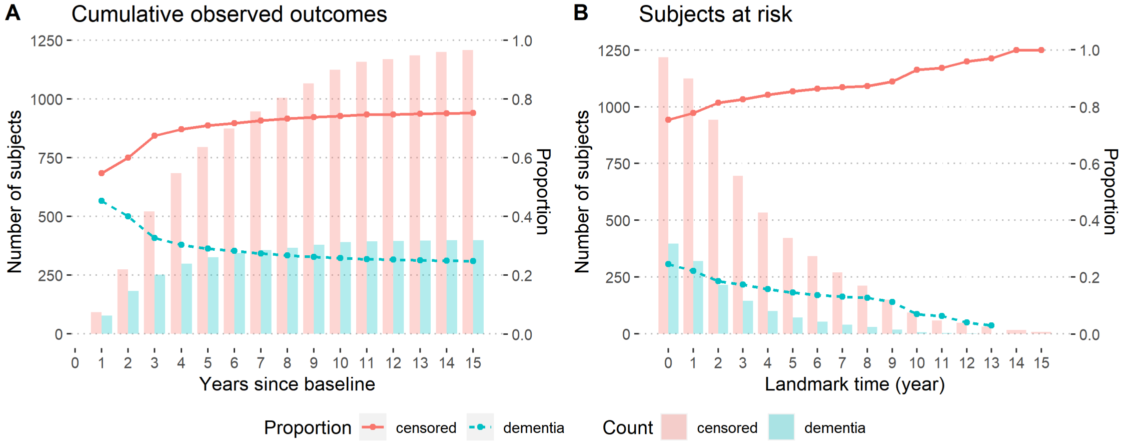


Figure 3.3: Panel A (left) shows the cumulative number of events and censored observations observed during the study. The overall incidence rate of dementia was approximately 25%. Panel B (right) shows the number of subjects still at risk after 0, 1, 2, ..., 15 years from baseline.

The age at enrollment ranged from 55.0 to 91.4 years, with an average of 73.2 years. For subjects diagnosed with dementia, the mean time to a dementia diagnosis was 3 years, whereas for censored subjects, the mean censoring time was 4.5 years. The age at which dementia diagnosis were made was 77.0 years on average. The eldest subject with dementia was diagnosed with dementia at 94.7 years old, after entering the study at 84.7 years old.

3.4.1 Baseline covariates

Based on the previous literature on dementia risk modelling, we consider the age, gender, baseline diagnosis, number of apolipoprotein $\varepsilon 4$ (APOE e4) allele and education received as baseline covariates to stratify the population. The summary of baseline characteristics of the subjects is tabulated in Table 3.5. Among the combined cohort from all ADNI phases consisting of 865 males and 750 females; 237 (27%) males and 161 (21%) females developed dementia.

At the time of enrollment i.e. baseline, there were 922 subjects diagnosed as mild cognitively impaired (MCI) and 693 subjects diagnosed as cognitively normal (CN). The former subgroup showed a higher incidence rate of dementia (39%) than the latter subgroup (5%). This difference is in line with our expectations, since the MCI is an intermediate / transitional stage between the cognitive decline of normal aging and the more severe cognitive decline in dementia.

The $\epsilon 4$ allele of the apolipoprotein gene is regarded as a strong genetic risk factor for the development of AD: the possession of one or two $\epsilon 4$ alleles is estimated to increase the risk of developing AD by 3 and more than 10 folds respectively (Suzuki et al., 2020). The ADNI study measured the number (either 0, 1, or 2) of $\epsilon 4$ allele possessed by a subject. A remarkable difference in incidence rate can be observed in the ADNI data:

- incidence rate for subjects without $\epsilon 4$ allele ($n = 941$) = 16%;
- incidence rate for subjects with one $\epsilon 4$ allele ($n = 552$) = 35%;
- incidence rate for subjects with two $\epsilon 4$ alleles ($n = 122$) = 44%.

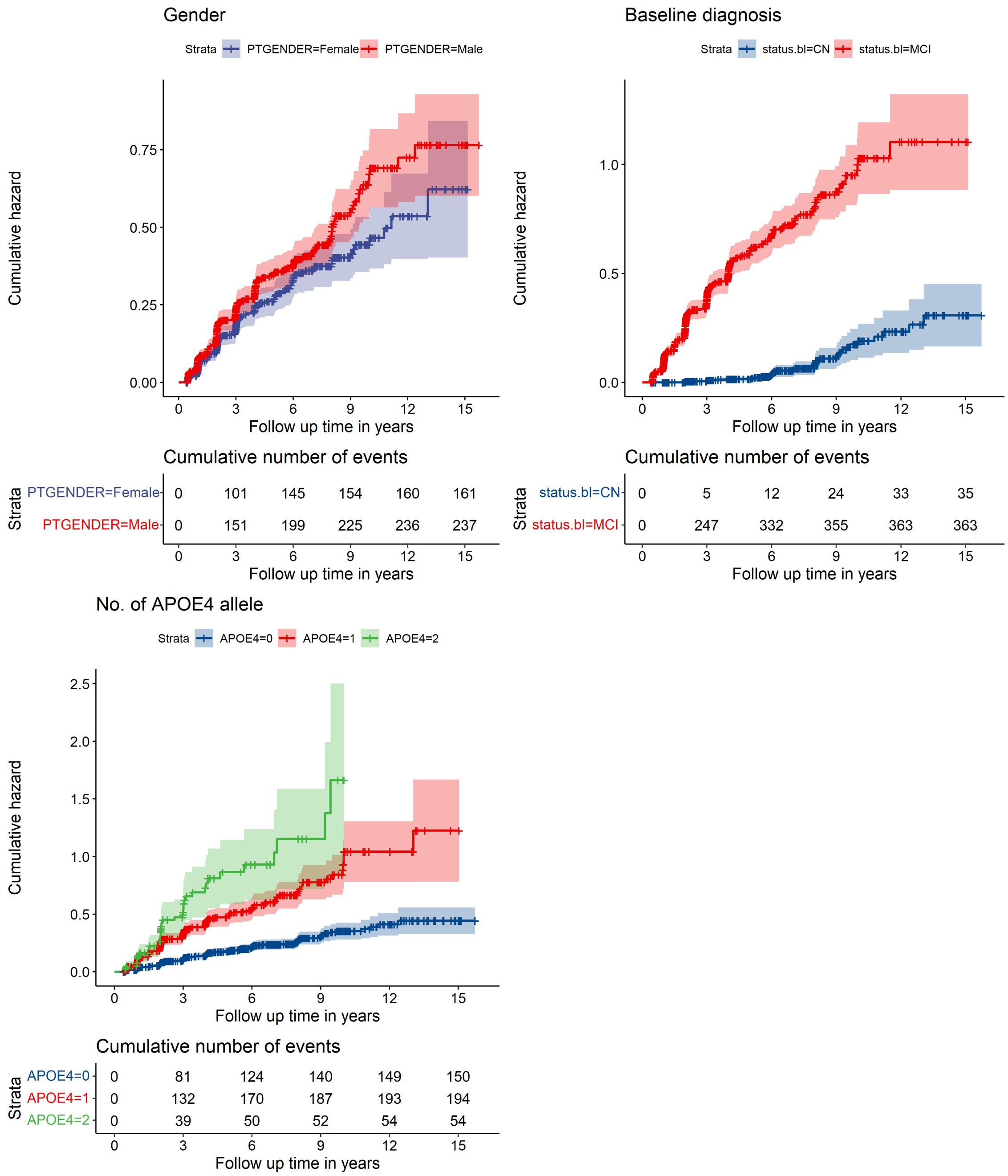
The cumulative events and cumulative hazards estimated by Kaplan-Meier estimator stratified by gender, baseline diagnosis and number of APOE e4 allele are illustrated in Figure 3.4.

It should be noted that the race and ethnicity in the study population are predominantly white (93%), therefore the models developed based on the ADNI data should be externally validated first to truly evaluate its generalizability.

Table 3.5: Descriptive Statistics of baseline characteristics ($N = 1615$)

	ADNI1	ADNIGO	ADNI2	ADNI3	Combined
	$N = 600$	$N = 114$	$N = 603$	$N = 298$	$N = 1615$
Diagnosis : CN	37% (220)	1% (1)	47% (282)	64% (190)	43% (693)
MCI	63% (380)	99% (113)	53% (321)	36% (108)	57% (922)
Age[1] (Years)	71.3 75.4 79.8	67.1 71.7 77.5	67.5 72.1 77.1	66.7 70.5 75.9	68.3 73.2 78.0
Sex : Male	60% (361)	54% (61)	51% (310)	45% (133)	54% (865)
Education	14 16 18	14 16 18	14 16 18	16 16 18	14 16 18
Ethnicity : Unknown	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Not Hisp/Latino	98% (586)	95% (108)	97% (583)	96% (286)	97% (1563)
Hisp/Latino	2% (14)	5% (6)	3% (20)	4% (12)	3% (52)
Race : Am Indian/Alaskan	0% (1)	1% (1)	0% (1)	0% (0)	0% (3)
Asian	2% (12)	1% (1)	1% (9)	1% (3)	2% (25)
Hawaiian/Other PI	0% (0)	0% (0)	0% (1)	0% (0)	0% (1)
Black	5% (29)	1% (1)	4% (26)	4% (11)	4% (67)
White	93% (557)	94% (107)	93% (559)	93% (276)	93% (1499)
More than one	0% (1)	4% (4)	1% (7)	3% (8)	1% (20)
Unknown	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Marital : Divorced	7% (41)	11% (13)	12% (73)	10% (30)	10% (157)
Married	76% (457)	77% (88)	72% (436)	80% (238)	75% (1219)
Never married	3% (17)	3% (3)	4% (27)	2% (7)	3% (54)
Widowed	14% (85)	9% (10)	11% (67)	8% (23)	11% (185)
Number of APOEe4 alleles : 0	56% (335)	58% (66)	59% (357)	61% (183)	58% (941)
1	36% (214)	35% (40)	33% (202)	32% (96)	34% (552)
2	8% (51)	7% (8)	7% (44)	6% (19)	8% (122)

[1] a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables. Numbers after proportions are frequencies.



Cumulative number of events

Strata	0	3	6	9	12	15
PTGENDER=Female	0	101	145	154	160	161
PTGENDER=Male	0	151	199	225	236	237

Follow up time in years

Cumulative number of events

Strata	0	3	6	9	12	15
status.bl=CN	0	5	12	24	33	35
status.bl=MCI	0	247	332	355	363	363

Follow up time in years

Cumulative number of events

Strata	0	3	6	9	12	15
APOE4=0	0	81	124	140	149	150
APOE4=1	0	132	170	187	193	194
APOE4=2	0	39	50	52	54	54

Follow up time in years

Figure 3.4: Cumulative hazard of dementia derived from the Kaplan–Meier method by gender (top right), baseline diagnosis (top right), and APOE e4 (bottom left).

3.4.2 Longitudinal covariates

The adnimerge dataset contains a large number of variables with repeated measurements: after screening, we identified 41 candidate longitudinal covariates. They comprise a wide variety of repeated measurements of cognitive, imaging, and biochemical markers. We will further discuss the final inclusion of these covariates in Section 3.5 as its missingness has an implication on the model estimation.

Here we present two examples of longitudinal covariates: (i) ADAS13, one of the cognitive assessments, and (ii) volume of middle temporal gyrus, one of the neuroimaging biomarkers, in Figure 3.5 to illustrate the different characteristics in the heterogeneous ADNI data. In general, we found the longitudinal data to be highly unbalanced, which requires flexibility when considering the modelling approach.

The Alzheimer's Disease Assessment Scale (ADAS) (Mohs et al., 1997; Kueper et al., 2018) was used to evaluate cognitive impairment in the assessment of AD. ADAS13 refers to the 13-items version of the ADAS cognitive subscale, which range from 0 to 85 and a higher score correspond to a worse performance. It involves both subject-completed tests and observer-based assessments to assess the cognitive domains of multiple cognitive domains including memory, language, praxis, orientation, executive functioning, and functional ability. The spaghetti plot of ADAS13 in Figure 3.5 highlights the repeated observation of ADAS13 in follow-up visits for a random subset of subjects. The subjects eventually developed into dementia tend to display higher ADAS13 scores at baseline and increased in follow-ups.

As one of the neuroimaging biomarkers in ADNI, the volume of middle temporal gyrus (MidTemp), which is a gyrus located on the temporal lobe of the brain and associated with certain cognitive domains, is measured by structural magnetic resonance imaging (sMRI). Some differences can be observed in the spaghetti plot of MidTemp when compared with that of ADAS13 (Figure 3.5): first, fewer repeated observations and trajectories could be seen in MidTemp; second, the difference between subjects with and without dementia in MidTemp is not pronounced as in ADAS13; last, the scale of measurement is 100 times greater in MidTemp than ADAS13, which suggests scaling of covariates might bring numerical stability to computation.

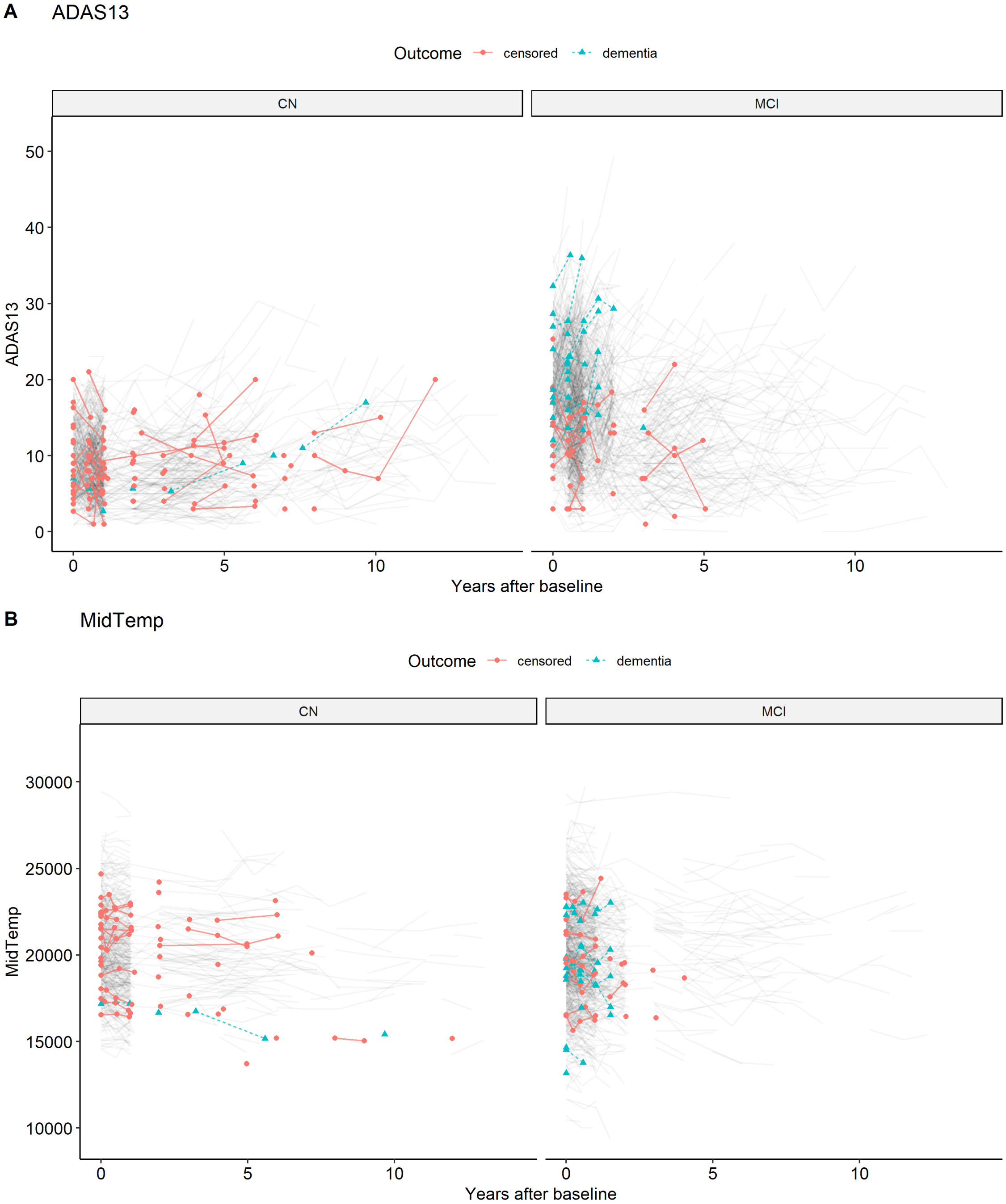


Figure 3.5: Spaghetti charts of longitudinal covariates ADAS13 (top) and middle temporal gyrus (MidTemp) (bottom) from 50 random subjects, grouped by baseline status CN (left) and MCI (right). Subjects with dementia are highlighted in blue; subjects with censoring are highlighted in red. Trajectories of remaining subjects are displayed in grey.

3.5 Missing data

Due to the data screening in Section 3.2, there is no missingness in the `surv` dataset that contains the survival time, the censoring status and the baseline covariates. On the other hand, the `long` dataset displays missingness patterns that differ across the longitudinal covariates as shown in Figure 3.6. This is mainly due to the difference in data collection protocols from different ADNI phases (see Figure 3.1).

Table 3.6 shows the the proportion of subjects without any observations for each of the 41 longitudinal covariates. We observe that the proportion ranges from 0 to 0.95. We exclude 20 covariates for which no information was available for more than 10% of the subjects. Only 14 covariates have at least one observation per subject. For the 21 longitudinal variables selected for modelling, the average number of observations per subject ranged between 2.88 and 3.35.

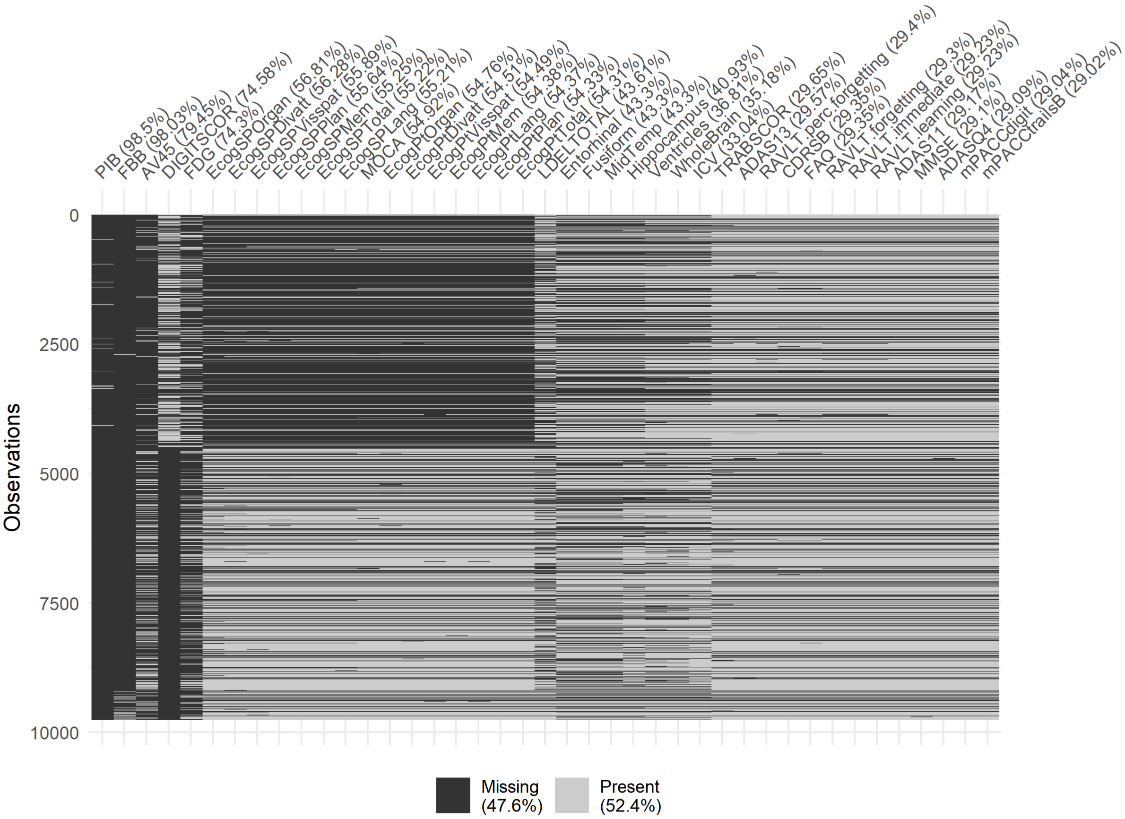


Figure 3.6: The missing values in longitudinal covariates at each visit from the `long` dataset are indicated in black. The overall percentage of missing values per covariate are displayed in brackets. 21 out of 41 longitudinal covariates have less than 50% missing values across all visits. The covariates PIB and FBB have the least measurements (98% missing values).

Table 3.6: Proportion of missing information in `long` dataset. Longitudinal covariates with up to 10% missingness were selected to be used as predictors of survival alongside the baseline covariates described in Section 4.1.

	Covariate	Proportion of subjects without any measurement	Average number of observation per subject	Included in model?
1	CDRSB	0.00	3.10	Yes
2	ADAS11	0.00	3.11	Yes
3	ADASQ4	0.00	3.11	Yes
4	MMSE	0.00	3.11	Yes
5	RAVLT.immediate	0.00	3.11	Yes
6	RAVLT.learning	0.00	3.11	Yes
7	RAVLT.forgetting	0.00	3.11	Yes
8	RAVLT.perc.forgetting	0.00	3.11	Yes
9	LDELTOTAL	0.00	2.13	Yes
10	FAQ	0.00	3.10	Yes
11	mPACCdigit	0.00	3.11	Yes
12	mPACCtrailsB	0.00	3.11	Yes
13	ADAS13	0.00	3.10	Yes
14	TRABSCOR	0.00	3.09	Yes
15	ICV	0.01	3.35	Yes
16	WholeBrain	0.02	3.27	Yes
17	Ventricles	0.02	3.18	Yes
18	Hippocampus	0.04	2.96	Yes
19	Entorhinal	0.05	2.88	Yes
20	Fusiform	0.05	2.88	Yes
21	MidTemp	0.05	2.88	Yes
22	FDG	0.29	1.29	No
23	EcogPtPlan	0.35	1.78	No
24	EcogPtOrgan	0.35	1.76	No
25	EcogPtTotal	0.35	1.78	No
26	MOCA	0.36	1.77	No
27	EcogPtMem	0.36	1.78	No
28	EcogPtLang	0.36	1.78	No
29	EcogPtVisspat	0.36	1.77	No
30	EcogPtDivatt	0.36	1.77	No
31	EcogSPLang	0.36	1.76	No
32	EcogSPMem	0.36	1.75	No
33	EcogSPPlan	0.36	1.74	No
34	EcogSPTotal	0.36	1.76	No
35	EcogSPDivatt	0.36	1.72	No
36	EcogSPVisspat	0.36	1.73	No
37	EcogSPOrgan	0.36	1.69	No
38	AV45	0.45	0.70	No
39	DIGITSCOR	0.65	1.33	No
40	FBB	0.91	0.10	No
41	PIB	0.95	0.07	No

3.6 Data transformation

As PRC rely on LMMs that assume normality and several longitudinal covariates are highly skewed, appropriate transformation was applied to reduce the skewness of these covariates. As shown in Table 3.7, the skewness ranged from -10.7 to 2.5. We transformed the covariates with skewness < -0.5 using cubic transformations and transform the covariates with skewness > 0.5 using $\log_{10}$ function. Finally, we scale the covariates to zero mean and unit variance to improve numerical stability in computation.

Table 3.7: Transformation of the longitudinal covariates, and skewness before and after the transformation

	Covariate	Skewness	Transformation	Skewness
1	RAVLT.perc.forgetting	-10.671	$(\text{RAVLT.perc.forgetting}+1650)^3$	-1.297
2	MMSE	-1.781	$(\text{MMSE})^3$	-0.999
3	mPACCtrailsB	-0.728	$(\text{mPACCtrailsB}+31)^3$	0.041
4	mPACCdigit	-0.688	$(\text{mPACCdigit}+29)^3$	0.151
5	RAVLT.forgetting	-0.347	-	-
6	Hippocampus	-0.127	-	-
7	LDELTOTAL	-0.124	-	-
8	Entorhinal	0.009	-	-
9	MidTemp	0.075	-	-
10	RAVLT.learning	0.093	-	-
11	WholeBrain	0.113	-	-
12	Fusiform	0.193	-	-
13	RAVLT.immediate	0.330	-	-
14	ADASQ4	0.442	-	-
15	ICV	0.556	$\log_{10}(\text{ICV})$	-0.061
16	ADAS13	0.772	$\log_{10}(\text{ADAS13}+1)$	-0.685
17	ADAS11	1.072	$\log_{10}(\text{ADAS11}+1)$	-0.487
18	Ventricles	1.292	$\log_{10}(\text{Ventricles})$	-0.139
19	CDRSB	1.364	$\log_{10}(\text{CDRSB}+1)$	0.524
20	TRABSCOR	1.970	$\log_{10}(\text{TRABSCOR}+1)$	-0.25
21	FAQ	2.515	$\log_{10}(\text{FAQ}+1)$	1.058

Chapter 4

Statistical modelling

In this Chapter, we describe several aspects involved in the implementation of the different prediction approaches compared in our study. First, in Section 4.1 we elaborate on the model building procedures including the model specification and model hyperparameters. Second, in Section 4.2 we describe the problem of dynamic prediction of survival. Then, we describe the interval validation procedure in Section 4.3 and specify performance measures for evaluation in Section 4.4. Lastly, in Section 4.5 we provide information on the implementation of the different methods using R.

4.1 Model development

Initially, we identified 46 candidate predictors, of which 5 are time-independent (Section 3.4.1) and 41 are longitudinal (Section 3.5) in the ADNI data. As discussed in Section 3.5 and shown in Table 3.6, we decided to retain as predictors 21 longitudinal covariates for which at least one measurement is available for 90% or more of the individuals, removing the remaining 20 covariates with more missingness and less repeated measurements.

This preliminary screening led to the selection of 26 variables, listed in Table 4.1, to be used as predictors of time to dementia. Steyerberg (2009) recommended that clinical prediction models for survival outcomes should be developed bearing in mind two rules of thumb: (i) at least 100 events; (ii) at least 10 events per variable (EPV) and preferably 20 if the event rate is lower than 20%. Under 10-fold cross validation and without landmarking, the training sets will contain $1615 \times 90\% \times 25\% \approx 363$ events on average, resulting in approximately 14 EPV. Note that the EPV is not a constant value in this study for two reasons: (i) the EPV decreases as the landmark time is increased; (ii) the EPV also depends on the methods used, for instance the MFPCCox could reduce the number of covariates fed to the Cox model and the PRC could roughly double the number covariates fed to the penalized Cox model, resulting in upward and downward adjustment to the EPV. As pointed out in Steyerberg (2009), medical prediction models that are constructed with $\text{EPV} < 10$ are commonly overfitted. But provided that the PRC already employs regularization and that all models were internally validated using repeated cross-validation, the risk of overfitting in lower EPV is probably alleviated.

Hereafter we describe some relevant implementation details for the four methods included in our comparison.

Table 4.1: List of covariates

Baseline covariate	**Description**	**Format**	**Type**
AGE	Age	Continuous	Background
APOE4	Number of APOEe4 alleles	Categorical	Genetic
PTEDUCAT	Education	Discrete	Background
PTGENDER	Sex	Categorical	Background
status.bl	Diagnosis at baseline	Categorical	Background
Longitudinal covariate	**Description**	**Format**	**Type**
ADAS11	ADAS[1]11	Continuous	Cognitive
ADAS13	ADAS[1]13 (including Delayed Word Recall and Number Cancellation)	Continuous	Cognitive
ADASQ4	ADAS[1]Delayed Word Recall	Discrete	Cognitive
CDRSB	CDR-SB[2]	Discrete	Cognitive
Entorhinal	UCSF[3]Entorhinal	Continuous	Imaging
FAQ	FAQ	Discrete	Cognitive
Fusiform	UCSF Fusiform	Continuous	Imaging
Hippocampus	UCSF Hippocampus	Continuous	Imaging
ICV	UCSF ICV	Continuous	Imaging
LDELTOTAL	Logical Memory - Delayed Recall	Discrete	Cognitive
MidTemp	UCSF Middle temporal gyrus	Continuous	Imaging
MMSE	MMSE	Discrete	Cognitive
mPACCdigit	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Digit Symbol Substitution	Continuous	Cognitive
mPACCtrailsB	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B	Continuous	Cognitive
RAVLT.forgetting	RAVLT Forgetting (trial 5 - delayed)	Discrete	Cognitive
RAVLT.immediate	RAVLT Immediate (sum of 5 trials)	Discrete	Cognitive
RAVLT.learning	RAVLT Learning (trial 5 - trial 1)	Continuous	Cognitive
RAVLT.perc.forgetting	RAVLT Percent Forgetting	Continuous	Cognitive
TRABSCOR	Trails B	Continuous	Cognitive
Ventricles	UCSF Ventricles	Continuous	Imaging
WholeBrain	UCSF WholeBrain	Continuous	Imaging

[1] ADAS: Alzheimer's Disease Assessment Scale

[2] CDR-SB: Clinical Dementia Rating scale Sum of Boxes

[3] UCSF: University of California, San Francisco

4.1.1 Penalized Cox model (with baseline measurements)

We specified a penalized Cox model as the baseline model (pCox-baseline) by simply using measurements at the baseline. In addition to the baseline covariates, we also included the baseline values of the longitudinal covariates to ensure fair comparison across methods. Under the assumption of missing at random, we employed mean imputation to impute missing values. The model was specified with 26 candidate predictors and regularized using the ridge penalty (2.2). The optimal ridge penalty was determined based on the λ_{min} i.e. the optimal value of penalty parameter λ giving minimum mean CV error under the inner cross validation (CV) procedure during model fitting.

4.1.2 Landmarking

The landmarking model (pCox-landmarking) was implemented using a penalized Cox model that included the same 26 variables as pCox-baseline. However, instead of using the baseline values, we employed the LOCF method for the 21 longitudinal covariates, using the last observation available before the landmark time.

4.1.3 Penalized Regression Calibration

The step 1 of the PRC involved fitting a LMM with random slope and intercept on baseline age for each of the 21 longitudinal covariates. The summaries of these trajectories were extracted in step 2 and used as predictors along with the 5 baseline covariates to fit the penalized Cox model in step 3. All predictors except baseline age were regularized with ridge penalty determined by the λ_{min}.

4.1.4 Multivariate Functional Principal Component Cox model

Differently from the other models, it was not possible to estimate MFPCCox using all 21 longitudinal covariates due to estimation problems caused by the following 7 variables: ICV, WholeBrain, Ventricles, Hippocampus, Entorhinal, Fusiform, and MidTemp. For this reason, such variables were not included as covariates, and MFPCCox was estimated using a total of 19 predictors (5 baseline and 14 longitudinal). Percentage of variation explained (PVE) to select the number principal components was set to 0.9. A higher PVE may lead to estimation problem, and a lower PVE may be too ineffective in approximating the trajectories in the longitudinal covariates.

4.2 Dynamic prediction of survival

In the context of dynamic prediction of survival, we are interested in the conditional survival probability $S(T > t_l + \Delta t \mid T > t_l)$ at given landmark time t_l over a prediction window $(t_l, t_l + \Delta t]$ where $t_l + \Delta t < t_{max}$, and t_{max} is the latest observation time in the data. Only (repeated) measurements before the landmark time are used to predict survival. As the landmark time progresses over time, more up to date information can be used for model fitting and to update predictions on survival. During model development and evaluation, the observations at or after the landmark time will be disregarded to avoid use of information obtained after the landmark time, and the survival model is estimated using only the subjects at risk at the landmark time.

Since all models in comparison employ Cox model to estimate the survival outcome, their predicted survival probabilities expressed in the general form:

$$\hat{S}_i(t|t_l) = \exp\left(-\int_0^t \hat{h}_0(z)e^{\hat{\eta}_{ij}}dz\right), \tag{4.1}$$

where $\hat{h}_0(z)$ is a nonparametric estimate of the baseline hazard function, and $\hat{\eta}_{ij}$ denotes the linear predictor of the model of which its composition depends on the method used.

For the comparison in this thesis, we choose to evaluate the models using landmark times $t_l = 2, 3, 4, 5, 6$ and prediction times $t' = t_l + 1, \ldots, t_{max} - 1, t_{max}$. Note that as the landmark time increases, the models will be trained using more repeated measurements, but number of subjects (at risk) in the data will decrease, as discussed in Figure 3.3.

4.3 Internal validation

Validation is the important process of evaluating the performance of a prediction model. According to the general framework for validation described by Steyerberg (2009), validation can be distinguished into three types:

- apparent validation is the evaluation when using the training data for testing. It leads to optimistically biased estimates of performance;
- internal validation determines the reproducibility of a prediction model for the setting of the underlying population for the data used for model development;
- external validation determines the generalizability of a prediction model for the populations that are plausibly related.

As we only considered the ADNI dataset in the scope, only internal validation will be carried out for this thesis.

There are two families of resampling validation techniques that can be used for internal validation: cross validation (CV) methods and bootstrap methods. Talyigás (2021) showed that repeated CV (RCV) and bootstrap methods generally perform better than simple CV and pooled CV in estimating the performance of prediction models. Hence, here we use RCV for internal validation.

We first describe the procedure of k-fold CV which the RCV is based on. In a k-fold CV, the full data will first be partitioned into k approximately equal subsets. Then, in each $i = 1, \ldots, k$ fold, the $k - 1$ subsets are used as training set and the remaining subset is used as testing set.

The estimated performance for k-fold CV is computed by averaging the performance evaluated on the k test sets. Given a data $(\mathbf{X}, y)$ of size n that are partitioned into k approximately equal disjoint sets $s_1, s_2, \ldots, s_k$ and a performance measure M, we define the k-fold CV estimated performance as follows:

$$\widehat{EP_{CV}} = \frac{1}{k} \sum_{u=1}^{k} M(y_{s_u}, r_{-s_u}(\mathbf{X}_{s_u})), \tag{4.2}$$

where r_{-s_u} denotes the predictive function trained on all data except in subset s_u, and y_{s_u} and $\mathbf{X}_{s_u}$ denote all the data in subset s_i. As CV performance can heavily dependent on the random partitioning of data, RCV can be employed to reduce the effect of such randomness.

The procedure of RCV is to repeat k-fold CV for l times, using a different random partitioning of data each time, as it is more robust to base a conclusion on multiple random partitioning than not. The estimated performance for RCV is computed by averaging the k-fold CV estimated performance $\widehat{EP_{CV}}$ described above over all repetitions, which is defined as follows:

$$\widehat{EP_{RCV}} = \frac{1}{lk} \sum_{u=1}^{l} \sum_{v=1}^{k} M(y_{s_{uv}}, r_{-s_{uv}}(\mathbf{X}_{s_{uv}})), \tag{4.3}$$

where $r_{-s_{uv}}$ denotes the predictive function trained on all data except in subset s_{uv}, and $y_{s_{uv}}$ and $\mathbf{X}_{s_{uv}}$ denote all the data in subset s_{uv}.

Besides, as random partitioning in RCV method may lead to very different proportion of events and censoring, we apply stratification to ensure such proportion is approximately equal between the training set and testing set.

To summarize, stratified 10-fold RCV with 10 repetitions is adopted for internal validation in this study. Within each fold, the training set and the testing set contains approximately $1,041$ and 116 non-overlapping subjects respectively. An identical set of random seeds is used for partitioning across all models to ensure fair comparison and reproducibility.

4.4 Performance measures

Under a validation process, the quality of the prediction model is quantified by one or more performance measures. A performance measure is also called the conditional expected error or accuracy in Hastie et al. (2009), considering that the expected performance is conditioned on a given training set consisting of a set of responses and covariates. The predictive performance of a risk prediction model can be considered in two important aspects (Steyerberg, 2009): first, discrimination to measure the model's ability to discriminate subjects between those who experienced the event of interest and those who did not; second, calibration to measure the agreement between the predicted and observed risks. Because good discrimination does not always imply good calibration, we are interested in assessing a model's predictive performance in both aspects. In the context of survival prediction, the common performance measures are either extensions of the proportion of variation explained R^2 (as in continuous response models) or extensions of sensitivity and specificity (as in binary response models).

We employ three methods to quantify predictive performance, namely the time-dependent area under the receiver operating characteristic (ROC) curve (tdAUC) (Heagerty et al., 2000), the concordance index (C index) (Harrell et al., 1996; Pencina and D'Agostino, 2004), and the Brier score (Graf et al., 1999; Schoop et al., 2008). The Brier score is an overall measure that considers both the discrimination and calibration of a model. The tdAUC and C index assess the discriminative ability in particular. While the Brier score and tdAUC depends on the prediction time, the C index only offers a single measure regardless of prediction time and is linked to the integral of tdAUC (Pencina and D'Agostino, 2004). Blanche et al. (2014) remarked that both AUC and Brier score complement each other for evaluating prediction performance. AUC has a convenient and easily understandable scaling as it does not depend on the cumulative incidence rate but only considers discrimination. Brier score is able to consider both calibration and discrimination (Steyerberg, 2009; Blanche et al., 2014) but the interpretation tends to be less direct as its scaling depends on the cumulative incidence rate. We introduce the formulations and properties of these performance measures below.

4.4.1 Time-dependent AUC

The area under the ROC curve (AUC) is a popular method in statistics and machine learning to quantify the discrimination ability of a model in binary classification context. It is based on sensitivity defined as $P(\hat{p}_i > c|Y_i = 1)$ and specificity $P(\hat{p}_i \leq c|Y_i = 0)$ where $\hat{p}_i$ is the estimated probability that $Y_i = 1$ according to a given model, and $c \in [0, 1]$ is a threshold for classifying the binary predicted outcome. The ROC curve represents the sensitivity and 1-specificity for all $c \in [0, 1]$.

In the context of survival prediction, Heagerty and Zheng (2005) proposed the time-dependent AUC as prediction accuracy summary for ROC curves based on extensions called time-specific incident sensitivity and dynamic specificity. We consider the survival time a time-varying binary outcome using the counting process representation $N_i^*(t) = \mathbb{1}(T_i \leq t)$. For incident sensitivity, the cases are said to be incident when $T_i = t$ i.e. $N_i^*(t) = 1$. For dynamic specificity, the controls are said to be dynamic as it considers subjects with $T_i > t$. The incident sensitivity and dynamic specificity are defined as follows:

$$\text{sensitivity}^I(c, t) = P(\eta_i > c \mid T_i = t) = P(\eta_i > c \mid dN_i^*(t) = 1), \tag{4.4}$$

and

$$\text{specificity}^D(c, t) = P(\eta_i \leq c \mid T_i > t) = P(\eta_i > c \mid N_i^*(t) = 0), \tag{4.5}$$

where η_i is the linear predictor.

The subjects at risk at time t are divided into two subsets with and without observed an event. The incident sensitivity measures the expected proportion of subjects with a marker greater than criterion c among the subset observed with event at time t. The dynamic specificity measures the expected proportion of subjects with a marker less than or equal to criterion c among the subset without observed an event before time t.

The incident/dynamic ROC curves can be defined as follows:

$$ROC_t^{I/D}(p) = TP_t^I\{[FP_t^D]^{-1}(p)\}, \tag{4.6}$$

where $p \in [0, 1]$ is the dynamic false-positive rate, c_p is the corresponding criterion that $p = 1 - \text{specificity}^D(c^p, t)$, true-positive rate function $TP_t^I = \text{sensitivity}^I(c, t)$, false-positive rate function $FP_t^D(c) = 1 - \text{specificity}^D(c, t) = p$.

From above, the tdAUC based on incident sensitivity and dynamic specificity can be obtained as:

$$AUC(t) = P(\eta_j > \eta_k \mid T_j = t, T_k > t). \tag{4.7}$$

By Heagerty and Zheng (2005), the estimator for true-positive function and false positive function are as follows:

$$\widehat{TP}_t^I(c) = \hat{P}(\eta_i > c \mid T_i = t) = \sum_k \mathbb{1}(\eta_k > c) \cdot \pi_k(\beta, t), \tag{4.8}$$

and

$$\widehat{FP}_t^D(c) = \hat{P}(\eta_i > c \mid T_i > t) = \sum_k \mathbb{1}(\eta_k > c) \cdot R_k(t+)/W^R(t+), \tag{4.9}$$

where β is the parameters estimated in Cox model, $R_k(t+) = \mathbb{1}(X_i \geq t+) = \lim_{\delta \to 0} R_k(t + |\delta|)$ is the at-risk indicator, $W^R(t+) = \sum_k R_k(t+)$ refers to the size of control set at time t i.e. subjects at risk excluding those who observe event at time t.

With the estimates $\hat{TP}_t^I(c)$ and $\hat{FP}_t^D(c)$, the tdAUC can be estimated as follows:

$$\widehat{AUC}(t) = \int \widehat{ROC}_t^{I/D}(p)\, dp. \tag{4.10}$$

The tdAUC ranges from 0 to 1; higher values of tdAUC correspond to discrimination. A tdAUC of 0.5 indicates that the model discriminates no better than a random prediction rule.

4.4.2 C index

Harrell et al. (1996) defines the concordance index, or C index, as the proportion of concordant pairs, based on the intuition that for a model that has good discrimination ability, the order in observed survival times between any two subjects should be in concordance with their predicted survival probabilities. The C index can be estimated using various approaches; in this thesis we follow the approach proposed by Pencina and D'Agostino (2004) below.

First, we denote $X_i : i = 1, 2, \ldots, n$ and $Y_i : i = 1, 2, \ldots, n$ as the observed survival time and the predicted survival probability for any subject. Considering any two pairs of observations $X_i, X_j : i \neq j$ and predictions $Y_i, Y_j : i \neq j$, a concordant pair means that $X_i < X_j$ and $Y_i < Y_j$, and a discordant pair means that $X_i < X_j$ and $Y_i > Y_j$. The predicted survival probability is used interchangeably with the predicted survival times as they remain a one-to-one correspondence (Harrell et al., 1996). The censoring in survival data also requires any two subjects to be first distinguished as an usable pair (i.e. either event vs event, or event vs non-event) and otherwise an unusable pair for comparison. The unconditional probability of concordance π_c and discordance π_d are defined as:

$$\pi_c = P(X_i < X_j \text{ and } Y_i < Y_j) + P(X_i > X_j \text{ and } Y_i > Y_j), \tag{4.11}$$

and

$$\pi_d = P(X_i < X_j \text{ and } Y_i > Y_j) + P(X_i > X_j \text{ and } Y_i < Y_j). \tag{4.12}$$

Then the probability of concordance is given by:

$$C = P((X_i < X_j \text{ and } Y_i < Y_j) \text{ or } (X_i > X_j \text{ and } Y_i > Y_j) \mid X_i \neq X_j) = \frac{\pi_c}{1-\pi_t} = \frac{\pi_c}{\pi_c + \pi_d}, \tag{4.13}$$

where $\pi_t = 1 - \pi_c - \pi_d$ is the probability of unusable pairs that cannot be compared as none of the subjects experienced any event.

Pencina and D'Agostino (2004) showed that if i, j above are interchangeable and $\{Y_i\}$ comes from a continuous distribution, then the expression in (4.13) can be simplified as:

$$C = \frac{P(X_i < X_j \text{ and } Y_i < Y_j)}{P(X_i < X_j)} = P(Y_i < Y_j | X_i < X_j). \tag{4.14}$$

To estimate the C index in a sample of subjects $i : 1, 2, \ldots, n$, we define c_h and d_h the number of concordant pairs and discordant pairs with respect to h-th subject as follows:

$$c_h = \sum_{h \neq j} c_{hj}, \tag{4.15}$$

and

$$d_h = \sum_{h \neq j} d_{hj}, \tag{4.16}$$

where c_{ij} is an indicator function when the i-j pair is concordant $c_{ij} = 1$ and otherwise $c_{ij} = 0$, similarly, d_{ij} is an indicator function when the i-j pair is discordant $d_{ij} = 1$ and otherwise $d_{ij} = 0$. Then the unbiased estimates of π_c and π_d are given by:

$$p_c = \frac{1}{n(n-1)} \sum_h c_h, \tag{4.17}$$

and

$$p_d = \frac{1}{n(n-1)} \sum_h d_h. \tag{4.18}$$

Hence, the C index is estimated as:

$$\hat{C} = \frac{p_c}{p_c + p_d}. \tag{4.19}$$

The C index ranges from 0 to 1, which a higher score represents better discrimination. $C \leq 0.5$ indicate that the model discriminates no better than a random prediction rule. Heagerty and Zheng (2005) demonstrated the connection between Harrell's C index introduced in Section 4.4.2 which the concordance is an integral of tdAUC over time t.

4.4.3 Brier score

Explained variation is a popular and simple direct measure to quantify the variability (information) in the data that can be explained by a model and can be used for continuous and binary outcomes. A model with better prediction accuracy results in smaller distances between predicted and observed outcomes. For binary outcomes, the Brier score is defined as the mean squared error between true outcomes Y_i and predictions P_i. Consider the dynamic survival prediction for landmark time s and prediction window t, Schoop et al. (2008); Blanche et al. (2014) defined the Brier score as:

$$BS(s,t) = \mathbb{E}\left[\Big((D(s,t) - \hat{S}(t|s)\Big)^2 \mid T > s\right], \tag{4.20}$$

where $\tilde{D}_i(s,t) = \mathbb{1}_{s<\tilde{T}_i \leq s+t}$ is an indicator function that equals to 1 when subject i experiences the event in $(s, s+t]$ and equals to 0 otherwise, $\hat{S}(t|s)$ is the survival prediction for landmark time s and prediction window t.

In practice, computation of the Brier score for survival data is more complex than this due to the presence of censoring. With censored data, the value of the Brier score cannot be directly computed, but it needs to be estimated. Graf et al. (1999) and Gerds and Schumacher (2006) proposed to use the inverse probability of censoring weight (IPCW) method, which involves the computation of subject-specific weight $\hat{W}_i(s,t)$, to account for the conditional probability that subject i is not censored in the interval (s,t):

$$\hat{W}_i(s,t) = \frac{\mathbb{1}_{\tilde{T}_i > s+t}}{\hat{G}(s+t|s)} + \frac{\mathbb{1}_{s<\tilde{T}_i \leq s+t}\,\Delta_i}{\hat{G}(\tilde{T}_i|s)}, \tag{4.21}$$

where

- $\mathbb{1}_{\tilde{T}_i > s+t}$ and $\mathbb{1}_{s<\tilde{T}_i \leq s+t}$ are indicator functions equal to 1 when the event is observed in the interval (s,t) and is 0 otherwise,
- $\Delta_i = \mathbb{1}_{T \leq C}$ is the indicator function which equals to 1 when the event happens before censoring C,
- $\hat{G}(u)$ is the censoring distribution estimated by the Kaplan-Meier estimator, and
- $\hat{G}(u|s) = \frac{\hat{G}(u)}{\hat{G}(s)}$ $\forall u > s$ estimates the conditional probability of not being censored at time u given that not being censored up to time s.

As such, the first component of 4.21 assigns a weight of $\frac{1}{\hat{G}(s+t|s)}$ for events with survival time $\tilde{T}_i > s+t$, and the second component assigns a weight of $\frac{1}{\hat{G}(\tilde{T}_i|s)}$ for events with survival time $s < \tilde{T}_i \leq s+t$.

Once the weights in 4.21 have been computed, the Brier score can be estimated as:

$$\widehat{BS}(s,t) = \frac{1}{n\hat{H}_{\tilde{T}}(s)} \sum_{i=1}^{n} \hat{W}_i(s,t)(\tilde{D}_i(s,t) - \hat{S}(t|s))^2, \tag{4.22}$$

where $\hat{H}_{\tilde{T}}(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\tilde{T}_i > s}$ estimates the probability of observing a subject at risk at s.

The Brier score ranges from 0 to 1, which a lower score represents better predictive performance. A Brier score of 0.25 indicate that the model predicts no better than a random prediction rule.

4.5 R packages and code

Computations in this study were performed using R version 4.1.3 (R Core Team, 2021). The R packages used to estimate the models considered in this thesis were:

- `glmnet` (available from CRAN) for the penalized Cox models that respectively use baseline covariates and the last observation before the landmark time as predictors;
- `pencal` (available from CRAN) for penalized regression calibration;
- the estimation of MFPCCox was based on the R scripts using simulated data available at `https://github.com/kan-li/MFPCCox`. Such scripts were developed for simulated data, and they had to be adapted so that they could be used on the ADNI data. The original functions that were reused without adaptations can be found in the script `function_MFPCCox.R`, whereas the adapted/additional scripts relevant to application to ADNI data can be found in `function_MFPCCox_exp.R` and `run_repCV_MFPCCox.R`.

The following packages were used to evaluate predictive performance:

- `survROC` (available from CRAN) for the tdAUC;
- `survcomp` (available from Bioconductor) for the C index;
- `pec` (available from CRAN) for the Brier score.

The R scripts implemented for this thesis are available at `https://github.com/freddy-feng/thesis_CompareSurvivalModels`. The model fitting and performance evaluation were performed using the Academic Leiden Interdisciplinary Cluster Environment (ALICE) of Leiden University.

Chapter 5

Results

In this Chapter, we report the results of the application of the 4 models described in Section 4.1 to the problem of predicting time to to dementia on the ADNI data for different landmark times. Table 5.1 shows a brief overview of the aforementioned methods.

The predictive performance of each model was evaluated considering 5 different landmark times $t_l \in \{2, 3, 4, 5, 6\}$; the optimism-corrected values of the performance measures, i.e. the tdAUC, the C index, and the Brier score, were estimated through a repeated 10-fold cross-validation with 10 repetitions. The tdAUC and the Brier score were evaluated every year after each landmark time, up to the 15 years from baseline. Table 5.2 shows how the number of subjects at risk and the event rate change with the landmark time. As the landmark time varies from 2 to 6 years after baseline, the number of subjects at risk reduce from 1,157 to 397 due to the progressive removal of subjects censored or diagnosed with dementia at or before the landmark times. Moreover, the event rate decreases from 18.6% to 13.6%.

The average number of repeated measurements for each longitudinal covariate at different landmark times are illustrated in Figure 5.1. After excluding measurements at or after the landmark times, most longitudinal covariates have at least 3 repeated measurements per subject on average. As the landmark time varies from 2 to 6, more repeated measurements could be utilized by PRC-LMM and MFPCCox, and pCox-landmarking would use the most updated measurements available before the landmark time.

Table 5.1: Overview of the models included in the comparison

	No. of covariates		Uses repeated
Model name	Baseline	Longitudinal	measurements for prediction
pCox-baseline[1]	5	21	No
pCox-landmarking[1]	5	21	No
PRC-LMM	5	21	Yes
MFPCCox	5	14	Yes

[1] pCox: penalized Cox model.

Table 5.2: Summary of observed events at different landmark times

Landmark time	Number of subjects at risk	Number of events	Event rate	Average time to dementia after landmark time
2	1157	215	0.186	3.3
3	842	146	0.173	3.4
4	634	100	0.158	3.4
5	494	72	0.146	3.3
6	397	54	0.136	3.0

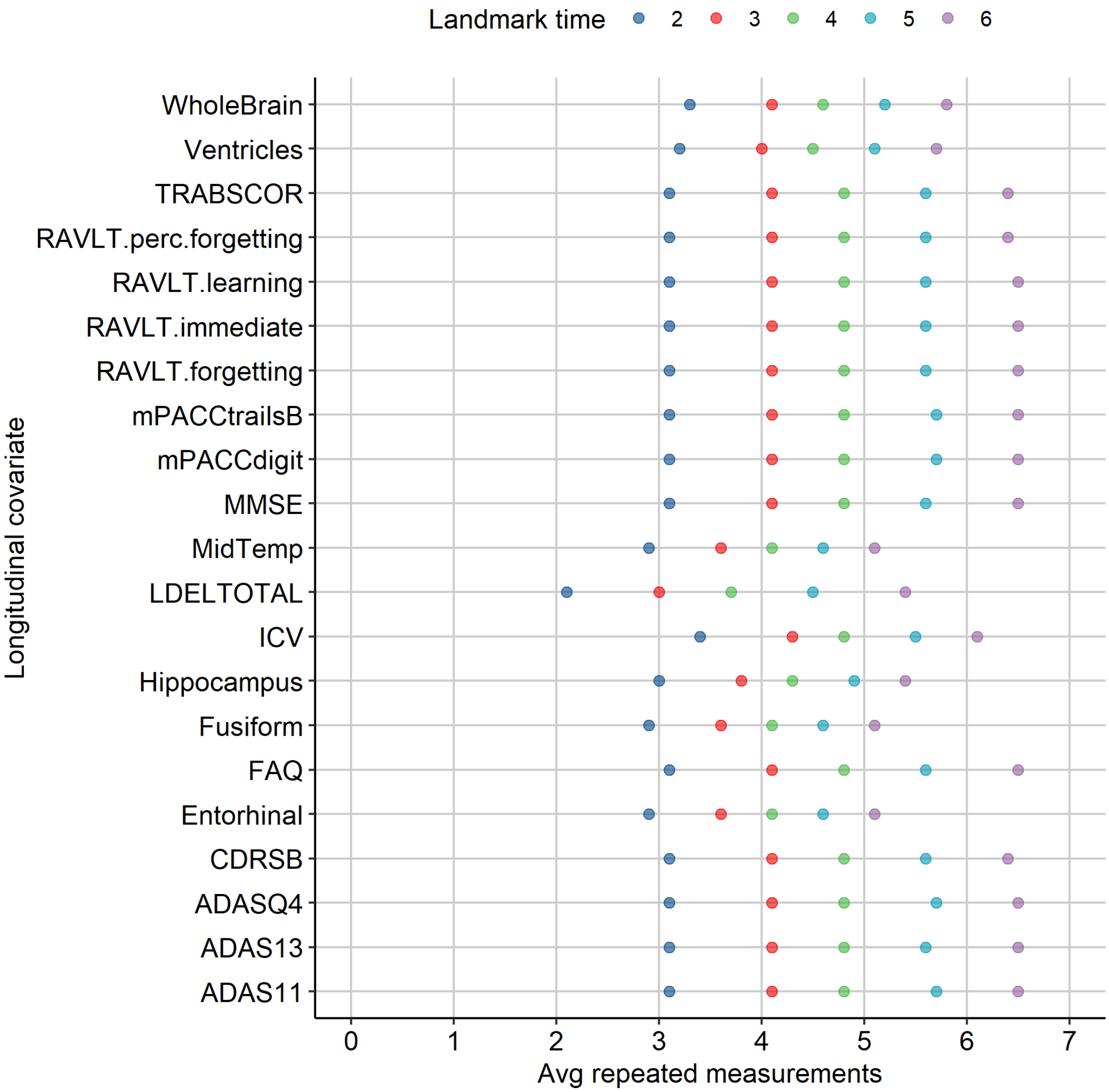


Figure 5.1: Average number of repeated measurements per subject at different landmark times.

5.1 Time-dependent AUC

The optimism-corrected estimates of the tdAUC at different landmark times are shown in Figure 5.2. We observe that the values of the tdAUC typically decrease with the landmark time, irrespective of the prediction method. pCox-landmarking is the best performing model according to this metric, as it outperforms other models at most prediction times and landmark times. It is followed by PRC-LMM, whose performance is rather similar to that of pCox-landmarking at landmark times 2, 3 and 6, and a bit more different at landmark times 4 and 5. MFPCCox ranks third and its performance dropped substantially as the landmark time increased. pCox-baseline consistently shows the worst performance at all landmark times, and its performance also worsens as the landmark time increased.

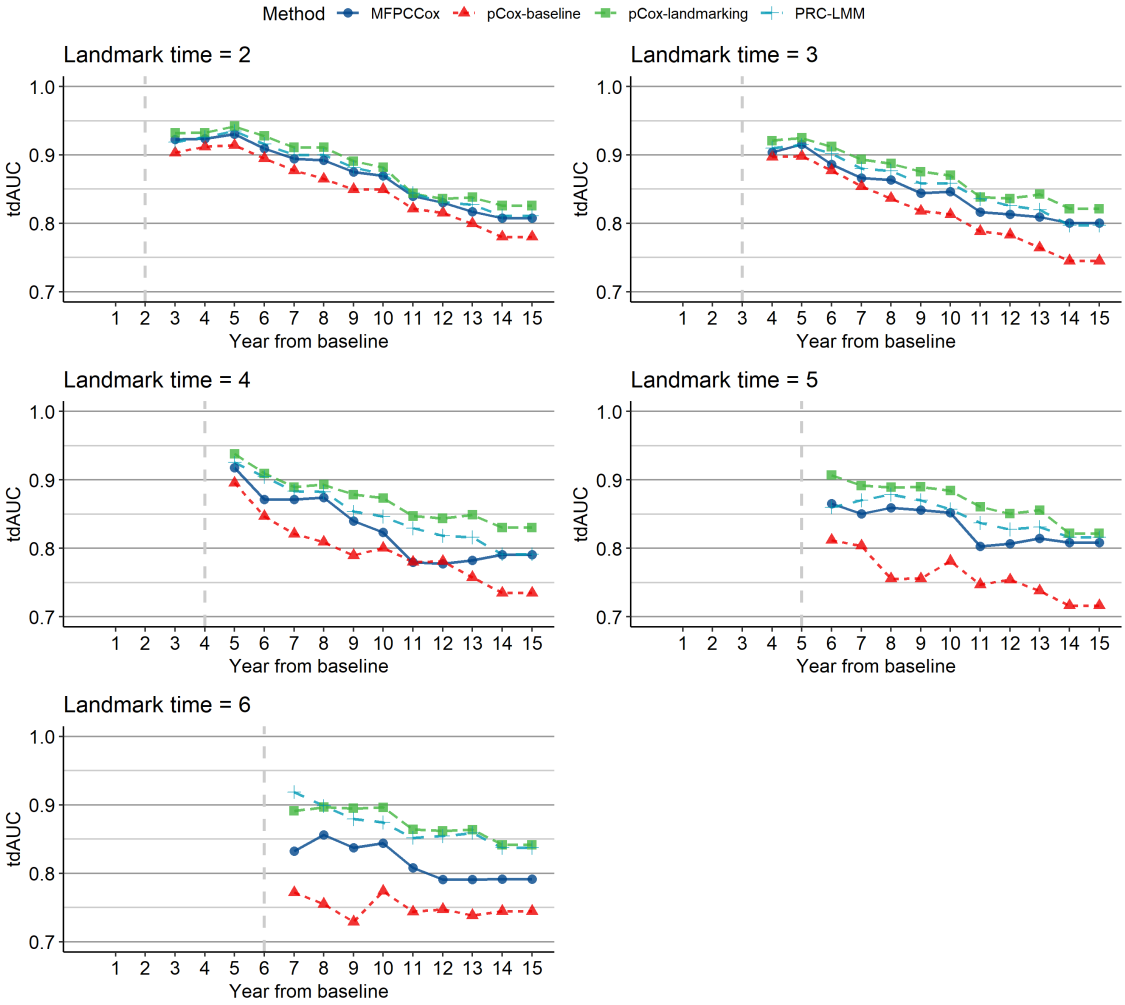


Figure 5.2: Cross-validated tdAUC estimates for the prediction of time to dementia at landmark times 2 to 6.

5.2 C index

The optimism-corrected estimates of the C index, averaged over 10 cross-validation replications, are compared in Table 5.3. pCox-landmark has the highest C index at all landmark times except at landmark time 6, where it is slightly below PRC-LMM. The performance of PRC-LMM is generally close to that of pCox-landmark. MFPCCox and pCox-baseline are slightly behind the best model at landmark time 2, where the difference is no greater than 0.025. Shifting the landmark time from 2 to 6, the performance of all models are generally reduced, but at a different rate. This effect is more pronounced in the pCox-baseline and MFPCCox. At landmark time 6, the difference between the pCox-baseline and the best model widens to 0.142, and the difference between the MFPCCox and the best model widens to 0.06. Figure 5.3 shows the distribution of the cross-validated C index as evaluated on each fold (i.e., before averaging over folds and replications). The box plots are grouped by models and arranged by ascending landmark times. We can observe that while the performance of pCox-landmarking and PRC-LMM does not vary a lot with the landmark time, the performance of MFPCCox and pCox-baseline clearly decreases as the landmark time increases. Moreover, for pCox-baseline and MFPCCox, the variance of the C index estimates clearly increases with the landmark time.

Table 5.3: Cross-validated C index at different landmark times

	Landmark time				
Method	2	3	4	5	6
MFPCCox	0.897	0.877	0.866	0.847	0.828
pCox-baseline	0.881	0.854	0.814	0.760	0.746
pCox-landmarking	0.905	0.893	0.887	0.885	0.884
PRC-LMM	0.901	0.881	0.872	0.867	0.888

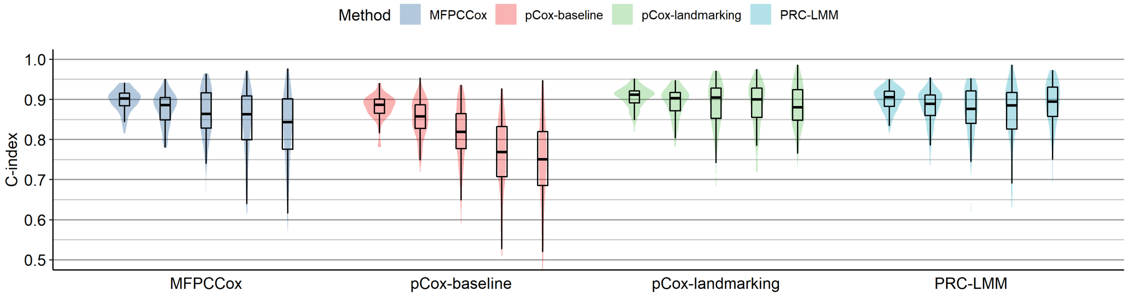


Figure 5.3: Distribution of the C index as evaluated on each CV fold, before aggregation over the 10 folds and the 10 replications of CV. The grouped box plots show the estimated C index evaluated at landmark time 2 (left) to 6 (right) for each method under 10-fold CV repeated for 10 times. The performance of pCox-landmarking appears to be the best at most landmark times, and is closely followed by the PRC-LMM. These two models clearly outperformed MFPCCox and pCox-baseline, especially at later landmark times. The decline in the C index with the landmark time is more severe in pCox-baseline and MFPCCox.

5.3 Brier score

Figure 5.4 shows the optimism-corrected estimates of the Brier score at the different landmark times. Displaying a different pattern from above results, the estimated Brier scores among pCox-landmarking, PRC, and MFPCCox model are very close at the first two landmark times. At landmark times 4, 5 and 6, the Brier score curves for these three models are very similar up until $t = 10$. On the other hand, the difference becomes more pronounced for prediction times beyond $t = 10$, when PRC outperforms pCox-landmarking and MFPCCox. pCox-baseline exhibits the worst Brier scores across all landmark times.

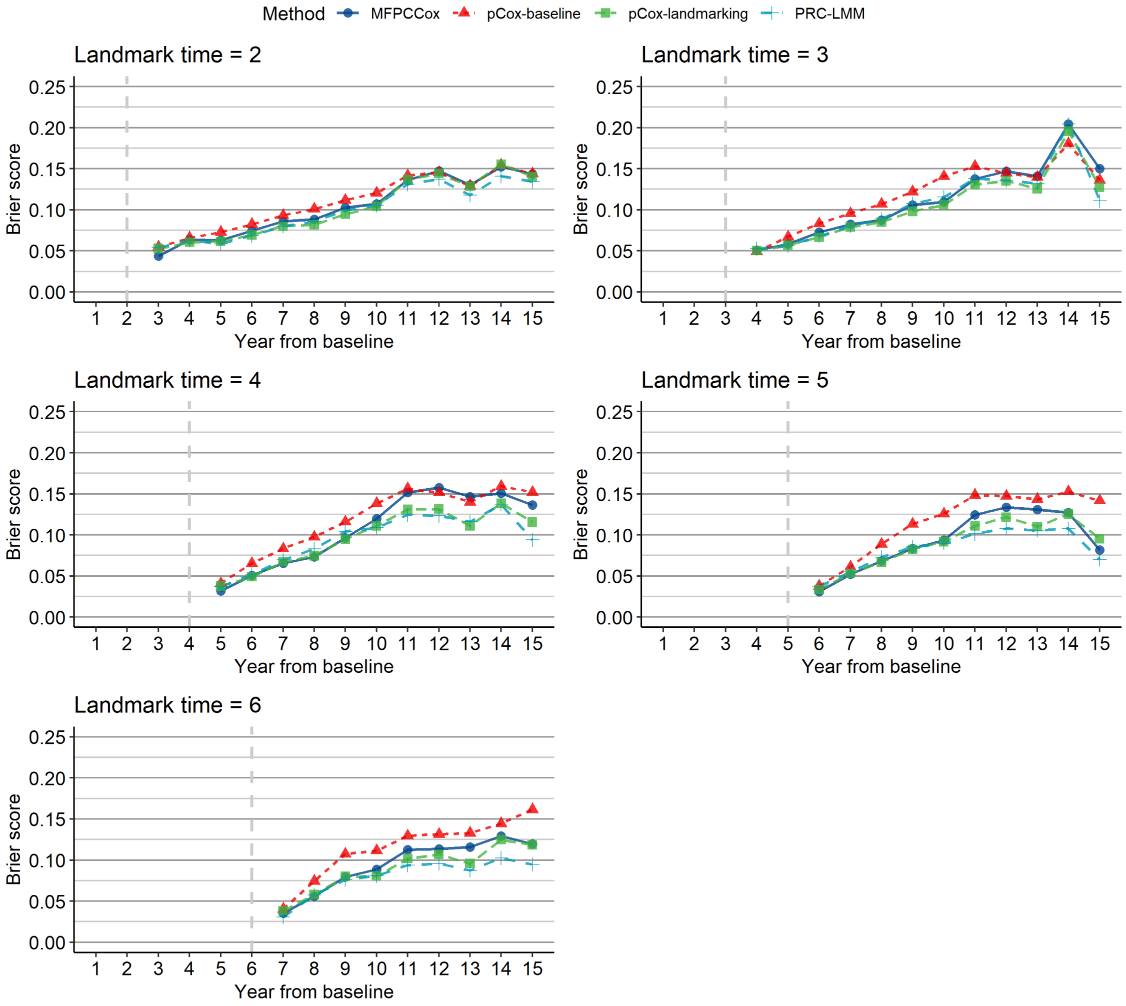


Figure 5.4: Cross-validated Brier score estimates for the prediction of time to dementia at landmark times 2 to 6.

5.4 Summary

When we look jointly at the results presented in Sections 5.1, 5.2 and 5.3, we can conclude that:

- pCox-landmarking is the best performing method when looking at the tdAUC and C index, and the second best method with respect to the Brier score;
- PRC-LMM is the second best performing method in terms of the tdAUC and C index, and the best method according to the Brier score;
- MFPCCox performed worse than pCox-landmarking and PRC-LMM irrespective of the performance measure considered;
- pCox-baseline was the worst performing method with respect to all metrics.

The difference in ranking between tdAUC / C index and Brier score may be due to the different aspect of prediction that the measure is evaluating, i.e. discrimination versus calibration.

Chapter 6

Discussion

6.1 Conclusions

The goal of this thesis was to perform an empirical comparison of four different statistical modelling methods that can be used to predict a survival outcome given numerous longitudinal covariates. These methods included (i) two novel methods, PRC and MFPCCox, developed to fully utilize the longitudinal data using different techniques (mixed models and MFPCA respectively), (ii) landmarking - a conventional approach that uses the last observations until landmark time, and (iii) the penalized Cox model only using the baseline measurements. We evaluated their predictive performance in the context of dynamic prediction of time to dementia using the ADNI data, which comprises a heterogeneous set of variables measured longitudinally.

In Chapter 2 we provided an overview on these modelling methods, highlighting their differences and similarities. In Chapter 3 we summarized the most important features of the ADNI data, and described practical issues such as data preprocessing and dealing with missing values. In Chapter 4 we elaborated on the modelling process and the comparison framework. We have chosen 5 baseline covariates and 21 longitudinal covariates as the candidate predictors. The developed models were evaluated based on the dynamic prediction of survival, where the predicted survival probabilities to develop dementia were evaluated conditionally on survival until a given landmark time. Three different performance measures, namely tdAUC, C index, and Brier score, were employed to assess the predictive performance of the different methods. We employed repeated cross-validation, stratified by event status (dementia and censoring), to compute optimism-corrected estimates of the tdAUC, C index and Brier score.

In Chapter 5 we compared the performance of the different models in predicting $P(T > t \mid T > t_l)$, where $t_l \in \{2, 3, 4, 5, 6\}$ and $t \in \{t_l + 1,\ t_l + 2, ..., 15\}$. The choice of the landmark time has some important implications: first, increasing the landmark time increases the number of available measurements that can be used to model the trajectories described by the longitudinal covariates in PRC and MFPCCox (Figure 5.1); similarly, this allows the landmark model to use more update measurements; second, increasing the landmark time effectively reduces the sample size available for model development and validation, due to the fact that as the landmark time is increased, more and more subjects will either be censored or have developed dementia before the landmark time (Table 3.3). To balance these two opposite effects, here we have considered a range of five landmark times.

The results of our comparison showed that the landmarking approach achieved the best performance in terms of tdAUC and C index, and it was closely followed by PRC (Figure 5.2 and Table 5.3). When looking at the Brier score, landmarking and PRC performed similarly for predictions up until $t = 10$, whereas for predictions after $t = 10$ PRC outperformed landmarking (Figure 5.4). MFPCCox was outperformed by landmarking and PRC with respect to all three performance measures. Lastly, the penalized Cox which only used the baseline measurements was the worst performing method according to all performance measures.

Overall, these results show that the first three models were all able to improve the prediction of time to dementia by exploiting the longitudinal information gathered at visits after baseline, thus showing the importance of collecting longitudinal measurements to improve the accuracy of risk predictions. The result also seemed to indicate that for the ADNI data a simpler approach such as landmarking may be enough to deliver accurate predictions, as we observed that more sophisticated approaches that model the evolution of the longitudinal covariates over time achieved only marginal (Brier score for PRC) or no improvement (MFPCCox) over landmarking.

On the discrimination performance (measured using the tdAUC and C index) we observed that as the landmark time increased, the performance of pCox-baseline and MFPCCox decreased noticeably, whereas the performance of landmarking and PRC decreased only slightly. A possible explanation for this result is the aforementioned availability of more repeated measurements for PRC and of more up to date measurements for landmarking for later landmark times.

On the calibration performance (measured using the Brier score) the penalized Cox model was once again the worst performing method; the results were less clear cut for other three methods. The performances of the landmarking, PRC, and MFPCCox were not so distinguishable at smaller landmark times or at prediction times up to $t = 10$. However, when the landmark time increased, PRC gradually outperformed landmarking and MFPCCox with for predictions from $t = 10$ onward. This may indicate that the rate of progression in the longitudinal covariates may improve the accuracy of long-term predictions.

Overall, the performance of PRC was quite close to that of the landmarking. PRC benefited from having more repeated measurements in the longitudinal covariates as seen in the Brier score. This result is in line with the expectation that more repeated measurements can improve the estimation of the LMMs and help to derive more accurate summaries of the longitudinal covariates.

On the contrary, MFPCCox performed worse than landmarking and PRC, which could be due to various reasons:

- due to estimation problems, MFPCCox was specified with 14 instead of 21 longitudinal covariates as other models, possibly losing part of the predictive information;
- MFPCCox models the trajectories of the longitudinal covariates with respect to follow-up time. Considering the fact that each subject entered the study at a different age, such specification seems arbitrary and could impair predictive patterns in the original data. In our opinion, a more reasonable choice would be to use the subjects' age as the time scale like in step 1 of PRC, however for MFPCCox this is problematic when baseline age differs across patients, because that would lead to estimation problems due to the presence of highly sparse matrices;

Despite these concerns, our results at least showed that the MFPCCox method could improve the predictive performance for dementia risk over a model that only uses baseline information.

Better tdAUC and Brier score were observed in this thesis when compared to the application of MFPCCox to the ADNI data presented in Li and Luo (2019), however the results are not comparable due to following reasons: (i) only data from the first phase of ADNI study (ADNI1) were used in the previous study, resulting in a much smaller dataset, (ii) only 6 longitudinal covariates were used in Li and Luo (2019), (iii) the MFPCCox in the previous study was developed based on repeated measurements before and after the landmark time, which is different from the dynamic prediction formulated in Section 4.2.

6.2 Limitations

A first limitation of the present study is that its results may not generalize to other datasets. To ensure a fairer comparison of prediction models, Boulesteix et al. (2008) recommended that the comparison should be based on at least two datasets. Due to data availability and time constraints, the methods were only compared on a single dataset. The conclusions of this thesis should thus be interpreted with caution, as it is solely based on the ADNI data and the performance between some models were really close. The so-called best method might be very sensitive to the characteristics of the dataset and the covariates used as predictors.

A subject that was not explored in this thesis is the use of different strategies to account for missing data in the ADNI dataset. As illustrated in Section 4.1, we chose to use complete case for baseline covariates in data screening which led to slightly loss of information. We used mean imputation for missing values when extracting cross-sectional measurements from the longitudinal covariates when they were used in the penalized Cox model and the landmarking model. Alternative approaches such as multiple imputation might be better to account for the data uncertainty.

In principle, patients may develop dementia between two visits, meaning that the survival outcome should be treated as interval-censored. The problem was simplified to be right-censoring in this thesis to ensure the problem formulation is compatible with the methods for comparison. As such, the predicted survival probabilities computed here should be considered as a upper bound estimate.

Studies involving an elder population group often comprise drop outs due to death, so ideally the competing risk should be taken into account in the risk modelling in these contexts. In the ADNI data, death contributed to a small portion of drop outs ($< 10\%$ of the subjects). Since the methods compared in this thesis do not account for competing risks, we treated deaths as censored observations.

Lastly, we should be cautious when interpreting the models' predicting ability reported in this thesis because they were developed based on data sampled from subpopulation that was predominantly white. The generalizability to underrepresented ethnic groups, is not known without proper external validation. But as our goal was methodological comparison instead of development of a clinical prediction model, this issue was not considered in this thesis.

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4).

Anderson, J. R., Cain, K. C., and Gelber, R. D. (1983). Analysis of survival by tumor response. *J Clin Oncol*, 1(11):710–719.

Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2014). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113.

Boulesteix, A.-L., Strobl, C., Augustin, T., and Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, 6:CIN.S408.

Bull, L. M., Lunt, M., Martin, G. P., Hyrich, K., and Sergeant, J. C. (2020). Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagnostic and Prognostic Research*, 4(1).

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dafni, U. (2011). Landmark analysis at the 25-year landmark point. *Circulation: Cardiovascular Quality and Outcomes*, 4(3):363–371.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*, 20(1):145–157.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics (Oxford, England)*, 5:329–340.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1).

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hou, X.-H., Feng, L., Zhang, C., Cao, X.-P., Tan, L., and Yu, J.-T. (2018). Models for predicting risk of dementia: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 90(4):373–379.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3).

Jiang, S., Xie, Y., and Colditz, G. A. (2021). Functional ensemble survival tree: Dynamic prediction of alzheimer's disease progression accommodating multiple time-varying covariates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):66–79.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kueper, J. K., Speechley, M., and Montero-Odasso, M. (2018). The Alzheimer's Disease Assessment Scale–cognitive subscale (ADAS-cog): Modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer's Disease*, 63(2):423–444.

Li, K. and Luo, S. (2019). Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, 38(24):4804–4818.

Licher, S., Leening, M. J., Yilmaz, P., Wolters, F. J., Heeringa, J., Bindels, P. J., Vernooij, M. W., Stephan, B. C., Steyerberg, E. W., Ikram, M. K., and and, M. A. I. (2019). Development and validation of a dementia risk prediction model in the general population: An analysis of three longitudinal studies. *American Journal of Psychiatry*, 176(7):543–551.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., and Rizopoulos, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 30(4):999–1014.

Mercer, J. (1909). XVI. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458):415–446.

Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., and Thal, L. J. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. *Alzheimer Disease and Associated Disorders*, 11 Suppl 2:S13–S21.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Pencina, M. J. and D'Agostino, R. B. (2004). OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123.

Putter, H. and Houwelingen, H. C. (2022). Landmarking 2.0: Bridging the gap between joint models and landmarking. *Statistics in Medicine*, 41(11):1901–1917.

Putter, H. and van Houwelingen, H. C. (2016). Understanding landmarking and its relation with time-dependent Cox regression. *Statistics in Biosciences*, 9(2):489–503.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610.

Signorelli, M., Spitali, P., Szigyarto, C. A.-K., and and, R. T. (2021). Penalized regression calibration: A method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*, 40(27):6178–6196.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5).

Steyerberg, E. (2009). *Clinical Prediction Models*. Springer.

Suzuki, K., Hirakawa, A., Ihara, R., Iwata, A., Ishii, K., Ikeuchi, T., Sun, C.-K., Donohue, M., and and, T. I. (2020). Effect of apolipoprotein e e4 allele on the progression of cognitive decline in the early stage of alzheimer's disease. *Alzheimer's &amp; Dementia: Translational Research &amp; Clinical Interventions*, 6(1).

Talyigás, G. (2021). A systematic comparison of methods for the validation of binary classification models. Master's thesis, Universiteit Leiden.

Tang, E. Y. H., Harrison, S. L., Errington, L., Gordon, M. F., Visser, P. J., Novak, G., Dufouil, C., Brayne, C., Robinson, L., Launer, L. J., and Stephan, B. C. M. (2015). Current developments in dementia risk prediction modelling: An updated systematic review. *PLOS ONE*, 10(9):e0136181.

Tang, E. Y. H., Robinson, L., and Stephan, B. C. M. (2017). Dementia risk assessment tools: an update. *Neurodegenerative Disease Management*, 7(6):345–347.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37.

Verweij, P. J. M. and Houwelingen, H. C. V. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436.

Weiner, M. W., Aisen, P. S., Jack, C. R., Jagust, W. J., Trojanowski, J. Q., Shaw, L., Saykin, A. J., Morris, J. C., Cairns, N., Beckett, L. A., Toga, A., Green, R., Walter, S., Soares, H., Snyder, P., Siemers, E., Potter, W., Cole, P. E., and and, M. S. (2010). The alzheimer's disease neuroimaging initiative: Progress report and future plans. *Alzheimer's & Dementia*, 6(3):202.

WHO (2021). *Global status report on the public health response to dementia.* World Health Organization.