



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Extended Cox Models with Time-Dependent Variables for Mortality Risk Analysis in Multimorbid Danish Adults

TESI DI LAUREA MAGISTRALE IN
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **Ange Dakouri**

Student ID: 214784

Advisor: Prof. Francesca Ieva

Co-advisors: Anders Stockmarr, Nikolaj Normann Holm, Ove Andersen

Academic Year: 2023-24

Abstract

Multimorbidity, defined as the coexistence of two or more chronic diseases, presents a growing challenge for healthcare systems. In this research, the mortality risk associated with five multimorbidity clusters identified in the Danish adult population Allergies (ALL), Chronic Heart Conditions (CHC), Hypercholesterolemia (CHL), Diabetes (DIA), and Musculoskeletal and Psychiatric Conditions (M-P) was assessed through the application of survival analysis.

Two extended Cox proportional hazards models were estimated using different time scales (time-on-study and attained age), allowing for a broader understanding of the impact of cluster membership on mortality, as well as the influence of sex and educational level.

The results suggest that the M-P cluster includes individuals with the highest relative risk of death, although the CHC cluster shows the highest frequency of transitions to death. Furthermore, individuals with higher educational attainment exhibit lower hazard ratios, indicating that healthier lifestyles and regular health check-ups may improve disease management and influence mortality risk through earlier detection.

Aware of the limitations of the estimated models, particular attention was paid to the interpretation of results. Additionally, multistate models are proposed as a potential future direction to study disease progression in greater depth.

These findings emphasize the importance of preventive strategies and targeted healthcare policies to reduce multimorbidity-related risks.

Keywords: Multimorbidity, Clusters, Cox Model, Time scale and Education.

Abstract in lingua italiana

La multimorbilità, definita come la coesistenza di due o più malattie croniche, rappresenta una sfida crescente per i sistemi sanitari. In questo studio, il rischio di mortalità associato a cinque cluster di multimorbilità identificati nella popolazione adulta danese Allergie (ALL), Condizioni Cardiache Croniche (CHC), Ipercolesterolemia (CHL), Diabete (DIA) e Condizioni Muscoloscheletriche e Psichiatriche (M-P) è stato analizzato mediante tecniche di survival analysis.

Sono stati stimati due modelli estesi di Cox con differenti scale temporali (tempo nello studio e età raggiunta), per comprendere più a fondo l'impatto dell'appartenenza al cluster sulla mortalità, nonché l'influenza del sesso e del livello di istruzione.

I risultati suggeriscono che il cluster M-P comprende gli individui con il rischio relativo di morte più elevato, sebbene il cluster CHC mostri la frequenza più alta di transizioni verso il decesso. Inoltre, gli individui con un livello di istruzione più alto presentano hazard ratio inferiori, indicando che uno stile di vita più sano e controlli medici regolari possono migliorare la gestione delle malattie e influenzare il rischio di mortalità attraverso una diagnosi precoce.

Consapevoli dei limiti dei modelli stimati, è stata posta particolare attenzione all'interpretazione dei risultati. Inoltre, si propone l'utilizzo di modelli multistato come possibile estensione futura per studiare in modo più approfondito l'evoluzione delle condizioni croniche.

Questi risultati sottolineano l'importanza di strategie preventive e politiche sanitarie mirate per ridurre i rischi associati alla multimorbilità.

Parole chiave: Multimorbilità, Cluster, Modello di Cox, Scala temporale e Livello d'Istruzione

Disclaimer

Originally titled "Survival Analysis and Clusters of Chronic Diseases: A Statistical Approach to Mortality Risk in Multimorbid Populations," this work was submitted to the Technical University of Denmark in February 2025.

Contents

Abstract	i
Abstract in lingua italiana	iii
Disclaimer	v
Contents	vii
1 Introduction	1
1.1 Objectives	1
1.2 Study Background and Population Characteristics	2
2 Data Preprocessing	7
2.1 Dataset preparation	7
2.2 Cluster assignment	10
2.3 Other variables of interest	10
3 Methods	13
3.1 Nonparametric Tests	13
3.1.1 Permutational t-test	13
3.1.2 Mann-Whitney test	14
3.2 Survival Analysis	15
3.2.1 The Risk Hazard function	15
3.2.2 Kaplan-Meier Curves	16
3.2.3 Kaplan-Meier curves with Age as time scale	17
3.2.4 The Log-Rank test	19
3.2.5 The Proportional Hazard Cox Model	20
3.2.6 The Hazard Ratios and their interpretations	22
3.2.7 Concordance Index definition	23
3.3 Model Selection Criteria	24

3.3.1	Likelihood Ratio Test (LRT)	24
3.3.2	Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)	25
3.4	Model Diagnostic	26
3.4.1	Schoenfeld residuals	26
3.4.2	Proportional Hazards (PH) Test based on Schoenfeld residuals . . .	27
3.5	Regression Splines	27
4	Exploratory Data Analysis	29
4.1	Profile of Individuals at the Time of Multimorbidity Onset	29
4.1.1	Individuals distributions among the clusters	30
4.2	Dead Individuals	32
4.2.1	Death distribution among the clusters	33
4.3	Censored Individuals	35
4.3.1	Subjects distribution among the clusters	35
4.4	Kaplan-Meier curves with Time of follow up as time scale	37
4.4.1	Interaction between EntryAge and Sex	39
4.5	Kaplan-Meier curves with Age as time scale	41
4.5.1	EntryAge Greater Than or Equal to 20	41
4.5.2	EntryAge Greater Than or Equal to 30	42
4.5.3	EntryAge Greater Than or Equal to 50	43
4.5.4	EntryAge Greater Than or Equal to 80	44
4.6	Studying the dynamics within the population	45
4.6.1	The most frequent trajectories	45
4.6.2	Relative Frequencies of Transitions Across Multimorbidity Clusters	46
5	Model Construction	51
5.1	First model: Time on study as timescale	51
5.1.1	Model Diagnostics and Model Selection	52
5.1.2	How the number of Degrees of Freedom for the spline terms were chosen	57
5.1.3	Interepretation of the estimated HR	59
5.2	Second Model: Age as timescale	64
5.2.1	Model Diagnostic and Model Selection	65
5.2.2	Choose the Numebers of the Degrees of freedom for the splines terms	68
5.2.3	Interpretation of the estimated HR	70
5.3	Discussion and possible ranking	74
5.3.1	First impression	74

5.3.2	Focus on M-P cluster	74
5.3.3	Interpreting the Relationship Between ALL and DIA Clusters . . .	77
5.3.4	Impact of Time Scale Choice: Time on Study vs. Age	78
5.3.5	Limitations of the Cox Models	80
5.3.6	Proposal of a Mortality Risk Hierarchy Among Clusters	81
5.4	The dynamics within each Education level	81
5.5	High Mortality in CHC Does Not Equate to the Highest Risk	83
6	Conclusions and future developments	85
	 Bibliography	 87
	 A Appendix A	 93
A.1	Exploratory Data Analysis	93
A.1.1	Permutational t-test	93
A.1.2	Transition Matrix construction	94
A.2	Model Construction	94
A.2.1	Cross-Validation to select the best number of degree of freedom . .	94
	 List of Figures	 97
	 List of Tables	 99
	 Acknowledgements	 101

1 | Introduction

1.1. Objectives

In recent decades, the progressive ageing of the population, combined with advances in the diagnosis and treatment of chronic diseases, has led to a significant increase in the prevalence of multimorbidity, defined as the co-occurrence of two or more chronic conditions in the same individual. This phenomenon poses a growing challenge for modern healthcare systems, which are historically structured around the management of single diseases [1] and [2].

Multimorbidity is not only associated with higher healthcare costs and increased complexity of care, but also with worse prognoses in terms of quality of life and mortality risk. Individuals affected by multiple chronic diseases are more likely to experience faster functional decline, more frequent hospitalisations, therapeutic complexity, and adverse health outcomes [3].

Despite growing attention, there remains a significant unmet methodological need: most existing studies describe disease combinations in a static or cross-sectional way, without taking into account the temporal evolution of chronic conditions and their dynamic interactions. Furthermore, only a limited number of studies adopt advanced survival models incorporating time-dependent covariates, which would allow for a more realistic and informative understanding of how multimorbidity affects mortality over time [2] and [3].

Another gap in the literature concerns the limited stratification of risk by socio-demographic variables such as educational attainment, which has been shown to influence disease awareness, treatment adherence, and healthcare access. These social determinants may impact not only the onset of multimorbidity but also survival outcomes [1].

The Danish healthcare system offers an ideal setting to address these challenges thanks to its nationwide population registers, which allow for comprehensive longitudinal tracking of health conditions, healthcare use, and socio-demographic characteristics at the individual level [4]. This study draws on data from over 2.6 million Danish citizens aged 18 or

older who became multimorbid between 1995 and 2018, enabling a rich and detailed investigation of disease trajectories and mortality.

The aim of this thesis is to analyse the mortality risk associated with five clinically meaningful multimorbidity clusters, by applying extended Cox proportional hazards models with time-dependent covariates, using both time-on-study and age as time scales. Specifically, the study aims to:

- Quantify relative hazards associated with each cluster.
- Evaluate the modifying effect of sex and educational level.
- Compare the performance of the models depending on the selected time scale.

The results are intended to provide evidence-based insights for public health interventions, supporting the identification of high-risk subpopulations and promoting more dynamic, stratified strategies for the management of multimorbidity in Denmark.

Furthermore, this work draws inspiration from [5]: the author have studied the effect of Heart Disease patients' co-occurring diseases on mortality, modelling their dynamically expanding disease portfolios while identifying interactions among the co-occurring diseases and socioeconomic and biological variables.

1.2. Study Background and Population Characteristics

This study builds upon [6] identified five multimorbidity clusters based on typical disease portfolios on the targeted studied population: Allergies (ALL), Chronic Heart Conditions (CHC), Hypercholesterolemia (CHL), Diabetes (DIA), and Musculoskeletal and Psychiatric Conditions (M-P).

The **K-means algorithm** [7], testing solutions from **1 to 10 clusters**. To ensure robustness against random initializations [8], **200 iterations** of K-means were performed for each cluster count, selecting the optimal configuration based on the **minimum within-cluster sum of squares (WCSS)**. If a stable minimum was not identified, additional runs were conducted.

The **Elbow method**, along with the **Calinski-Harabasz index** [9], the **Silhouette score** [10], and a metric referred to as **Rim data frequency**, was used to determine the optimal number of clusters.

Rim data are defined as individuals whose cluster assignment in the $k + 1$ solution is

neither matched to their cluster in the k solution nor associated with the newly introduced cluster. These observations suggest instability in the clustering structure, as they indicate inconsistent allocation across successive clustering solutions.

To illustrate the progression as the number of clusters increases, and to justify the optimal selection, the authors paired the clusters obtained with $k+1$ clusters to those obtained with k clusters, for $k = 1$ to 9. Specifically, they selected the configuration of k clusters among the $k+1$ clusters that minimized the sum of the Euclidean distances between the centroids of the new clusters and those of the previous k -cluster solution. In this framework, k clusters are matched to the prior configuration, while the remaining unmatched cluster is defined as the *new cluster*.

Individuals falling into clusters that were not matched to their previous assignment and were not part of the *new cluster* were counted as *rim data* for the k -cluster solution. Rim data are considered undesirable, as they indicate erratic cluster allocation. In contrast to traditional internal validation metrics such as the Within-Cluster Sum of Squares (WCSS), the Caliński-Harabasz index, and the silhouette score, rim data frequency incorporates information from two consecutive clustering solutions rather than a single one.

Several internal validation metrics were employed by the original authors to assess the appropriate number of clusters, within a clinically meaningful upper bound of ten. Both the elbow method applied to the Within-Cluster Sum of Squares (WCSS) and the Caliński-Harabasz index suggested that a reasonable choice lies around five clusters. However, the lack of a clear elbow in the WCSS curve indicated that the mathematically optimal number of clusters might in fact be higher than what is clinically feasible. This interpretation was supported by the silhouette score, which showed a steady increase with the number of clusters up to the upper limit considered.

To further inform the decision, the authors examined the progression of clustering solutions across increasing values of k , with specific attention to the presence of *rim data* which presence was interpreted as an indicator of instability in the clustering structure.

The authors observed that although cluster configurations for lower values of k appeared relatively stable typically involving a simple subdivision of one cluster into two as k increased some minor reallocations still occurred, particularly among multimorbid individuals. These reallocations were quantified through the frequency of rim data, and the analysis showed that the lowest such frequency (excluding configurations deemed irrelevant due to excessive within cluster variation) was found when five clusters were used. This finding was consistent across both the multimorbid population and the broader group of individuals with at least one chronic condition.

On the basis of these combined criteria the elbow method, Caliński-Harabasz index, silhouette score, and rim data frequency the five cluster solution was deemed the most appropriate.

A disease portfolio is a medical profile documenting the presence of diagnosed chronic conditions in an individual. In the context of our study, since patients exhibit multimorbidity, each individual has at least two chronic diseases. Over time, additional diagnoses may be recorded, altering the composition of the disease portfolio. Each disease portfolio is assigned to the cluster that best represents it, based on the type and combination of conditions present.

Information about chronic conditions and socioeconomic characteristics (age, gender and maximum educational attainment), were extracted from national registers:

- The Danish National Patient Registry [11];
- The Danish Psychiatric Central Research Register [12];
- The Danish National Health Service Registry [13];
- The Danish Population Education Register [14].

The National registers do not comprise direct information about the type of chronic conditions diagnosed in the primary sector (e.g. General practitioners, Municipal healthcare services and Private specialists). To have the correct information on chronic conditions for the total population, the authors used diagnostic algorithms developed by the Research Center for Prevention and Health at Glostrup University Hospital for the 16 selected chronic conditions, using information from registers including data from both primary and secondary (e.g., hospitals and specialists) healthcare sectors [15] and [16]. Further details about diagnostic algorithm in [6].

The 16 chronic conditions considered in a subject's portfolio include Allergies, Anxiety, Back Pain, Cancer, Heart Disease, Chronic Obstructive Pulmonary Disease (COPD), Dementia, Depression, Diabetes, Hypercholesterolemia, Hypertension, Osteoarthritis, Osteoporosis, Schizophrenia, Joint Disease, and Stroke. It is important to note that cluster labels reflect predominant or absolute trends rather than strictly similar conditions. For instance, an individual classified in the Musculoskeletal and Psychiatric Conditions (M-P) cluster may have neither a musculoskeletal nor a psychiatric condition but instead conditions such as Cancer, COPD, or a combination of Stroke and Hypertension. All the details about the clusters follows (note that in this research the Anxiety disease is not considered):

Presence (%) of the chronic conditions in each clusters:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Allergies	100.00	14.04	13.71	12.64	0.0
Anxiety	0.38	0.11	0.11	0.13	0.30
Back Pain	13.08	8.88	7.67	6.89	20.53
Cancer	7.49	9.33	8.11	6.85	18.49
Heart Disease	5.01	100.00	0.00	0.00	1.62
COPD	26.05	19.06	12.00	10.41	23.27
Dementia	1.29	3.37	2.24	1.84	4.52
Depression	20.30	12.40	13.81	12.40	27.97
Diabetes	4.20	24.40	0.00	100.00	1.15
Hypercholesterolemia	9.25	78.89	100.00	81.98	0.08
Hypertension	44.31	76.29	86.00	83.99	71.67
Osteoarthritis	11.82	11.80	11.47	9.61	24.21
Osteoporosis	11.45	12.89	12.77	6.52	28.17
Joint Disease	2.02	3.23	1.67	1.92	4.75
Schizophrenia	4.53	2.21	2.16	3.45	5.73
Stroke	2.93	11.83	12.75	6.08	6.32

Table 1.1: Chronic disease presence (in %) in each identified cluster

Sociodemographic:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Age (years) males	58.20	69.77	67.13	64.26	64.60
Age (years) females	58.45	73.97	69.76	66.15	66.91
Age (years) all	58.36	71.50	68.57	65.11	66.09
Male frequency (%)	35.59	58.71	45.43	55.25	35.52
Female frequency (%)	64.41	41.29	54.57	44.75	64.48

Table 1.2: Sociodemographic info in each identified cluster

Education attainment (%):

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
No Education (≤ 10 years)	19.70	36.07	31.07	32.03	30.01
Short Education (10–14 years)	50.40	44.06	48.78	49.26	48.00
Medium Education (15–16 years)	5.89	9.61	10.70	9.90	11.34
Long Education (≥ 17 years)	11.73	6.55	7.62	5.89	7.60

Table 1.3: Education info in each identified cluster

2 | Data Preprocessing

2.1. Dataset preparation

The data used in this study originates from a cross sectional design study of all individuals aged 18 years and older who lived in Denmark and become multimorbid between January 1st in year 1995 and December 31st 2018, counting 2.694.143 individuals. All the information about chronic conditions and socioeconomic characteristics (age, gender and maximum educational attainment), were extracted from national registers [11], [12], [13] and [14].

To perform the analysis, the first step is to create a long-format dataset, where the starting time corresponds to the date when an individual becomes multimorbid. It is important to explain that in this format, each row represents a new disease portfolio, where one or more new diseases are diagnosed, potentially leading to a change in the subject's cluster assignment. Therefore, it is possible to have multiple rows corresponding to the same individual.

Initially the raw dataset, available through a collaboration between the Technical University of Denmark and the Department of Clinical Research at Hvidovre Hospital, contains the following variables:

- **SubjectID**: Unique identifier for each individual;
- **KOEN (Sex)** : Equal to 1 for males and 2 for females;
- **FOEG DAG**: Birth date;
- **DODDATO**: Death date (NA if the subject is still alive);
- **StartObsDate**: Start of the observation period (either the date the subject turned 18 or immigrated to Denmark);
- **EndObsDate**: End of the observation period (due to death, emigration, or the end of follow-up);

- **15 Disease Variables:** Dates of diagnosis for specific diseases (NA if the disease is absent).

While the final dataset contains the followings:

- **SubjectID:** Unique identifier for each individual;
- **Sex:** Equal to 1 for males and 2 for females;
- **Age:** Individual's age at the beginning of the time span;
- **Calendar Time:** Date marking the start of the time span;
- **Time Init:** Time (in years) at the beginning of the time span (set to 0 when the subject becomes multimorbid);
- **Time End:** Time (in years) at the end of the time span;
- **Event:** Equal to 0 if the subject was censored or did not experience the event, and 1 if the patient died;
- **15 Indicator Variables:** Binary variables indicating the presence of specific diseases during the time span.

About the censored individual

In survival analysis, a censored individual has the characteristic to have the unknown exact time of the event of interest (e.g., death, disease progression). This can occur in the following cases:

- When the event has not been observed by the end of the study period.
- The individual is lost to follow-up.
- The event is only partially recorded within a given time frame.

Censoring is a key consideration in survival analysis, as it ensures that incomplete observations are properly accounted for in statistical models.

For each individual i , let T_i^* be the non-negative random variable denoting the failure time and C_i be a random variable that denotes the time at which a censoring mechanism kicks in. What we actually observe in time-to-event studies is the *failure time* that is either the event time T_i^* or, whichever is smaller, the censoring time C_i :

$$T_i = \min(T_i^*, C_i) \tag{2.1}$$

In addition, we usually get information on whether T_i is an actual event time or a censored observation, defining an indicator random variable δ_i for non-censoring:

$$\delta_i = \begin{cases} 1 & \text{if } T_i^* \leq C_i \\ 0 & \text{if } T_i^* > C_i \end{cases} \quad (2.2)$$

Hence the observed data consist of pairs (T_i, δ_i) for each individual i .

There are three main types of censoring:

- **Right Censoring:** Occurs when an individual does not experience the event before the study ends or is lost to follow-up. The only known information is that the event, if it occurs, happens after the last recorded observation time.
- **Left Censoring:** Occurs when the exact time of the event is unknown, but it is known to have occurred before the subject entered the study. This can happen, for example, when a disease is already present at the time of first observation, but the precise onset date is unknown (e.g. HIV check-up);
- **Interval Censoring:** Occurs when the exact time of the event is unknown, but it is known to have occurred within a specific time interval. This often happens in medical studies where patients are examined at discrete time points, and the event is detected at some intermediate visit without knowing the precise timing (e.g. first HIV check-up is negative while the second is positive).

In our study, since death can occur after the end of the observation period, we are dealing with a case of right censoring.

2.2. Cluster assignment

To complete the dataset and proceed with the analysis, each disease portfolio was assigned to the cluster that best described it. The cluster centroids were based on those presented in Table 1.1, which originated from the reference study [6]. Since the portfolios are represented as binary vectors indicating the presence (1) or absence (0) of each condition, it was necessary to map the centroids into the same subspace of \mathbb{R}^{15} , where each component ranges from 0 to 1. This ensures that both the centroids and the portfolios lie in the same space, allowing for consistent and robust distance computations.

Let $\mathbf{X} \in \mathbb{R}^{15}$ be the vector representing a subject's disease portfolio, and let $\mathbf{C}_i \in \mathbb{R}^{15}$ (for $i = 1, \dots, 5$) denote the centroids. The Euclidean distance (L2 norm) between \mathbf{X} and centroid \mathbf{C}_i is computed as:

$$\|\mathbf{X} - \mathbf{C}_i\|_2 = \sqrt{\sum_{j=1}^{15} (x_j - c_{ij})^2}$$

where x_j is the j -th component of the subject's portfolio, and c_{ij} is the corresponding value in centroid i . The subject is assigned to the cluster whose centroid yields the smallest L2 distance to their portfolio vector.

Once this assignment is completed, the resulting cluster label is added as a new variable to the dataset.

2.3. Other variables of interest

During the analysis other variables were included or constructed, and then involved during the modeling stage:

- **EntryDate**, the date when the individual becomes multimorbid (in day format). This variable is included in the Age time scale Cox model to account for potential differences in healthcare conditions over time. The probability of entering the study in 2000 may differ from that in 2010 due to advancements in medical treatments, changes in healthcare policies, or improvements in disease detection and management. By incorporating EntryDate, the model captures the possible impact of the period in which an individual enters the study.
- **EntryAge**, the age in which the subject becomes multimorbid. Instead, this variable is included in the time on study as time scale Cox model to

consider the Age in which the individual enters in the study. This can have an important impact on the mortality.

- **Age End:** Individual's age at the end of the time span.

This variable is fundamental to fit the Age time scale Cox model.

- **Education:** Maximum education level achieved: Long, Medium, Short, None and Missing (if there is no information).

This variable is important as it can explain differences in mortality risk not only from a biological perspective but also from a socioeconomic standpoint. Note: the results related to Missing education are not considered since they don't provide additional information.

3 | Methods

Before delving into the exploratory data analysis, the methodological pipeline and results, is important to introduce several key concepts, including the structure of the Cox model, the role of covariates, and the interpretation of hazard ratios. These elements are crucial for understanding the insights gained from the research.

3.1. Nonparametric Tests

3.1.1. Permutational t-test

Let X and Y be two random variables representing two independent populations, we would like to test if there is difference in term of distrubution. In other words verify the following **hypothesis**:

$$H_0 : X \stackrel{d}{=} Y \quad (\text{i.e., the two distributions are the same})$$

$$H_1 : X \stackrel{d}{\neq} Y \quad (\text{i.e., the two distributions are different}).$$

Since we don't assume Gassuian distribution of the data we consider the permutational t-test is a variant of the parametric version [17]. The class of the permutational test belongs to the family of **Likelihood Invariant Transformation** of the dataset: under the hypothesis H_0 , the observation of the both groups are exchangeable(reflection, regrouping,...), meaning every possible reassignment of values is equally likely.

The **test statistic** T , in this case $T = |\text{Mean}(X) - \text{Mean}(Y)|$ to have more robust approach, is computed on the observed data. The **critical region** for the test is defined as the set of the extreme values of T , correponding to the sigificance **level** α .

Since computing all permutaion can be computationally intensive, a **Monte Carlo** approach is othen used:

- Compute the test statistic T_0 on the observed data;

- Generate B random permutations of the dataset:
 - Shuffle the data to create a permuted dataset;
 - Compute T_b for each permutation;
 - Store the values of T_b ;
- Estimate the p -value as the proportion of the test statistics T_b that exceed T_0 .

This approach provides a valid approximation of the exact permutation test and is widely used when exact computation is infeasible. The R code used to perform the test is provided in Appendix A.

3.1.2. Mann-Whitney test

To assess whether the population represented by X is in distributionally lower than another represented by Y the Mann-Whitney U test is an appropriate nonparametric method [18]. This test compares the distributions of two independent groups without assuming normality of the data.

The **null hypothesis** H_0 and **alternative hypothesis** H_1 are defined as:

$$H_0 : X \stackrel{d}{=} Y$$

$$H_1 : \mathbb{P}(X < Y) > 0.5$$

where H_0 asserts that the distributions of the two groups are the same (i.e., the probability of X being smaller than Y is 0.5), and H_1 asserts that the distribution of X is shifted to the left of Y .

Test Statistic:

The Mann-Whitney U statistic is calculated as follows:

$$U_1 = n_X n_Y + \frac{n_X(n_X + 1)}{2} - R_X$$

$$U_2 = n_X n_Y + \frac{n_Y(n_Y + 1)}{2} - R_Y$$

where:

- n_X and n_Y are the sample sizes for the two groups respectively;
- R_X and R_Y are the sums of the ranks for the two groups.

The final U statistic is:

$$U = \min(U_1, U_2)$$

For large sample sizes, the distribution of U can be approximated by a normal distribution with mean $\mu_U = \frac{n_X n_Y}{2}$ and variance $\sigma_U^2 = \frac{n_X n_Y (n_X + n_Y + 1)}{12}$.

Critical Region:

- The critical region is defined as the set of values of U such that the probability of obtaining a value of U under null hypothesis is less or equal than α ;
- If U falls in the critical region, the null hypothesis is rejected in favour of the alternative hypothesis.

3.2. Survival Analysis

3.2.1. The Risk Hazard function

Given a continuous random variable T , which represents the survival time, and the instantaneous time t , the hazard risk function $h(t)$ describes the instantaneous rate at which an event occurs at a specific time t , conditional on the individual having survived up to that time. The mathematical formulation follows here [19]:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Here the main features:

- is not a probability so can be larger than 1;
- it is always nonnegative, that is, equal to or greater than zero;
- it has no upper bound.

Relationships with the Survival Function

Let $f(t)$ denote the probability density function of T , $F(t)$ the distribution function such that $F(t) = \Pr(T \leq t)$, and $S(t)$ the survival function, where $S(t) = 1 - F(t)$.

Starting from the definition hazard function $h(t)$ above and apply the Bayesian rule we have here [19]:

$$\lim_{\Delta t \rightarrow 0} \frac{P(T \in [t, t + \Delta t) \cap T \geq t)}{P(T \geq t) \cdot \Delta t},$$

then play with the intersection of sets:

$$= \lim_{\Delta t \rightarrow 0} \frac{P(T \in [t, t + \Delta t))}{\Delta t} \cdot \frac{1}{P(T \geq t)}$$

$$= \lim_{\Delta t \rightarrow 0} \frac{\int_t^{t+\Delta t} f(u) du}{\Delta t} \cdot \frac{1}{P(T > t)},$$

then by the Fundamental Theorem of Calculus:

$$= \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t)$$

The last equality follows from the fact that $-f(t)$ is the derivative of $S(t)$.

In summary $h(t) = -\frac{d}{dt} \ln S(t)$. While $S(t)$ in function of $h(t)$ can be expressed in the following formula:

$$S(t) = \exp\left(-\int_0^t h(u) du\right), \quad t \geq 0$$

with the boundary condition $S(0) = 1$. The term inside the exponential is the cumulative risk hazard function $H(t)$.

3.2.2. Kaplan-Meier Curves

In the context of multimorbidity, understanding the survival probabilities of subjects over time provides critical insights into the progression and outcomes of multiple coexisting conditions. The Kaplan-Meier estimator is a valuable tool for this purpose, allowing us to estimate survival functions even when individuals have different follow-up durations or when some outcomes are unknown due to censoring. Essentially, this method divides time into intervals and calculates survival probabilities at each step, accounting for both events and censored data. By applying this non-parametric approach, it becomes possible to evaluate factors that influence survival outcomes and identify disparities between subjects groups.

Survival Function and Kaplan-Meier estimator

Let T be a random variable representing survival time, with probability density function $f(t)$ and cumulative distribution function $F(t) = \mathbb{P}(T \leq t)$. The **survival function** is

defined as the complement of the cumulative distribution function:

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t).$$

Formally the **Kaplan-Meier estimator** can be defined as follows:

$$\hat{S}(t) = \prod_{i:t_i^* \leq t} p_i = \prod_{i:t_i^* \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where:

- i = failure (event) index, $i \in \{1, \dots, I\}$;
- I = total number of individuals with events;
- $0 < t_1^* < \dots < t_I^* < \infty$ = observed ordered times of deaths;
- p_i = conditional probability of surviving time t_i^* ;
- n_i = number of subjects alive just before t_i^* , i.e., number of subjects at risk at time t_i^* ;
- d_i = number of observed events at t_i^* .

So the definition of $\hat{S}(t)$ as a product limit holds.

3.2.3. Kaplan-Meier curves with Age as time scale

In survival analysis, the choice of time scale can substantially influence the interpretation of results. In this study, both *time-on-study* (i.e., time since entry into the cohort) and *attained age* are used as time scales to estimate survival functions. While time-on-study is the conventional choice in many clinical studies, age is a biologically meaningful time scale, particularly relevant when studying mortality. Using age as the time scale allows for direct comparison of individuals at the same age and better captures the age-dependent baseline hazard. Therefore, survival curves are computed under both time scales to investigate how the choice of time origin affects survival estimates and to explore the main characteristics of the study population from different temporal perspectives.

Formally, let A a random variable that denote the age at death, and define the survival function as

$$S(a) = \mathbb{P}(A > a),$$

where a is the attained age. This function expresses the probability that an individual

survives beyond age a , allowing survival to be studied as a function of biological age rather than follow-up time [19].

Correspondingly, the hazard function with age as the time scale is defined as

$$h(a) = \lim_{\Delta a \rightarrow 0} \frac{\mathbb{P}(a \leq A < a + \Delta a \mid A \geq a)}{\Delta a},$$

which represents the instantaneous risk of death at age a , given survival up to that age.

Under the Cox proportional hazards framework, the hazard function for individual i with covariates \mathbf{X}_i is modeled as

$$h_i(a) = h_0(a) \exp(\boldsymbol{\beta}^\top \mathbf{X}_i),$$

where $h_0(a)$ is the unspecified baseline hazard as a function of age, and $\boldsymbol{\beta}$ are the regression coefficients. In this formulation, the model estimates how covariates affect the hazard of death, assuming proportionality of hazards with respect to age.

The concept of Left truncation

Left truncation occurs when a subject is not observed from the start of the risk period (time 0), but instead enters the study at a later time point t_0 . If the individual experiences the event of interest before t_0 , they are excluded from the study. However, if the event occurs after t_0 , they are included in the analysis under the condition that they were at risk but unobserved until t_0 .

In the context of this study, where age is used as the time scale, t_0 corresponds to the subject's age at entry into the cohort. Since individuals enter the study at different ages, left truncation must be accounted for to avoid biased estimation. Specifically, the risk set at each age includes only those individuals who were under observation and still at risk at that age. Proper handling of left truncation ensures that survival and hazard estimates are not distorted by the staggered entry of individuals into the cohort.

There are two types of left truncation at t_0 :

- **Type 1:** The individual experiences the event before t_0 , and therefore is excluded from the study. This can lead to a selective survival bias if, for example, the exposure under investigation causes individuals to die before entering the study. Such bias may result in an underestimation of the exposure effect.
- **Type 2:** The subject survives beyond t_0 ($t > t_0$), allowing their survival time to

be observed. In this case, a prerequisite for the subject's inclusion in the study is survival until t_0 .

It is important to differentiate left truncation from left censoring, as they are often confused. Left censoring occurs when a subject is included in the study and is known to:

- be event-free at time 0;
- be at risk for the event after time 0;
- have experienced the event before a specific time t , though the exact timing of the event remains unknown.

The type of risk sets

When time-on-study is used as the time scale, the risk sets are organized based on follow-up times. This results in a **closed cohort**, where the size of the risk set, initially containing all the individuals, always decreases over time. At each failure time, the risk set includes only those subjects who have neither experienced the event nor been censored prior to that time. The key feature of a closed cohort is that the risk set at any later time is always a subset of the risk set at earlier times.

By contrast, in an **open cohort**, individuals can enter the study at different calendar times and at different ages. When age is used as the time scale, this means that individuals become at risk only from their age at entry onward. As described by [19], this situation induces left truncation, since individuals who experienced the event of interest before entering the study (i.e., before their entry age) are not observed and therefore excluded from the analysis.

To handle left truncation appropriately, survival analysis methods must ensure that an individual is included in the risk set only from the age at which they actually entered the study. This adjustment is essential to avoid biased estimation of survival and hazard functions. For example, if an individual enters the cohort at age 65 and dies at age 70, their contribution to the risk set begins at age 65, not at time zero. By incorporating left truncation in this way, we ensure that survival estimates accurately reflect the risk over attained age, conditional on being under observation.

3.2.4. The Log-Rank test

During the analysis is important to spot differences between the survival curves in order to find important insights. In this situation the Log-Rank test is the right tool. The

rigorous definition of the test follows: given k survival curves, the *Log-Rank test* is used to assess whether there are significant differences in the survival distributions of these groups over time [19].

Statistics:

The *exact log-rank statistic* is distributed as a χ^2 with $k - 1$ degrees of freedom:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{V_i},$$

where:

- O_i is the observed number of events in group i ;
- E_i is the expected number of events in group i .
- V_i is the variance of the difference between the observed and the expected number of events in group i .

While the *approximate log-rank statistic* is:

$$\sum_i \frac{(O_i - E_i)^2}{E_i},$$

Given the large number of variables and multiple groups under consideration, the approximate log-rank statistic provides a sufficiently reliable result while being computationally more efficient.

3.2.5. The Proportional Hazard Cox Model

The **Cox Proportional Hazards (PH) model** is a widely used regression model in survival analysis. It assumes that the hazard function at any given time t for an individual is proportional to a baseline hazard, with the effect of covariates on the hazard function being constant over time [19]. The model can be expressed as:

$$h(t|X_i) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{X}_i),$$

where:

- $h(t|X_i)$ is the hazard function at time t for the i -th individual with covariate vector

X_i ;

- $h_0(t)$ is the baseline hazard function, and assumes nonnegative values;
- X_i is the covariate vector for the i -th individual;
- β is the vector of regression coefficients, which describe the relationship between the covariates and the hazard function.

Remarks:

- The use of the exponential in the formula assures to not have negative value since the scalar product $\beta^\top \mathbf{X}_i$ can be negative;
- The Cox model is a **semiparametric regression** since the baseline risk $h_0(t)$ is an unspecified function;
- The estimation of the regression coefficients is performed by maximizing the **Partial Likelihood**, which considers explicitly only the probabilities of subjects who experience the event and is based on the relative ordering of observed failure times within the risk set (which consider both death and censored subjects). Mathematically is defined as follows [19]:

$$L(\beta) = \prod_{i=1}^I \frac{\exp(\beta^\top \mathbf{X}_i)}{\sum_{j \in R(t_i)} \exp(\beta^\top \mathbf{X}_j)},$$

where:

- I is the number of observed events (e.g., deaths);
- \mathbf{X}_i is the vector of covariates for the individual experiencing the event at time t_i ;
- β is the vector of regression coefficients to be estimated;
- $R(t_i)$ is the risk set at time t_i , consisting of individuals still at risk just before t_i .

The core assumption of the Cox model is the **proportional hazards** assumption, which means that the ratio of hazards between any two individuals is constant over time and is solely determined by the covariates.

Extension with Time-Dependent Covariates

In the extended version of the Cox model, time-dependent covariates are included to allow the effect of covariates to vary over time. This extension is useful when the influence of a covariate on the hazard may change as time progresses [19]. The modified model can be written as:

$$h(t|X_i(t)) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{X}_i(t))$$

Where:

- $X_i(t)$ is now a vector of time-dependent covariates for the i -th individual, meaning that the values of the covariates can change as the individual progresses through time.

In this case, the proportional hazard assumption does not hold because the hazard ratio depends on time.

3.2.6. The Hazard Ratios and their interpretations

The **Hazard Ratio (HR)** is a key metric in survival analysis, particularly within the Cox Proportional Hazards model. It quantifies the relative risk of an event occurring between two groups with different covariate values, while holding other factors constant. In this study, the HR is essential as it allows us to rank the different clusters in terms of their relative mortality risk. Furthermore, it provides a clearer understanding of the differences between sexes and varying levels of education, thereby enhancing our comprehension of the factors influencing mortality risk.

The Hazard Ratio (HR) for a particular covariate X_j is mathematically expressed as:

$$HR_i = \frac{h(t|X_1^*, X_2^*, \dots, X_i^*, \dots, X_p^*)}{h(t|X_1, X_2, \dots, X_i, \dots, X_p)} = \exp(\beta_i),$$

where:

- β_i is the coefficient corresponding to the covariate X_j in the Cox model;
- $X = (X_1, X_2, \dots, X_p)$ represents the vector of covariate values for the **reference group** (e.g., individuals without the condition, or individuals with baseline values for each covariate);
- $X^* = (X_1^*, X_2^*, \dots, X_p^*)$ represents the vector of covariate values for the **group of**

interest (e.g., individuals with the condition, or individuals whose covariate values are altered).

In this formulation the difference stays in the values of X_i and X_i^* while the other factors are kept **constant**.

Moreover, the HR can be computed simultaneously for multiple covariates. When this is the case, the HR represents the combined effect of all included covariates on the hazard, while still controlling for the influence of other factors. This allows for a more nuanced understanding of how multiple factors (e.g., sex, education, and comorbidities) interact and affect mortality risk. The ability to assess the effect of multiple covariates is particularly valuable in complex datasets, where several variables may influence the outcome.

3.2.7. Concordance Index definition

The **concordance index (C-index)** is a measure of how well a model's predicted risk scores agree with the observed outcomes in survival analysis. It quantifies the discriminatory ability of the model by comparing pairs of individuals and assessing whether the model correctly ranks them in terms of their predicted survival times.

Formally, the C-index is defined as the proportion of all comparable pairs of individuals whose predicted survival times are in concordance with their actual survival times. A pair is considered comparable if the survival times of both individuals are distinct. If the model predicts a higher risk (shorter survival time) for the individual who dies first in a comparable pair, the pair is said to be concordant; otherwise, it is discordant. The C-index ranges from 0.5 (no discrimination, similar to random guessing) to 1.0 (perfect discrimination).

Mathematically, the C-index can be expressed as:

$$C = \frac{1}{N_{\text{comparable}}} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \mathbb{I}(\text{sign}(R_i - R_j) = \text{sign}(S_i - S_j))$$

Where:

- $N_{\text{comparable}}$ is the total number of *comparable* pairs, defined as the number of pairs (i, j) such that:
 - The survival times S_i and S_j are distinct (i.e., $S_i \neq S_j$);
 - Neither i nor j is censored or at least one of them is not censored, ensuring

that the pair can be meaningfully compared.

- R_i and R_j are the predicted risk scores for individuals i and j
- S_i and S_j are the observed survival times for individuals i and j ;
- $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition inside is true, and 0 otherwise.

A higher C-index indicates better model performance in distinguishing between individuals at different risks of the event (e.g., death).

3.3. Model Selection Criteria

Model selection is a critical step in statistical modeling aimed at identifying a model that balances complexity and explanatory power. In this study, criteria such as the Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) are employed to evaluate whether the inclusion of additional covariates significantly improves model performance. This balance is particularly important because the final model is intended for practical use in the medical field, where clear communication of results to clinicians and other healthcare professionals is essential. While more complex models may achieve better fit, maintaining interpretability ensures that the model's predictions are trustworthy and actionable in real-world decision-making.

3.3.1. Likelihood Ratio Test (LRT)

The LRT compares the goodness-of-fit between a reduced model and a more complex, nested full model by evaluating their respective log-likelihoods.

Hypotheses:

H_0 : The reduced model with parameter vector $\boldsymbol{\beta}_{\text{reduced}}$ is sufficient;

H_1 : The full model with parameter vector $\boldsymbol{\beta}_{\text{full}}$ provides a better fit.

Test Statistic:

The LRT is based on the log-likelihood function $\ell(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ represents the vector of model coefficients. The test statistic is given by:

$$D = 2 \left(\ell(\hat{\boldsymbol{\beta}}_{\text{full}}) - \ell(\hat{\boldsymbol{\beta}}_{\text{reduced}}) \right),$$

where:

- $\ell(\hat{\boldsymbol{\beta}}_{\text{reduced}})$ is the maximized log-likelihood of the reduced model;
- $\ell(\hat{\boldsymbol{\beta}}_{\text{full}})$ is the maximized log-likelihood of the full model.

Under the null hypothesis, D follows an asymptotic chi-squared distribution:

$$D \sim \chi_k^2$$

where k is the difference in the number of parameters between the full and reduced models (degrees of freedom).

3.3.2. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

When comparing non-nested models, model performance is often evaluated using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), defined as:

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}) + 2p$$

$$BIC = -2\ell(\hat{\boldsymbol{\beta}}) + p \log(n)$$

where:

- $\ell(\hat{\boldsymbol{\beta}})$ is the log-likelihood of the estimated model.
- p is the number of estimated parameters in the model.
- n is the number of observations.

Both criteria penalize model complexity, but BIC applies a stronger penalty for additional parameters, making it more conservative in model selection. A lower AIC or BIC value indicates a better-fitting model.

3.4. Model Diagnostic

While model selection criteria such as the Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) help identify the best-fitting model, it is also essential to assess whether the underlying assumptions of the Cox Proportional Hazards (PH) model hold. This is done through model diagnostics, which include tests for the proportional hazards assumption and residual analysis

3.4.1. Schoenfeld residuals

Schoenfeld residuals represent the difference between the observed covariate and the expected given the risk set at that time. They should be flat, centered about zero. In principle, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

Since the model includes time-dependent covariates, the choice of residuals is crucial for proper diagnostics. Given the structure of the model, Schoenfeld residuals are the most appropriate, as they allow for assessing the proportional hazards assumption even in the presence of time-varying effects. These residuals can highlight covariates that change effect over time, guiding potential stratification or interaction terms to improve model fit.

For an individual i who fails at time t_i , the Schoenfeld residual for the j -th covariate is defined as the difference between the observed covariate value and the expected value of that covariate over the risk set at t_i . Formally, it is given by:

$$r_{ij} = x_{ij} - \frac{\sum_{\ell \in R(t_i)} x_{\ell j} \exp(\boldsymbol{\beta}^\top \mathbf{X}_\ell)}{\sum_{\ell \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{X}_\ell)},$$

- x_{ij} is the value of the j -th covariate for the subject failing at time t_i .
- $R(t_i)$ denotes the risk set at t_i .
- $\boldsymbol{\beta}$ is the vector of estimated regression coefficients.

This formulation and its diagnostic use are described in detail in [20]. The definition is consistent with [21].

3.4.2. Proportional Hazards (PH) Test based on Schoenfeld residuals

The proportional hazards test based on Schoenfeld residuals evaluates the correlation between these residuals and time to assess the proportional hazards (PH) assumption. Specifically, the test examines whether the Schoenfeld residuals exhibit a systematic trend over time.

The hypotheses are as follows:

- H_0 : The Schoenfeld residuals are independent of time, indicating that the proportional hazards assumption holds.
- H_1 : The Schoenfeld residuals display a time-dependent pattern, indicating a violation of the proportional hazards assumption.

A significant test result, commonly defined as a p-value less than 0.05, suggests that the PH assumption is violated for at least one covariate in the model [22].

3.5. Regression Splines

To capture potential nonlinear relationships between continuous covariates and the log hazard function, spline functions were incorporated into the Cox proportional hazards model. Traditional Cox models assume a linear effect of each covariate on the log hazard, which may fail to reflect more complex associations present in real data. By modeling covariates using spline transformations—specifically, piecewise polynomial functions joined smoothly at predefined knots—the model gains the flexibility to represent a wider range of functional forms. This approach allows the hazard to vary smoothly with the covariate, without requiring a priori specification of a particular nonlinear shape [23].

A degree- d spline is defined as a piecewise polynomial of degree d , constrained to have continuous derivatives up to order $d - 1$ at each knot—the values that divide the domain into intervals. This smoothness ensures gradual transitions between segments while enabling the model to capture intricate covariate effects. In model diagnostics, spline-based representations are particularly valuable for revealing nonlinear trends and identifying deviations from assumed relationships, thereby improving model interpretability in complex settings such as the medical field.

The *cubic spline* (degree 3) is a piecewise cubic polynomial defined on different intervals of `EntryAge`, where each interval is separated by a *knot*. The function is smooth across

these intervals, meaning that the spline is continuous and has continuous first and second derivatives at the knots, ensuring no unexpected changes in the fitted curve.

The mathematical representation can be written as:

$$S(z) = \sum_{k=1}^K \theta_k B_k(Z)$$

where:

- $S(z)$ is the spline function at any point Z ;
- θ_k are the spline coefficients;
- $B_k(z)$ are the *basis functions*, which are the piecewise polynomials that define the shape of the spline.

The flexibility of a spline model depends on the number and placement of knots. More knots increase the flexibility of the function, allowing it to capture finer details in the data, but excessive knots can lead to overfitting. The number of degrees of freedom (df) in a spline model is directly related to the number of knots and the order of the polynomial used in each segment.

For a cubic spline with M internal knots, the degrees of freedom are approximately:

$$\text{df} \approx M + 3$$

where the additional 3 degrees of freedom account for order of the polynomials and boundary constraints.

To improve the stability of estimates and reduce the risk of overfitting at the boundaries of the covariate range, **natural splines** were used. Natural splines are a type of restricted cubic spline that impose additional boundary constraints by forcing the function to be linear beyond the outermost knots. This constraint enhances numerical stability and interpretability, which is particularly important in survival analysis, where the behavior of the hazard function at the extremes of the covariate distribution can strongly influence the overall model fit and predictions.

It is important to note that the degrees of freedom correspond to the number of basis functions used in the spline representation, as each basis function contributes one degree of freedom to the model.

4 | Exploratory Data Analysis

This chapter presents the exploratory analysis conducted to better understand the characteristics of the study population. In particular, the main objectives are the followings:

- Identify potential patterns.
- Examine relationships between covariates and survival outcomes.
- Study the evolution of the individuals' conditions.

The insights gained through this preliminary analysis help to give a clear direction to the modeling choices, such as variable selection, potential non-linear effects, and the appropriate specification of time scales.

4.1. Profile of Individuals at the Time of Multimorbidity Onset

The dataset presents a population of individuals who are 18 years old or older. 46% of them are males while the mean entry age is 61.20 years (Standard Deviation 15.61 years).

The distribution of the Entry Age is shown in Figure 4.1:

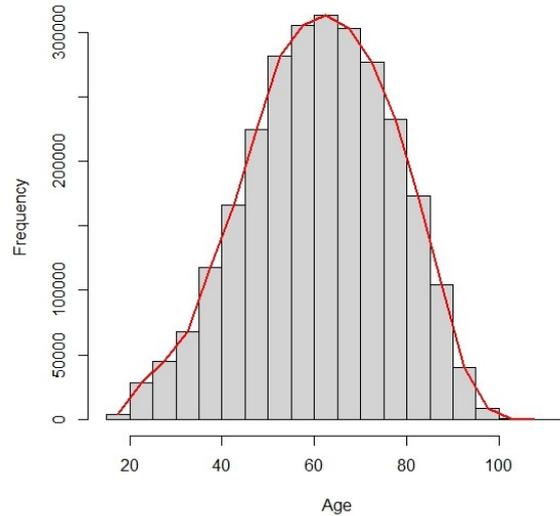


Figure 4.1: Entry Age of the study population

This figure illustrates that, on average, this phenomenon primarily affects individuals at an advanced age. However, due to the high variability, there are notable cases of individuals becoming multimorbid during early adulthood, even is less frequent. This highlights that everyone, regardless of age, can potentially be affected [24],[25].

Another key aspect is the difference in Entry Age between Men and Women with respectively the mean of 60.90 years (Std. 14.70 years) and 61.46 years (Std. 16.35 years).

To confirm the difference in distributions, the nonparametric version of the t -test is applied which gave a p -value $\leq 10^{-16}$: the hypothesis that the mean Entry Age is equal between males and females is rejected.

4.1.1. Individuals distributions among the clusters

A fundamental objective of this analysis is to identify the cluster to which an individual belongs at the time they become multimorbid. Recall that the clusters identified in [6] are: Allergies (ALL), Chronic Heart Conditions (CHC), Hypercholesterolemia (CHL), Diabetes (DIA), and Musculoskeletal and Psychiatric Conditions (M-P). These clusters capture distinct patterns of chronic disease combinations based on individual disease portfolios. As an individual's health status evolves over time, their disease portfolio and consequently their cluster assignment may change.

The clusters were derived using the K-means algorithm, and the optimal number of clus-

ters was determined through a combination of metrics: the Elbow method, the Calinski–Harabasz index, the Silhouette Score, and the Rim Data frequency. All methodological details regarding the clustering procedure are described in [6].

Table 4.1 presents the number of individuals assigned to each cluster (the corresponding % respect to the entire study population) and the associate mean Entry Age when they become multimorbid:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Number of Individuals	764,948 (28.39%)	402,266 (14.93%)	277,112 (10.29%)	187,964 (6.98%)	1,063,653 (39.48%)
Mean Age	53.98	69.40	61.48	59.50	63.52

Table 4.1: Distribution of subjects across clusters when they become multimorbid

The M-P cluster is the largest, with over 1 million citizens. Its mean entry age, 63.52 years, typically corresponds to the phase of life marking the end of a Danish citizen’s professional career. In contrast, the CHC cluster has a mean entry age of 69.40 years, aligning with the elderly phase of life, suggesting that multimorbidity often manifests later for these subjects.

A different pattern is observed in the Allergies (ALL) cluster, where multimorbidity tends to occur earlier in life, with a mean entry age of 53.98 years. This highlights how multimorbidity onset can vary significantly across clusters, reflecting differing risk factors and underlying conditions.

To provide a clearer summary of the disease composition within each cluster at the time individuals become multimorbid, Table 4.2 reports the ranking of conditions based on their prevalence. A lower rank indicates a higher frequency of the condition within the corresponding cluster. This ranking facilitates comparison across clusters and highlights which diseases are most commonly observed at the onset of multimorbidity in each group:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Allergies	1	15	11	13	14.5
Back Pain	4	9	4	6	3
Cancer	7	7	8	4	5
Heart Disease	10	1	14.5	15	13
COPD	5	4	9	9	7
Dementia	15	12	13	12	10
Depression	3	11	5	5	2
Diabetes	11	5	14.5	1	12
Hypercholesterolemia	9	3	1	3	14.5
Hypertension	2	2	2	2	1
Osteoarthritis	6	10	6	8	6
Osteoporosis	8	8	7	14	4
Joint Disease	14	14	12	11	11
Schizophrenia	13	13	10	10	9
Stroke	12	6	3	7	8

Table 4.2: Ranking of diseases prevalence when individuals become multimorbid

Consistent with the findings in [6], hypertension is highly prevalent across all clusters, suggesting that it may play a significant role in influencing survival outcomes among multimorbid individuals.

The Table 4.3 shows the distribution of women across the clusters:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
% of Women	61.9%	41.6%	47.8%	40.5%	58.2%
Women mean Age	53.64	72.93	62.75	60.75	64.14
Men mean Age	54.53	66.89	60.31	58.65	62.64

Table 4.3: Sex Distribution and Mean Entry Age

From the table, it is evident that women tend to become multimorbid at a later age than men and are most frequently assigned to the ALL cluster. Additionally, the M-P cluster also contains a high number of individuals.

4.2. Dead Individuals

Understanding the main patterns among individuals who died and how these may have changed since the onset of multimorbidity is important to gain insight into potential

factors that may contribute to mortality in multimorbid individuals. First of all the mean Death Age is 76.60 years (Standard Deviation of 12.01 years).

The histogram in Figure 4.2 illustrates the age distribution of deceased individuals:

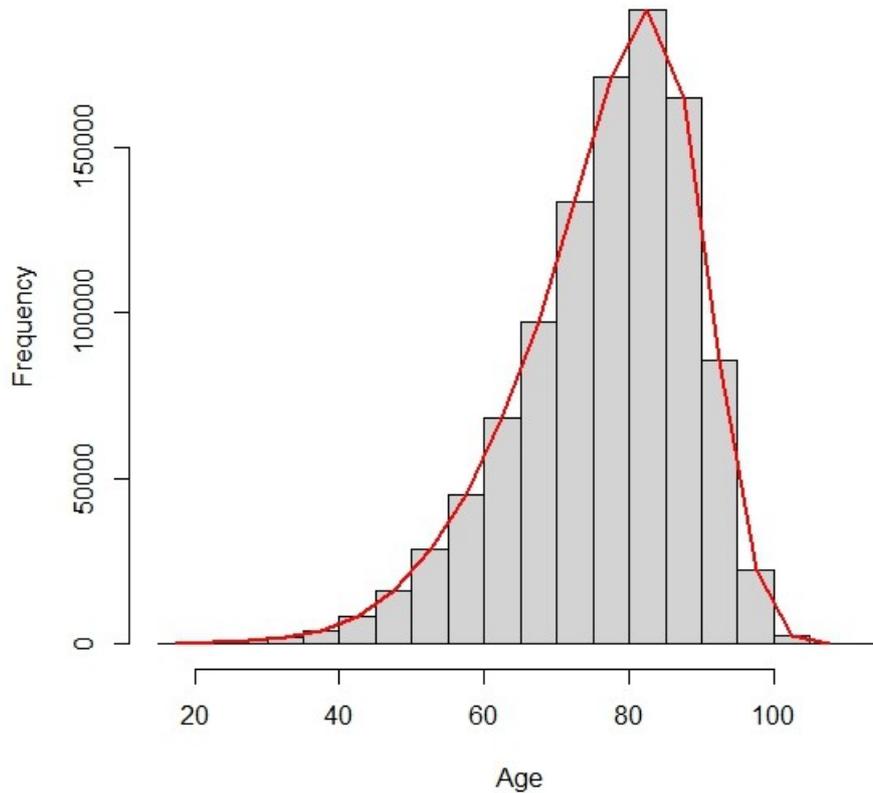


Figure 4.2: Age distribution of individuals when the event happens

The men, in mean, die at 74.42 years (Standard Deviation of 11.85 years), while women at 78.53 years (Standard Deviation of 11.83 years).

To confirm these observations, the nonparametric version of the t -test with a p -value $\leq 10^{-16}$: the hypothesis that the mean Death Age is equal between males and females is rejected.

4.2.1. Death distribution among the clusters

In Table 4.4, the number of individuals (and the % respect to all subjects who died):

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Number of Individuals	196,591 (18.89%)	396,308 (38.08%)	66,752 (18.89%)	64,156 (6.41%)	318,025 (30.55%)

Table 4.4: Distribution of subjects across clusters when the event happens

In this case the cluster that contains more died people is CHC followed by M-P. Instead, the table 4.5 show the ranking of conditions based on their prevalence. A lower rank indicates still a higher frequency of the condition within the corresponding cluster. In this case the ranking facilitates comparison across clusters and highlights which diseases might have a decisive role of the death of a subject:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Allergies	1	9	6	9	14
Back Pain	11	13	11	12	9
Cancer	4	7	3	4	2
Heart Disease	3	1	14.5	15	11
COPD	6	5	7	7	5
Dementia	10	12	10	10	7
Depression	7	10	8	6	4
Diabetes	13	6	14.5	1	13
Hypercholesterolemia	15	3	1	3	15
Hypertension	2	2	2	2	1
Osteoarthritis	9	11	9	11	8
Osteoporosis	5	8	5	8	3
Joint Disease	14	15	13	14	12
Schizophrenia	12	14	12	13	10
Stroke	8	4	4	5	6

Table 4.5: Ranking of diseases prevalence when individuals die

Among individuals in the CHC cluster, the most frequently observed conditions are Heart Disease, Hypertension, and Hypercholesterolemia, which is consistent with the finding that the majority of deaths occurred in this group. Interestingly, in the M-P cluster, Osteoporosis appears to play a potentially important role in determining mortality, which is a somewhat unexpected result.

To gain insight into how the patterns leading to death may differ between men and women, the distribution of female individuals across clusters was examined, as shown in Table 4.6.

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
% of Women	60.7%	46.6%	54%	43.9%	58%

Table 4.6: Sex distribution across the clusters

Allergies in combination with Hypertension and Heart Disease might play an important role to determine the death of women. But also the psycophysioical condition (e.g. Cancer or Depression) might be a significant factor.

4.3. Censored Individuals

To understand how the cluster patterns of censored individuals differ from those observed at the time of multimorbidity onset or among individuals who died, the same analytical approach was applied. The mean age of censored individuals at the end of follow-up was 59.35 years, with a standard deviation of 14.86 years. The men have a mean age of 59.29 years (Standard Deviation 14.08 years), women 59.40 years (Standard Deviation: 15.45 years).

4.3.1. Subjects distribution among the clusters

In the table 4.7, the numbers of individuals in each cluster (and the % respect to all subjects who "survived"):

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Number of Individuals	528,310 (31.97 %)	282,736 (17.11 %)	330,903 (20.03 %)	170,159 (10.30 %)	340,203 (20.59 %)

Table 4.7: Distribution of subjects across clusters when the censoring happens

Most of the individuals who remained alive at the end of follow-up belong to the ALL cluster, suggesting that the combinations of the conditions associated with this group may be less severe and therefore less likely to significantly reduce life expectancy.

Also in this case, the table 4.8 show the ranking of conditions based on their prevalence:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
Allergies	1	4	3	4	15
Back Pain	4	8	7	7	2
Cancer	8	12	10	9	7
Heart Disease	10	1	14.5	15	11
COPD	6	7	9	8	6
Dementia	15	14	13	14	12
Depression	3	9	6	5	3
Diabetes	13	5	14.5	1	13
Hypercholesterolemia	9	2	1	2	14
Hypertension	2	3	2	3	1
Osteoarthritis	5	6	4	6	4
Osteoporosis	7	11	8	11	5
Joint Disease	14	13	12	13	10
Schizophrenia	11	15	11	12	8
Stroke	12	10	5	10	9

Table 4.8: Ranking of diseases prevalence when individuals are censored

The most prevalent conditions in the ALL cluster are Allergies, Hypertension, and Depression a combination that may not be particularly severe or life-threatening. A notable difference compared to individuals from the same cluster who died is that Heart Disease does not appear among the top three most frequently observed conditions in the surviving group.

The distribution of women across the clusters in each cluster are shown in the following table:

	Cluster ALL	Cluster CHC	Cluster CHL	Cluster DIA	Cluster M-P
% of Women	64.4%	39.7%	55.6%	45.7%	59.2%

Table 4.9: Sex distribution and average age at the censoring of subjects

Most of the women who remained alive at the end of follow-up belong to the ALL cluster, although a significant proportion are also found in the M-P cluster. The top three conditions in this group Hypertension, Back Pain, and Depression suggest a burden that combines both physical and mental health issues. In contrast, for men, heart related conditions appear to play a more prominent role.

4.4. Kaplan-Meier curves with Time of follow up as time scale

Estimating the probability of survival based on factors such as age and sex is essential for understanding how the risk of death evolves over time. This information helps identify individuals at higher risk and supports the development of targeted interventions. In addition, survival analysis can reveal patterns and insights that may inform subsequent mathematical modeling. From a healthcare management perspective, these insights can support more efficient hospitalization strategies, helping to avoid system overload and reduce unnecessary costs.

First Case: Optimal cutpoint of EntryAge range

To analyze the relationship between the EntryAge and survival outcomes, the optimal cutpoint was determined by splitting the EntryAge variable at different values and comparing the survival distributions of the resulting groups. The log-rank test was used to assess whether the survival curves of these groups significantly differed.

The log-rank test statistic was then computed, and the cutpoint that resulted in the most significant difference in survival (the smallest p-value) was selected as the optimal cutpoint.

This method is useful to create categories within the continuous variable in a way that reflects meaningful differences in survival, enhancing the interpretability of survival models and highlighting significant risk factors. In the analysis the cutpoint that gives the most useful information correspond to 69.39 Years Old.

In fig. 4.3 the two curves corresponding to the individuals who becomes multimorbid before and after to 69.39 are shown:

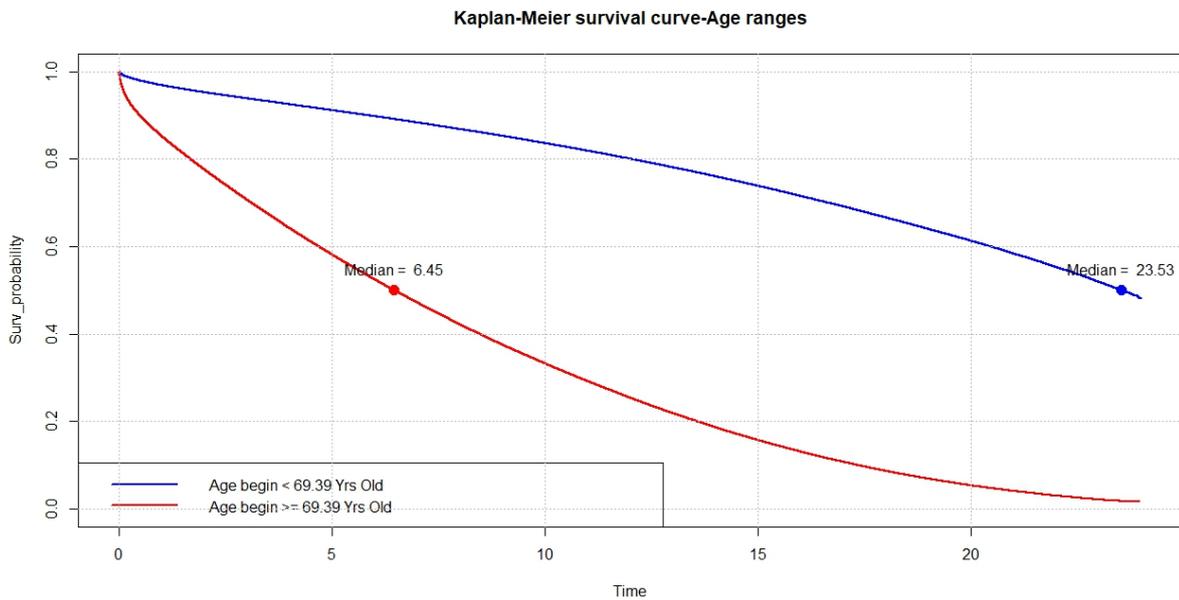


Figure 4.3: Kaplan-Meier curves with Optimal cut point of EntryAge range

Older citizens (represented by the red line) exhibit a significantly higher risk, with survival probability dropping below 50% after 6.45 years. In contrast, younger citizens generally experience greater longevity, with a median survival time of 23.53 years covering nearly the entire follow-up period.

Second case: EntryAge range cut by hand

In the second case there the EntryAge range is divided in four part correspondig to different life stages:

- Young (18-29 Years old);
- Aldut (30-49 Years old);
- Senior (50-79 Years old);
- Old (80+ Years old);

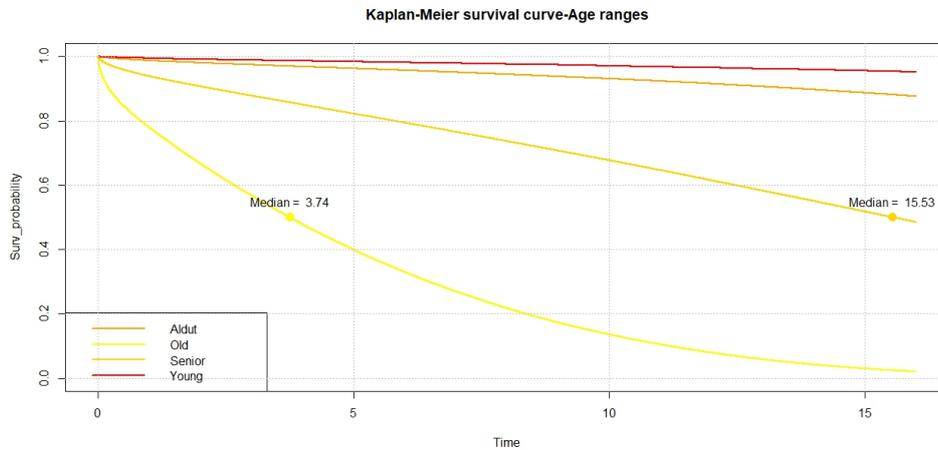


Figure 4.4: Kaplan-Meier curves with EntryAge range cut by hand

Who become muktimorbid in the first stages of the life (Youngs and Alduts) do not experience a significant decreasing of the life expectancy: In fact, the graph does not show a medium survival time for them. Conversely, the situation is much worse for who become multimorbid later.

4.4.1. Interaction between EntryAge and Sex

To understand if there is another factor that can affect the mortality, an interaction between the EntryAge and Sex is performed. The Figure 4.12 shows the plot of the situation:

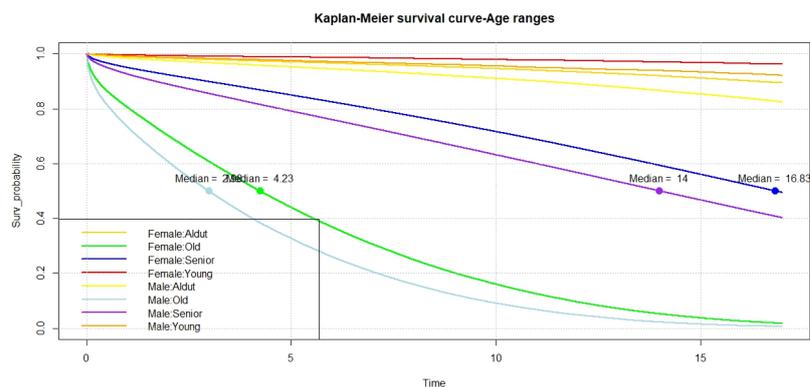


Figure 4.5: Kaplan-Meier curves with interaction between EntryAge and Sex

For each EntryAge category, the survival curve for females consistently lies above that of males, indicating that being female may act as a protective factor. This suggests that Sex is an important variable influencing mortality risk.

To spot eventual distributional similarities between the curves the Log-Rank test is performed for the all described cases as shown on 4.10:

Case	<i>p</i> -value
Optimal EntryAge split	10^{-16}
Manual EntryAge split	10^{-16}
EntryAge–Sex interaction	10^{-16}

Table 4.10: *p*-values from the Log-Rank test for each case

The *p*-value is consistently equal to 10^{-16} , which led to the conclusion that the curves are difference in distribution between each other.

4.5. Kaplan-Meier curves with Age as time scale

Since the previous section clearly showed that the EntryAge variable is an important factor influencing mortality, it is worthwhile to explore this aspect further by analyzing survival curves using age as the timescale. The key difference compared to using time-on-study as the timescale is that, when using age, individuals are compared only with others of similar age. This approach provides a biologically meaningful perspective and may reveal age-specific mortality patterns that are not apparent when using follow-up time alone.

4.5.1. EntryAge Greater Than or Equal to 20

The survival curves illustrating the survival probability as a function of age considering the individuals who become multimorb from 20 years old are shown in Figure 4.13:

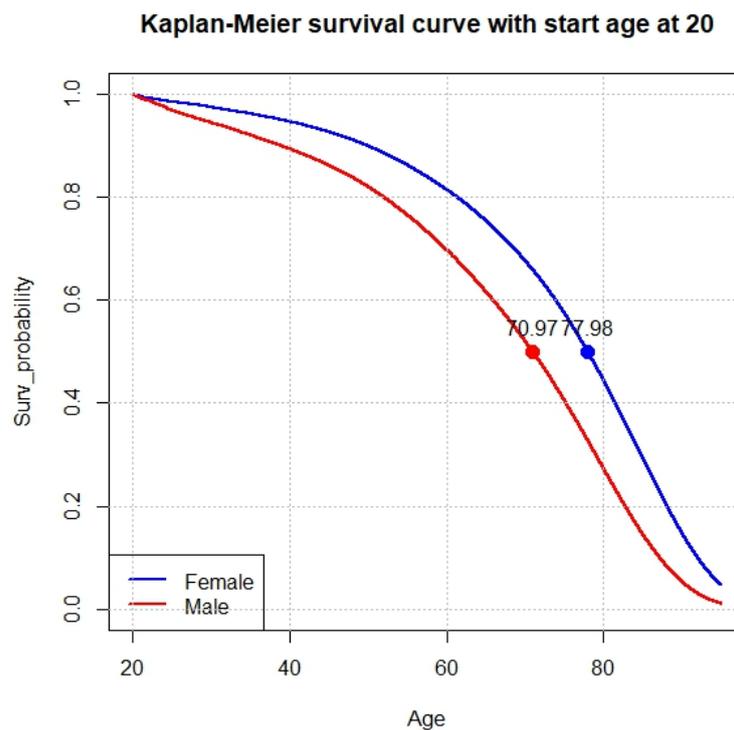


Figure 4.6: Survival curves with EntryAge Greater Than or Equal to 20

The illustrated curves reinforce the insights derived from the previous analyses: young multimorbid individuals exhibit a low risk of mortality. Specifically, the median survival age is 70.97 for men and 77.98 for women, highlighting that women consistently have a higher probability of living longer.

4.5.2. EntryAge Greater Than or Equal to 30

The survival curves of individuals who become multimorbid after 30 years old are shown as follows:

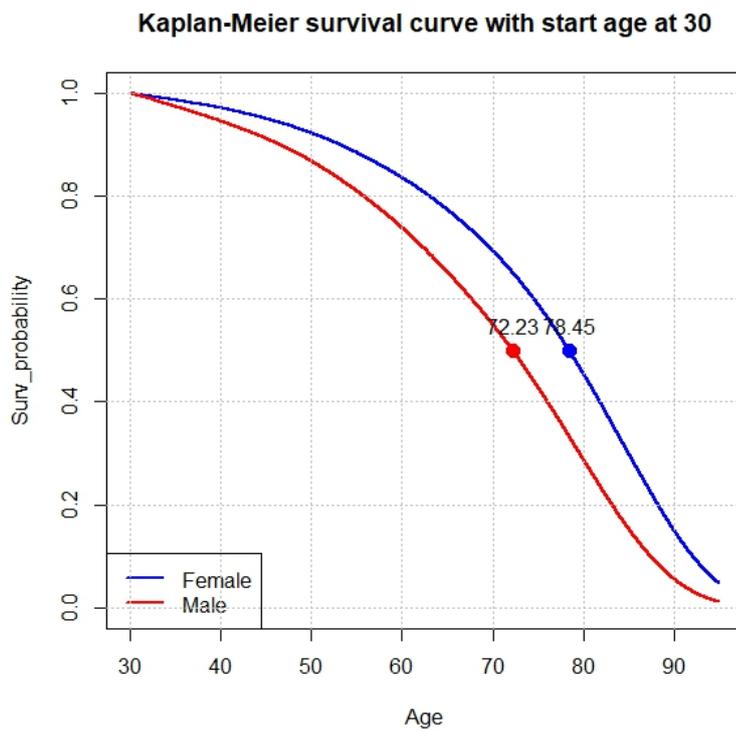


Figure 4.7: Survival curves with EntryAge Greater Than or Equal to 30

The Figure 4.14 shows a pattern similar to the previous case: both sexes have a median survival time exceeding 70 years. In practical terms, an individual who becomes multimorbid in their 30s has roughly the same life expectancy as someone who develops multimorbidity in their 20s. During these stages of life, biological functions remain robust, so the presence of chronic diseases does not significantly impact overall health at these ages.

4.5.3. EntryAge Greater Than or Equal to 50

In following picture, it is possible to observe that the curves start to exhibit a different shape respect to the previous two cases: after the age of 60, the relation between the survival probability and the age becomes linear. This provides further evidence that developing multimorbidity at a later stage in life is associated with more severe consequences:

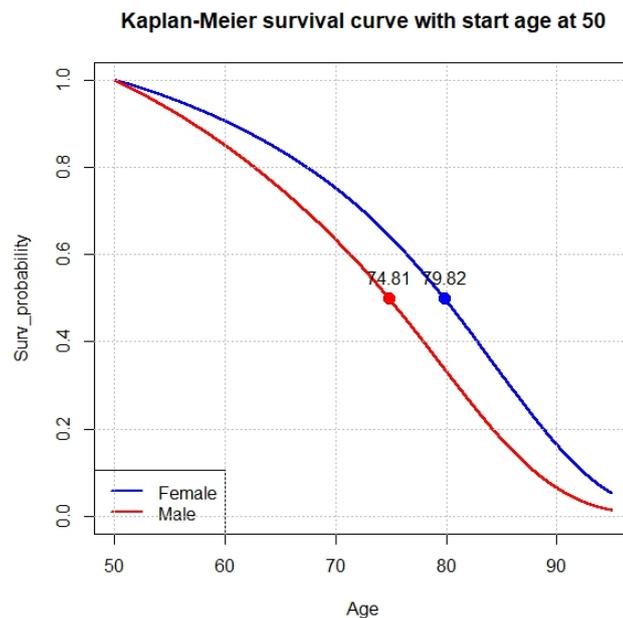


Figure 4.8: Survival curves with EntryAge Greater Than or Equal to 50

4.5.4. EntryAge Greater Than or Equal to 80

In Figure 4.16 the curves are notably steep, with the median age occurring approximately six years after the individuals entered the study. Once again, this highlights that for elderly citizens, the risk of death is significantly higher:

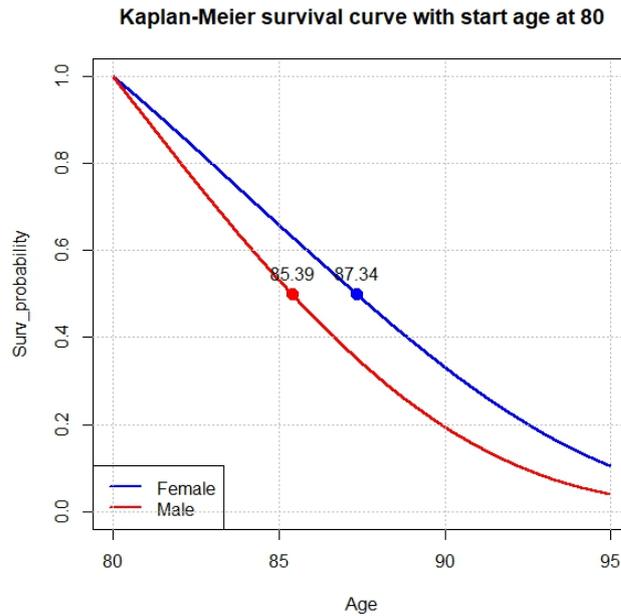


Figure 4.9: Survival curves with EntryAge Greater Than or Equal to 80

4.6. Studying the dynamics within the population

After examining the distribution of clusters within the study population and identifying factors that may influence mortality, it is important to analyze how frequently individuals change cluster membership over the course of the study. This dynamic perspective not only helps identify which clusters may be more critical in terms of risk, but also provides insight into the progression toward the endpoint. Furthermore, tracking transitions between clusters can offer valuable intuition about how multimorbidity develops depending on an individual's initial disease profile.

4.6.1. The most frequent trajectories

Understanding the most frequent sequences observed in the study population is essential for capturing how multimorbidity evolves over time at the individual level. The table 4.11 and table 4.12 present the most common sequences, along with the corresponding percentages of the total population that followed each trajectory. Since this analysis is intended to be informative, a minimum threshold of 0.18% (approximately 4,849 individuals) was applied to retain only the most representative trajectories. This allows for a clearer overview of the most common patterns while filtering out extremely rare transitions that may not be relevant for descriptive purposes.

Sequences	% of individuals
M-P	23.18 %
ALL	21.56 %
CHC	12.87 %
CHL	7.28 %
DIA	4.53 %
M-P → CHL	4.25 %
M-P → CHC	3.75 %
M-P → ALL	3.55 %
ALL → CHL	2.75 %
ALL → CHC	1.92 %
DIA → CHC	1.61 %
CHL → CHC	1.54 %
M-P → DIA	1.04 %
CHL → DIA	0.97 %
ALL → DIA	0.92 %

Table 4.11: Most frequent observed trajectories

Sequences	% of individuals
CHC → ALL	0.87 %
M-P → CHL → CHC	0.86 %
CHC → M-P	0.56 %
ALL → CHL → CHC	0.52 %
M-P → CHL → DIA	0.50 %
M-P → ALL → CHL	0.39 %
CHC → ALL → CHC	0.36 %
ALL → CHL → DIA	0.35 %
M-P → CHC → M-P	0.31 %
M-P → DIA → CHC	0.26 %
M-P → ALL → CHC	0.25 %
M-P → CHC → ALL	0.20 %
DIA → ALL	0.20 %
CHL → ALL	0.19 %
CHL → DIA → CHC	0.18 %

Table 4.12: Most frequent observed trajectories

The first observation is that the majority of subjects remain in the same cluster throughout the entire study period. For the more complex trajectories, 15.36 % of the individuals transition out of the M-P cluster. Similarly, 6.46 % leave the ALL cluster, again primarily transitioning to the CHC or CHL clusters. In most cases (24.12 %), the CHC cluster represents the final cluster in many trajectories, indicating that this cluster is associated with a higher frequency of death and is primarily composed of older individuals. The M-P and ALL clusters might also be associated with significant mortality risk. This is suggested by the fact that subjects tend to remain in these clusters throughout the study period (23.18 % and 21.56%). It is likely that the diseases linked to these clusters have such a substantial impact on individuals' health that other potential conditions or effects become negligible in comparison.

4.6.2. Relative Frequencies of Transitions Across Multimorbidity Clusters

After extracting the most frequent multimorbidity sequences, our goal was to compute the relative sequences of observing transitions between distinct clusters or to death. To focus the analysis on actual changes in multimorbidity profiles, we first formalize the process of collapsing repeated cluster assignments and then compute transition frequencies based on the resulting sequences.

Raw Sequences

Let N be the total number of individuals. For each individual $i \in \{1, \dots, N\}$, we observe a sequence of cluster assignments over time:

$$C_{i,1}, C_{i,2}, \dots, C_{i,T_i}$$

where T_i is the number of observed time points for individual i , and each $C_{i,t} \in \mathcal{C} \cup \{\text{DEATH}\}$, with:

$$\mathcal{C} = \{\text{ALL}, \text{CHC}, \text{CHL}, \text{DIA}, \text{M-P}\}, \quad \text{DEATH representing an absorbing death state.}$$

Collapsing Same-Cluster Observations

To focus only on actual transitions between different states, we collapse consecutive identical values of $C_{i,t}$. This results in a reduced sequence:

$$\tilde{C}_{i,1}, \tilde{C}_{i,2}, \dots, \tilde{C}_{i,K_i}$$

where $K_i \leq T_i$ is the number of retained (non-redundant) time points for individual i . Formally, this sequence is defined as:

$$\tilde{C}_{i,1} := C_{i,1}$$

and for $k > 1$,

$$\tilde{C}_{i,k} := C_{i,t_k} \quad \text{where} \quad t_k = \min\{t > t_{k-1} \mid C_{i,t} \neq C_{i,t_{k-1}}\}$$

The process continues until no further changes in cluster assignment occur.

Valid Transitions

Based on the collapsed sequences, we define the set of valid transitions:

$$\mathcal{T}_{\text{valid}} := \left\{ (\tilde{C}_{i,k}, \tilde{C}_{i,k+1}) \mid i \in \{1, \dots, N\}, k \in \{1, \dots, K_i - 1\} \right\}$$

Each element of this set corresponds to a transition between two different states (i.e., between clusters or from a cluster to death).

Absolute Transition Frequencies

For each ordered pair $(a, b) \in (\mathcal{C} \cup \{\text{DEATH}\}) \times (\mathcal{C} \cup \{\text{DEATH}\})$ with $a \neq b$, define the absolute transition frequency as:

$$n_{a \rightarrow b} := |\{(c_1, c_2) \in \mathcal{T}_{\text{valid}} \mid c_1 = a, c_2 = b\}|$$

Let n_a be the total number of outgoing transitions from state a :

$$n_a := \sum_{\substack{b \in \mathcal{C} \cup \{\text{DEATH}\} \\ b \neq a}} n_{a \rightarrow b}$$

Relative Transition Frequencies

The relative transition frequency from state a to state b is then given by:

$$f_{a \rightarrow b} := \begin{cases} \frac{n_{a \rightarrow b}}{n_a}, & \text{if } n_a > 0 \\ 0, & \text{otherwise} \end{cases}$$

Here are some important remarks to justify the steps leading to the construction of the transition matrix:

- In this context the relative frequencies of the transitions in which the belonging cluster does not change is equal to 0.
- The dataset involves non-informative censoring, meaning censoring occurs either when the follow-up period ends or when a patient emigrates, regardless of their current cluster.
- Given the large number of subjects and observed transitions, the Law of Large Numbers ensures that the computed frequencies provide a reliable approximation of the true underlying probabilities describing population dynamics.

The R code is showed in the Appendix A.

With these considerations, we introduce the matrix. Given the state space

$$S = \{\text{ALL}, \text{CHC}, \text{CHL}, \text{DIA}, \text{M-P}, \text{DIED}\},$$

the following matrix represents the relative frequencies of transitioning between clusters

or dying:

$$P = \begin{bmatrix} 0.000 & 0.180 & 0.270 & 0.090 & 0.000 & 0.460 \\ 0.090 & 0.000 & 0.000 & 0.000 & 0.070 & 0.840 \\ 0.060 & 0.370 & 0.000 & 0.270 & 0.010 & 0.300 \\ 0.090 & 0.420 & 0.000 & 0.000 & 0.080 & 0.400 \\ 0.160 & 0.170 & 0.200 & 0.060 & 0.000 & 0.410 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

Here, each row corresponds to the current cluster (e.g., ALL, CHC, etc.), and each column represents the next cluster or the "DIED" state. Each row of the resulting matrix sums to 1 and represents the relative frequency distribution of transitions from a given cluster or to death.

From the above matrix, it is possible to observe several consistencies with the sequence rankings. For instance, subjects in the Allergies cluster (ALL) never transition to the M-P cluster, as the associated probability is 0. Similarly, individuals in the CHC cluster do not transition to the CHL or DIA clusters, and they exhibit the highest likelihood of dying. On the other hand, CHL subjects predominantly transition to either the CHC or DIA clusters. Lastly, individuals in the M-P cluster are more likely to move to the ALL, CHC, or CHL clusters, highlighting significant movement between these clusters.

If we consider CENSORED as an additional absorbing state and all the transitions that lead to it, the state spaces becomes $S = \{\text{ALL, CHC, CHL, DIA, M-P, DIED, CENSORED}\}$, and the matrix becomes as follows.

$$P = \begin{bmatrix} 0.000 & 0.079 & 0.123 & 0.040 & 0.002 & 0.205 & 0.551 \\ 0.056 & 0.000 & 0.000 & 0.000 & 0.046 & 0.524 & 0.374 \\ 0.023 & 0.148 & 0.000 & 0.108 & 0.002 & 0.121 & 0.598 \\ 0.043 & 0.204 & 0.000 & 0.000 & 0.040 & 0.195 & 0.518 \\ 0.113 & 0.116 & 0.141 & 0.040 & 0.000 & 0.285 & 0.305 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

The internal dynamics remain confirmed: for example, for a subject in CHC cluster, the frequency of transitioning to the death state remains the highest, followed by the frequency of transitioning to the Allergies cluster.

5 | Model Construction

After have explored the data, two extended Cox Proportional Hazards models were estimated: one using time-on-study as the timescale, and the other using attained age. This dual approach provides complementary perspectives on mortality risk, enabling a more nuanced understanding of the differences across clusters and allowing for more comprehensive conclusions.

In addition, the variable related to education level was included in the models to assess the impact of sociodemographic factors on mortality, and to explore how these may interact with disease patterns within the clusters.

In general, when comparing models, we will distinguish between two cases. If the models are nested, the Likelihood Ratio Test (LRT) is used. Conversely, if the models are not nested, model performance is assessed using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

5.1. First model: Time on study as timescale

The first model that was estimated is the following one:

$$h(t | \mathbf{X}(t), Z) = h_0(t) \exp \left(\boldsymbol{\beta}^\top \mathbf{X}(t) + \sum_{k=1}^K \theta_k B_k(Z) \right),$$

where

- $h(t | \mathbf{X}(t), Z)$ is the hazard function at time t given possibly time-dependent covariates $\mathbf{X}(t)$ and continuous covariate Z ;
- $h_0(t)$ is the baseline hazard function;
- $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))^\top$ is the vector of (possibly time-varying) covariates;
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ are their regression coefficients;
- $B_k(Z)$, for $k = 1, \dots, K$, are spline basis functions evaluated at Z ;

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$ are spline coefficients.

In $X_i(t)$ includes covariates such as Sex, Education, and Cluster membership. While Z is EntryAge modeled with Natural Spline. Additionally, pairwise interaction terms between Sex, Education, and Cluster were included in the model to account for potential effect modification. The AIC and BIC associated to this model are 28174215 and 28174737.

5.1.1. Model Diagnostics and Model Selection

First of all we verify if the estimated model meets the Proportional Hazards (PH) assumption. To answer this question first we computed the relative Schoenfeld residuals. The following plots illustrate the trends of the residuals for several covariates:

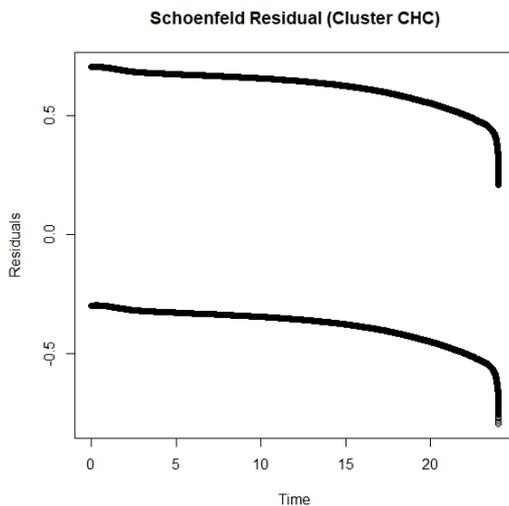


Figure 5.1: Residuals for Cluster CHC Covariate

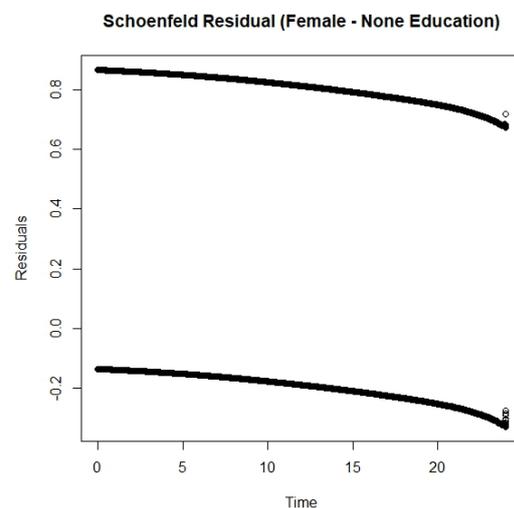


Figure 5.2: Residuals for the interaction Female and None Education Covariate

The residuals do not appear to be randomly distributed around zero, as would be expected if the model assumptions were met. This observation suggests a non-negligible likelihood that the proportional hazards assumption may be violated.

To further investigate this issue, we explored how the estimated coefficients (β) change over time. If a covariate effect is not constant throughout the follow-up period, this provides additional evidence against the validity of the proportional hazards assumption.

The following pictures describes the effects of the coefficient of covariates included in the model:

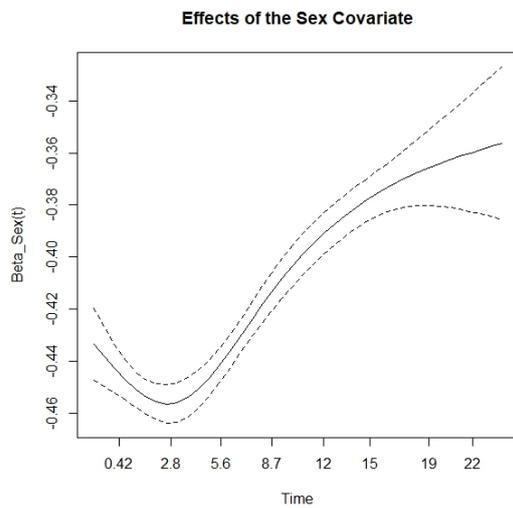


Figure 5.3: Sex covariate effect

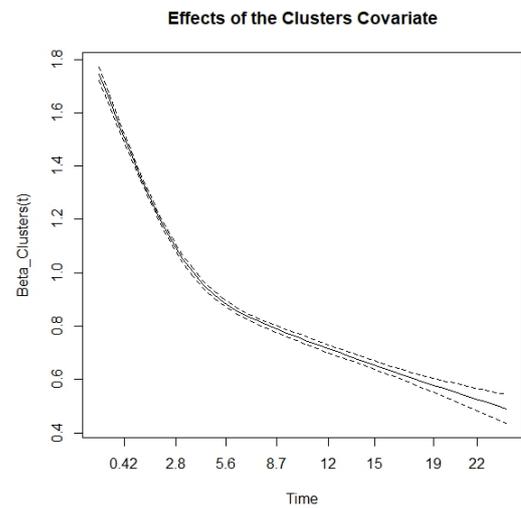


Figure 5.4: Clusters covariate effects

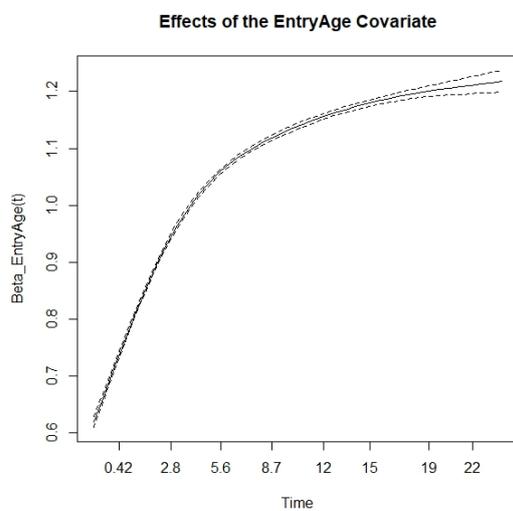


Figure 5.5: EntryAge covariate effects

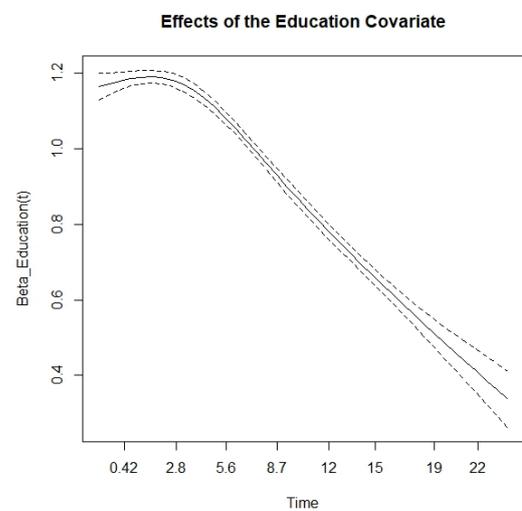


Figure 5.6: Education covariate effects

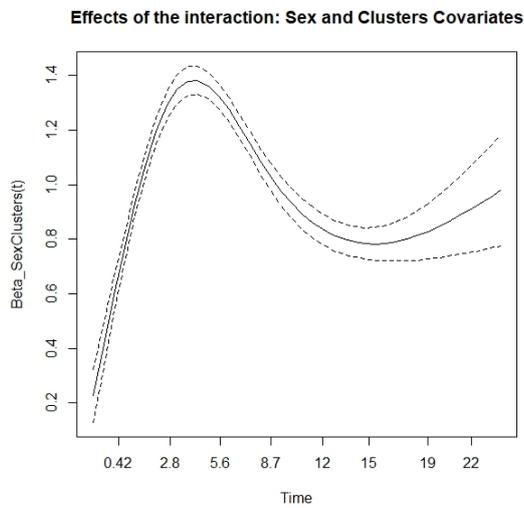


Figure 5.7: Effects of the interaction between Sex and Clusters Covariates

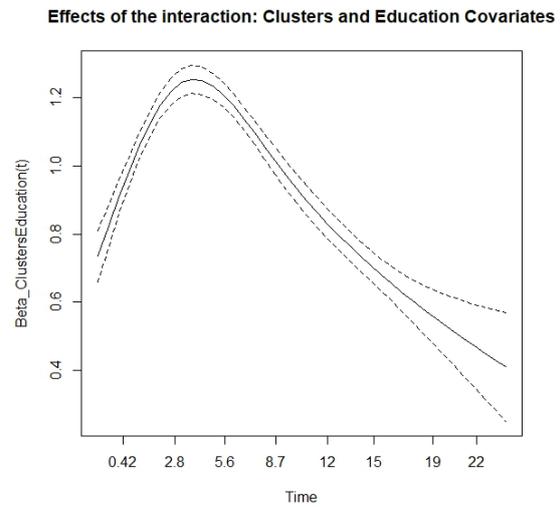


Figure 5.8: Effects of the interaction between Clusters and Education Covariates

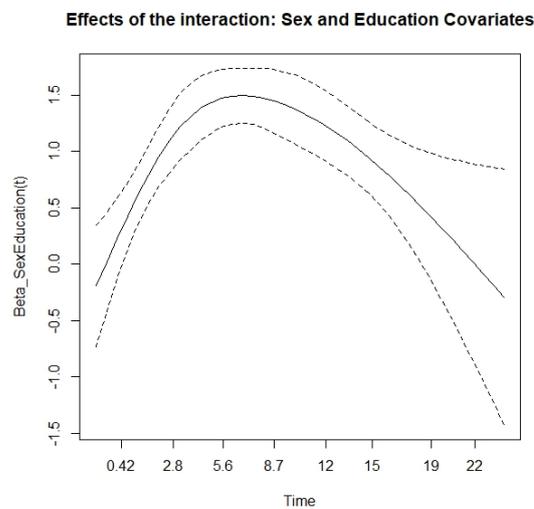


Figure 5.9: Effects of the interaction between Sex and Education Covariates

All covariate effects exhibit clear non-constant trends over time, indicating that the hazard ratios vary significantly during follow-up. Therefore, the proportional hazards (PH) assumption is violated as confirmed in table 5.1:

Covariate	Statistic value	Degree of freedom	P-value
Sex	797	1	10^{-16}
Clusters	6811	4	10^{-16}
EntryAge	12198	11	10^{-16}
Education	417	4	10^{-16}
Sex:Clusters	2130	4	10^{-16}
Clusters:Education	7257	16	10^{-16}
Sex:Education	1224	4	10^{-16}
Global Model	25161	44	10^{-16}

Table 5.1: PH test for final model with Time on study as timescale

To address this issue, first of all we asked what matters the most in this research and the answer was a correct and realistic interpretation of the studying population and the assessment of the mortality of each cluster. Then several strategies were considered, in particular stratification on categorical variables such as Sex, Cluster, or Birth Cohort. While stratification can effectively relax the PH assumption by allowing each stratum to have its own baseline hazard function, this approach reduces interpretability because it estimates separate baseline hazards $h_{0,s}(t)$ for each category s . Formally, the stratified Cox model is defined as:

$$h(t | \mathbf{X}, S = s) = h_{0,s}(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{X}),$$

where

- $h(t | \mathbf{X}, S = s)$ is the hazard function at time t for an individual with covariates \mathbf{X} in stratum s ;
- $h_{0,s}(t)$ is the baseline hazard specific to stratum s ;
- \mathbf{X} is the vector of covariates excluding the stratifying variable S ;
- $\boldsymbol{\beta}$ is the vector of regression coefficients assumed common across strata.

The main limitation of this approach is that absolute risks cannot be directly compared between strata, since each stratum has its own baseline hazard. For example, absolute risk comparisons between males and females are not meaningful. Moreover, the effects of non-stratified covariates are assumed to be constant across strata, precluding interactions between stratified and non-stratified variables. This constraint may mask important heterogeneity and limit the discovery of significant insights. It is important to remember

that a key assumption of the stratified Cox model is the absence of interaction between stratified and non-stratified covariates [19].

As an alternative to stratification, the use of time-varying covariates and smooth functional terms via splines was considered more appropriate. This approach allows for modeling non-proportional hazards within a unified framework and enables direct estimation of interpretable, time-dependent hazard ratios.

Additionally, the idea of partitioning the time axis i.e., defining time intervals over which the proportional hazards assumption approximately holds was considered. This so-called piecewise Cox approach [23] involves fitting separate Cox models over pre-specified time domains. However, this strategy increases model complexity and may lead to difficulties in interpretation and loss of power due to reduced sample size per interval.

Finally, Accelerated Failure Time (AFT) models represent a parametric alternative that models the survival time directly, assuming a log-linear relationship with covariates. While AFT models can accommodate non-proportional hazards implicitly, they rely on strong distributional assumptions and are not easily extended to allow for time-varying effects or flexible interactions. Given the goal of estimating dynamic hazard ratios in a semi-parametric framework, AFT models were not pursued.

Given the large size of the dataset, relying solely on formal statistical tests to assess the Proportional Hazards (PH) assumption may lead to misleading conclusions. As noted by [23], with substantial sample sizes, even minor deviations from proportionality often negligible in practical terms can result in statistically significant test results. This sensitivity arises because the tests are powered to detect even trivial violations of the PH assumption. Therefore, it is crucial to complement statistical testing with graphical diagnostics and substantive domain knowledge. Visual inspections, such as plots of scaled Schoenfeld residuals or time-dependent coefficient estimates, can help determine whether the deviations are meaningful for interpretation. In this context, a violation of the PH assumption does not necessarily undermine the model's interpretability or usefulness, particularly if the effect is stable over most of the follow-up period or if deviations are of minor practical importance.

For these reasons, modeling time-dependent effects using splines, while avoiding stratification and interval partitioning, was considered to offer the best trade-off between flexibility, interpretability, and generalizability.

The main limitation of not stratifying the model is that the estimated hazard ratios may be biased, as they do not capture instantaneous, time-varying effects. Therefore,

the computed hazard ratios should be interpreted as mean effects over the study period, weighted by the risk sets at each time.

Aware of these trade-offs, the model was fitted without stratification, while carefully interpreting the results with these limitations in mind.

5.1.2. How the number of Degrees of Freedom for the spline terms were chosen

To account for the potential nonlinear relationship between the continuous covariate EntryAge and the log hazard, we incorporated a spline basis expansion into the linear predictor of the Cox model. Let $Z \equiv \text{EntryAge}$ and $\{B_k(Z)\}_{k=1}^K$ denote a spline basis of dimension K . The hazard function becomes:

$$h(t \mid \mathbf{X}(t), Z) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{X}(t) + \boldsymbol{\theta}^\top \mathbf{B}(Z)),$$

where $\mathbf{B}(Z) = (B_1(Z), \dots, B_K(Z))^\top$ and $\boldsymbol{\theta} \in \mathbb{R}^K$ are the spline coefficients to be estimated.

This modification does not alter the structure of the Cox model but introduces additional parameters to estimate via the Partial Likelihood. The standard Cox Partial Likelihood for I observed events, time events t_i , and covariates $\mathbf{X}_i(t)$ and Z_i is given by:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^I \frac{\exp(\boldsymbol{\beta}^\top \mathbf{X}_i(t_i) + \boldsymbol{\theta}^\top \mathbf{B}(Z_i))}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{X}_j(t_i) + \boldsymbol{\theta}^\top \mathbf{B}(Z_j))},$$

where $R(t_i)$ denotes the risk set at time t_i .

The introduction of the spline basis increases the flexibility of the model but also its complexity. To prevent overfitting and ensure interpretability, a careful selection of the spline dimension K was performed. The procedure consisted of two steps:

1. **Model Selection via Information Criteria.** A grid search over values of K was conducted, where K represents the number of spline basis functions included in the model. For each candidate value of K , an intermediate Cox proportional hazards model was fitted (it contains only the Sex, Cluster and EntryAge variables and there is no interaction), and the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were computed from the corresponding log-partial likelihoods. The optimal number of basis functions was then selected as the smallest K

for which the AIC and BIC reached a minimum or exhibited a plateau, indicating diminishing returns from increased complexity. The initial range of K values, from 5 to 20, was chosen based on practical recommendations from [26], who suggests starting with a moderate number of knots or basis functions—commonly between 3 and 5 and extending this range as needed to balance model flexibility and the risk of overfitting. This choice reflects a conservative approach to capture nonlinear effects without excessive model complexity, considering the large sample size of the dataset, which supports more flexible modeling.

2. **Cross-Validation.** To validate the spline complexity selected via AIC and BIC, and to assess the model’s ability to generalize, a k -fold cross-validation procedure was applied. For each candidate number of spline basis functions, the model was repeatedly trained on $k - 1$ folds and evaluated on the remaining fold. The concordance index (C-index), a standard measure of predictive discrimination in survival models, was computed on each test fold and then averaged across all folds. The final number of basis functions was chosen as the one maximizing the cross-validated C-index, ensuring a balance between flexibility and predictive stability.

This methodology ensures that the spline-expanded Cox model is parsimonious while adequately capturing the nonlinear effects of EntryAge. Moreover, since the spline enters linearly in the linear predictor, standard inference methods based on the Cox model remain applicable [23].

The resulting AIC and BIC trends are presented in fig. 5.10 and fig. 5.11:

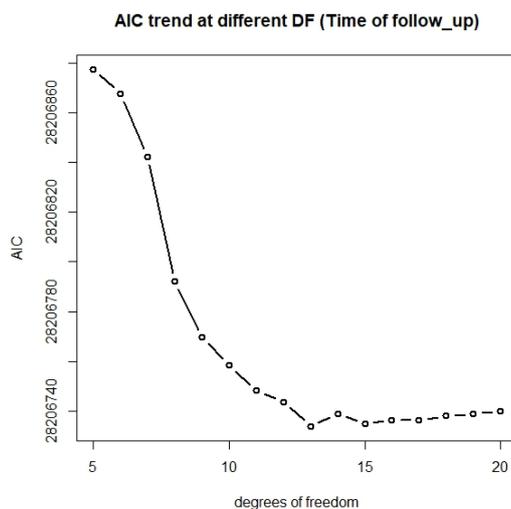


Figure 5.10: AIC trend with Time on study model

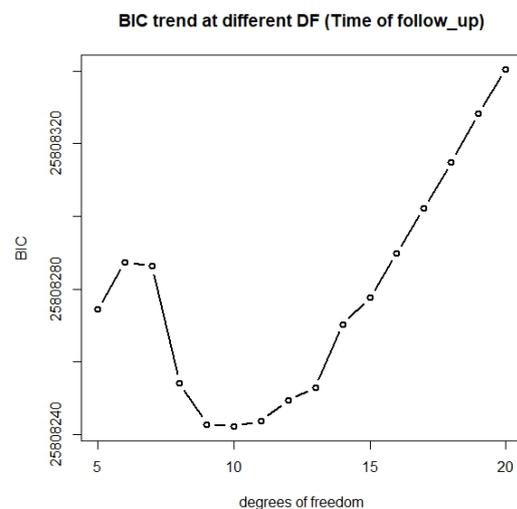


Figure 5.11: BIC trend with Time on study model

The AIC trend decreases until reaching a minimum at 13 degrees of freedom, after which it fluctuates slightly before increasing. In contrast, the BIC index exhibits a more erratic pattern: it initially increases, then decreases, and subsequently rises again. Based on these trends, the optimal choice appears to be 11 degrees of freedom, as it provides a relatively low AIC while avoiding a substantial increase in BIC.

To further validate this selection, Cross-Validation was performed for models with degrees of freedom ranging from 5 to 11, identifying the model with the highest Concordance Index (C-index). The results confirmed that 11 degrees of freedom was the best choice, yielding a C-index of 0.79. This suggests that the model has good discriminatory ability, meaning that in 79% of randomly chosen pairs, the model correctly assigns a higher risk to the individual who experiences the event first.

The **selected model** achieves an AIC of 28206748 and a BIC of 28206938.

The code outlines the procedure for selecting the best degrees of freedom (df) based on the AIC, BIC, and Cross-Validation results can be found in the Appendix A.

It is important to note that when constructing the train and test sets, a sampling method is used to ensure that the sets contain both censored and deceased individuals. This is crucial for preventing bias and ensuring that the models are tested on a representative sample of the population, rather than being skewed by sets with only censored or deceased subjects.

5.1.3. Interepretation of the estimated HR

In the following pages, we present the estimated Hazard Ratios (HRs) of each cluster relative to the reference cluster **Allergies (ALL)**, along with their 95% Confidence Intervals (CIs), separately for each Education level and for both Females and Males. Where confidence intervals are large, overlapping each other or included the null value 1, indicating non-significant differences, additional pairwise comparisons between clusters were performed to establish a clear and statistically supported ranking of mortality risk. Subsequently, focusing on the two clusters with the highest mortality risk (M-P and CHC), further analyses compared the effect of Education levels within each cluster, to assess how sociodemographic factors influence mortality risk in these high-risk groups.

The results are presented in tabular format, ensuring clarity and ease of interpretation.

- Long Education
 - Rank for Males: M-P, CHC, DIA, ALL and CHL;

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.714	1.665	1.764	1.989	1.930	2.049
CHL	0.675	0.648	0.703	0.726	0.696	0.757
DIA	1.183	1.129	1.239	1.248	1.190	1.310
M-P	1.902	1.848	1.957	1.775	1.723	1.827

Table 5.2: HR interval estimates for Long Education

– Rank for Females: CHC, M-P, DIA, ALL and CHL;

Since the confidence intervals are not too accurate, additional comparisons are performed to enhance reliability: additional HRs are provided without considering ALL as the reference. In this case, these comparisons **confirm** the ranking from Table table 5.2:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P vs CHC	1.110	1.080	1.140	0.892	0.868	0.917
M-P vs CHL	2.818	2.711	2.930	2.445	2.349	2.544
M-P vs DIA	1.608	1.537	1.682	1.422	1.357	1.489
CHC vs CHL	2.539	2.442	2.640	2.740	2.632	2.852
CHC vs DIA	1.449	1.385	1.515	1.593	1.520	1.669
DIA vs CHL	1.752	1.662	1.849	1.720	1.628	1.817

Table 5.3: Additional Clusters comparison within Long Education

- **Medium Education:**

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.753	1.708	1.799	2.035	1.982	2.089
CHL	0.698	0.674	0.724	0.751	0.725	0.779
DIA	1.162	1.118	1.209	1.227	1.178	1.278
M-P	1.922	1.873	1.972	1.794	1.748	1.840

Table 5.4: HR interval estimates for Medium Education

- Rank for Males: M-P, CHC, DIA, ALL and CHL;
- Rank for Females: CHC, M-P, DIA, ALL and CHL;

Also in this case other comparison are performed giving **consistency** to the ranking from table 5.4:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P vs CHC	1.096	1.071	1.122	0.881	0.861	0.902
M-P vs CHL	2.752	2.662	2.845	2.387	2.308	2.469
M-P vs DIA	1.653	1.593	1.717	1.462	1.406	1.520
CHC vs CHL	2.510	2.428	2.595	2.708	2.618	2.802
CHC vs DIA	1.508	1.453	1.565	1.659	1.596	1.724
DIA vs CHL	1.664	1.592	1.740	1.633	1.559	1.710

Table 5.5: Additional Clusters comparison within Medium Education

- Short Education:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.581	1.562	1.601	1.835	1.812	1.859
CHL	0.638	0.627	0.649	0.686	0.674	0.698
DIA	1.054	1.035	1.073	1.112	1.090	1.134
M-P	1.751	1.729	1.773	1.634	1.614	1.655

Table 5.6: HR interval estimates for Short Education

- Rank for Males: M-P, CHC, DIA, ALL and CHL;
- Rank for Females: CHC, M-P, DIA, ALL and CHL;

In this case, the confidence intervals are accurate, indicating that the estimates are more reliable.

- None Education:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.341	1.326	1.357	1.557	1.541	1.573
CHL	0.533	0.523	0.542	0.573	0.564	0.582
DIA	0.862	0.847	0.878	0.910	0.895	0.925
M-P	1.416	1.399	1.434	1.322	1.308	1.336

Table 5.7: HR interval estimates for None Education

- Rank for Males: M-P, CHC, DIA, ALL and CHL;
- Rank for Females: CHC, M-P, DIA, ALL and CHL;

Also in this case, the confidence intervals are accurate: the estimates are still reliable.

Below is the table summarizing the cluster rankings by Sex and Education Level:

Education Level	Male	Female
Long	M-P, CHC, DIA, ALL, CHL	CHC, M-P, DIA, ALL, CHL
Medium	M-P, CHC, DIA, ALL, CHL	CHC, M-P, DIA, ALL, CHL
Short	M-P, CHC, DIA, ALL, CHL	CHC, M-P, DIA, ALL, CHL
None	M-P, CHC, ALL, DIA, CHL	CHC, M-P, ALL, DIA, CHL

Table 5.8: Summary of clusters by Sex and Education level

The main distinction emerging from the results is that the model identifies the M-P cluster as the most hazardous for males, whereas the CHC cluster appears to be the most detrimental for females. For both sexes, the ALL and CHL clusters consistently exhibit the lowest levels of mortality risk.

Difference between levels of Education

To identify potential differences in mortality risk across educational levels, we compute additional hazard ratios (HRs) by comparing different levels of education within the same cluster. Given that M-P and CHC have been identified as the two most hazardous clusters in previous findings, our analysis focuses primarily on them.

The **Long Education** is considered as reference:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P Medium	1.076	1.047	1.106	1.074	1.043	1.107
M-P Short	1.288	1.260	1.317	1.239	1.209	1.270
M-P None	1.464	1.432	1.496	1.477	1.442	1.513
CHC Medium	1.090	1.062	1.118	1.088	1.053	1.124
CHC Short	1.291	1.265	1.317	1.242	1.209	1.276
CHC None	1.538	1.508	1.570	1.552	1.512	1.594

Table 5.9: Comparisons within Cluster but different Education Levels

For both sexes, individuals with higher Education are associated with lower hazard risk, possibly because they undergo regular check-ups and adopt healthier lifestyles. However, it is also possible that they are more likely to lead a sedentary lifestyle and this might be a factor that can affect their life expectancy.

5.2. Second Model: Age as timescale

The second estimated model, which uses age as the timescale, provides an additional perspective by accounting for biological aging processes that influence mortality. This approach is crucial for offering more comprehensive answers to the research question.

The second estimated model is the following:

$$h(a | \mathbf{X}(a), Z) = h_0(a) \exp \left(\boldsymbol{\beta}^\top \mathbf{X}(a) + \sum_{k=1}^K \theta_k B_k(Z) \right),$$

where

- $h(a | \mathbf{X}(a), Z)$ is the hazard function at the age a given possibly time-dependent covariates $\mathbf{X}(a)$ and continuous covariate Z ;
- $h_0(a)$ is the baseline hazard function;
- $\mathbf{X}(a) = (X_1(a), \dots, X_p(a))^\top$ is the vector of (possibly time-varying) covariates;
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ are their regression coefficients;
- $B_k(Z)$, for $k = 1, \dots, K$, are spline basis functions evaluated at Z ;
- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$ are spline coefficients.

In $X_i(a)$ includes covariates such as Sex, Education, and Cluster membership. While Z is EntryDay modeled with Natural Spline. Additionally, pairwise interaction terms between Sex, Education, and Cluster were included in the model to account for potential effect modification. The AIC and BIC associated to this model are 27511475 and 27512104.

5.2.1. Model Diagnostic and Model Selection

The diagnostic approach applied to this model is identical to that used for the first estimated model. Let's start looking the Schoenfeld residual on function of the age:

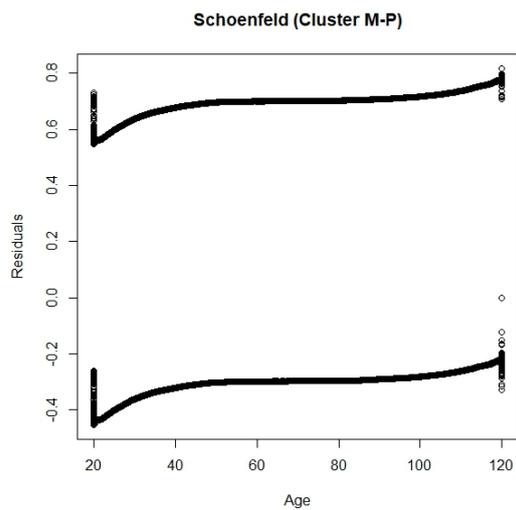


Figure 5.12: Residuals for M-P Cluster Covariate

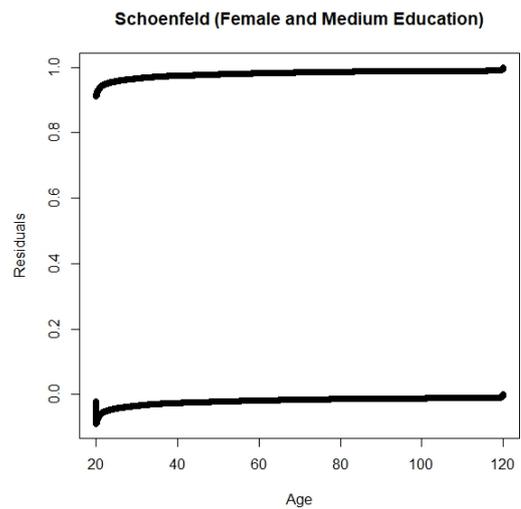


Figure 5.13: Residuals for interaction Sex Female and Medium Education Covariate

The residuals still show clear trend around 0. So it's most likely that the model does not meet the PH assumption. So the effects of the coefficient are taken in consideration.

The following plots show the effects of each coefficient during the time of the observation:

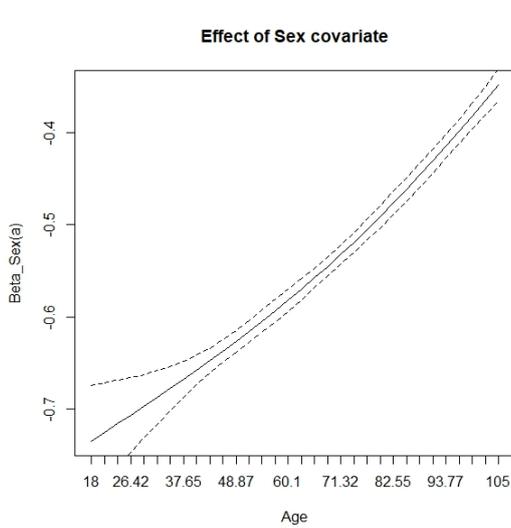


Figure 5.14: Effects of Sex Covariate

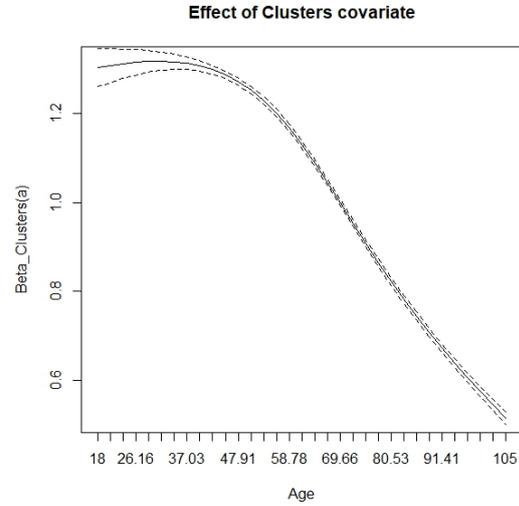


Figure 5.15: Effects of Clusters Covariate

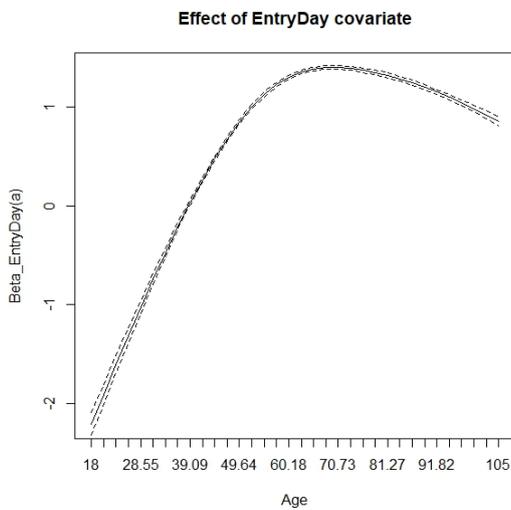


Figure 5.16: Effects of EntryDay Covariate

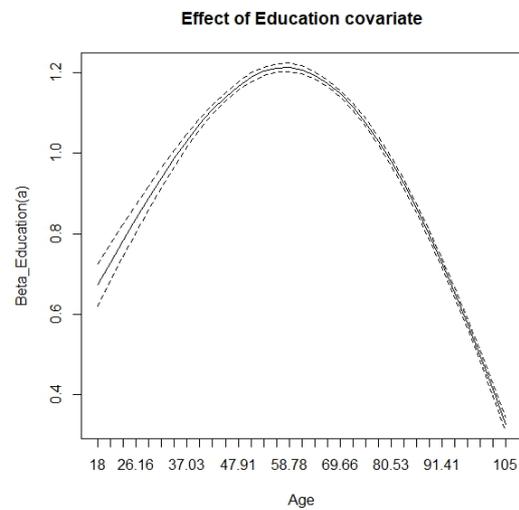


Figure 5.17: Effects of Education Covariate

From the analysis of Sex and Clusters covariates, a clear pattern emerges, with one showing an increasing trend and the other a decreasing one over time. In contrast, EntryDay and Education exhibit significant variability.

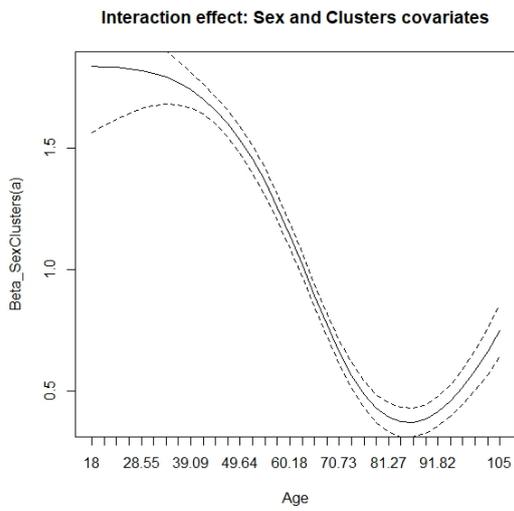


Figure 5.18: Effects of the interaction between Clusters and Sex Covariates

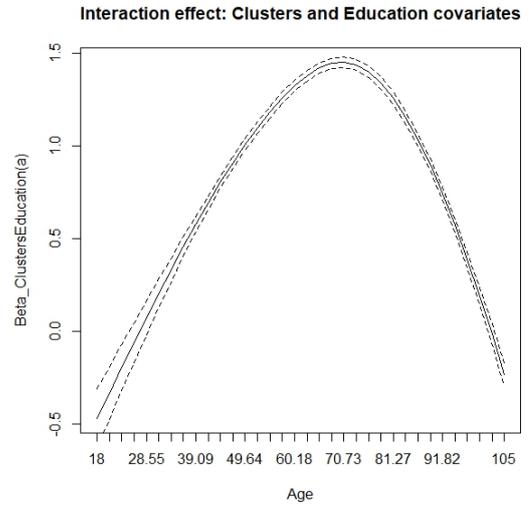


Figure 5.19: Effects of the interaction between Clusters and Education Covariates

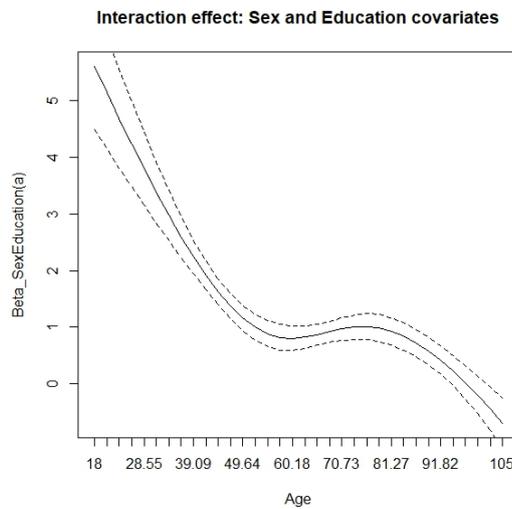


Figure 5.20: Effects of the interaction between Sex and Education Covariates

The effects of the interaction are still very variables. Only in the one that involves Sex and Education Covariates is possible to observe a decreasing pattern. Also for the second model, violations of the proportional hazards (PH) assumption are evident as confirmed from the result of the statistical test in table 5.10 . However, the considerations discussed for the first model remain valid. Since the primary objective of this work is the interpretation of mortality patterns, and being aware of the diagnostic outputs, we decided to proceed with the interpretation of the results, while carefully acknowledging

the limitations introduced by the assumption violations.

Covariate	Statistic value	Degree of freedom	P-value
Sex	272	1	10^{-16}
Clusters	18891	4	10^{-16}
EntryDay	29887	20	10^{-16}
Education	13880	4	10^{-16}
Sex:Clusters	9394	4	10^{-16}
Clusters:Education	28626	16	10^{-16}
Sex:Education	6765	4	10^{-16}
Global Model	63347	53	10^{-16}

Table 5.10: PH test for final model with Age as timescale

5.2.2. Choose the Numbers of the Degrees of freedom for the splines terms

Similarly, to determine an appropriate level of flexibility for the spline term modeling EntryDay, an intermediate model comprising only Sex, Cluster, and EntryDay as covariates, without interaction terms was fitted using spline bases with degrees of freedom ranging from 5 to 20 [26]. For each candidate model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were computed from the partial likelihood to assess the trade-off between model fit and complexity. This procedure helped identify a suitable range of degrees of freedom that avoided both underfitting and overfitting. Subsequently, to validate and refine the choice, k -fold cross-validation was performed, selecting the number of basis functions that maximized the average concordance index (C-index) across folds.

In figure 5.21 and 5.22 there are the resulting trends:

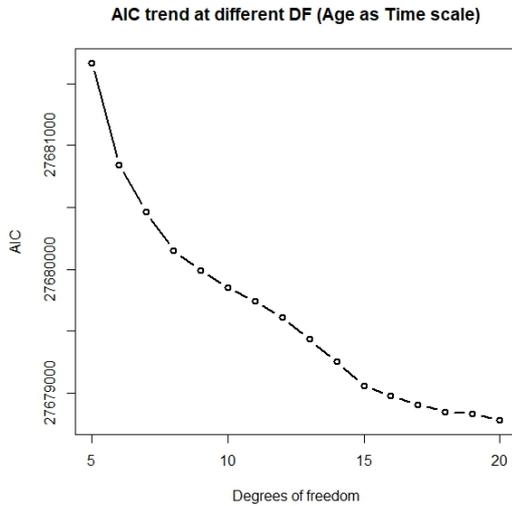


Figure 5.21: AIC trend with Age model

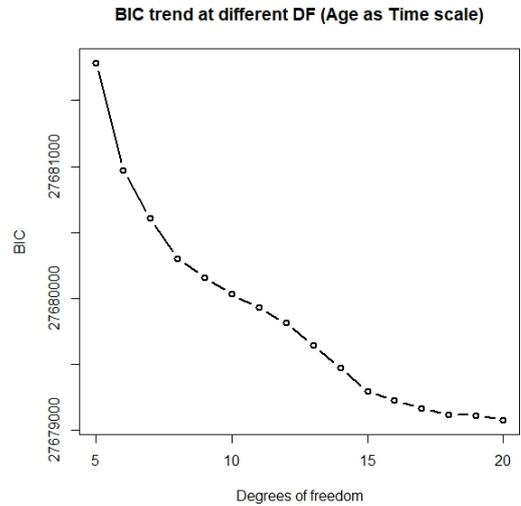


Figure 5.22: BIC trend with Age Model

In this case, both curves exhibit a clear pattern: they consistently decrease, reaching their minimum value at 20 degrees of freedom. To ensure robustness in the selection process, the model is re-evaluated across the same range of degrees of freedom using a Cross-Validation approach. The optimal choice remains 20 degrees of freedom, yielding a C-index of 0.65, indicating a moderate discriminative ability. The selected model has an AIC of 27678784 and BIC of 27679080.

5.2.3. Interpretation of the estimated HR

In the next tables, instead, the HRs estimates of the model with Age as timescale are shown for the same cases of the previous model. The cluster **Allergies (ALL)** is considered always as the reference and the same approach of the first estimated model is followed:

- **Long Education:**

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.327	1.289	1.366	1.521	1.477	1.567
CHL	0.658	0.632	0.685	0.665	0.639	0.695
DIA	0.968	0.924	1.014	0.977	0.931	1.025
M-P	2.790	2.711	2.872	2.569	2.495	2.645

Table 5.11: HR interval estimates for Long Education with Age as timescale

An important observation from the table is that, for the DIA cluster, the confidence interval includes 1 for both males and females. This indicates that the difference in mortality risk between DIA and the reference cluster is not statistically significant. From a clinical perspective, this suggests that individuals in the DIA cluster do not exhibit a markedly different mortality risk compared to those in the Allergies group, making the distinction between these clusters less pronounced in terms of survival outcomes.

- Rank for Males: M-P, CHC, ALL, DIA and CHL;
- Rank for Females: M-P, CHC, ALL, DIA and CHL;

In the same way of Table 5.2 the confidence intervals are wide, so additional comparisons are performed, without considering cluster Allergies as reference, to enhance reliability which **confirm** the ranking from Table 5.11:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P vs CHC	2.790	2.711	2.872	1.689	1.643	1.736
M-P vs CHL	4.239	4.077	4.407	3.855	3.704	4.013
M-P vs DIA	2.883	2.756	3.016	2.631	2.511	2.756
CHC vs CHL	2.016	1.939	2.096	2.283	2.193	2.376
CHC vs DIA	1.371	1.311	1.434	1.557	1.487	1.632
DIA vs CHL	1.470	1.394	1.551	1.466	1.387	1.548

Table 5.12: Additional Clusters comparison for Long Education

- **Medium Education:**

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.343	1.309	1.378	1.540	1.500	1.580
CHL	0.710	0.685	0.735	0.719	0.693	0.745
DIA	0.960	0.923	0.999	0.969	0.930	1.009
M-P	2.813	2.742	2.887	2.591	2.525	2.658

Table 5.13: HR interval estimates for Medium Education with Age as timescale

Similarly, the confidence interval for female subjects in the DIA cluster includes 1, indicating that it is not possible to determine whether this cluster poses a significantly higher mortality risk compared to ALL. For male subjects, the situation is at the threshold of statistical significance, suggesting a borderline distinction in mortality risk.

- Rank for Males: M-P, CHC, ALL, DIA and CHL;
- Rank for Females: M-P, CHC, ALL, DIA and CHL;

Also in this case other comparison are performed, without ALL as reference, giving consistency to the ranking from table 5.13:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P vs CHC	2.095	2.047	2.143	1.683	1.644	1.722
M-P vs CHL	3.963	3.833	4.097	3.605	3.485	3.728
M-P vs DIA	2.931	2.823	3.042	2.674	2.572	2.780
CHC vs CHL	1.892	1.830	1.956	2.142	2.071	2.216
CHC vs DIA	1.399	1.348	1.452	1.589	1.529	1.652
DIA vs CHL	1.352	1.293	1.414	1.348	1.288	1.411

Table 5.14: Additional Clusters comparison for Medium Education

- Short Education:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.220	1.205	1.235	1.399	1.381	1.417
CHL	0.661	0.649	0.672	0.669	0.657	0.681
DIA	0.897	0.881	0.913	0.905	0.888	0.923
M-P	2.602	2.569	2.635	2.396	2.366	2.426

Table 5.15: HR interval estimates for Short Education with Age as timescale

- Rank for Males: M-P, CHC, ALL, DIA and CHL;
- Rank for Females: M-P, CHC, ALL, DIA and CHL;

In this case, the confidence intervals are accurate indicating that the estimates are more reliable.

- None Education

Cluster	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
CHC	1.069	1.057	1.081	1.225	1.212	1.238
CHL	0.581	0.570	0.591	0.588	0.579	0.597
DIA	0.797	0.783	0.811	0.804	0.791	0.818
M-P	2.229	2.202	2.257	2.053	2.031	2.075

Table 5.16: HR interval estimates for None Education with Age as timescale

- Rank for Males: M-P, CHC, ALL, DIA and CHL;
- Rank for Females: M-P, CHC, ALL, DIA and CHL;

Also in this case, the confidence intervals are accurate: the estimates are still reliable.

Below is the table summarizing the cluster rankings by Sex and Education Level:

Education Level	Male	Female
Long	M-P, CHC, ALL, DIA, CHL	M-P, CHC, ALL, DIA, CHL
Medium	M-P, CHC, ALL, DIA, CHL	M-P, CHC, ALL, DIA, CHL
Short	M-P, CHC, ALL, DIA, CHL	M-P, CHC, ALL, DIA, CHL
None	M-P, CHC, ALL, DIA, CHL	M-P, CHC, ALL, DIA, CHL

Table 5.17: Summary of clusters by Sex and Education level of models with Age as timescale

The results obtained using the second model show a stable pattern across subgroups: the ranking of clusters is consistent across all educational levels and both sexes, with the M-P cluster emerging as the most hazardous and the CHL cluster as the least risky.

Difference between levels of Education

Also in this case is important to spot potential differences in mortality risk across educational levels in order to verify the results provided from Table 5.9 is consistent. The focus is still on M-P and CHC clusters since they results still the most two dangerous clusters and **Long Education** is still be considered as reference level:

Clusters	Males			Females		
	HR	CI Inf	CI Sup	HR	CI Inf	CI Sup
M-P Medium	1.011	0.984	1.039	1.037	1.007	1.068
M-P Short	1.180	1.155	1.207	1.187	1.158	1.216
M-P None	1.469	1.438	1.502	1.555	1.518	1.593
CHC Medium	1.015	0.990	1.041	1.041	1.008	1.076
CHC Short	1.164	1.141	1.188	1.170	1.139	1.202
CHC None	1.482	1.452	1.512	1.567	1.527	1.609

Table 5.18: Comparisons within Cluster but different Education Levels

This table confirms the previous findings: individuals with higher education levels exhibit higher life expectancy. This supports the intuition made for Table 5.9.

5.3. Discussion and possible ranking

5.3.1. First impression

Examining the results of the model using time of follow-up as the time scale, it is clear that, for male individuals, the M-P cluster consistently exhibits the highest risk, followed by CHC, ALL, DIA, and CHL. Although an exception is noted among subjects with no formal education where the DIA cluster proves more hazardous than ALL but remains less hazardous than M-P and CHC. The overall trend indicates that the M-P cluster is the most dangerous. For female subjects, the general pattern is similar; however, the CHC cluster appears as the most hazardous for them.

When considering the model with age as the time scale, the ranking becomes even more consistent across both sexes: M-P, CHC, ALL, DIA, CHL. This consistency reinforces the conclusion that the M-P cluster is the most hazardous, particularly since the model with age as the time scale provides a superior fit. In my opinion, the evidence strongly supports the view that the M-P cluster, which always appears at the top of the risk ranking, represents the greatest threat.

Both models also reveal significant differences in risk across educational levels. Specifically, when focusing on the two most hazardous clusters (M-P and CHC), there is little change in risk between Medium and Long education. However, a marked increase in risk is observed when moving from Long education to Short (or None) education. These findings suggest that higher educational attainment likely linked to improved socioeconomic conditions contributes to better health outcomes through healthier lifestyle choices, increased self-care, and enhanced access to medical screenings. Conversely, individuals with lower educational attainment may experience poorer living standards and a greater likelihood of underdiagnosed conditions, thereby increasing their overall health risks.

5.3.2. Focus on M-P cluster

The M-P cluster presents a particularly high mortality risk due to the coexistence of severe **physical and mental health** conditions. The combination of diseases affecting both body and mind such as osteoporosis, osteoarthritis, depression, and schizophrenia results in a profound **deterioration** in quality of life and life expectancy. Moreover, the presence

of hypertension as a **silent killer** further exacerbates mortality risks. Understanding this cluster's vulnerability requires an integrated approach that considers not only biomedical factors but also lifestyle influences, healthcare accessibility, and societal attitudes toward mental health.

The Role of Mental Health in the M-P Cluster

Mental health conditions are a defining characteristic of the M-P cluster, making it particularly high-risk. Depression and schizophrenia can significantly impact a patient's ability to manage chronic physical illnesses, adhere to treatments, and maintain a healthy lifestyle. In Denmark, mental health disorders contribute significantly to disease burden and mortality rates (see the articles [27] and [28]). Several factors exacerbate these risks:

- **Underreporting and late diagnosis**, especially among men who are less likely to seek mental health care, leading to delayed treatment and increased suicide risk.
- **Mental disorders reduce mobility and increase frailty**: Depression is associated with reduced physical activity, which accelerates osteoporosis and cardiovascular disease.
- **The stigma surrounding psychiatric conditions**: Although Denmark has made progress in mental health awareness, stigma still prevents adequate screening and interventions.

Lifestyle Factors in Denmark

Denmark is often recognized for its strong **work-life balance** policies and active lifestyle culture. However, certain **paradoxes**, supported by [29] and [30], contribute to health risks for individuals in the M-P cluster:

- **Sedentary Work**: While many Danes engage in physical activities, a growing percentage of the workforce is involved in prolonged sedentary work, which exacerbates osteoporosis, hypertension, and cardiovascular disease.
- **Social Isolation and Mental Health Decline**: Patients suffering from depression and schizophrenia often experience **social withdrawal**, leading to **increased physical inactivity**, which in turn accelerates frailty and mortality risks.
- **Seasonal Affective Disorder (SAD) and Its Impact on Chronic Conditions**: The long Danish winters contribute to seasonal depression, which can further reduce physical activity levels and health monitoring behaviors.

Hypertension: The Silent Killer in the M-P Cluster

Among the most critical risk factors for mortality in the M-P cluster is hypertension, which is often **underdiagnosed** in individuals with mental health conditions respect to the ones with diabetes who are more likely to go on regular check-ups as suggested from [31]. Several mechanisms explain why hypertension becomes particularly lethal in this group:

- **Mental illnesses interfere with blood pressure regulation:** Depression and anxiety are associated with increased **blood pressure variability**, making hypertension harder to control.
- **Hypertension is less likely to be detected in individuals with psychiatric disorders:** Patients presenting with depression or osteoporosis might not undergo routine cardiovascular screening, leading to missed diagnoses and inadequate treatment.
- **Delayed diagnosis increases the risk of cardiovascular events:** Without proper monitoring, hypertension can progress silently, increasing the likelihood of strokes, heart attacks, and organ failure.

The Need for Integrated Healthcare Approaches

Addressing the mortality risks in the M-P cluster requires **multidisciplinary** interventions. Some key strategies include:

- **Improved Screening for Mental and Cardiovascular Diseases:** Routine health check-ups should include mental health assessments, particularly for older patients and men who are less likely to seek help.
- **Promoting Active Lifestyles:** Encouraging physical activity especially among individuals with osteoporosis and depression can help mitigate mobility loss and cardiovascular risks.
- **Enhancing Mental Health Awareness and Reducing Stigma:** Strengthening community support and awareness programs can help improve early diagnosis and treatment adherence.
- **Increasing Awareness of Hypertension in High-Risk Groups:** Healthcare professionals should emphasize the importance of regular blood pressure monitoring, particularly for patients with coexisting mental health disorders, where hypertension is often overlooked.

5.3.3. Interpreting the Relationship Between ALL and DIA Clusters

The relationship between the Allergies (ALL) and Diabetes (DIA) clusters exhibits an interesting pattern that varies depending on the chosen time scale. These findings provide insights into how multimorbidity and education levels influence mortality risk.

Follow-Up Time Model: DIA More Hazardous than ALL

- DIA is associated with a higher mortality risk than ALL;
- The mortality risk between the two clusters is significantly different across all levels of education;
- The risk difference is stronger in individuals with higher education (20%) and lower (10-15%) in those with less education;
- Possible explanations:
 - **Earlier diagnosis and better management:** Individuals with higher education are more likely to undergo regular medical check-ups, leading to longer survival with diabetes but also more recorded complications;
 - **Delayed detection in lower education groups:** Some individuals might die before diabetes is diagnosed, leading to a potential underestimation of its true risk in this group.

Age-Based Model: ALL More Hazardous than DIA

- ALL appears more hazardous than DIA;
- Subjects with long education and females with medium education do not exhibit a significant difference in mortality risk;
- The difference is minimal in highly educated individuals (<5%) but increases (20%) in those without formal education.
- Possible explanations:
 - **Chronic inflammation and respiratory conditions in the ALL cluster:** Some allergic conditions (e.g., asthma, autoimmune diseases) may increase long-term mortality risk, especially if poorly managed.
 - **Healthcare disparities:** Lower-educated individuals may face greater barriers

ers to accessing effective treatment, increasing long-term risks.

Medical and Public Health Implications

- **Diabetes presents an immediate mortality risk**, particularly in higher-educated individuals, likely due to longer survival with complications;
- **Allergic conditions may contribute to long-term mortality**, especially in lower-educated groups, emphasizing the need for improved management of chronic inflammation and respiratory diseases;
- **Education influences mortality risk differently for each cluster**, highlighting disparities in early diagnosis, treatment access, and healthcare engagement;
- **The choice of time scale (Follow-Up vs. Age) alters the perceived severity of disease clusters**, reinforcing the importance of studying both short-term and long-term multimorbidity outcomes.

According to the study [32] allergic conditions and type 1 diabetes may interact through the immune system. The research suggests that allergic diseases activate a different immune response pathway than type 1 diabetes, potentially reducing the likelihood of both conditions occurring in the same individual. However, this does not imply a direct prevention mechanism, but rather an observed inverse association between the two diseases.

From a healthcare perspective, managing these conditions remains essential to prevent complications. Even if allergies and diabetes show opposing immune patterns, chronic inflammation from poorly managed allergies could still contribute to overall health risks, just as diabetes-related immune dysfunction might affect allergic reactions. These insights highlight the need for early diagnosis and tailored healthcare strategies to minimize long-term risks.

5.3.4. Impact of Time Scale Choice: Time on Study vs. Age

Performance Comparison: Favoring the Age-Based Model

Among the two time scale choices, performance metrics strongly favor the age-based model, as indicated by significantly lower AIC and BIC values. This suggests that modeling survival based on biological aging provides a more accurate representation of disease progression and mortality risk.

Strengths of the Age-Based Model

- **Age as a Key Factor in Multimorbidity Research:** Given that age is a fundamental determinant of health, using it as a time scale aligns naturally with the study's goal of understanding long-term health trajectories. Chronic diseases tend to progress over time, and the risk of mortality is inherently age-dependent.
- **Incorporating Left Truncation:** By applying left truncation the comparisons are made only between subjects of similar age. This prevents bias that could arise from excluding older individuals who have survived longer despite multimorbidity.
- **Capturing Long-Term Effects:** Since the model aligns risk assessment with aging, it is particularly useful for studying long-term disease progression, rather than just the duration of follow-up in the study.

Limitations of the Age-Based Model

- **Variation in Multimorbidity Duration:** A major limitation of this approach is that it does not differentiate between individuals who have been multimorbid for years and those who became multimorbid just a month ago at the same age. This could lead to potential biases, as the duration of multimorbidity might influence mortality risk independently of chronological age.

Advantages of the Time-on-Study Model

- **Standardized Study Entry:** This model ensures that all subjects enter the study at the same time, allowing for a more controlled comparison of survival outcomes, particularly when evaluating the impact of external factors such as healthcare interventions or policy changes.

Limitations of the Time-on-Study Model

- **Potential Oversimplification of Biological Aging:** In this approach, EntryAge is modeled as a covariate, meaning that when comparing patients, the assumption is that EntryAge's effect on mortality is fully captured by the model. However, this could lead to underestimation of biological aging effects, which are crucial in long-term survival analysis.

Balancing Both Approaches

Although the age-based model is a more appropriate choice in this study, it is essential to consider how long subjects have been multimorbid. Without this consideration, there is a risk of assuming an unrealistic uniformity among individuals of the same age. Future analyses should integrate multimorbidity duration to refine risk estimates and enhance clinical applicability.

5.3.5. Limitations of the Cox Models

In both the estimated model the PH assumption is violated, so is important not rely solely on the interval estimates. This suggests that:

- The Cox model captures the global effect of cluster over the full observation period, but it is not sufficient to explain how risk evolves over time.
- The impact of multimorbidity clusters on mortality likely changes dynamically, meaning that a confidence interval does not fully describe the varying risk.
- The observed PH violation could be due to underlying transitions between disease clusters over time, which the Cox model does not explicitly capture.

Possible solutions to solve these critical point

While the Cox model provides valuable insights into overall survival trends, it is not fully adequate for explaining the dynamic changes in mortality risk associated with multimorbidity progression. This limitation arises because:

- Although time-dependent covariates and interactions capture some variability, they do not track individual transitions between multimorbidity clusters.
- the estimated HRs represents an average effect over the study period, rather than explicitly modeling how risk evolves at different time points.

Given these constraints, **Multi-State Models (MSMs)** offer a more suitable approach. MSMs allow for:

- Explicit modeling of transitions between multimorbidity clusters, rather than assuming static cluster membership.
- Dynamic estimation of hazard ratios, which change as individuals move through different disease states.
- A better representation of chronic disease evolution, making them more appropriate

for studying multimorbidity.

5.3.6. Proposal of a Mortality Risk Hierarchy Among Clusters

Based on the results and their interpretation, the most appropriate ranking of clusters appears to be: M-P, CHC, ALL, DIA, and CHL, with the caveat that it is not possible to precisely determine whether ALL or DIA represents a higher mortality risk. The M-P cluster consistently exhibits the highest risk scores in almost all cases, except in the model using follow-up time as the timescale and focusing on female individuals. The CHC cluster is generally ranked second and, as previously observed in the exploratory data analysis, includes the highest number of deceased individuals. The CHL cluster is consistently identified as the least risky across all scenarios and models. Finally, the relative risk between ALL and DIA appears to depend on the chosen time scale, suggesting that their interpretation is sensitive to the modeling framework.

5.4. The dynamics within each Education level

To assess whether the level of education influences cluster transition dynamics and mortality, we computed relative frequencies that describe the relative frequencies of moving between clusters or reaching the "DIED" state.

Let us define the state space as $S = \{\text{ALL}, \text{CHC}, \text{CHL}, \text{DIA}, \text{M-P}, \text{DIED}\}$, where each row represents the current cluster (e.g., ALL, CHC, etc.), while each column corresponds to the next cluster or the "DIED" state. This structure enables the analysis of individuals' likelihood of transitioning between different multimorbidity clusters or progressing to mortality, offering insights into the role of education in these dynamics.

- Long Education:

$$P = \begin{bmatrix} 0.000 & 0.187 & 0.388 & 0.093 & 0.005 & 0.327 \\ 0.117 & 0.000 & 0.000 & 0.000 & 0.080 & 0.803 \\ 0.076 & 0.397 & 0.000 & 0.254 & 0.005 & 0.268 \\ 0.111 & 0.426 & 0.000 & 0.000 & 0.081 & 0.382 \\ 0.221 & 0.139 & 0.274 & 0.054 & 0.000 & 0.312 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

- Medium Education:

$$P = \begin{bmatrix} 0.000 & 0.195 & 0.384 & 0.116 & 0.005 & 0.300 \\ 0.108 & 0.000 & 0.000 & 0.000 & 0.087 & 0.805 \\ 0.076 & 0.382 & 0.000 & 0.281 & 0.006 & 0.255 \\ 0.116 & 0.444 & 0.000 & 0.000 & 0.085 & 0.355 \\ 0.228 & 0.139 & 0.279 & 0.062 & 0.000 & 0.292 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

- Short Education:

$$P = \begin{bmatrix} 0.000 & 0.187 & 0.364 & 0.119 & 0.004 & 0.325 \\ 0.095 & 0.000 & 0.000 & 0.000 & 0.079 & 0.826 \\ 0.065 & 0.366 & 0.000 & 0.293 & 0.006 & 0.269 \\ 0.096 & 0.439 & 0.000 & 0.000 & 0.085 & 0.380 \\ 0.196 & 0.145 & 0.278 & 0.070 & 0.000 & 0.311 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

- None Education:

$$P = \begin{bmatrix} 0.000 & 0.197 & 0.252 & 0.083 & 0.006 & 0.462 \\ 0.082 & 0.000 & 0.000 & 0.000 & 0.075 & 0.843 \\ 0.043 & 0.365 & 0.000 & 0.249 & 0.007 & 0.336 \\ 0.081 & 0.426 & 0.000 & 0.000 & 0.083 & 0.411 \\ 0.147 & 0.179 & 0.206 & 0.062 & 0.000 & 0.406 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

The relative frequencies corresponding to different education levels do not reveal significant differences in the probabilities of transitioning from one cluster to another. This suggests that an individual's disease trajectory is primarily influenced by biological factors rather than social or lifestyle determinants. However, a significant increase in frequencies of death is observed when moving from a medium or high education level to a low or no education level. This finding is particularly relevant as it further highlights the strong association between low education and poor lifestyle choices factors that contribute to an accelerated mortality process.

5.5. High Mortality in CHC Does Not Equate to the Highest Risk

In our analysis, survival models consistently identify the M-P cluster as the most dangerous, indicating that patients in this group experience the highest relative hazard risk. This suggests that the psychophysical conditions characterizing the M-P cluster lead to a rapid deterioration in health outcomes. In contrast, the transition frequency matrices computed in Exploratory Data Analysis and during the investigation within each educationlevel, reveal that the CHC cluster exhibits the highest proportion of transitions to death. This apparent discrepancy likely reflects differences in the underlying population characteristics: the CHC cluster is predominantly composed of older individuals, which naturally increases the observed frequency of death, whereas the M-P cluster, despite a lower absolute number of deaths, shows a more pronounced impact on survival when evaluated in relative terms. In summary, although the CHC cluster shows the highest frequency of transitions to death due to its older demographic, the M-P cluster remains the most hazardous overall because of its severe influence on the quality of life and accelerated health decline. This integrated perspective underscores the importance of targeted interventions for patients in the M-P cluster.

6 | Conclusions and future developments

This study contributes to the understanding of multimorbidity and its impact on mortality by applying Survival Analysis on the targeted population. The findings highlight that the Musculoskeletal and Psychiatric Conditions (M-P) cluster carries the highest mortality risk, despite the Chronic Heart Conditions (CHC) cluster exhibiting the highest transition frequency to death. This suggests that while CHC patients are generally older and more likely to die in the short term, M-P conditions significantly deteriorate daily life quality, potentially leading to increased long-term mortality risk.

Additionally, the study reveals the relationship between education and mortality risk: individuals with higher educational attainment exhibit lower hazard ratios. This may be explained by greater access to healthcare, more frequent medical check-ups, and better disease awareness, leading to earlier diagnoses of severe conditions that ultimately influence survival statistics. This finding underscores the importance of considering healthcare access and diagnostic practices when interpreting mortality risk across socioeconomic groups.

By adopting age as the time scale, the survival models provide a clearer risk stratification than models using follow-up time, reinforcing the importance of considering both biological aging and disease progression in multimorbidity risk assessment.

Looking ahead, future research could benefit from multistate models, which would allow for a dynamic analysis of transitions between multimorbidity clusters over time. By modeling the progression of chronic conditions before mortality, these models can offer deeper insights into which transitions carry the highest risk and how different patient subgroups evolve within the healthcare system.

Ultimately, these findings emphasize the need for integrated healthcare approaches focused on early diagnosis, lifestyle interventions, and improved accessibility to medical care. Advanced statistical techniques can assume an important role for future studies to continue exploring multimorbidity patterns and evaluate their impact on mortality within

the Danish population.

Bibliography

- [1] C. Boyd and M. Fortin, “Future of multimorbidity research: how should understanding of multimorbidity inform health system design?” *Public Health Reviews*, 2010. [Online]. Available: <https://doi.org/10.1007/BF03391611>
- [2] K. Barnett, S. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, “Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study.” *The Lancet*, 2012. [Online]. Available: [https://doi.org/10.1016/S0140-6736\(12\)60240-2](https://doi.org/10.1016/S0140-6736(12)60240-2)
- [3] C. Violan, Q. Foguet-Boreu, G. Flores-Mateo, C. Salisbury, J. Blom, and M. Freitag, “Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies.” *PLOS ONE*, 2014. [Online]. Available: <https://doi.org/10.1371/journal.pone.0102149>
- [4] L. Thygesen and A. Ersbøll, “When the entire population is the sample: strengths and limitations in register-based epidemiology.” *European Journal of Epidemiology*, 2014. [Online]. Available: <https://doi.org/10.1007/s10654-013-9873-0>
- [5] N. N. Holm, H. D. Annw Frølich, K. Dalhoff, H. G. Juul-Larsen, O. Andersen, and A. Stockmarr, “Co-occurring diseases and mortality in patients with chronic heart disease modeling their dynamically expanding disease portfolio: A nationwide register study,” unpublished. [Online]. Available: <https://preprints.jmir.org/preprint/57749>
- [6] A. Stockmarr and A. Frølich, “Clusters from chronic conditions in the danish adult population,” *PLOS ONE*, 2024. [Online]. Available: <https://doi.org/10.1371/journal.pone.0302535>
- [7] J. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 1979, 28(1), 100–108. [Online]. Available: <https://doi.org/10.2307/2346830>
- [8] D. Ketchen and C. Shook, “The application of cluster analysis in strategic management research: an analysis and critique.” *Strategic management journal*, Tech. Rep., 1996, 17(6), 441–458.

- [9] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis.” *Communications in Statistics*, Tech. Rep., 1974, 3(1):1–27.
- [10] R. P. Silhouettes, “A graphical aid to the interpretation and validation of cluster analysis,” *Computational and Applied Mathematics*, 1987, 20: 53–65. [Online]. Available: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [11] E. Lynge, J. Sandegaard, and M. Rebolj, “The danish national patient register,” *Scan J Public Health*, 2011, 39 (Suppl 7), 30–3. [Online]. Available: <https://doi.org/10.1177/1403494811401482>
- [12] O. Mors, G. Perto, and P. Mortensen, “The danish psychiatric central research register,” *Scan J Public Health*, 2011, 39 (Suppl 7), 54–57. [Online]. Available: <https://doi.org/10.1177/1403494810395825>
- [13] H. Kildemoes, H. Sørensen, and J. Hallas, “The danish national prescription registry,” *Scan J Public Health*, 2011, 39 (7 Suppl), 38–41. [Online]. Available: <https://doi.org/10.1177/1403494810394717>
- [14] N. Olivarius, H. Hollnagel, A. Krasnik, P. Pedersen, and H. Thorsen, “The danish national health service register. a tool for primary health care research,” *Danish medical bulletin*, 1997, 44(4):449–53. [Online]. Available: PMID:9377908
- [15] M. L. Schiøtz, A. Stockmarr, D. Høst, C. Glümer, and A. Frølich, “Social disparities in the prevalence of multimorbidity – a register-based population study,” *BMC Public Health*, 2017. [Online]. Available: <https://doi.org/10.1186/s12889-017-4314-8>
- [16] K. M. Robinson, C. L. Lau, M. Jeppesen, A. B. Vind, and C. Glümer, “Kroniske sygdomme—hvordan opgøres kroniske sygdomme (chronic conditions—how to assess chronic conditions),” Region Hovedstaden, Research Centre for Prevention and Health, Glostrup, Denmark, Tech. Rep., 2011, evaluation and Analysis Model Project under the Chronic Disease Programme.
- [17] A. Pini, “Permutation tests for univariate and multivariate data,” Politecnico di Milano, Tech. Rep., 2018.
- [18] H. Mann and D. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Mathematical Statistics*, 1947, 18(1), 50-60. [Online]. Available: <http://dx.doi.org/10.1214/aoms/1177730491>
- [19] D. G. Kleinbaum and M. Klein, *Survival Analysis, A Self-Learning Text*, Third Edition. Springer, 2012. [Online]. Available: <https://link.springer.com/book/10.1007/978-1-4419-6646-9>

- [20] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000. [Online]. Available: 10.1007/978-1-4757-3294-8
- [21] D. Schoenfeld, “Partial residuals for the proportional hazards regression model,” *Biometrika*, 1982, 69(1): 239–241. [Online]. Available: <https://doi.org/10.1093/biomet/69.1.239>
- [22] P. Grambsch and T. Therneau, “Proportional hazards tests and diagnostics based on weighted residuals,” *Biometrika*, 1994, 81(3): 515–526. [Online]. Available: <https://doi.org/10.1093/biomet/81.3.515>
- [23] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, ser. Statistics for Biology and Health. Springer, 2000. [Online]. Available: <https://link.springer.com/book/10.1007/978-1-4757-3294-8>
- [24] N. Shiff, A. Voineskos, M. Ferro, K. Bennett, A. Levinson, and P. Kurdyak, “Physical–mental multimorbidity in youth: Prevalence, clinical burden, and risk for mental illness,” *The Canadian Journal of Psychiatry*, 2024. [Online]. Available: <https://doi.org/10.1177/07067437241226998>
- [25] M. Leppänen, P. Hovi, E. Kajantie, J. Koponen, M. Lahti, M. Ojaniemi, T. Tammelin, T. Tikanmäki, K. Heinonen, R. Pyhälä, K. Räikkönen, S. Andersson, S. Heinonen, M. Gissler, and P. Hovi, “Prematurity and multimorbidity in adolescence and early adulthood: A nationwide cohort study,” *PLOS ONE*, 2022. [Online]. Available: <https://doi.org/10.1371/journal.pone.0261952>
- [26] F. J. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed., ser. Springer Series in Statistics. Cham: Springer, 2015. [Online]. Available: <https://link.springer.com/book/10.1007/978-3-319-19425-7>
- [27] P. Kugathasan, B. Stubbs, J. Aagaard, S. Jensen, T. M. Laursen, and R. Nielsen, “Increased mortality from somatic multimorbidity in patients with schizophrenia: a danish nationwide cohort study,” *ACTA PSYCHIATRICA SCANDINAVICA*, 2029. [Online]. Available: <https://doi.org/10.1111/acps.13076>
- [28] T. Willadsen, V. Siersma, D. Nicolaisdóttir, R. Køster-Rasmussen, D. Jarbøl, S. Reventlow, M. S.W, and N. de Fine Olivarius, “Multimorbidity and mortality: A 15-year longitudinal registry-based nationwide danish population study,” *Journal of Comorbidity*, 2018, doi: 10.1177/2235042X18804063. PMID: 30364387; PMCID: PMC6194940. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6194940/>

- [29] K. Overgaard, P. G. Aagaard, A. Grøntved, K. Nielsen, I. K. Dahl-Petersen, and M. Aadahl, “Sedentary behaviour in Denmark is growing and is a possible independent risk factor,” *Ugeskrift for Læger*, 2013, article in Danish. Accessing the full text may require institutional credentials or a ResearchGate account. PMID: 24629196. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/24629196/>
- [30] K. Laugesen, L. M. Baggesen, S. A. J. Schmidt, M. M. Glymour, M. Lasgaard, A. Milstein, H. T. Sørensen, N. E. Adler, and V. Ehrenstein, “Social isolation and all-cause mortality: a population-based cohort study in Denmark,” *Scientific Reports*, 2018, 10.1038/s41598-018-22963-w. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/29549355/>
- [31] T. Sehestedt, H. Ibsen, and T. Jørgensen, “Awareness, treatment and control of hypertension in Denmark. The Inter99 study,” *Blood Pressure*, 2009. [Online]. Available: <https://doi.org/10.1080/08037050701428307>
- [32] K. Engkilde, T. Menné, and J. D. Johansen, “Inverse relationship between allergic contact dermatitis and type 1 diabetes mellitus: a retrospective clinic-based study,” *Diabetologia*, 2006. [Online]. Available: <https://doi.org/10.1007/s00125-006-0162-2>
- [33] N. N. Holm, A. Frølich, O. Andersen, H. G. Juul-Larsen, and A. Stockmarr, “Longitudinal models for the progression of disease portfolios in a nationwide chronic heart disease population,” *PLOS ONE*, 2023. [Online]. Available: <https://doi.org/10.1371/journal.pone.0284496>
- [34] A. Frølich, N. Ghith, M. Schiøtz, R. Jacobsen, and A. Stockmarr, “Multimorbidity, healthcare utilization and socioeconomic status: A register-based study in denmark,” *PLOS ONE*, 2019. [Online]. Available: <https://doi.org/10.1371/journal.pone.0214183>
- [35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R, Second Edition*. Springer, 2021. [Online]. Available: <https://www.statlearning.com/>
- [36] O. Aalen, “Nonparametric estimation of partial transition probabilities in multiple decrement models,” *The Annals of Statistics*, 1978. [Online]. Available: <https://doi.org/10.1214/aos/1176344198>
- [37] O. Aalen, Ørnulf Borgan, and H. K. Gjessing, *Survival and Event history Analysis: A Process Point of View*. Springer, 2008. [Online]. Available: <https://link.springer.com/book/10.1007/978-0-387-68560-1>
- [38] M. Bland, *An Introduction to Medical Statistics - 4th Edition*. Oxford

- University Press, 2015. [Online]. Available: <https://global.oup.com/academic/product/introduction-to-medical-statistics-9780199589920?cc=us&lang=en&>
- [39] M. Greenwood, “The natural duration of cancer,” *Reports on Public Health and Medical. Her Majesty’s Stationery Office*, 1926. [Online]. Available: <https://doi.org/10.1136/bmj.2.3320.266>
- [40] D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data 2nd Edition*. Wiley-Interscience, 2008. [Online]. Available: <https://www.wiley.com/en-us/Applied+Survival+Analysis%3A+Regression+Modeling+of+Time-to-Event+Data%2C+2nd+Edition-p-9780471754992>
- [41] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data 2nd Edition*. Wiley, 2002. [Online]. Available: <https://www.wiley.com/en-us/The+Statistical+Analysis+of+Failure+Time+Data%2C+2nd+Edition-p-9781118032985>
- [42] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *Journal of American Statistical Association*, 1958. [Online]. Available: <https://doi.org/10.2307/2281868>
- [43] N. Mantel, “Evaluation of survival data and two new rank order statistics arising in its consideration,” *Cancer Chemotherapy Reports*, 1966, pMID: 5910392. The article is not available. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/5910392/>
- [44] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society*, 1972. [Online]. Available: <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- [45] ———, “Partial likelihood,” *Biometrika*, 1975. [Online]. Available: <https://doi.org/10.2307/2335362>
- [46] F. Pesarin and L. Salmaso, *Permutation Tests for Complex Data*. Wiley, 2010. [Online]. Available: <https://onlinelibrary.wiley.com/doi/book/10.1002/9780470689516>
- [47] OpenAI, “Chatgpt,” 2024/2025, large language model. [Online]. Available: <https://openai.com>
- [48] P. F. Antonietti, S. Bonetti, A. Gruttadauria, G. Mescolini, and A. Zingaro, “A LaTeX template for MSc Thesis submissions to Politecnico di Milano (PoliMi) - School of Industrial and Information Engineering,” 2021. [Online]. Available: <https://it.overleaf.com/latex/templates/classical-format-thesis-scuola-di-ingegneria-industriale-e-dellinformazione-politecnico-di-milano-dkmvtndqkyxg>

A | Appendix A

In this appendix, we provide additional details on selected portions of the code used in the analysis.

A.1. Exploratory Data Analysis

A.1.1. Permutational t-test

```

1 perm_t_test=function(x,y,iter){
2   T0=abs(mean(x)-mean(y)) # define the test statistic
3   T_stat=numeric(iter) # a vector to store the values of each iteration
4   x_pooled=c(x,y) # pooled sample
5   n=length(x_pooled)
6   n1=length(x)
7   pb=progress::progress_bar$new(total=B, format = " Processing [:bar] :
      percent eta: :eta")
8   set.seed(seed)
9   for(perm in 1:iter){ # loop for conditional MC
10    # permutation:
11    permutation <- sample(1:n)
12    x_perm <- x_pooled[permutation]
13    x1_perm <- x_perm[1:n1]
14    x2_perm <- x_perm[(n1+1):n]
15    # test statistic:
16    T_stat[perm] <- abs(mean(x1_perm) - mean(x2_perm))
17    pb$tick()
18  }
19
20  # p-value
21  p_val <- sum(T_stat>=T0)/iter
22  return(p_val)
23 }
24 x.1 <- age_male
25 x.2 <- age_female
26 B <- 1000

```

```
27 p.value <- perm_t_test(x.1,x.2,iter=B)
```

A.1.2. Transition Matrix construction

```
1 Transitions <- All_patients %>%
2   group_by(SubjectID) %>%
3   mutate (NextCluster = ifelse(!is.na(lead(Clusters)),lead
4     (Clusters),
5     ifelse(Event == 1, "DIED", "CENSORED")))%>%
6   filter(Clusters!= NextCluster & NextCluster != "CENSORED
7     ")
8
9 Transitions_table <- Transitions%>%
10  ungroup()%>%
11  count(Clusters, NextCluster)%>%
12  group_by(Clusters)%>%
13  mutate(Probability = n/sum(n))
14
15 Transition_Matrix <- Tranbsition_table %>%
16  pivot_wider(names_from = NextCluster, values_from
17    = Probability, values_fill = 0) %>%
18  column_to_rownames("Clusters")%>%
19  as.matrix()
20
21 Transition_Matrix <- as.data.frame(Transition_Matrix)
22 colnames(Transition_Matrix) <- c("ALL", "CHC", "CHL", "DIA", "M-P", "
23   DIED")
24 Transition_Matrix[6,] <- c(0,0,0,0,0,1)
25 rownames(Transition_Matrix) <- colnames(Transition_Matrix)
```

A.2. Model Construction

A.2.1. Cross-Validation to select the best number of degree of freedom

```
1 dfs <- c(5:11) # Degrees of freedom to test
2 c_index_per_df <- numeric(length(dfs))
3 n_iterations <- 100
4 unique_ids <- unique(dataset$SubjectID)
```

```
5 pb = progress::progress_bar$new(total = length(dfs),
6                               format = "Processing [:bar]
7                               :percent eta: :eta")
8 for (df in dfs) {
9   c_index_values <- numeric(n_iterations)
10
11  for (i in 1:n_iterations) {
12    # Split data
13    set.seed(2025)
14    train_ids <- sample(unique_ids, size = 0.8 * length(unique_ids))
15    train_data <- dataset[dataset$SubjectID %in% train_ids, ]
16    test_data <- dataset[!dataset$SubjectID %in% train_ids, ]
17
18    # Fit the model with a specific df
19    model <- coxph(Surv(Time_init, Time_end, Event) ~ Sex + Clusters +
20                  ns(EntryAge, df = df), data = train_data)
21
22    # Predict risk scores
23    risk_scores <- predict(cox_model, newdata = test_data, type = "lp")
24
25    # Calculate the C-index
26    c_index <- survConcordance(Surv(Time_init, Time_end, Event) ~ risk_
27                              scores)
28
29    # Store the C-index
30    c_index_values[i] <- c_index$c.index
31  }
32
33  # Average C-index for this df
34  c_index_per_df[df - 4] <- mean(c_index_values)
35 }
36
37 # Find the best df
38 best_df <- dfs[which.max(c_index_per_df)]
39 print(paste("Best df:", best_df))
```


List of Figures

4.1	Entry Age of the study population	30
4.2	Age distribution of individuals when the event happens	33
4.3	Kaplan-Meier curves with Optimal cut point of EntryAge range	38
4.4	Kaplan-Meier curves with EntryAge range cut by hand	39
4.5	Kaplan-Meier curves with interaction between EntryAge and Sex	39
4.6	Survival curves with EntryAge Greater Than or Equal to 20	41
4.7	Survival curves with EntryAge Greater Than or Equal to 30	42
4.8	Survival curves with EntryAge Greater Than or Equal to 50	43
4.9	Survival curves with EntryAge Greater Than or Equal to 80	44
5.1	Residuals for Cluster CHC Covariate	52
5.2	Residuals for the interaction Female and None Education Covariate	52
5.3	Sex covariate effect	53
5.4	Clusters covariate effects	53
5.5	EntryAge covariate effects	53
5.6	Education covariate effects	53
5.7	Effects of the interaction between Sex and Clusters Covariates	54
5.8	Effects of the interaction between Clusters and Education Covariates	54
5.9	Effects of the interaction between Sex and Education Covariates	54
5.10	AIC trend with Time on study model	58
5.11	BIC trend with Time on study model	58
5.12	Residuals for M-P Cluster Covariate	65
5.13	Residuals for interaction Sex Female and Medium Education Covariate	65
5.14	Effects of Sex Covariate	66
5.15	Effects of Clusters Covariate	66
5.16	Effects of EntryDay Covariate	66
5.17	Effects of Education Covariate	66
5.18	Effects of the interaction between Clusters and Sex Covariates	67
5.19	Effects of the interaction between Clusters and Education Covariates	67
5.20	Effects of the interaction between Sex and Education Covariates	67

5.21 AIC trend with Age model	69
5.22 BIC trend with Age Model	69

List of Tables

1.1	Chronic disease presence (in %) in each identified cluster	5
1.2	Sociodemographic info in each identified cluster	5
1.3	Education info in each identified cluster	6
4.1	Distribution of subjects across clusters when they become multimorbid . . .	31
4.2	Ranking of diseases prevalence when individuals become multimorbid . . .	32
4.3	Sex Distribution and Mean Entry Age	32
4.4	Distribution of subjects across clusters when the event happens	34
4.5	Ranking of diseases prevalence when individuals die	34
4.6	Sex distribution across the clusters	35
4.7	Distribution of subjects across clusters when the censoring happens	35
4.8	Ranking of diseases prevalence when individuals are censored	36
4.9	Sex distribution and average age at the censoring of subjects	36
4.10	p -values from the Log-Rank test for each case	40
4.11	Most frequent observed trajectories	45
4.12	Most frequent observed trajectories	46
5.1	PH test for final model with Time on study as timescale	55
5.2	HR interval estimates for Long Education	60
5.3	Additional Clusters comparison within Long Education	60
5.4	HR interval estimates for Medium Education	61
5.5	Additional Clusters comparison within Medium Education	61
5.6	HR interval estimates for Short Education	62
5.7	HR interval estimates for None Education	62
5.8	Summary of clusters by Sex and Education level	63
5.9	Comparisons within Cluster but different Education Levels	63
5.10	PH test for final model with Age as timescale	68
5.11	HR interval estimates for Long Education with Age as timescale	70
5.12	Additional Clusters comparison for Long Education	71
5.13	HR interval estimates for Medium Education with Age as timescale	71

5.14	Additional Clusters comparison for Medium Education	72
5.15	HR interval estimates for Short Education with Age as timescale	72
5.16	HR interval estimates for None Education with Age as timescale	72
5.17	Summary of clusters by Sex and Education level of models with Age as timescale	73
5.18	Comparisons within Cluster but different Education Levels	73

Acknowledgements

Francesca Ieva, Internal Supervisor,
MOX - Modeling and Scientific Computing laboratory, Department of Mathematics,
Politecnico di Milano, Milan, Italy;
Associate Head of Human Technopole-Health Data Science Center, Milan, Italy;

Anders Stockmarr, Co-Supervisor,
DTU Compute - Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Konges Lyngby, Denmark;

Nikolaj Normann Holm, Co-Supervisor,
DTU Compute - Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Konges Lyngby, Denmark;

Ove Andersen, Co-Supervisor,
Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark;
Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre,
Hvidovre, Denmark;
Emergency Department, Copenhagen University Hospital Amager and Hvidovre,
Hvidovre, Denmark;

I would like to express my sincere gratitude to Prof. Anders Stockmarr and Postdoc Nikolaj Normann Holm for entrusting me with this project and for their invaluable support, insightful feedback, and availability throughout this journey. I am also deeply grateful to Prof. Ove Andersen for his contributions to the inception of this research. Additionally, I would like to thank Prof. Francesca Ieva for her willingness to serve as my Internal Supervisor.

I also acknowledge the assistance of ChatGPT, an AI language model developed by OpenAI, which helped me refine my ideas, clarify concepts, and improve the structure of my work. While all interpretations and conclusions remain my own, the interaction with this tool provided useful insights during both the analysis and writing process.

Finally, I extend my heartfelt thanks to my family for their unwavering support throughout my studies, making this achievement possible.