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

Monday 15th May, 2023

Abstract

Test test

Contents

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.

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 healtspan 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 healtspan 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 infomration 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).

Data

In this chapter we introduce the data. Section 0.1 offers a general description of the data used in this paper. The subsequent section 0.2 outlines the selection process of diseases includes in the target. In particular, in section 0.3 we highlight our approach to defining disease incidence, and in section 0.4 we discuss the feature selection process. Lastly, in section 0.5 we discuss how missing data is handled.

0.1 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.

0.2 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 disease. 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). 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%)

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%

0.3 Disease incidence ruling

The Lifelines cohort consists of 3 main assessments, and 4 intermediate assesments. Information on disease presence and development is collected through questionnaires. Baseline assessment 1a contains information on prevalence of disease. Given that this study aims to investigate the dynamics of factors contributing to disease incidence, participants who have disease prevalence at or prior to baseline are excluded from analysis.

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 4, and an overview of the measurements collected during the clinical visits can be found in Table 5 in Appendix 1.

The follow-up structure of the data and the target requires a custom disease incidence ruling approach. Not all participants have participated in every assessment, and for some participants disease development information is 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, custom rules of exlusion and time-dependent covariate selection are required. As mentioned, participants who have disease prevalence at or prior to baseline are excluded from analysis. 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?'*. Therefore, disease incidence time is inherently conditional on what assessments a participant took part in.

0.4 Feature selection and preprocessing

After feature selection and preprocessing...

0.5 Missing data

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 chronic disease. Survival analysis entails the estimation of a *survival function*, denotes as $S(t)$, which represents the percentage of participants for which the event has not occurred up to and including time t. A survival curve may be interpreted as the individual probability of living past time t without onset of a first chronic disease. Consequently, a survival curve is defined as a collection of *hazard rates h(t)*, where *h(t)* is the probability of the event occuring at time t given that it has not occured up to time *t-1*. The effect of covariates on time-to-event can be assessed by comparing survival curves for different covariates.

This section provides a comprehensive overview of the *Cox Proportional Hazard*(CPH) model, which is widely recognized as the most commonly used model in multivariate time-to-event analysis. In addition, this section highlights several extensions and specifications of the CPH model, such as the inclusion of time-dependent covariates and the choice of time-scale.

0.6 Cox Proportional Hazard Model

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

$$\underbrace{h(t \mid x)}_{\text{hazard}} = \overbrace{b_0(t)}^{\text{baseline hazard}} \underbrace{\exp\overbrace{\left(\sum_{i=1}^{n} b_i \left(x_i - \overline{x_i}\right)\right)}^{\text{log-partial hazard}}}_{\text{partial hazard}}$$

According to this specification, the log-hazard of an individual is a linear function of their covariates x 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 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.

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

The CPH model can be extended in such a way that it can incorporate covariates $x_i(t)$ that change over time. This characteristic is of great importance to clinical follow-up studies, as datasets usually include both baseline (time-independent) and time-dependent covariates. Mathematically, the ECM is represented by:

$$h(t \mid x(t)) = \overbrace{b_0(t)}^{\text{baseline}} \underbrace{\exp\overbrace{\left(\sum_{i=1}^{n} \beta_i \left(x_i(t) - \overline{x_i}\right)\right)}^{\text{log-partial hazard}}}_{\text{partial hazard}}$$

The package that is most commonly used for survival analysis in R is *Survival*. To use this package, the dataset must be organized in a *long* format, where each individual is represented by a unique *project_pseudo_id* variable and each time interval is identified by *start* and *stop* variables. The interval is considered to be open on the left and closed on the right, implying that any variable linked with an interval is presumed to be measured at the *stop* time. Furthermore, both baseline and follow-up covariates are included, and a time-dependent binary *event* variable which indicates whether or not an interval ends in the occurence of the event of interest. However, prediction of survival curves and time-to-event is not trivial when dealing with time-varying covariates, because this would require knowledge on the covariate values beyond the observed times.

Time-varying covariate values would be taken as the values at the start of each interval.

0.8 Censoring and choice of time-scale

Inherent to making inferences about duration is the presence of censoring and the choice of time-scale in the ECM. Censoring is especially common in clinical studies where event status updates and covariates are collected during periodic follow-up assessments. 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.

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?'. This question will inherently result in interval-censoring, because it does not provide the specific incidence time of disease.

There are several techniques on how to fit a ECM with interval-censored data.

Left-censored individuals are not taken into account, because it is impossible to evaluate the association of time-varying covariates with chronic disease incidence when the event has already occured.

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.

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.

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.

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.

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.

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

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.

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.

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).

Appendix

I Lifelines data overview

Table 4: 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 (cite)
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 5: 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