Healthy Ageing

Data, Literature Review, Methodology

by

Fenna de Meijier

Student ID: 603450

Thesis supervisor: Prof. Patrick Groenen

Co-reader: TBD

Master of Science in Econometrics & Operations Research

Erasmus School of Economics

Erasmus Universiteit Rotterdam

June 2023

Contents

1

Literature Review

As human life expectancy continuously increases, healthy ageing has become an important topic in geriatric research. The World Health Orginization defines healthy ageing as the process of developing and maintaining the functional ability that enables well-being in older age (**?**). Hence, healthy ageing is not only about living a longer life, but about maintaining good physical and mental health, independence and social participation in later life. Global health estimates confirm that the last century saw an increase in healthy life expectancy, also defined as healthspan, but that this trend has not kept pace with the increase in lifespan. The delineation between healthspan and lifespan calls for researchers to identify determinants of healthy ageing and develop interventions that promote it. In this section, we will review the literature on previous ageing research, and discuss various methods that have been applied to assess (quantify?) the impact of risk factors on age progression. Next, we overview healthspan research and introduce chronic disease development as a proxy for ageing. Then we dicuss application on survival modelling techniques in the field of geriatric research and we consider some main features. Finally, gaps in previous research will be discussed and we will identify areas of further research.

Ageing is an complex process that has been studied extensively in recent years. Previous studies have deployed diverse modelling techniques that aim to capture and quantify the impact of various factors associated with ageing. **?** propose a low-dimensional representation in the form of a 'frailty index', defined by the proportion of accumulated health deficits, to quantify ageing. Specifically, the concept of biomarkers of age was introduced by **?** in the 1980s, and is based on the assumption that there exist biological parameters that better measure the rate of ageing than chronological age. Since then, many papers have been published that identify biomarkers, such as telomere length (**?**) or DNA methylation (DNAm) patterns **?**, that assess the biological age of an individual. Alternatively, with AI methods advancing and computational power increasing, we saw the emergence of advanced deep learning

based approaches. For example, **?** suggest a neural network based model that uses physical, biological and demographical variables and can simulate high-dimensional individual trajectories of health and survival. Similarly, **?** present a deep learning prediction model based on electronic medical records that can accurately predict biological age (as measured by telomere length). They also state that individuals with large discrepancy between their chronological age and their predicted biological age are at higher risk for age-related health problems, and that they have higher systolic blood pressure, higher cholesterol, liver damage and anemia. Altogether, there exists extensive literature on biological age and markers of biological ageing. Nonetheless, the recent years saw a shift from longetivity research to healthspan research, fuelled by the societal need to not only extend the years of life but also to improve the quality of those years.

Healthspan research aims at identifying factors that are associated with the development of major diseases that drive morbidity and mortality. Given that ageing is the single most important risk factor for chronic disease accumulation, and therefore for end-of-healthspan, it is a promising target for the development of interventions that increase risilience to functional decline (**?**). Multimobidity refers to the co-occurence of two or more chronic conditions in an individual (**?**), and is associated with a broad range of behavioural and physiosocial factors. In particular, a number of lifestyle risk factors, such as smoking, obestity and unhealthy diet predispose to multimorbidity (**?**). Moreover, it is well estabislished that there is an association between socioeconomic status and multimorbidity (**?**). The premise that ageing is amongst the underlying mechanisms that drive development of multimorbidity is based on several studies that address this topic. A study by **?** discuss how mechanisms of age-related inflammation lead to functional decline and the development of chronic disease. Chronic inflammation is associated with a wide range of chronic diseases, including diabetes, cardiovascular disease, kidney disease, Alzheimer's disease, and cancer. Although accute inflammation is a required natural response of the body to defend itself against microbial infection, evidence suggests that the mechanisms responsible for regulating inflammation become dysregulated as a result of ageing (**?**). Dietary interventions, such as caloric restriction and increased intake of saturated fatty acids, have been proposed to deactivate the inflammasome and improve healthspan. Other research proposes a set of objective healthy ageing indicators, including tests of grip strength, walking speed, chair rising and standing balance to capture physical function at an individual level associated with specific health outcomes. Their findings are summarized by **?**, and indicate that lower performance on these tests is associated with higher risk of cardiovascular disease, dementia and loss of independence. The diversity of the aforemetioned risk factors for chronic disease development and indicators for decreased healthspan emphasize the need for a multifactorial approach to healthspan research. Fortunately, modern longitudinal cohort studies that include large arrays of environmental, sociodemographic, and socioeconomic data are becoming

more publicly available in recent years. They are particularly suited to investigate age-related chronic disease development and multifactorial dynamics controlling the ageing process. Specifically, based on the assumption that ageing is the underlying process that drives chronic disease, data from large clinical cohorts is exploited to investigate healthspan or incidence of chronic disease as a proxy for ageing.

In a medical context, finding prognostic markers associated with a time-to-event outcome in the form of disease onset or incidence is often of interest, to help clinicians with decision making. Several studies deploy survival-based risk models on clinical multifactorial data to reveal determinants of health outcomes, such as healthspan. A model that is regularly used for this purpose is the Cox Proportional Hazard Model. For example, **?** find novel biomarkers that associate a healthy lifestyle score to all-cause mortality, cardiovascular disease and cancer risk by deployment of a Cox regression model. Similarly, **?** study the incidence of coronary heart disease, type 2 diabetes, atrial fibrillation, breast cancer and prostate cancer in relation to polygenic risk score derived from genomic information using a Cox proportional hazard approach. **?** build a Cox-Gompertz proportional hazard model to predict the age at the end of healthspan depending on a set of demographic and genetic variables. They define healthspan as an integrated quantity, based on the incidence of cancer, dementia, COPD, congestive heart failure and diabetes; chronic diseases that follow Gompertz dynamics. In line with this healthspan approach, **?** use the first incidence of either myocardinal infarction, heart failure stroke, dementia, hip fracture, cancer, or death as the target in their Cox proportional hazard model. They find 8 single nucleotide polymorphisms (SNPs) that predict risk of major disease, and evaluate candidate genes for ageing by genome-wide association study (GWAS). However, the aforemetioned methods do not exploit the longitudinal nature of many clinical studies, where periodic follow-up beyond baseline produces updated biomarker information that can improve inference and risk prediction. Such dynamic survival models can incorporate time-varying covariates or account for time-varying effects, and play a vital role in individualized clinical decision making. This is the type of model that we will consider in this paper.

Several clinical studies have used dynamic Cox models to investigate the relationship between time-varying covariates or coefficients, and disease outcomes. Inclusion of time-varying elements in a Cox model entails relaxation of the proportional hazard assumption, and is usually modelled using time-dependent Cox models or joint modeling of longitudinal and survival data (**?**). For example, **?** construct a dynamic Cox model through the landmarking approach and identify dynamic effects of treatment, albumin, creatinine, calcium, hematocrit and hemoglobin on amyotrophic lateral sclerosis (ALS) survival. They find that their dynamic approach better reflects the condition changes of patients in real time. In a paper by **?**, a time-varying Cox model is used to examine exposure to ambient particulate matter and incidence of hypertension to underline the effectiveness if air pollution mitigation to reduce the risk of cardiovascular disease. Similarly, **?** find that the neutrophil to lymphocyte ratio has a significant

time-varying effect and therefore use the extended Cox model to capture these biomarker changes during hospitalization on the rate of death of COVID-19 patients. Altogether, time-dependent variations of the Cox model are widely used in longitudinal studies, to capture the dynamic effect of covariates or the effect of dynamic covariates on health outcomes.

Inherent to using clinical assessment data in a suvival model to study healthspan, is the presence of censoring and the choise of time-scale. In short, censoring refers to incomplete information about the event of interest for some individuals in the study. It occurs when the event of interest does not take place during the study period or when the precise event time is unknown due to periodic follow-up, leading to right- and interval-censoring respectively. Left-censoring occurs when an event has already happened to an inividual before the start of the study. Generally, a survival model evaluates the association of *current* covariate values with the log hazard of an event *at that time*. Therefore, especially when unpredictable time-varying covariates are included in the analysis, left-censored individuals cannot be used to evaluate the covariate-event association. A problem that often arises when applying an extended Cox model with time-varying covariates to health data are interval-censored survival times. This type of survival times cannot be handled by conventional partial likelihood method to estimate the coefficients. **?** consider a maximum penalised likelihood approach that allows for partially interval-censored survival times, where a penalty function is used to regularise the baseline hazard estimate. Similarly, **?** propose an inverse probability weight to select event time pairs in the Cox proportional hazard model with interval censored data. Other approaches include middle point imputation or multiple imputation. It is recognized that such approaches can lead to bias. Moreover, when interval censored data is analyzed as as right-censored data, this can lead to significant bias in hazard ratio estimation (**?**). Another specification of a survival model is the choise of time-scale. In medical cohort studies, the choice of *time = 0* can either be start-of-study or age (time-since-birth). The specification is important because it determines which individuals are at risk at what time, and which individuals contribute to the likelihood function at a particular event time for estimating the coefficients. Typically in cohort studies, time-on-study is used in Cox regression models, adjusting for age as a covariate (**?**). Nonetheless, **?** propose to use age as the time-scale in Cox regression on data from a healthy population. They state that calender effects, for example due to medical advances, can be overcome by birth cohort stratification. Though being slightly more computationally intensive, using age as time-scale is less restrictive and more meaningful than using time-on-study as time-scale. Altogether, especially in medical studies that periodically monitor participants' biomarkers, taking interval-censoring and choise of time-scale into consideration is very important. Both specifications affects the hazard function and the interpretation of the estimated coefficients.

In conclusion, the field of ageing research has made significant progress in understanding the complex process of ageing and its impact on healthspan. Previous studies have explored various methods to

quantify ageing and identify biomarkers that better measure the rate of ageing than chronological age. The shift from longetivity research to healthspan research highlights the importance of not only extending the years of life but also improving the quality of those years. Multimorbidity, the co-occurrence of multiple chronic conditions, has emerged as a key area of interest, as ageing is a major risk factor for chronic disease accumulation. Moreover, the availability of large clinical cohorts with longitudinal data has facilitated the application of survival modeling techniques to study healthspan. Dynamic survival models, such as time-dependent Cox models, have been succesfully deployed to capture the time-varying effects of covariates and provide a more accurate representation of the ageing process. Nonetheless, the growing magnitude and longitudinality of cohort studies offer the potential to further enhance our understanding of the biology of healthspan and ageing. By leveraging the extensive data available in these studies, future research can delve deeper into the determinants of healthy ageing, develop interventions to promote it, and ultimately contribute to improving the overall well-being of older adults.

2

Data

In this chapter we introduce the data. Section **??** offers a general description of the data used in this paper. The subsequent section **??** outlines the selection process of diseases included in the target. In section **??**, an overview of the collected information is provided. Next, in section **??** we highlight our approach to defining disease incidence, and in section **??** we discuss the feature selection process. Lastly, in section **??** we discuss how missing data is handled.

2.1 Lifelines

The study was conducted with data from the Lifelines cohort, which is a large multi-generational study based in the northern part of the Netherlands (**?**). It was established by the UMCG in 2006, and primarily aims at gaining insight into the interactions between environmental, phenotypic and genotypic risk factors that affect the development of chronic diseases and healthy ageing. At baseline, data were collected for 167,729 participants ranging in age from 6 months to 93 years. The study involves regular physical examinations, cognitive tests, lung function and electrocardiogram (ECG), and extensive questionnaires completed every 5 years at a Lifelines location. In addition, participants complete follow-up questionnaires approximately every 1.5 years, providing insight into changes in behavior over time. Exclusion criteria include severe psychiatric or physical illnes, a limited life expectancy (< 5 years) or insufficient proficiency of the Dutch language. The data is provided by the University Medical Centre Groningen and the Lifelines research office and can be accessed via a secure Linux environment running on the high-performance cluster of the UMCG. All participants signed an informed consent form before participation. Moreover, the Lifelines Cohort Study is conducted according to the principles of the Declaration of Helsinki and in accordance with research code of the UMCG.

2.2 Healthspan

The target of the time-to-event analysis conducted in this paper is the incidence age of first disease from a shortlift of selected diseases. This incidence age is defined as the healthspan, or disease-free survival time of an individual. The diseases on the shortlist are Chronic Obstructive Pulmonary Disease (COPD), stroke, diabetes, dementia and cancer. The diseases on the shortlist are selected based on a number of criteria.

Firstly, the diseases are selected based on their chronic nature, their impact on an individual's ability to function, and their relatively equal effect on Health-related Quality of Life (HRQoL). They are highly associated with mortality and have a high risk factor attribution according to the Global Burden of Disease (**?**). All selected diseases adhere to this criterium, as can be seen in table **??**. This table contains the metrics as well as an upper and lower confidence limit. The Global Burden of Disease study is led by the Institute of Health Metrics and Evaluation at the University of Washington, and is the most comprehensive observational epidemiological study to date. It tracks mortality and morbidity in 204 countries and is an important tool for understanding the changing health challenges that exist in the world.

Table 1: Global Burden of Disease 2019

	Percentage attributable of total deaths	Risk factor attribution
Stroke	11.59% (10.78% - 12.22%)	84.96% (81.16% - 88.93%)
Diabetes	2.74% (2.58% - 2.87%)	100%
COPD	5.8% (5.19 %- 6.27%)	79.15% (76.00% - 82.08%)
Neoplasms	17.83% (16.87 %- 18.55%)	44.16% (41.04% - 48.15%)
Alzheimer's disease and other dimentias	2.87% (0.70% - 7.51%)	31.08% (20.17% - 44.20%)

Secondly, the diseases are selected on their association with age and their prevalence and incidence numbers. The prevalence and incidence numbers of the diseases in Lifelines can be found in table **??**. Moreover, all selected diseases strongly correlate with age. (correlatie coefficienten toevoegen).

Table 2: Disease prevalence before start of study and incidence rate during study

	Prevalent cases	Incidence percentage
Stroke	1034	0.49%
Diabetes	3527	1.91%
COPD	7770	2.12%
Cancer (all types)	6628	2.63%
Dementia	18	0.10%

Lastly, in the selection process, several clinical experts from the UMCG have been consulted. They agreed to the specified shortlist of diseases and actknowledge the power of using the defined healthspan target to study ageing.

2.3 Data overview

The Lifelines cohort consists of 3 main assessments, and 4 intermediate assessments. Information on disease presence and development is collected through questionnaires. Baseline assessment 1a contains information on presence of a disease before start of study. Follow-up assessments 1b, 1c, 2a, 3a and 3b contain information on disease development since the last time a Lifelines questionnaire was filled in. This structure allows for determination of between what ages a disease has developed, based on the assessments that an individual participated in. Besides disease presence and development information, Lifelines contains extensive information on demographics, lifestyle, psychosocial aspects and haematological and biochemical measures. The majority of the data is collected during the baseline assessment, referred to as assessment 1a. The subsequent main assessments, 2a and 3a, primarily contain follow-up information, which overlaps significantly with the baseline assessment 1a. The intermediate assessments, 1b, 1c, 2b, and 3b, include a smaller subset of information. An overview of the number of variables and overlapping variables before feature selection is provided in Table **??**:

Table 3: Overview of variables and overlap with baseline of all assessments

Assessment	Nr of columns	Nr of overlapping columns with 1a
1a (*)	2062	2062
1b	120	109
1c	118	107
2a	982	743
2b	43	39
3a	1063	802
3b	86	76
2a + 3a (**)	1374	980

* baseline assessment

** merged with inner join

An overview of data that is collected through questionnaires and clinical visits can be found in the data catologue of Lifelines. An overview of the information collected in the baseline questionnaire can be found in Table **??**, and an overview of the measurements collected during the clinical visits can be found in Table **??** in Appendix **??**.

2.4 Disease incidence ruling and Censoring

In clinical studies where event status updates and covariates are collected during periodic follow-up assessments, censoring is very common. Generally, there are three variations; left-, right- and interval-censoring. Participants who have not developed a chronic disease before the end of the follow-up period are labeled as right-censored. Right censoring can either occur due to end-of-study or due to loss-to-follow-up. When a participant enters the study with a disease of interest already present, this is a case of left-censoring. Interval-censoring occurs when the event occurs inbetween two clinical assessments, and the exact time of incidence is unknown. It is assumed that censoring are non-informative about the event, regardless of the type of censoring. Left-censored individuals are not taken into account in the analysis, because it is impossible to evaluate the association of time-varying covariates with chronic disease incidence when the event has already occured. Moreover, the follow-up questions with which disease incidence is derermined are of the form: *'Did the health problems listed below start since the last time you filled in a Lifelines questionnaire?'*. This question will inherently result in interval-censoring, because it does not provide the specific incidence time of disease. In addition, this question makes that disease incidence time is conditional on what assessments a participant took part in.

The follow-up structure of the data and the target requires a custom disease incidence ruling and covariate matching approach. Not all participants have participated in every assessment, and for some participants disease development information or covartiates are missing for some assessments. Moreover, besides a set of constant covariates, there are measures that will vary over time and consequently over assessments. The effect of both the constant and time-varying variables on the outcome will be assessed in this paper. In order to do so, given the aforementioned missingness of data and censoring, custom rules of exlusion and disease incidence determination are required. The dataset schema required for the time-varying Cox model is the *long* format. This schema contains one row per successive assessment set, including an ID, left (exclusive) timepoint, right (exclusive) timepoint, explanatory variables and an event indicator. The explanatory data is linked to the left timepoint, or the left clinical assessment of the set. The event indicator is linked to the right timepoint. This means that the explanatory data that is collected at a particular assessment, is linked to the time between that assessment and the next assessment, and is associated with the event occuring between those assessments or not. This coding scheme assumes tht there is no interval-censoring. Furthermore, both age and time-on-study are included in the initial survival set, as well as an assessment and assessment difference indicator. This particular data structure allows for time-varying covariates. For example, the following example survival table in table **??** tracks thee individuals:

Table 4: Long format example

id	*start_age*	*stop_age*	*start*	*stop*	*var1*	*var2*	*ass_start*	*ass_stop*	*ass_diff**	*event*
1	54	56	0	2	1	0.1	1a	1b	1	0
1	56	57	2	3	1	0.2	1b	1c	1	0
1	57	59	3	5	1	0.4	1c	2a	1	0
2	26	27	0	1	0	0.4	1a	1b	1	0
2	27	30	1	4	0	0.2	1a	1c	2	1
3	69	70	0	1	0	0.3	1a	1b	1	0
3	70	74	1	5	0	0.4	1b	1c	1	1

* based on assessment sequence 1a, 1b, 1c, 2a, 3a, 3b

In this dataset, *var1* is a constant variable and *var2* is a time-varying variable. Given this format, individuals who have only participated in assessment 1a are excluded. Furthermore, participants that have a row where *ass_diff** is larger than 2 are excluded. Including these participants would introduce too much uncertainty about the association between disease development and the time-varying covariates. Lastly, participants with too many missing variable are excluded, but this is discussed in section **??**.

2.5 Feature selection and preprocessing

Given the high-dimensional nature of Lifelines, variable selection is a fundamental step in the modelling process. A parsimonous model will increase interpretation, and is therefore preferred. Nonetheless, we want to take into account as much information as possible, as Lifelines encompasses a wide range of potential predictors. To achieve this, features are selected based on existing literature and their relevance to research hypotheses. Additionally, dimensionality reduction techniques are employed to further reduce the feature space, as elaborated in the Methodology (section **??**). Moreover, in preprocessing, the time-variablity of the variables is analysed, and decisions are made on whether to allow a feature to be dynamic or not.

In this study, features are selected based on a number of criteria. Predictors that are known to be associated with the risk of developing the diseases specified in the target are included in the analysis. The exclusion of such important factors will introduce omitted variable bias. Omitted variable bias entails that the model falsely attributes the effect of the omitted factors to the included variables. However, a certain amount of omitted variable bias is inevitable, as Lifelines is not an exhaustive study. Moreover, the presence of a feature in hypothesis automatically guarantees the inclusion of this feature in the analysis set. For example, it is hypothesized that the total amount of stress from long-term/chronic stressors experienced by the participant influences the development of disease. This factor is therefore included in the analysis set.

In preprocessing, the time-variability of variables is assessed, and dynamic or static transformations are performed accordingly. The only variable that is static in Lifelines is gender. However, there are more variables that we take into account as static in analysis. Whenever a variable barely changes over time, it is set to a single value per individual by means of mean imputation. Moreover, when the dynamic nature of a variable is not of interest, a static tranformation is applied, and the variable is considered a constant throughout analysis. In the end, this method will result in a set of time-varying covariates, and a set of constant baseline covariates. It is essential that the time-varying variables are available in as many assessments as possible, as the applied method does not allow for empty values. Given that the explanatory data is linked to the left timepoint, and that assessment 2b does not contain disease development information, the maximum number of time-varying variables depends on the overlap of assessments 1a, 1b, 1c, 2a and 3a. However, when a certain variable is missing or empty in one of the follow-up assessments, the information will be imputated by carrying forward the most recent available information. This technique is called the last observation carried forward (LOCF) method. For example, when a participants' body mass index (BMI) is available in assessments 1a, 1c and 2a, the BMI of

assessment 1b and 3a will equal the BMI measured in assessment 1a and 2a respectively. This method, as described in **?**, is selected because it safeguards the sample size and it is easily implementable. In order to fit the model that is applied in this study knowledge of all selected features at every potential event time is required. The LOCF technique prevents exclusion of participants for which a part of this information is missing, and therefore reduces selection bias. Selection bias, which is a side effect of visit-specific missingness, occurs when the population of participants with incomplete information is different from the population of participants with complete information. The downside of this method is that it also introduces bias by discretizing and pulling forward of continuously changing features. However, we conclude that the advantages associated with the LOCF method outweigh the limitations. Hence, this method is applied in order to maximize the sample size and allow more boundless inclusion of time-dependent effects.

After feature selection and preprocessing, the remaining variable set will regularized further by the application of dimensionality reduction techniques. Hereafter they will be assessed on their significance and effect on the outcome.

2.6 Missing data

Missing data is a prevalent phenomenon in clinical follow-up studies, and refers to the absence of information for certain features or observations for participants in the study. Missing data can arise due to various reasons such as participant attrition, loss to follow-up, or incomplete data collection. The occurrence of missing data poses challenges for researchers and can potentially introduce biases and affect the validity and reliability of study findings.

In clinical follow-up studies, missing data can be categorized into different types (**?**). Missing Completely at Random (MCAR) refers to missingness that occurs randomly and independently of any observed or unobserved data. It is the considered to be the strongest and most unrealistic type of missingness. Missing at Random (MAR) implies that the probability of missingness is dependent on observed data, but not on the unobserved data. Lastly, Missing Not at Random (MNAR) occurs when the probability of missingness is related to both observed and unobserved data, leading to potentially biased estimates. In this study, it is assumed that the missing data is MAR. Under this assumption, bias can be mitigated by applying appropriate imputation techniques.

To address the issue of missing data in this study, various methods can be applied. These include complete case analysis, where only participants with complete data are included in the analysis. However, this results in a loss of information and a reduced sample size. Another approach is to perform imputation, which involves filling in missing values using statistical techniques. As introduced in section **??**, missing

constant data is imputed by means of mean imputation, and missing time-varying data is imputed by means of the LOCF technique. However, to minimize potential bias, participants who have more than 70% imputed data are excluded. More sophisticated imputation methods such as multiple imputation or maximum likelihood estimation are not considered in this study, as the aim of this study is to study factors associated to healthspan, and not to study data imputation techniques.

Altoghether, it is important to note that while these methods help mitigate biases associated with MAR data, they rely on the assumption that the missingness mechanism is correctly specified. Moreover, mean imputation and LOCF are not the optimal imputation methods for MAR data, so biases may still persist. However, these methods do address the missing data in this study, and they aim to reduce bias by optimizing the sample size. This will in turn help to draw accurate conclusions regarding the association between features of interest and human healthspan.

3

Methodology

The main objective of multivariate survival modelling is to understand and quantify factors that influence the time until an event occurs. In this study, the event of interest is the time to onset of a first disease from a shortlift of selected diseases. There are various types of multivariate survival models, including (semi-)parametric statistical appoaches and several machine- or deep-learning approaches. In this paper, we will focus mainly on the semi-parametric Cox Proportional Hazard Model, and its dynamic extension.

Let T be a random response variable representing time, then the survival function of a population is defined as $S(t) = P(T > t)$. $S(t)$ represents the probability of not experiencing the event up to and including time t, or surviving past time t. On the other hand, the harzard rate. $h(t)$, is the instantaneous risk of an event occuring at time t given that it has not occured up to time *t-1*. Mathematically, the hazard rate can be defined as the limit of the conditional probability of the event occurring within the infinitesimal interval $(t, t + \delta t)$ given that $T > t$, divided by the infinitesimal interval length δt. This can be expressed as:

$$h(t) = \lim_{\delta t \to 0} \frac{\mathrm{P}(t \leq T \leq t + \delta t \mid T > t)}{\delta t}$$

Additionally, the cumultive hazard function, $H(t)$ represents the accumulated risk up to time t, and is defined as: $H(t) = \int_0^t h(u)du$. knowledge of any two of these functions enables the computation of the third function, as $S(t) = exp(-H(t))$ and $h(t) = \frac{-S'(t)}{S(t)}$. A thorough understanding of these functions is essential for the remainder of this section.

In the first part of this section, a comprehensive overview of the Cox Proportional Hazard (CPH) model and its time-varying variant, the Extended Cox model (ECM), is provided. In the second part, we discuss model estimation, dimsensionality reduction and model evaluation techniques.

3.1 Cox Proportional Hazard Model

The CPH model is a *regression* model that attempts to model the hazard rate $h(t|\mathbf{Z})$ as a function of time t and the vector of covariates $\mathbf{Z}$. Mathematically, the CPH is represented by:

$$\underbrace{h(t \mid \mathbf{Z})}_{\text{hazard rate}} = b_0(t)\exp(\beta\mathbf{Z}) = \overbrace{b_0(t)}^{\text{baseline hazard}} \underbrace{\exp\left(\sum_{j=1}^{n}\beta_j\mathbf{Z}_j\right)}_{\text{partial hazard}}$$

According to this specification, the log-hazard of an individual is a linear function of their covariates $\mathbf{Z}$ and a population-based baseline hazard $b_0(t)$. Note that the only time-component in this model is the baseline hazard. The partial hazard, which is dependent on the subject specific covariates, is the time-invariant scaling factor that either inflates or deflates the baseline hazard. This also implies that survival curves can never cross each other. The baseline hazard can be estimated using different methods, such as *Breslow*.

A fundamental assumption of the standard CPH is that the hazard ratio adheres to the *proportional hazard assumption*. This assumption implies that the hazard ratio is constant over time for all levels of the covariates. The hazard ratio in a CPH model can be presented by:

$$\frac{h(t \mid \mathbf{Z} = \mathbf{z})}{h(t \mid \mathbf{Z} = 0)} = \exp(\beta\mathbf{z})$$

The hazard ratio depends on covariates $z_1, ..., z_p$, but is independent of time t. It is a measure of the relative effect of a particular covariate on the hazard rate. It quantifies the ceteris paribus change in hazard rate when there is a unit change of a particular predictor covariate. Hence, a hazard ratio that is equal to 1 indicates that the covariate has no effect on the hazard rate. A hazard ratio greater than 1 implies an increased risk, and a hazard ratio lower than 1 implies a decreased risk of an event with a unit change of the predictor. The proportional hazard assumption should be tested and handled if violated. Violation results in biased and unreliable results, and can lead to misinterpretation of factors that influence survival. There are several approaches which address violation of the proportional hazard assumption, such as stratification or the use of time-dependent covariates. In stratified proportional hazard models, seperate Cox models are fit on different groups, which allows these groups to have a different baseline hazard. Another approach is to allow for time-varying covariates in a Cox model. This model, hereafter referred to as the *Extended Cox Model* (ECM), allows the hazard ratios to vary over time and provides a more accurate assessment of the impact of time-varying covariates on the event of interest. This is the model that will be considered in this paper.

3.2 Extended Cox Model (Cox's Time-varying Model)

The CPH model can be extended in such a way that it can incorporate covariates $Z_i(t)$ that change over time. This extension is possible because of the way the Cox model works: the current covariate values of the participant who had the event are compared to the current covatiate values of the participants who were at risk at that time. It is of great importance to clinical follow-up studies to be able include information that changes with time, as datasets usually include both baseline (time-independent) and time-dependent covariates. Mathematically, the ECM is represented by:

$$h(t \mid \mathbf{Z}(t)) = b_0(t) \exp(\beta \mathbf{Z}(t))$$

In this formula, $\mathbf{Z}(t)$ is a vector of covariates, of which at least one changes over time. For example, $Z_i(t) = const$ represents a constant characteristic of a participant, such as gender. On the other hand, $Z_i(t) = \sum_{j=1}^{n_i} z_{ij} \mathbb{I}_{[t_{i,j-1}, t_{ij})}(t)$ represents a piecewise constant process that gradually updates values over assessments. Examples of such variables are cholesterol concentration and smoking frequency. Whenever there is a time-interactive component added to the traditional Cox model, the proportional hazards assumption is violated. The hazard ratio for two different participants is time-dependent:

$$\frac{h(t \mid \mathbf{Z_1}(t))}{h(t \mid \mathbf{Z_2}(t))} = \frac{b_0(t)}{b_0(t)} \cdot \frac{\exp(\beta \mathbf{Z_1}(t))}{\exp(\beta \mathbf{Z_2}(t))} = \exp(\beta(\mathbf{Z_1}(t) - \mathbf{Z_2}(t)))$$

Note that the interpretation of the estimated coefficient of the constant covariate remains the same as in the CPH model; they represent the change in the hazard ratio associated with a unit change in the covariate. In contrast, the coefficients of the time-dependent covariates represent the instantaneous change in hazard ratio as a result of a unit change in the covariate *at a particular timepoint.*

A practical disadvantage of the ECM is that prediction of survival curves and individual survival times is not trivial. This would require knowledge about the covariate values beyond the observed times, and these are not available.

The package that is most commonly used for survival analysis is *Survival* in R, and *Lifelines* in Python. *Lifelines* fits the Extended Cox Model model using iterative gradient descent. Model estimation is elaborated on in the next section.

3.3 Model estimation

In the Extended Cox Model, the regression coefficients β are estimated with the partial likelihood method. This likelihood is constructed with the observed event times and knowledge of the order in which the

events occur. Suppose there are N individuals and D distinct event times. Let $(t_1, ..., t_D)$ be the D ordered, distinct event times, assuming that there are no tied event times. For any timepoint $t \geq 0$, the risk set that defines the set of individuals at risk of an event at time t is $R(t) := \{i \mid t_i \geq t\}$. Furthermore, let i_j denote the identity of the participant experiencing the event at time t_j, and H_j the history of the dataset up to the j-th event time. Inference is made via the partial-likelihood:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} P(i_j|H_j) = \prod_{j=1}^{d} \frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}_{i_j}(t_j))}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i(t_j))}$$

The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the partial likelihood function $L(\boldsymbol{\beta})$. Moreover, the function can be altered to suit the case of tied event times. The partial likelihood function has some nice properties, and is equal to the likelihood function used in the CPH model with time-constant covariates. The only difference is that the values of (Z) will vary over time and thus over risk sets. Conveniently, the partial likelihood function can be estimated with an unspecified baseline hazard, as it does not depend on b_0. The partial likelihood function also solely depends on the order of the events, and not on the specific event times. And lastly, right-censored individuals are only take into account in the risk set; resulting in an elegant incorporation of censored participants.

In Pythons' *Lifelines*, an iterative gradient descent algorithm is deployed to find the optimal coefficients of the ECM. The negative log-partial likelihood function, which is the objective function that is minimized in the optimization task is:

$$-\ell(\boldsymbol{\beta}) = -\sum_{j=1}^{d} \left[\boldsymbol{\beta}^T \mathbf{Z}_{i_j}(t_j) - \log \left(\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i(t_j)) \right) \right]$$

Note that this function should be continuous, differentiabel and convex. In most cases, the negative log-partial likelihood is a strictly convex function, and thus has a unique local optimum. In the case of convergence issues, *Lifelines* takes extra effort to help by giving specific warnings.

By taking the derivative with respect to β, we derive the *gradient* of the objective function:

$$\nabla(-\ell(\boldsymbol{\beta})) = -\sum_{j=1}^{d} \left[\mathbf{Z}_{i_j}(t_j) - \frac{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i(t_j)) \mathbf{Z}_i}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta}^T \mathbf{Z}_i(t_j))} \right]$$

The local minimum of the objective function is iteratively found by exploiting the gradient of the function around $\boldsymbol{\beta}$. Depending on the sign of the gradient, the approximated coefficients are updated in the direction that decreases the objective function $-\ell(\boldsymbol{\beta})$. An arbitrary starting point $\boldsymbol{\beta_n}$ is identified, which is sufficiently close to the minimum of $-\ell(\boldsymbol{\beta})$, a better approximation of $\boldsymbol{\beta}$ is computed by iterating

over this formula:

$$\boldsymbol{\beta_{n+1}} = \boldsymbol{\beta_n} - \alpha\Delta(-\ell(\boldsymbol{\beta}))$$

Gradient descent is parametric, as the learning rate (α) specifies the magnitude of the iterative steps. The magnituse of alpha should be selected carefully to guarantee successful optimization. A too small α results in slow convergence, while a too large α may overshoot the minimum, causing the algorithm to diverge. Nevertheless, a small learning rate provides more precise updates, as the iterative steps are smaller. Optimal learning rate selection involves finding a trade-off between convergence speed, stability, and avoiding local optima. Gradient descent is the method that is applied by *Lifelines* the the negative log-partial likelihood to find the ECM's regression coefficients.

3.4 Regularization

As introducted in the Data section, the feature space of Lifelines is not only reduced by means of feature selection, but also by regularization. The regularization techniques considered is elastic net, which is a combination of Lasso and Ridge regularization. Given the partial-likelihood equation in **??**, the classical estimation case is when there are more participants than predictors ($N \gg p$). Based on the large number of participants in Lifelines, the classical case will most probably apply, and estimation of the coefficients will be trivial. However, we do take into account dimensionality reduction techniques that are proposed to combat the case when $p > N$. In such cases, there exists a collinearity problem when applying partial-likelihood estimation to the Cox model, which sends all the estimated coefficients to $\pm\infty$. To reduce the feature space and avoid collinearity issues, elestic net regularization is applied to the ECM used in this paper.

Proposed by **?**, elastic net combines the strengths of both the Lasso (L1) and Ridge (L2) penalties by incorporating a convex combination of the two. This allows for the simultaneous selection of relevant variables and shrinkage of coefficients. On the one hand, the Least Absolute Shrinkage Selector Operator (Lasso) adds a penalty equal to the absolute value of the magnitude of the coefficients multiplied by the L1-ratio. The mathematical form of the L1 penalty is: $\textit{L1-ratio} \cdot ||\boldsymbol{\beta}||_1$. This penalty eliminates some of the coefficients by reducing them to zero, hereby reducing the number of features. However, Lasso does not work well with collinearity issues. It tends to randomly select either of the collinear covariates, which could lead to the elimination of relevant information. On the other hand, Ridge regularization adds the squared magnitude of the coefficients to the loss function multiplied by the L2-ratio. This penalty looks as follows: $\textit{L2-ratio} \cdot ||\boldsymbol{\beta}||_2^2$. This results in some coefficients that shrink towards. However, unlike Lasso regularization, Ridge will not shrink the coefficients to zero exactly, so these coefficients

will not be eliminated. This could be disadvantageous when the objective is to reduce the feature space. Elastic net combines the strengths of Lasso and Ridge by simultaneously perform feature selection and coefficient regularization. In *Lifelines*, the mathematical form of the Elastic Net regularization is: $\frac{1}{2}penalizer((1 - L1\text{-}ratio) \cdot ||\boldsymbol{\beta}||_2^2 + L1\text{-}ratio \cdot ||\boldsymbol{\beta}||_1)$. Note that $L2\text{-}ratio = 1 - L1\text{-}ratio$ and that *penalizer* and *L2-ratio* are parameters that can be specified when fitting the *CoxTimeVaryingFitter*. Increasing the *penalizer* will result in a stronger elastic net effect, and a more sparse solution. By increasing the *L1-ratio*, we can make the elastic net behave more like a Lasso penalty. Parameters for the elastic net are optimized by cross-validation or grid search, by optimizing for a specific evaluation metric. This evaluation metric is elaborated on in the next section.

3.5 Model evaluation

The performance of the ECM that is deployed in this paper is assessed by time-dependent evaluation metrics, proposed by **?**. In many medical scenarios, decision-making involves using updated patient information to forecast transitions in health status, such as, in our case, disease development. The objective is to utilize the patient's clinical characteristics to estimate the risk of an adverse event within a specific time frame and identify individuals at a high risk of experiencing such an event in the near future. As Lifelines contains participants' data of clinical follow-up assessments, the aim is to find prognostic factors that contribute to the high-risk profile of participants. Such participants can consequently be selected as candidates for more frequent monitoring of these prognostic factors. In this section, we elaborate on the performance evaluation method that is deployed to support this objective. The ECM's time-varying performance is evaluated by the area under the time-dependent receiver operating characteristic (ROC) curve (AUC). A fundamental concept of the ECM is the time-varying risk set of participants. At any assessment time, the set of participants that are disease-free and thus at risk of the event can be divided into cases (participants who will experience the event) and controls (participants who have not yet experienced the event). Moreover, cases can be divided into incident and cumulative cases; the latter being our specification of interest. Cumulative cases are participants that develop disease over a specified period of time. The goal of the ECM is to be able to best distinguish cases from controls at any time point. However, the prognostic information of participants changes over time, and thus this discriminative accuracy changes over time. This requires the traditional diagnostics of sensitivity and specificity to be extended to a time-varying variant. By convention, larger values of markers are assumed to more indicative of disease development in the future. The sensitivity of a marker is the probability that it is positive in the case of disease development. This equals the true positive rate (TPR) in the context of classification errors. Specificity is the probability that the marker is negative or small in the case that

disease is not developed. This equals one minus the false positive rate (1 - FPR). In the cumulative case with dynamic controls, these metrics are evaluated within some fixed time frame. The outcome is dichotomized at time t, and cumulative cases C are defined as participants who have experienced the event before time t, and dynamic controls D are participants who remain disease-free beyond time t. Let T denote survival time, ans s the start time of the interval of interest. Then, for a specific threshold C and for a continuous marker M, the time-dependent sensitivity and specifitiy is :

$$\begin{aligned} \text{sensitivity}^C(c|\text{start} = s, \text{stop} = t) &= P(M > c|T \geq s, T \leq t) \\ \text{specificity}^D(c|\text{start} = s, \text{stop} = t) &= P(M \leq c|T \geq s, T > t). \end{aligned}$$

Based on these definitions, the time-dependent AUC for cumulative cases and dynamic controls is definited by:

$$AUC^{C/D}(s,t) = P(M_j > M_k|T_j \geq s, T_j \leq t, T_k \geq s, T_k > t)$$

The $AUC^{C/D}(s,t)$ is the probability that a random participant who is a cumulative case has a larger marker value than a randomly selected dynamic control, under the assumption that both subjects are event free at time s.

Instead of individual marker values https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0057-6Sec3 https://academic.oup.com/bioinformatics/article/25/14/1775/225393

Bibliography

Bansal, A. and Heagerty, P. J. (2018). A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. *Medical Decision Making*, 38(8):904–916.

Bo, Y., Guo, C., Lin, C., Chang, L.-y., Chan, T.-C., Huang, B., Lee, K.-P., Tam, T., Lau, A. K., Lao, X. Q., et al. (2019). Dynamic changes in long-term exposure to ambient particulate matter and incidence of hypertension in adults: a natural experiment. *Hypertension*, 74(3):669–677.

Bonaccio, M., Di Castelnuovo, A., Costanzo, S., De Curtis, A., Persichillo, M., Cerletti, C., Donati, M., de Gaetano, G., Iacoviello, L., and sani Study Investigators, M. (2019). Impact of combined healthy lifestyle factors on survival in an adult general population and in high-risk groups: prospective results from the moli-sani study. *Journal of internal medicine*, 286(2):207–220.

Brüünsgaard, H. and Pedersen, B. K. (2003). Age-related inflammatory cytokines and disease. *Immunology and Allergy Clinics*, 23(1):15–39.

Canchola, A. J., Stewart, S. L., Bernstein, L., West, D. W., Ross, R. K., Deapen, D., Pinder, R., Reynolds, P., Wright, W., Anton-Culver, H., et al. (2003). Cox regression using different time-scales. *Western Users of SAS Software. San Francisco, California.*

Epel, E. S., Merkin, S. S., Cawthon, R., Blackburn, E. H., Adler, N. E., Pletcher, M. J., and Seeman, T. E. (2009). The rate of leukocyte telomere shortening predicts mortality from cardiovascular disease in elderly men. *Aging (Albany NY)*, 1(1):81.

Farrell, S., Mitnitski, A., Rockwood, K., and Rutenberg, A. D. (2022). Interpretable machine learning for high-dimensional trajectories of aging health. *PLOS Computational Biology*, 18(1):e1009746.

Geraili, Z., Hajian-Tilaki, K., Bayani, M., Hosseini, S. R., Khafri, S., Ebrahimpour, S., Javanian, M., Babazadeh, A., and Shokri, M. (2022). Evaluation of time-varying biomarkers in mortality outcome in covid-19: an application of extended cox regression model. *Acta Informatica Medica*, 30(4):295.

Global burden of disease (2019). Global burden of disease. https://www.thelancet.com/gbd.

Goldberg, E. L. and Dixit, V. D. (2015). Drivers of age-related inflammation and strategies for healthspan extension. *Immunological reviews*, 265(1):63–74.

Heller, G. (2011). Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime data analysis*, 17:373–385.

Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):1–20.

Huang, B., Geng, X., Yu, Z., Zhang, C., and Chen, Z. (2023). Dynamic effects of prognostic factors and individual survival prediction for amyotrophic lateral sclerosis disease. *Annals of Clinical and Translational Neurology*.

Kom, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American journal of epidemiology*, 145(1):72–80.

Kuh, D., Karunananthan, S., Bergman, H., and Cooper, R. (2014). A life-course approach to healthy ageing: maintaining physical capability. *Proceedings of the Nutrition Society*, 73(2):237–248.

Lifelines (2023). Lifelines. https://www.lifelines.nl/.

Liu, Y. and Craig, B. A. (2006). Incorporating time-dependent covariates in survival analysis using the lvar method. *Statistics in medicine*, 25(10):1729–1740.

Mack, C., Su, Z., and Westreich, D. (2018). Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide.

Marmot, M. (2005). Social determinants of health inequalities. *The lancet*, 365(9464):1099–1104.

Mars, N., Koskela, J. T., Ripatti, P., Kiiskinen, T. T., Havulinna, A. S., Lindbohm, J. V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature medicine*, 26(4):549–557.

Mitnitski, A. B., Mogilner, A. J., and Rockwood, K. (2001). Accumulation of deficits as a proxy measure of aging. *TheScientificWorldJournal*, 1:323–336.

Niccoli, T. and Partridge, L. (2012). Ageing as a risk factor for disease. *Current biology*, 22(17):R741–R752.

Sprott, R. L. (2010). Biomarkers of aging and disease: introduction and definitions. *Experimental gerontology*, 45(1):2–4.

Sun, X. and Chen, C. (2010). Comparison of finkelstein's method with the conventional approach for interval-censored data analysis. *Statistics in Biopharmaceutical Research*, 2(1):97–108.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. (2009). Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363.

Walter, S., Atzmon, G., Demerath, E. W., Garcia, M. E., Kaplan, R. C., Kumari, M., Lunetta, K. L., Milaneschi, Y., Tanaka, T., Tranah, G. J., et al. (2011). A genome-wide association study of aging. *Neurobiology of aging*, 32(11):2109–e15.

Wang, Z., Li, L., Glicksberg, B. S., Israel, A., Dudley, J. T., and Ma'ayan, A. (2017). Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *Journal of biomedical informatics*, 76:59–68.

Webb, A. and Ma, J. (2023). Cox models with time-varying covariates and partly-interval censoring–a maximum penalised likelihood approach. *Statistics in medicine*, 42(6):815–833.

Wendel-Vos, G. W., Schuit, A. J., Saris, W. H., and Kromhout, D. (2003). Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. *Journal of clinical epidemiology*, 56(12):1163–1169.

Wikström, K., Lindström, J., Harald, K., Peltonen, M., and Laatikainen, T. (2015). Clinical and lifestyle-related risk factors for incident multimorbidity: 10-year follow-up of finnish population-based cohorts 1982–2012. *European journal of internal medicine*, 26(3):211–216.

World Health Organization (2019). Global health estimates : Life expectancy and leading causes of death and disability. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

Zenin, A., Tsepilov, Y., Sharapov, S., Getmantsev, E., Menshikov, L., Fedichev, P. O., and Aulchenko, Y. (2019). Identification of 12 genetic loci associated with human healthspan. *Communications biology*, 2(1):41.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., and Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7).

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

A

Appendix

I Lifelines data overview

Table 5: Overview of subcategories in the baseline questionnaires for adults (18 years and older)

Subcategory	**Examples**
GENERAL INFORMATION	
Demographics	Age, nationality, marital status
Family composition	Children
Employment	Income, function, unemployment
Education	Highest level of education
HEALTH	
Health status	Prevalence of disease, disabilities, disorders
Medical treatment	Medication, doses
Healthcare use	Contact with healthcare professionals
Questions for females	Number of pregnancies, age at menopause
Birth and development	Birthweight, birth defects, breastfeeding
LIFESTYLE & ENVIRONMENT	
Physical activity	SQUASH **?**
Nutrition	Diet score, alcohol abuse
Smoking	Past and current smoking activity and frequency
Activities	Volunteer work, sleep, hobbies
Physical environment	home environment, pets
PHYSIOLOGICAL VARIABLES	
Quality of Life	Health Related Quality of Life survey (RAND-36)
Health perception	Fitness rate
Personality	Neuroticism-Extroversion-Openness Personality Inventory (NEO), Anxiety Sensitivity Index (ASI)
Stress	List of Threatening Experiences (LTE), Longterm Difficulties Inventory (LDI)
Social support, independence	Social production function (SPF-IL)

Table 6: Overview of haematological and biochemical measures

Subcategory	**Measure**
BLOOD	
Haematology	Haemoglobin
	Haematocrit
	Leukocytes and differentiation
	Thrombocytes
Diabetes	Glucose
	HbAIc
Lipids	Total cholesterol
	HDL-cholesterol
	LDL-cholesterol
	Triglycerides
	Apolipoprotein A1
	Apolipoprotein B100
Electrolytes	Sodium
	Potassium
	Calcium
	Phosphorus
Renal function	Creatinine
	Urea
	Uric acid
Liver and inflammation	Aspartate aminotransferase
	Alanine aminotransferase
	Alkaline phosphatase
	Gamma glutamyl transferase
	Albumin
	High sensitivity C-reactive protein
Thyroid function	Thyroid stimulating hormone
	Free T4
	Free T3
URINE	
Morning sample	Albumin
	Creatinine
24-h collection sample	Albumin
	Creatinine