Healthy Ageing

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

May 2023

Abstract

Test test

Contents

1 Introduction **3**

2 Literature Review **5**

3 Data **10**
3.1 Lifelines . . . 10
3.2 Healthspan . . . 11
3.3 Data overview . . . 12
3.4 Disease incidence ruling and Censoring . . . 13
3.5 Feature selection and preprocessing . . . 15
3.6 Missing data . . . 15
3.7 Censoring and choice of time-scale . . . 15

4 Methodology **16**
4.1 Cox Proportional Hazard Model . . . 17
4.2 Extended Cox Model (Cox's Time-varying Model) . . . 18
4.3 Model estimation . . . 18
4.4 Model evaluation . . . 19

A Appendix **24**
I Lifelines data overview . . . 25

1

Introduction

Ageing is defined as the process of time-dependent functional decline of biological organisms and manifests itself on a variety of physiological scales López-Otín et al. (2013). It is the result of accumulation of cellular and molecular damage over time, and is the highest risk factor for susceptibility to diseases and ultimately death. The progression of ageing relies on the balance between exposure and resilience to damaging factors, which are both subject to the heterogeneity of individuals. Huge differences in human lifespan suggest that there exist underlying differences in ageing processes and that the passage of time is not the ideal measure of ageing speed Sprott (2010). Consequently, a person's chronological age is merely a proxy for the rate of ageing, but not a reliable reflection of general health status. Instead, ageing can be quantified through the identification and measurement of biomarkers in the form of biological age. Predicting and identifying factors that influence biological age has rapidly gained popularity in the field of ageing research over the years. Given that human life expectancy is continuously increasing, but not in parallel with increased healthy lifespan, understanding the biological cause of ageing and increasing health in elderly has become one of the most urgent societal and scientific challenges of today.

Previous studies on the progression of ageing have deployed diverse modelling techniques that aim to capture the impact of health conditions on age. The more simple approaches include integrating multiple variables into a low-dimensional representation of age, such as a frailty index defined by the proportion of accumulated health deficits Mitnitski et al. (2001). Alternatively, multivariate linear regression techniques have been explored to develop formulas Levine (2013) and prediction models Bae et al. (2008) for biological age, as well as a DNA methylation (DNAm) ageing clock Horvath (2013). Other papers employ principle component analysis (PCA) to assess biological age, in which the 1st principle component score is transformed into biological age using a T-score transformation Nakamura et al. (1988),Park et al. (2009). However, given increased computational power and availability of large medical datasets,

there is a potential to advance our understanding of the complex and high-dimensional ageing process even more Farrell et al. (2021). This observation fuelled the emergence of advanced deep learning based approaches, such as a deep hematological aging clock and a deep learning prediction model of biological age based on electronic medical records Wang et al. (2017).

Nevertheless, there is a lot of ground to be gained as Such techniques take chronological age as the dependent variable and label the predicted variable as the biological age. By design, all of such supervised biological age prediction techniques aim to minimize the difference between chronological age and predicted biological age of an individual. Approach suffers from a theoretical contradiction.

ith advent algorithms that employ deep learning, through non-linear transformations, it is now possible to better handle data incompleteness, inaccuracy, and scalability to learn from EMR. For instance, DeepPatient uses denoising Autoencoder (dAE) t

– deep ageing clocks

However, there is a lot of ground to be gained as

–¿ complexity non-linear process

The merging of multiple measures into a single biomarker of may prove useful in both biological research - to study how lifestyle, environment and evolution impact ageing speed - as well as in public health research or clinical practice - to indentify individuals at increased risk of disease.

Advanced analytical methodologies in pattern recognition and computational learning, as Machine Learning approaches, can also be employed to explore factors associated with the metric of health. Even so, the aging process is complex and has multiple interacting physiological scales from the molecular to cellular to whole tissues. In the face of this complexity, we can significantly advance our understanding of aging with the use of computational models that simulate realistic individual trajectories of health as well as mortality. T

with artificial intelligence methods advancing and large data sets becoming more publicly available, there is an opportunity to deepen the understanding of multiple underlying mechanisms that influence the rate at which people age.

which makes ageing research high-dimensional and complex.

2

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 (World Health Organization, 2019). 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. Mitnitski et al. (2001) 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 Sprott (2010) 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 (Epel et al., 2009) or DNA methylation (DNAm) patterns Horvath (2013), 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, Farrell et al. (2022) 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, Wang et al. (2017) 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 (Niccoli and Partridge, 2012). Multimobidity refers to the co-occurence of two or more chronic conditions in an individual (Valderas et al., 2009), 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 (Wikström et al., 2015). Moreover, it is well estabislished that there is an association between socioeconomic status and multimorbidity (Marmot, 2005). 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 Goldberg and Dixit (2015) 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 (Brüünsgaard and Pedersen, 2003). 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 Kuh et al. (2014), 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, Bonaccio et al. (2019) 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, Mars et al. (2020) 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. Zenin et al. (2019) 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, Walter et al. (2011) 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 (Zhang et al., 2018). For example, Huang et al. (2023) 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 Bo et al. (2019), 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, Geraili et al. (2022) 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. Webb and Ma (2023) 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, Heller (2011) 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 (Sun and Chen, 2010). 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 (Canchola et al., 2003). Nonetheless, Kom et al. (1997) 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.

3

Data

In this chapter we introduce the data. Section 3.1 offers a general description of the data used in this paper. The subsequent section 3.2 outlines the selection process of diseases included in the target. In section 3.3, an overview of the collected information is provided. Next, in section 3.4 we highlight our approach to defining disease incidence, and in section 3.5 we discuss the feature selection process. Lastly, in section 3.6 we discuss how missing data is handled.

3.1 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 (Lifelines, 2023). 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.

3.2 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 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). Furthermore, selection criteria include that they are highly associated with mortality and have a high risk factor attribution according to the Global Burden of Disease (Global burden of disease, 2019). 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. Lastly, in the selection process, several clinical experts from the UMCG have been consulted.

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.

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

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

3.3 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. On the other hand, 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 3:

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 5, and an overview of the measurements collected during the clinical visits can be found in Table 6 in Appendix A.

3.4 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 4 tracks thee individuals:

Table 4: Long format example

id	*start_age*	*stop_age*	*start*	*stop*	*var1*	*var2*	*assessment_start*	*assessment_stop*	*assessment_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 *assessment_diff** is 3 are excluded. Including these participants would introduce too much uncertainty about when 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 3.6.

3.5 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 there preferred.

3.6 Missing data

3.7 Censoring and choice of time-scale

Inherent to

The Lifelines dataset has three main assessments. Chronic disease presence and incidence is self-reported in questionnaires. For each condition, presence is reported at the baseline assessment 1a, and incidence is reported at follow-up assessments 2a and 3a. The question on disease follow-up is formulated as 'did the health problems listed below start since the last time you filled in the lifelines questionnaire?'.

Although survival analysis is designed to deal with estimation of right-censored data, it is important to understand the risks of underestimating the true survival time when ignoring right-censored individuals.

Another challenge inherent to duration is the choice of time-scale, especially in the analysis of large-scale health surveys where the interest is to find the association of risk factors with the development of a disease. In this case, either the time since the baseline survey (time-on-study), or age can be used as time-scale. Kom et al. (1997) Kom et al. (1997) propose that age, with stratification for birth cohort effects, is the appropriate time scale in Cox proportional hazard regression models that analyze health data from longitudinal studies.

4

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 (what techniques we applied, the steps I took?) and model evaluation techniques.

4.1 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.

4.2 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.

4.3 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 $(\tau_1, ..., \tau_D)$ 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 τ_j, and H_j the history of the dataset up to the j-th event time. The partial likelihood function can then be written as:

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

The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the likelihood function $L(\boldsymbol{\beta})$. This 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 likelihood function can be estimated with an unspecified baseline hazard, as it does not depend on b_0. The 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.

4.4 Model evaluation

Bibliography

Bae, C.-Y., Kang, Y. G., Kim, S., Cho, C., Kang, H. C., Yu, B. Y., Lee, S.-W., Cho, K. H., Lee, D. C., Lee, K., et al. (2008). Development of models for predicting biological age (ba) with physical, biochemical, and hormonal parameters. *Archives of gerontology and geriatrics*, 47(2):253–265.

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.

Farrell, S., Stubbings, G., Rockwood, K., Mitnitski, A., and Rutenberg, A. (2021). The potential for complex computational models of aging. *Mechanisms of Ageing and Development*, 193:111403.

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.

Levine, M. E. (2013). Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 68(6):667–674.

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

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–1217.

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.

Nakamura, E., Miyao, K., and Ozeki, T. (1988). Assessment of biological age by principal component analysis. *Mechanisms of ageing and development*, 46(1-3):1–18.

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

Park, J., Cho, B., Kwon, H., and Lee, C. (2009). Developing a biological age assessment equation using principal component analysis and clinical biomarkers of aging in korean men. *Archives of gerontology and geriatrics*, 49(1):7–12.

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).

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 Wendel-Vos et al. (2003)
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