FROM MEDICAL CHARTS TO MARKET FORCES: HOW HEALTH TYPES SHAPE LABOUR MARKET OUTCOMES

Gavin J. Qu

University of Essex Economics: MSc. Dissertation

September 5th, 2024

I would like to express my sincere gratitude to Dr. David Zentler-Munro for his invaluable guidance and insightful advice throughout the development of this MSc dissertation.

Abstract

I attempt to illustrate the complex relationship between chronic illness and labour market outcomes, employing unsupervised machine learning techniques on the UK Household Longitudinal Study. I identify three distinct health types: the "Resilient," the "Early Onset Frail," and the "Accelerated Decline," each charting a unique course through health and economic landscapes. My analysis reveals striking disparities in employment, earnings, and education across these health trajectories, illustrating how individual health patterns may shape economic futures. Moving beyond traditional demographic factors, this paper demonstrate that health types significantly enhance predictions of frailty and labour market performance. The findings carry profound implications for retirement policies and inequality mitigation strategies, suggesting a need for tailored economic approaches that account for diverse health journeys.

1. Introduction

The intersection of health and labour economics has long captivated researchers and policymakers alike. This paper attempts to answer a fundamental question in health and labour economics: are individuals characterized by fundamentally different health types, and what is the association between any such health types and labour market outcomes? Furthermore, I investigate whether health-related factors can serve as predictors of earning potential. These inquiries extend beyond academic interest, carrying profound implications for individuals, healthcare systems, and economic policy.

Using an unsupervised machine learning approach pioneered in De Nardi et al (WP, 2024) and drawing from the rich dataset of the Understanding Society: UK Household Longitudinal Study (UKHLS), I aim to uncover patterns and relationships that illuminate the complex interplay between health trajectories and economic outcomes. Our analysis focuses particularly on chronic illnesses and substantial difficulties caused by health problems or disabilities, seeking to identify distinct health types. Through this investigation, I aspire to contribute to a more nuanced understanding of health-related economic disparities and inform policies that can better address the diverse needs of individuals across different health trajectories.

To address these questions, our study is structured around three key areas of investigation:

1. Frailty Trajectories: I examine the frailty trajectories associated with different health types. By employing unsupervised clustering techniques on longitudinal health data, I identify distinct patterns in how individuals' health evolves over time, particularly focusing on the progression of chronic illnesses and disabilities. This analysis reveals three primary health types, each with unique implications for long-term health outcomes.

2. Employment Dynamics: I explore the employment dynamics across these health types, considering factors such as earnings and employment status. This section illuminates the association between health trajectories and labour market participation, income levels, and human capital accumulation.

3. Predictive Power: I assess the explanatory power of health types in predicting frailty and, by extension, labour market outcomes. By comparing models with and without health type dummy variables, I demonstrate a substantial improvement in the adjusted R-squared value, underscoring the significance of considering heterogeneous health trajectories in labour economic analyses.

Through these three investigations, I aim to provide a comprehensive understanding of how chronic illnesses and health trajectories shape economic outcomes. Our findings offer valuable insights that can inform both policy decisions and further research area of labour economics.

2. Data Overview

In this analysis, I employ an unsupervised machine learning approach to address fundamental questions in labour economics. I begin by outlining the methodology and data preparation techniques used prior to the algorithmic modeling. Our primary data source is the Understanding Society: UK Household Longitudinal Study (UKHLS), a comprehensive longitudinal panel study spanning 13 years. This study serves as a continuation of the British Household Panel Study (BHPS), which ran from 1991 to 2008.

I extracted relevant health variables in a long panel format, focusing on two key aspects of health deficits: chronic illnesses and substantial difficulties caused by health problems or disabilities (“Main Survey Variable: Disdif6 Sight,” n.d.). Chronic illnesses in the survey include diagnosed conditions such as asthma, arthritis, cancer, and high blood pressure. While my analysis follows

that of De Nardi et al. (2024) , though I place greater emphasis on chronic illnesses (Borella et al., n.d.). The difficulties/disability variables, in contrast, are not necessarily diagnoses but rather descriptions of challenges in daily activities, such as "difficulty with moving objects" or "difficulty with daily tasks."

The focus on chronic illness is crucial due to its pervasive impact on individuals and healthcare systems. The National Institutes of Health (NIH) defines chronic conditions as persistent health issues lasting more than three months, encompassing a range of ailments including arthritis, asthma, cancer, diabetes, and certain viral diseases (Bernell and Howard 2016). The World Health Organization categorizes these as non-communicable diseases, characterized by their long-lasting nature and slow progression, primarily including cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes (“Noncommunicable Diseases,” n.d.). By examining chronic illnesses, this paper address a significant aspect of public health that has profound implications for labour market participation and economic outcomes. The enduring nature of these conditions necessitates a deeper understanding of how they shape individuals' health trajectories and, consequently, their long-term economic prospects.

In the UKHLS dataset, health deficits are recorded as 1 (present/diagnosed), 0 (not present), or -9 (refusal/not applicable). For our analysis, I converted any value other than 1 or 0 to Python's undefined data format – "NaN" (not a number). To account for the nuanced recording methods in the UKHLS survey, I created a new "healthcond" variable that aggregates all variations of the "hcond" variables. This new variable operates under the assumption that once a person is diagnosed with a chronic illness, such as asthma, they continue to have it until the end of the survey or their death. In total, our analysis incorporates 16 variables related to chronic illnesses and 11 difficulty/disability variables across 13 waves of data.

```
Total rows: 533476
Total columns: 92
healthcond1: 1s: 67349, 0s: 443318, NaNs: 22809
healthcond2: 1s: 82987, 0s: 427680, NaNs: 22809
healthcond3: 1s: 3841, 0s: 506826, NaNs: 22809
healthcond4: 1s: 11233, 0s: 499434, NaNs: 22809
healthcond5: 1s: 13856, 0s: 496811, NaNs: 22809
healthcond6: 1s: 12429, 0s: 498238, NaNs: 22809
healthcond7: 1s: 10312, 0s: 500355, NaNs: 22809
healthcond8: 1s: 4186, 0s: 506481, NaNs: 22809
healthcond9: 1s: 6133, 0s: 505149, NaNs: 22194
healthcond10: 1s: 19378, 0s: 491289, NaNs: 22809
healthcond11: 1s: 9672, 0s: 500995, NaNs: 22809
healthcond12: 1s: 10158, 0s: 500509, NaNs: 22809
healthcond13: 1s: 26378, 0s: 484289, NaNs: 22809
healthcond14: 1s: 36020, 0s: 474647, NaNs: 22809
healthcond15: 1s: 5674, 0s: 504993, NaNs: 22809
healthcond16: 1s: 100244, 0s: 410423, NaNs: 22809
```

Variable Name	Description	Type
hcond1	Asthma	Health Problem
hcond2	Arthritis	Health Problem
hcond3	Congestive heart failure	Health Problem
hcond4	Coronary heart disease	Health Problem
hcond5	Angina	Health Problem
hcond6	Heart attack or myocardial infarction	Health Problem
hcond7	Stroke	Health Problem
hcond8	Emphysema	Health Problem
hcond9	Hyperthyroidism or an over-active thyroid	Health Problem
hcond10	Hypothyroidism or an under-active thyroid	Health Problem
hcond11	Chronic bronchitis	Health Problem
hcond12	Any kind of liver condition	Health Problem
hcond13	Cancer or malignancy	Health Problem
hcond14	Diabetes	Health Problem
hcond15	Epilepsy	Health Problem
hcond16	High blood pressure	Health Problem
disdif1	Mobility (moving around at home and walking)	Difficulty
disdif2	Lifting, carrying or moving objects	Difficulty
disdif3	Manual dexterity (using your hands to carry out everyday tasks)	Difficulty
disdif4	Continence (bladder and bowel control)	Difficulty
disdif5	Hearing (apart from using a standard hearing aid)	Difficulty
disdif6	Sight (apart from wearing standard glasses)	Difficulty
disdif7	Communication or speech problems	Difficulty
disdif8	Memory or ability to concentrate, learn or understand	Difficulty
disdif9	Recognising when you are in physical danger	Difficulty
disdif10	Your physical co-ordination (e.g. balance)	Difficulty
disdif11	Difficulties with own personal care	Difficulty

These 27 variables are used to construct our key variable of interest – "frailty". The frailty variable provides a snapshot of an individual's health at a given point in time. It is calculated by dividing the number of an individual's health deficits by the total number of possible health deficits. This approach yields a normalized measure ranging from 0 to 1, where 0 represents perfect health (no deficits) and 1 represents the most extreme poor health (all deficits present or death).

For instance, an individual with no health deficits will have a frailty score of 0, while an individual who dies will have a frailty score of 1 in the corresponding wave, as death represents the ultimate manifestation of poor health. It's important to note that death data is available only for existing participants, totaling 4,444 individuals across the UKHLS survey lifespan. The wave data continues to be recorded even after an individual's death, maintaining the continuity of the dataset.

The calculation of the frailty variable involves a nuanced approach to handling missing data:

1. If all health conditions are recorded as NaN (Not a Number) across all waves for an individual, the frailty is set to NaN for all waves of that person.

2. For individuals with at least some health data, frailty is calculated for each wave, treating NaN values as 0 in the calculation. This approach assumes that missing data indicates the absence of a health deficit, which may introduce a slight bias towards lower frailty scores in cases of incomplete data.

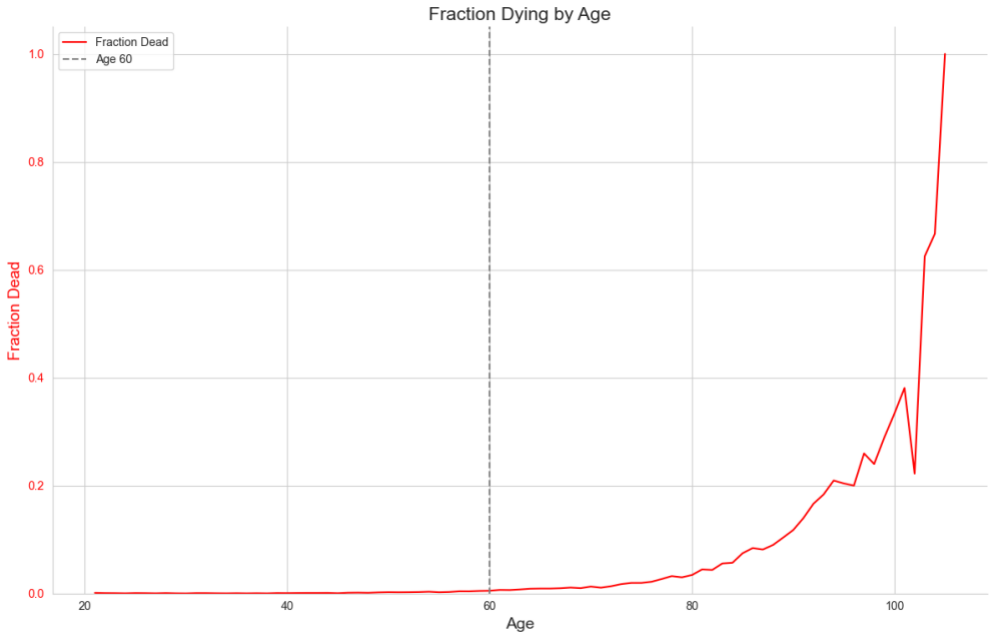
3. If an individual has no health data across all waves, their frailty is set to NaN for all waves, effectively excluding them from analyses that require this variable.

The summary statistic and distribution of the frailty data are shown:

```
Frailty Summary:
count    529303.000000
mean          0.057325
std           0.116342
min           0.000000
25%           0.000000
50%           0.037037
75%           0.074074
max           1.000000
```

```
0.00    0.000000
0.25    0.000000
0.50    0.037037
0.75    0.074074
0.99    0.481481
Name: frailty, dtype: float64
```

To better illustrate the age characteristic of frailty at this earlier stage of the analysis, two binned scatter plots are used with age on the x-axis. I plotted the points after taking the mean of the total frailty in each respective age group.



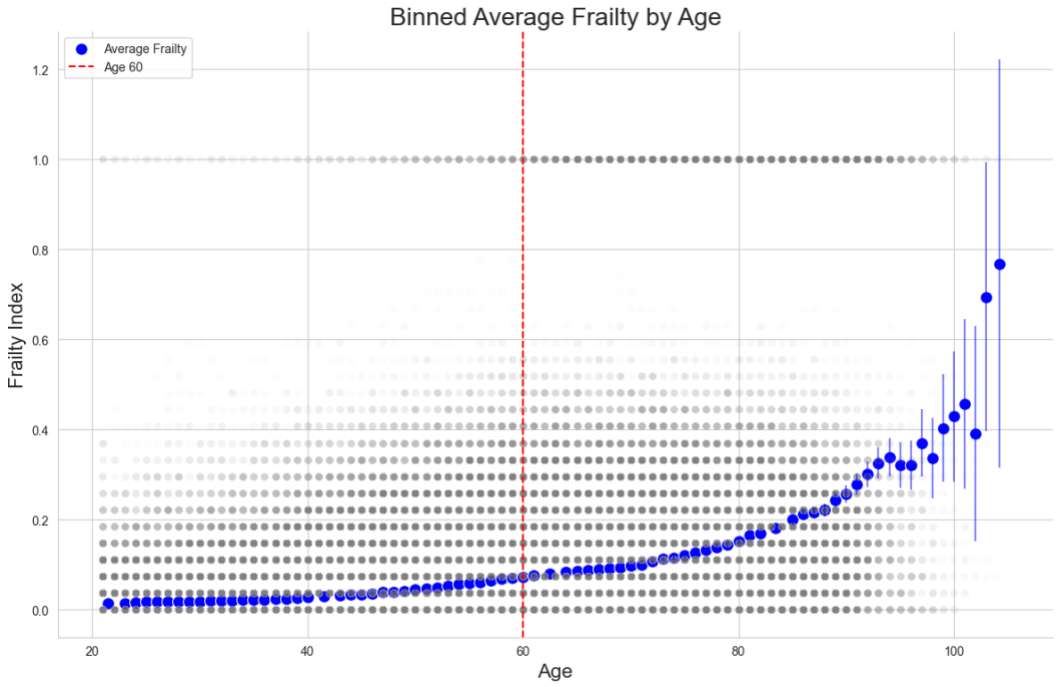
The survey assigns a frailty score of 1 to all individuals above 105 years old. This decision is based on two factors:

1. The extreme rarity of survival beyond this age.
2. Research indicating that the risk of mortality plateaus after age 105, rather than continuing to increase exponentially as it does from age 60 onwards.

This approach ensures that our frailty measure accounts for the unique health characteristics of extreme longevity while maintaining consistency in our data analysis. It also aligns with current understanding of mortality patterns in very advanced age groups. (“News: Age 105? Then You’ve a Better Chance... (The Guardian) - Behind the Headlines - NLM,” n.d.).

	age_dv	frailty
count	489471.000000	489471.000000
mean	50.981478	0.060537
std	17.326131	0.117902
min	21.000000	0.000000
25%	37.000000	0.000000
50%	50.000000	0.037037
75%	64.000000	0.074074
max	105.000000	1.000000

Number of deaths (frailty == 1): 4180



The data reveals a clear upward trajectory in frailty after age 60, which can be attributed to dramatic age-related changes in the human immune system. This observation aligns with recent findings published in Nature Aging (Shen et al. 2024), which document significant alterations in immune cell populations and functions with advancing age. The study highlights immunological shifts that may contribute to increased vulnerability to infections, chronic diseases, and overall health decline in older adults. These findings provide a potential explanation for the upward trend in health issues post-60 observed in our data, linking biological mechanisms to population-level health trajectories.

3. K-means Clustering

To identify distinct patterns in health trajectories based on frailty, I employ k-means clustering, a widely used unsupervised machine learning technique (MacQueen 1967). K-means clustering partitions n observations into k clusters, where each observation belongs to the cluster with the nearest mean (cluster centroid), serving as a prototype of the cluster (Hartigan and Wong 1979). As Hastie et al. note, "The K-means algorithm is simple and fast, which accounts for its popularity" (Hastie, Tibshirani, and Friedman 2009). The frailty is first turned into a vector for those aged 50 to 60 with a 2-year step size due to data availability. Essentially, I are clustering individuals into groups that are as similar as possible in terms of their health trajectories (not just levels at a fixed age).

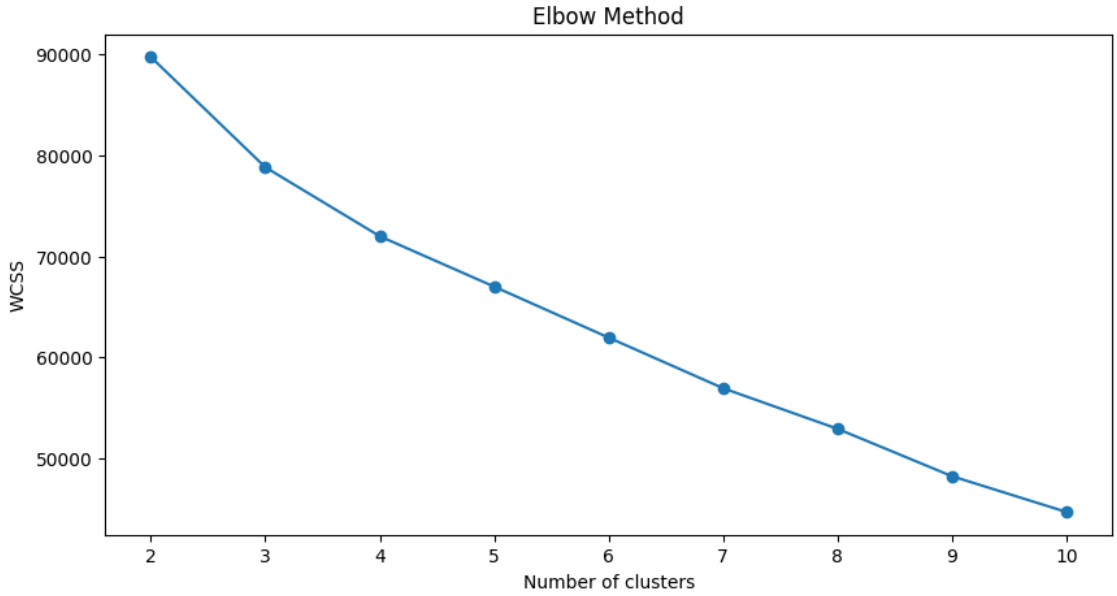
It is crucial to identify the optimal number of clusters (k), and I use two methods common in the literature to inform this decision: the elbow method and the silhouette method.

3.1. Elbow Method

The elbow method plots the within-cluster sum of squares (WCSS) against the number of clusters. As k increases, the WCSS typically decreases, with the rate of decrease often showing a clear "elbow" point. This point, where additional clusters yield diminishing returns in reducing WCSS, suggests an optimal k value (Kodinariya and Makwana 2013). The method provides a visual tool for determining the most efficient number of clusters, balancing model complexity with explanatory power.

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

The elbow method showed a noticeable "elbow" at k=3, suggesting diminishing returns for higher k values. Basically, adding more clusters does not significantly reduce inertia:



3.2 Silhouette Method

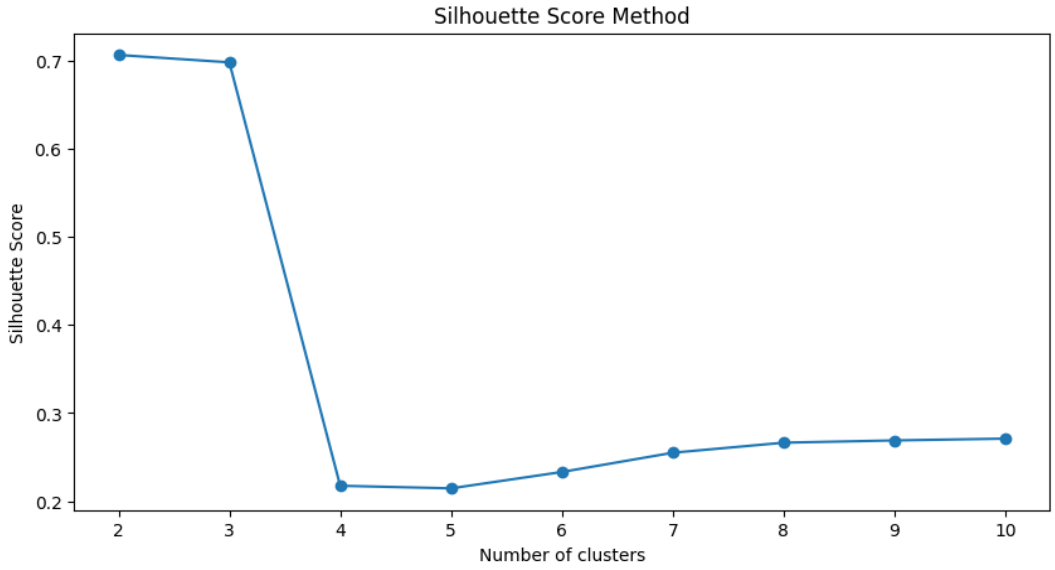
The silhouette method measures how similar an object is to its own cluster compared to other clusters (Rousseeuw 1987). Silhouette scores range from -1 to 1, with higher values indicating better-defined clusters. By plotting average silhouette scores against different k values, I can

identify the number of clusters that maximizes the score, suggesting optimal cluster separation and cohesion.

$$s(i) = \begin{cases} 0 & |C_I| = 1 \\ \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} & \text{otherwise} \end{cases}$$

In our analysis, the silhouette method typically produced a concave shape, with relatively higher scores for lower k values. Based on these results, I selected k=3 as an optimal midpoint between 2 and 4 clusters. This approach aligns with the principles discussed in Athey and Imbens who emphasize the importance of balancing model complexity with explanatory power in machine learning applications to economics (Athey, 2019)

Athey and Imbens (2019) highlight that while increasing the number of clusters can reduce within-cluster variance, it may lead to overfitting and reduced interpretability. They suggest that in economic applications, where the goal is often to uncover meaningful patterns rather than merely optimize predictive accuracy, a more parsimonious clustering solution may be preferable. Our choice of k=3 strikes a balance between capturing heterogeneity in the data and maintaining interpretable, economically significant clusters.



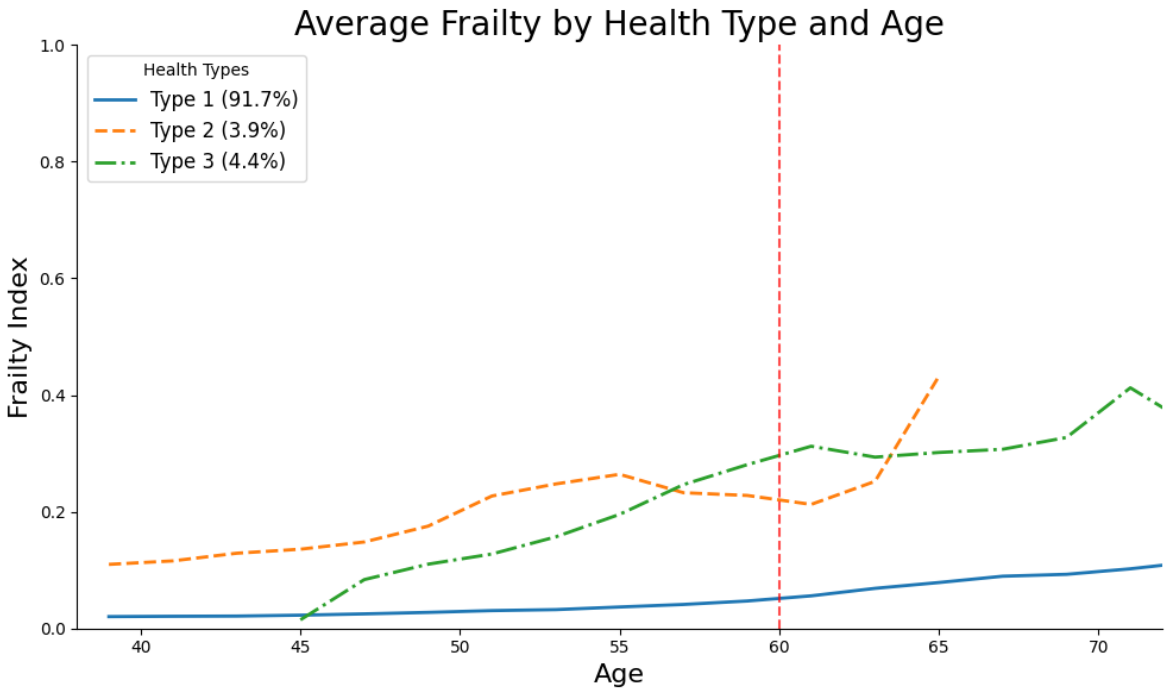
Based on the results from both the elbow and silhouette methods, and considering the interpretability of the resulting clusters in the context of health trajectories, I selected k=3 for our subsequent analysis. This choice strikes a balance between statistical indicators and practical significance in our economic health study.

Selecting three clusters provides a manageable framework for interpretation while capturing meaningful variations in health trajectories. This approach allows for nuanced analysis without over-segmenting the data, which could potentially lead to less robust or interpretable results. The three-cluster solution offers a comprehensive yet accessible way to categorize health trajectories, facilitating clear insights into the relationship between health patterns and economic outcomes.

4. Health Types

After the K-means clustering algorithm assigned four distinct clusters to the eligible individuals aged 50-60, I observed the following heterogeneity in health trajectories. Our analysis revealed three distinct health types, each characterized by different patterns of frailty progression and mortality risks.

Our analysis revealed a striking disparity in the distribution of health types within the sample population. Type I emerged as the dominant trajectory, accounting for 91.7% of the sample, while Types II and III represented only 3.9% and 4.4%, respectively. This marked imbalance suggests the presence of distinct underlying characteristics or risk factors associated with the less common health trajectories, warranting further investigation.




```
Number of unique individuals in each cluster:
Cluster 1: 18477 (92.26%)
Cluster 2: 742 (3.70%)
Cluster 3: 809 (4.04%)

Age range: 38.0 to 72.0
```

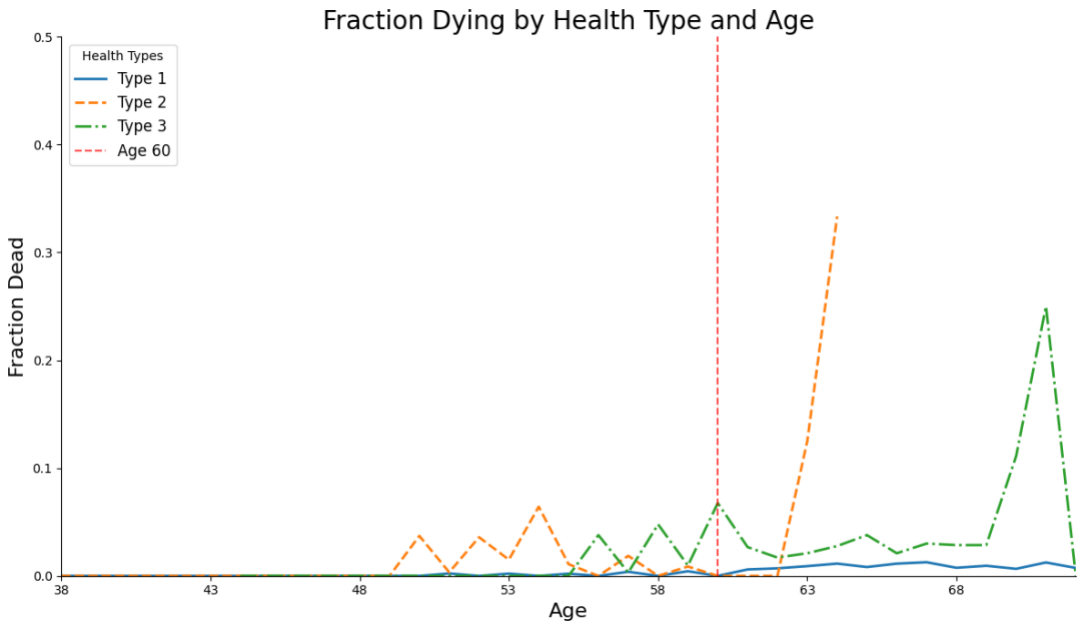
At age 40, the starting point of our data, I observed different initial frailty levels across the health types:

1. Type I ("Resilient Health Type"): Starts with the lowest frailty and shows a very gradual, steady increase over time, maintaining the lowest frailty index throughout the entire age range. This type represents the majority of the population.
2. Type II ("Early Onset Frailty Type"): Begins with the highest initial frailty and experiences a more pronounced increase, especially after age 60.
3. Type III ("Accelerated Decline Type"): Starts between Types I and II but rapidly increases, surpassing Type II around age 55. It shows some fluctuations but generally maintains higher frailty levels in later years.

These distinct trajectories highlight the heterogeneity in health patterns across the population and underscore the importance of considering these differences in health and economic analyses.

By age 70, I observe a widening gap between the "resilient health type" and the other two types. This suggests that initial health trajectories can significantly influence long-term health outcomes, at least when considering these specific health deficits. The divergence in frailty indices between the majority (Type I) and the minority (Types II and III) is particularly noteworthy and could have important implications for healthcare planning and policy.

This pattern of diverging health trajectories is consistent with the ongoing debate in bioscience regarding the interplay of genetics and environment in determining health outcomes (Ryff and Singer 2005). The fact that a large majority of the population follows the more favorable Type I ("resilient health type") trajectory, while a small percentage experience much steeper increases in frailty, suggests that both inherent factors and potentially modifiable environmental or lifestyle factors may be at play.



A notable vertical line at age 60 indicates a significant transition point, this is related to typical retirement age and immunological factors noted in gerontology literature.

Similar to the “vigorous resilient” type found in De Nardi’s paper, type I the “resilient health type” follows similar trajectories. The clear distinction originates from type II and type III, type II begins with higher frailty than type III, but their trends converge after age 55 and briefly crosses over at age 60. These starting points and health progressions are crucial in identifying an individual’s labour market and mortality.

Among those health types, the fraction of death also follows a trajectory where type II and type III are more affected. Although it follows a cyclical pattern, those two types have consistent higher death counts which further verifies the statement that vulnerable health types do exist consistently, as it is seen in the following table:

Death Counts and Percentages by Cluster:

Cluster	Total Count	Death Count	Death Percentage
1	166435.0	429.0	0.26%
2	7109.0	99.0	1.39%
3	7919.0	163.0	2.06%
Total	181463.0	691.0	3.71%

5. Employment Dynamics

The next question we can ask is how health types differ when it comes to employment status and earning. Is health type a good predictor for labour market participation? Are there clear correlations between health types and sociodemographic features such as sex, education and earnings?

The answer seems to be affirming our hypothesis that health trajectory is indeed correlated with labour market participation and average earnings. Both type II and type III reflect higher

unemployment rate than the UKHLS sample average, type II more than double the sample's 4.1% unemployment rate at 8.63%.

It's important to note that the "inactive" category includes students, those who are sick or disabled, individuals engaged in family care, and various types of retirees. Given that our cohort spans ages 38-72 and retired individuals are counted as inactive, these figures clearly indicate that:

1. Type I corresponds to the vigorous health group
2. Types II and III correspond to different degrees of frail health

These results underscore the significant impact of health trajectories on labour market outcomes, with the healthier Type I individuals showing better employment prospects compared to the more vulnerable Types II and III. This pattern suggests a strong link between health status and economic participation, highlighting the potential long-term economic consequences of different health trajectories.

```
Employment status distribution by cluster:
   Cluster  Employed (%)  Unemployed (%)  Inactive (%)
0        1         72.23            3.78         23.99
1        2         37.85            8.63         53.52
2        3         26.89            6.74         66.37

Overall employment status distribution:
lf_stat_label
Employed      68.92
Inactive      26.99
Unemployed     4.10
```

To further elucidate the relationship between health types and labour market outcomes, I examine employment rates and average earnings by cluster and age. This analysis is particularly relevant in the context of increasing retirement ages across developed countries, driven by fiscal deficits ("France Pension Reforms: Macron Signs Pension Age Rise to 64 into Law" 2023).

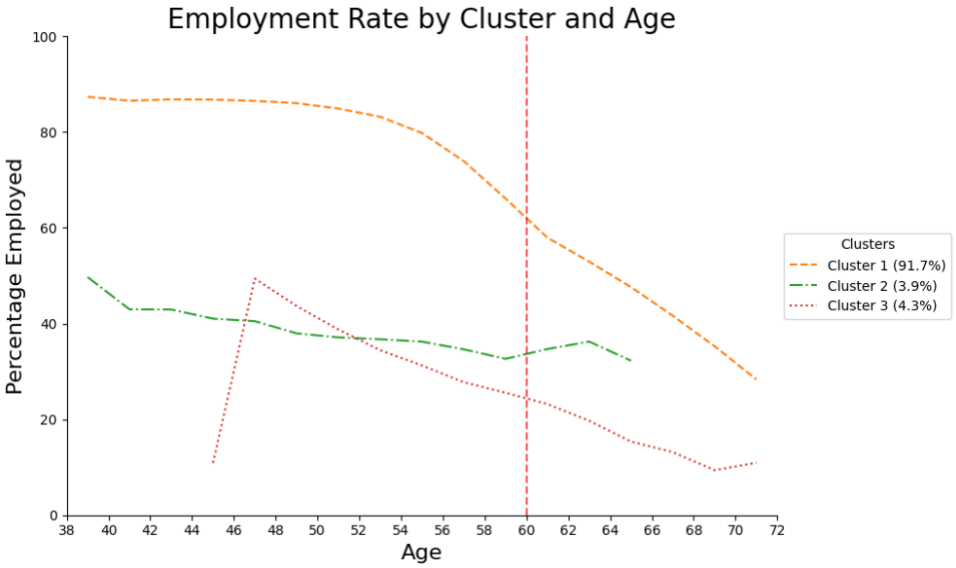
While our primary focus is on health dynamics in middle-aged and older adults, understanding unemployment patterns in this cohort has significant implications for policy and economic planning.

Our findings reveal:

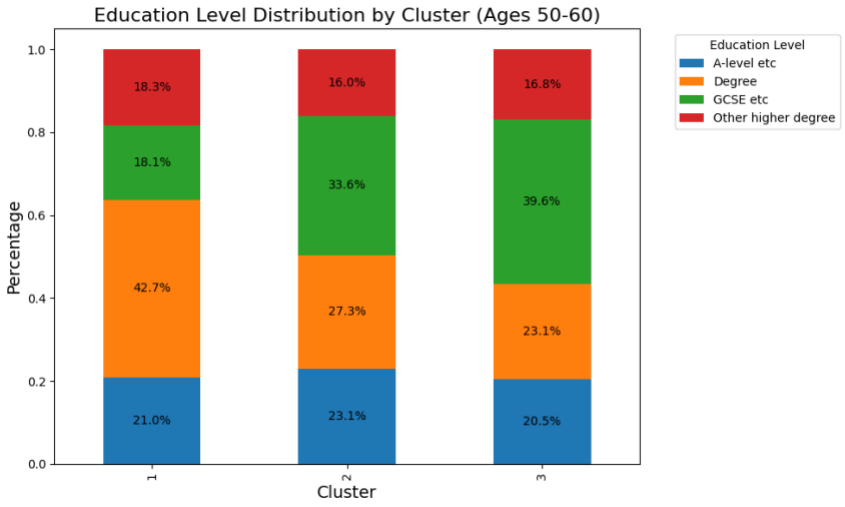
1. Employment Rates:
    - Types II and III (frail health groups) show lower employment rates both before and after the typical retirement age of 60-67 ("What Are the Average Retirement Ages Around the World?," n.d.).
2. Average Hourly Earnings:
    - Type I: £13.11 (above sample average)
    - Sample average: £12.95
    - Type II: £8.86 (31.6% below average)
    - Type III: £9.84 (24.0% below average)

These disparities in earnings are striking, with Types II and III earning approximately 25% less than the sample average. While post-retirement individuals in the sample may influence these figures, the distinct characteristics of the three health types persist across age groups.

This analysis underscores the profound impact of health trajectories on both employment prospects and earning potential. The consistent pattern of lower employment rates and earnings for the frail health groups (Types II and III) suggests a strong link between health status and economic outcomes throughout working life and into retirement age. These findings highlight the need for further research into the causal relationships between health types and retirement decisions, as well as potential policy interventions to address these disparities.


Employment Rate by Cluster and Age
Percentage Employed
Age
Clusters
Cluster 1 (91.7%)
Cluster 2 (3.9%)
Cluster 3 (4.3%)


Average Earnings by Cluster and Age with 95% Confidence Intervals
Average Gross Pay (£ per hour)
Age
Clusters
Cluster 1 (95.7%)
Cluster 2 (2.4%)
Cluster 3 (1.9%)


Education Level Distribution by Cluster (Ages 50-60)
Percentage
Cluster
Education Level
A-level etc
Degree
GCSE etc
Other higher degree
18.3%
18.1%
42.7%
21.0%
16.0%
33.6%
27.3%
23.1%
16.8%
39.6%
23.1%
20.5%

The education level distribution across the three health clusters for individuals aged 50-60 reveals significant variations in educational attainment, with important implications for health outcomes and socioeconomic status:

1. University Degree Attainment:
   - Type I (Cluster 1): 42.7%
   - Type II (Cluster 2): 27.3%
   - Type III (Cluster 3): 23.1%
2. A-level Attainment:
   - Consistent across all types at approximately 20-23%

The most striking feature is the markedly higher proportion of degree-holders in Type I compared to Types II and III. This sharp contrast suggests a strong association between higher education and more favorable health trajectories.

Interestingly, the consistency in A-level attainment across all types indicates that the main educational divergence occurs in the progression to university rather than in secondary school achievement. This observation prompts several considerations:

1. Potential barriers to higher education for individuals in Types II and III
2. Differing career paths that may influence health outcomes
3. The role of higher education in shaping long-term health trajectories

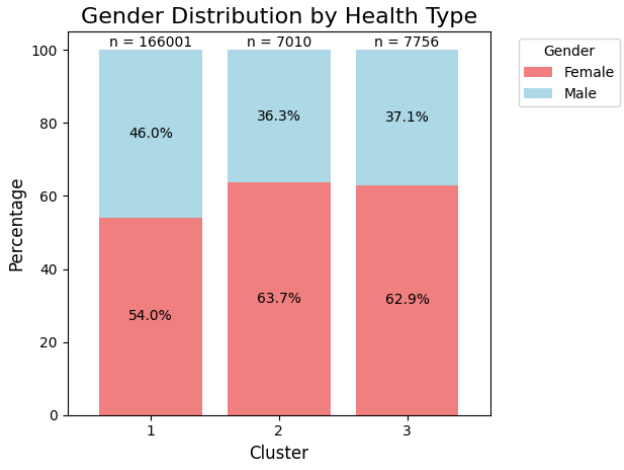
These findings underscore the complex interplay between education, health, and socioeconomic status. They highlight the need for further investigation into the causal relationships between

educational attainment and health outcomes, as well as potential interventions to address educational disparities that may contribute to divergent health trajectories.

These findings carry significant implications for our understanding of the relationship between education, health, and economic outcomes. The higher prevalence of degree-holders in Type I suggests a strong association between higher education and better health outcomes. This relationship may be attributed to increased health literacy, better access to healthcare resources, and lifestyle factors commonly associated with higher education. These factors likely contribute to the more favorable health trajectories observed in this group.

Conversely, the lower proportion of degree-holders in Types II and III may indicate potential barriers to higher education for these groups or reflect differing life circumstances that impact both educational opportunities and health trajectories. This observation raises important questions about the underlying factors that contribute to these disparities and their long-term consequences.

Importantly, individuals in Type I appear to benefit from a virtuous cycle, enjoying advantages in both health and labour market outcomes. This suggests that educational attainment may serve as a key factor in determining long-term health and economic prospects. The compounding nature of these advantages underscores the potential far-reaching impacts of educational disparities.



The gender distribution across health types reveals notable disparities. Types II and III exhibit a distinct female skew, with 63.7% and 62.9% female representation respectively. In contrast, Type I closely mirrors the UK population average, showing a more balanced distribution of 54% female and 46% male. This alignment of Type I with national demographics underscores its representation of typical health trajectories. The pronounced female majority in Types II and III suggests potential gender-specific influences on health outcomes or healthcare utilization, particularly for conditions associated with these more vulnerable health types. These disparities warrant further investigation into gender-related factors that may contribute to divergent health trajectories and their subsequent economic implications.

```
Gender Distribution Summary:
Sex      Female   Male
Cluster
1         89665  76336
2          4466   2544
3          4877   2879

Percentages:
Sex         Female       Male
Cluster
1        54.014735  45.985265
2        63.708987  36.291013
3        62.880351  37.119649
```

The gender imbalance observed in the latter clusters should be interpreted with caution, considering the substantial difference in cohort sizes. Types II and III each comprise approximately 800 individuals, making some degree of gender disparity statistically expected. In contrast, Type I represents a much larger sample of about 18,000 individuals, providing a more robust basis for gender distribution analysis. This significant disparity in sample sizes between Type I and Types II/III underscores the need for careful interpretation of the observed gender imbalances. While the trends are noteworthy, the smaller cohort sizes in Types II and III may

amplify random variations, potentially overstating the extent of gender-specific health trajectories in these groups.

6. Modeling Health Dynamics with Adjusted R-Squared

```
Basic Model Adjusted R-squared: 0.05257041726561085
Basic Model + Health Types Adjusted R-squared: 0.2596651838746622
   df_resid         ssr  df_diff  ss_diff             F  Pr(>F)
0   98740.0  244.649921      0.0      NaN           NaN     NaN
1   98738.0  191.169020      2.0  53.4809  13811.330684     0.0
Heteroscedasticity test p-value: 0.0
Durbin-Watson statistic: 0.6907970816210688
Variance Inflation Factors:
         features       VIF
0       Intercept       NaN
1  type_1[T.True]  2.176709
2  type_2[T.True]  2.181668
3          age_dv  1.020648
4          female  1.140247
5       education  1.180762
6          hrgpay  1.318884

Basic Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                 frailty   R-squared:                       0.053
Model:                             OLS   Adj. R-squared:                  0.053
Method:                  Least Squares   F-statistic:                     1371.
Date:                 Tue, 03 Sep 2024   Prob (F-statistic):               0.00
Time:                         13:26:32   Log-Likelihood:             1.5615e+05
```

To further investigate the importance of health types, I conducted a comprehensive regression analysis using frailty as a dependent variable. Our approach employed two distinct models: a baseline model incorporating fundamental demographic and labour market explanatory variables such as gender, earnings, education, and age, and an extended model that integrated our newly identified health types (derived from post-clustering groups). This comparative analysis aimed to quantify the additional explanatory power of health types in predicting labour market outcomes, beyond what is captured by traditional demographic variables. By juxtaposing these models, I

sought to isolate whether the health types tell us anything that traditional demographic or labour market variables cannot.

$Basic\ Model$: $[f_t = aXt + fage(t) + \epsilon_t]$

Basic Model + Health Types: $[f_t = aXt + fage(t) + \sum_{n=1}^{2} a_n D_n + \epsilon_t]$

To analyze the impact of health types on frailty, I employed two regression models. Our basic model is represented by the first equation, where f_t is the frailty index, X_t includes control variables such as gender, education, and hourly pay, and f_age(t) accounts for age effects. I then extended this to incorporate health types where D_n are dummy variables for health Types I and II, with Type III as the reference category. This approach allows us to quantify the additional explanatory power of health types beyond basic demographic and economic factors in predicting frailty outcomes.

I then used Ordinary Least Squares (OLS) regression with frailty as the dependent variable. The basic model included age, gender, education, and hourly pay (hrgpay) as independent variables. The extended model added dummy variables for health types (using Type III as the reference category).

6.1 R-Squared Statistics:

- Basic Model Adjusted R-squared: 0.0525
- Extended Model (with Health Types) Adjusted R-squared: 0.2597

The inclusion of health types substantially improved the model's explanatory power, increasing the adjusted R-squared from 5.25% to 25.97%. This significant improvement suggests that health types are crucial factors in explaining variations in frailty.

6.2 F-test for Model Comparison:

An F-test was conducted to compare the basic and extended models:

F-statistic: 13811.33, p-value < 0.001

The highly significant F-test result ($p < 0.001$) confirms that adding health types significantly improves the model fit, supporting the importance of including health types in our analysis.

6.3 Heteroscedasticity: The Breusch-Pagan test p-value is less than 0.001. The low p-value indicates the presence of heteroscedasticity. To address this, I will employ robust standard errors in our final model estimation.

6.4 Autocorrelation: The Durbin-Watson statistic is 0.6908. This value suggests positive autocorrelation in the residuals, which is expected given the longitudinal nature of our data. I will address this using clustered standard errors at the individual level.

6.5 Multicollinearity: The Variance Inflation Factors (VIF) for all variables were below 2.2, indicating that multicollinearity is not a significant concern in our model.

Our statistical analysis provides strong evidence that health types are significant predictors of frailty, even after controlling for demographic factors. The substantial improvement in model fit when including health types (increase in adjusted R-squared from 5.25% to 25.97%) underscores the importance of considering heterogeneous health trajectories in labour market analyses.

The significant F-test result further supports the relevance of our health type classification. However, the presence of heteroscedasticity and autocorrelation in our model necessitates cautious interpretation of coefficient estimates and the use of robust estimation techniques.

The low VIF values suggest that our health type classifications provide distinct information not captured by traditional demographic variables, validating our clustering approach.

While our model demonstrates the importance of health types, the relatively low overall R-squared (25.97%) indicates that there are other significant factors influencing frailty not captured in our current model, including but not limited to initial conditions of health. Future research should explore additional variables, such as lifestyle factors, occupational characteristics, or more detailed health indicators, to further improve the model's explanatory power.

The imbalanced distribution of health types, while reflective of real-world health disparities, may affect the precision of our estimates for the less common types. Future studies with larger samples or oversampling of less common health trajectories could provide more robust estimates for these groups.

Overall, our statistical analysis supports the inclusion of health types as a crucial factor in understanding frailty and, by extension, labour market outcomes. This approach offers a more nuanced view of health trajectories and their economic implications, potentially informing more targeted policy interventions and healthcare strategies.

7. Conclusion and Further Research

In our study, I used UKHLS data on chronic illnesses and daily ailments to identify 3 health types: The "resilient health type" (type I) is considered to be the common type accounting for 92% of the clustering sample, the "early onset frailty type" (type II) is suffering from higher initial frailty but a steadier decline after age 60. While the "accelerated decline type" (type III) begins with decent health similar to the Resilient Type but experiences a rapid decline in health starting the middle age and retirement age.

The identification of these distinct health trajectories within an economic sample has significant implications for retirement policy and economic inequality. Our analysis of employment rates, educational attainment, and earnings across these health types reveals substantial disparities. The resilient health type consistently demonstrates higher employment rates and earnings, as well as a higher proportion of degree holders. In contrast, the early onset frailty and accelerated decline types show lower labour market participation and earnings, suggesting a potential cycle of health-related economic disadvantage.

These findings underscore the need for nuanced policy approaches that account for diverse health trajectories. Traditional retirement policies based on a uniform retirement age may inadequately address the needs of individuals in the Early Onset Frailty or Accelerated Decline groups, who may require earlier interventions or more flexible work arrangements. Moreover, the stark differences in economic outcomes across health types highlight the potential for health-driven economic inequality, emphasizing the importance of targeted healthcare and social support initiatives.

Further research should delve deeper into the underlying biological factors that determine an individual's health type. The UKHLS dataset offers opportunities using biomarkers, genetic data, and proteomic information. Incorporating these biological indicators as propensity score variables could provide insights beyond the impact of age-related factors, potentially uncovering "economic markers" of an individual's health trajectory. Such markers could serve as early warning signs, allowing for preemptive interventions in both healthcare and economic planning. Additionally, longitudinal studies tracking individuals' transitions between health types over time could offer valuable insights into the malleability of these trajectories and the effectiveness of various interventions. Exploring the interplay between occupation types, lifestyle factors, and

health trajectories could further illuminate the complex relationship between work, health, and economic outcomes.

Finally, I observed the effect of health type as exogenous variables with large explanatory power in regard to predicting one's frailty. Although as De Nardi et al. mentioned, the formation of health types are certainly the result of earlier build-ups before adulthood (Borella et al., 2024). The combined effects should be manifested well beyond our analysis through significant disparities in retirement savings, lifetime earnings, and marriage patterns across health types. This underscores the importance of considering health trajectories in economic models and policy formulation, as they not only influence individual well-being but also have far-reaching implications for societal productivity, healthcare costs, and economic inequality. Understanding these factors would be tremendously beneficial to policymakers and individuals alike.

References

Athey, Susan. 2019 "Machine Learning Methods That Economists Should Know About | Annual Reviews." Accessed August 24, 2024. https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080217-053433.

Bernell, Stephanie, and Steven W. Howard. 2016. "Use Your Words Carefully: What Is a Chronic Disease?" *Frontiers in Public Health* 4 (August):159. https://doi.org/10.3389/fpubh.2016.00159.

Borella, Margherita, Francisco A Bullano, Mariacristina De Nardi, Benjamin Krueger, and Elena Manresa. n.d. "NBER WORKING PAPER SERIES."

"France Pension Reforms: Macron Signs Pension Age Rise to 64 into Law." 2023, April 14, 2023. https://www.bbc.com/news/world-europe-65279818.

Hartigan, J. A., and M. A. Wong. 1979. "Algorithm AS 136: A K-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1): 100–108. https://doi.org/10.2307/2346830.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Kodinariya, Trupti, and Prashant Makwana. 2013. "Review on Determining of Cluster in K-Means Clustering." *International Journal of Advance Research in Computer Science and Management Studies* 1 (January):90–95.

MacQueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of the Fifth Berkeley Symposium on Mathematical*

Statistics and Probability, Volume 1: Statistics, 5.1:281–98. University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

"Main Survey Variable: Disdif6 Sight (Apart from Wearing Standard Glasses)." n.d. *Understanding Society* (blog). Accessed August 26, 2024. https://www.understandingsociety.ac.uk/documentation/mainstage/variables/disdif6/.

"News: Age 105? Then You've a Better Chance... (The Guardian) - Behind the Headlines - NLM." n.d. NCBI. Accessed August 26, 2024. https://www.ncbi.nlm.nih.gov/search/research-news/1887.

"Noncommunicable Diseases." n.d. Accessed September 3, 2024. https://www.who.int/health-topics/noncommunicable-diseases#tab=tab_1.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (November):53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Ryff, Carol D., and Burton H. Singer. 2005. "Social Environments and the Genetics of Aging: Advancing Knowledge of Protective Health Mechanisms." *The Journals of Gerontology: Series B* 60 (Special_Issue_1): 12–23. https://doi.org/10.1093/geronb/60.Special_Issue_1.12.

Shen, Xiaotao, Chuchu Wang, Xin Zhou, Wenyu Zhou, Daniel Hornburg, Si Wu, and Michael P. Snyder. 2024. "Nonlinear Dynamics of Multi-Omics Profiles during Human Aging." *Nature Aging*, August, 1–16. https://doi.org/10.1038/s43587-024-00692-2.

"What Are the Average Retirement Ages Around the World?" n.d. US News & World Report. Accessed August 30, 2024. https://money.usnews.com/money/retirement/articles/what-are-the-average-retirement-ages-around-the-world.