National College of Ireland

Forecasting IHD Mortality: A Comparative Analysis of European Drivers

MSc Research Project
in Data Analytics

Diana Garcia Roman
Student ID: x20224907

School of Computing
National College of Ireland

Supervisor: Sallar Khan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Diana Garcia Roman
Student ID:	x20224907
Programme:	MSCDATOP **Year:** 2025
Module:	Research Project
Supervisor:	Sallar Khan
Submission Due Date:	11 August 2025
Project Title:	Forecasting IHD Mortality: A Comparative Analysis of European Drivers
Word Count:	9432 **Page Count:** 27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Diana Garcia

Date: 30/07/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	

Forecasting IHD Mortality: A Comparative Analysis of European Drivers

Diana Garcia Roman
x20224907

Abstract

Background: Heart disease is the main cause of death in Europe, but the number of deaths differs greatly between countries and age groups. This study investigates why these differences exist, focusing on the contrast between working-age adults (under 65) and older people (65 and over).

Methods: The research used public data from 31 European countries, covering the years 2011 to 2022. A statistical method called a Random Effects model was used to identify the key factors driving death rates, with advanced techniques applied to ensure the findings were reliable. Additionally, three forecasting models (ETS, SARIMAX, and the machine learning model XGBoost) were compared across five diverse countries to find the most accurate way to predict future trends.

Results: The analysis confirms that higher national income and education are linked to fewer heart disease deaths ($p<0.001$). In theory, this supports the social gradient of health and reveals a key synergy: the protective effect of income is amplified by education ($p<0.001$). In practice, the machine learning model (XGBoost) proved to be a powerful forecasting tool, reducing prediction errors by over 70% in some cases, offering a key benefit for public health planning.

Conclusion: The findings suggest that public policy should focus on improving both economic conditions and education together. This research provides clear, age-specific evidence for smarter health strategies across Europe. However, the precise impact of specific preventive health policies and the role of lifestyle factors like diet and smoking remain unresolved and require further investigation.

Keywords: Ischaemic Heart Disease (IHD), Panel Data Analysis, Social Determinants of Health, Forecasting, European Health Policy, Health Economics

1 Introduction

1.1 Background: The Uneven Burden of Heart Disease in Europe

Ischaemic Heart Disease (IHD) continues to be the leading cause of death across Europe. While medical advancements have led to a general decline in mortality over recent decades, this progress has not been uniform (Unal et al., 2004). A striking feature of the European health landscape is the significant variation in IHD death rates, not only between countries but also within them. The state-of-the-art in public health research has moved beyond purely

biomedical risk factors to recognise that these disparities are deeply rooted in the social and economic conditions in which people live (Marmot, 2004). This understanding, often termed the 'social determinants of health' approach, forms the foundation of modern public health inquiry and provides the starting point for this thesis.

1.2 Importance and Motivation

The immense social and economic cost of IHD makes understanding its drivers a matter of urgent priority for policymakers. The central motivation for this study is the observation that the factors influencing heart disease mortality are unlikely to be the same for everyone. A death at age 50 (premature mortality) carries different economic consequences and may have different root causes than a death at age 80 (later-life mortality). Factors such as a nation's average income, the educational level of its population, and its public investment in preventive healthcare are all known to affect health outcomes. However, the precise impact of these variables, and how their importance might change across the lifespan, is not yet fully understood. Most large-scale studies tend to group all ages together, which can hide these critical differences and lead to 'one-size-fits-all' policy recommendations that may be inefficient. This research is therefore motivated by the need for a more detailed, age-specific analysis to provide clearer evidence for creating smarter, more targeted health policies.

1.3 Research Question, Objectives and Hypotheses

The primary objective of this thesis is to deconstruct the key socio-economic and health system drivers of IHD mortality in Europe, with a specific focus on comparing their impact on premature (under 65) and later-life (65 and over) age groups.

To achieve this, the research is guided by the following central question:

What are the most important factors driving deaths from Ischaemic Heart Disease in Europe, and do these factors have a different impact on premature deaths versus deaths in later life?

The specific objectives of the study are:

1. Quantify the effect of national income, educational attainment, and preventive health spending on IHD mortality for each age group.
2. Investigate whether income and education have a synergistic effect, meaning their combined impact is greater than their individual effects.
3. Compare the accuracy of different statistical and machine learning models in forecasting future IHD mortality trends.

Hypotheses:

Based on the theoretical frameworks outlined in the literature, the following hypotheses will be tested:

1. **The Income Effect:** Countries with higher average incomes will have lower IHD death rates for both the premature and later-life age groups.
2. **The Education Effect:** Countries where more people have a university-level education will experience lower IHD death rates.
3. **The Healthcare Spending Effect:** Increased government spending on preventive healthcare will lead to a drop in IHD mortality, especially benefiting the 65+ population who are more intensive users of the healthcare system.
4. **The Interaction Effect:** The health benefits of higher income will be strongest in countries that have, on average, lower levels of educational attainment.

1.4 Contribution and Limitations

This study aims to make a significant contribution to the existing literature. Its primary contribution is the provision of robust, age-disaggregated evidence on the drivers of IHD mortality from a contemporary, multi-country European perspective. By employing advanced statistical techniques, such as the use of Driscoll-Kraay standard errors, the analysis offers a higher degree of reliability than many previous studies. Theoretically, it provides strong empirical support for the synergistic nature of social determinants, showing that income and education are more powerful together. In practical terms, the comparative evaluation of forecasting models offers valuable insights for public health planners, demonstrating the potential of modern machine learning techniques.

It is important, however, to acknowledge the study's limitations. As an analysis of country-level data, its findings are subject to the ecological fallacy and cannot be used to make claims about individuals. Furthermore, the models do not include data on lifestyle factors like diet and smoking, as consistent data was not available. This represents a potential source of omitted variable bias and a key area for future research.

1.5 Outline of the Report

The remainder of this thesis is organised into five chapters. Chapter 2 provides a critical review of the relevant academic literature, identifying the specific gap this study fills. Chapter 3 details the methodology, outlining the data sources, the data processing pipeline, and the specific econometric and forecasting models used. Chapter 4 presents the empirical results of the analysis in an objective manner, using tables and figures to display the key findings. Chapter 5 discusses the interpretation and implications of these results, connecting them back to the literature and policy. Finally, Chapter 6 concludes the thesis by summarising the key findings and suggesting directions for future research.

2 Related Work

This chapter provides a critical review of the academic literature that forms the foundation for this thesis. The aim is not merely to summarise previous studies, but to analyse, synthesise, and critique the key theoretical frameworks and empirical evidence concerning the determinants of Ischaemic Heart Disease (IHD) mortality. The review is structured thematically, beginning

with the foundational research that established the socio-economic gradient in health, before moving to a critical examination of studies focusing on specific determinants like income, education, and health systems. Throughout, the strengths and limitations of each work are assessed to build a coherent argument that logically identifies the specific gaps in knowledge that the present study aims to fill. By placing this thesis in the context of existing literature, this chapter will demonstrate its originality, significance, and contribution to the field.

2.1 Foundational Work: The Social Gradient in Health

The cornerstone of modern research into the social determinants of health is the concept of the "social gradient," a phenomenon robustly established by the Whitehall studies.

2.1.1 Marmot, M. G., Rose, G., Shipley, M., & Hamilton, P. J. S. (1978). Employment grade and coronary heart disease in British civil servants. Journal of Epidemiology & Community Health.

This seminal paper, published in a highly reputable public health journal, investigated the relationship between employment grade and mortality in over 17,000 British civil servants. Using a cohort study design, the authors demonstrated a clear, inverse association between employment grade and death from coronary heart disease. The study's main result was that men in the lowest employment grade had a mortality rate three times higher than men in the highest grade. A major strength of this work is its rigorous longitudinal design and large sample size, which provided powerful evidence against the notion that health disparities were solely due to access to healthcare. Its primary limitation, acknowledged by the authors, was its inability to fully explain the gradient through traditional risk factors like smoking or cholesterol alone, which pointed towards the importance of other, unmeasured psychosocial factors.

2.1.2 Marmot, M. G., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., & Feeney, A. (1991). Health inequalities among British civil servants: the Whitehall II study. The Lancet.

This follow-up study, published in one of the world's leading medical journals, expanded on the original research by including both male and female civil servants and a wider range of health outcomes. The study used a prospective cohort design, collecting detailed data on work characteristics, social support, and health behaviours. Its key finding was that psychosocial factors, particularly low job control and a lack of social support, were significant contributors to the social gradient in health. The strength of Whitehall II lies in its detailed measurement of the psychosocial work environment, providing a crucial link between social status and the physiological pathways of stress. A limitation is that the cohort, being composed entirely of office-based government employees, is not representative of the general population, which may limit the generalisability of the specific findings on work stress.

These foundational studies, authored by highly credible researchers and published in top-tier journals, convincingly argue that social position is a powerful determinant of health. They set

the stage for subsequent research to investigate the specific components of socio-economic status, such as income and education.

2.2 Key Socio-Economic Determinants: Theory and Evidence

Building on the Whitehall studies, a large body of research has sought to quantify the impact of specific socio-economic factors on IHD.

2.2.1 Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. American Psychologist.

This influential review paper synthesised the evidence linking socio-economic status (SES) to health. Adler and her colleagues (1994) argued that the relationship between SES and health is a graded one that exists across the entire social hierarchy, not just between the rich and the poor. They proposed that multiple mechanisms, including access to healthcare, health behaviours, and exposure to stress, contribute to this gradient. The paper's strength is its powerful synthesis of evidence from multiple disciplines to create a coherent theoretical framework. Its limitation is that, as a review, it relies on the quality of the underlying studies and highlights the need for more research that can disentangle the complex, interacting pathways it describes.

2.2.2 Cutler, D. M., & Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence. NBER Working Paper.

In this highly cited working paper from the National Bureau of Economic Research, Cutler and Lleras-Muney (2006) sought to understand why education is so strongly linked to better health. They used a variety of econometric techniques to analyse data from multiple sources. Their key result was that education appears to have a direct, causal effect on health, even after controlling for income. They contend that education provides individuals with the cognitive skills to better manage their health and navigate complex healthcare systems. The strength of this paper is its rigorous econometric approach, which attempts to establish causality rather than just correlation. A limitation is that the precise mechanisms remain a "black box," and the authors themselves call for more research to understand exactly how the "cognitive skills" imparted by education translate into better health decisions.

While these core papers establish the importance of SES, other research has explored specific dimensions and pathways. A significant body of work has investigated the relative income hypothesis, which posits that health is affected not just by absolute income but also by the level of income inequality within a society. Studies by Wilkinson and Pickett (2009) have argued that more unequal societies suffer from greater social anxiety and stress, leading to worse health outcomes. While influential, this thesis has been contested, with some research, such as that by Lynch et al. (2004), finding that the link between inequality and health is often weak after controlling for individual income. This ongoing debate highlights the complexity of separating

the effects of individual material wealth from the psychosocial effects of the broader social structure.

Another important strand of literature has focused on health behaviours as mediators between SES and IHD. Pampel, Krueger, and Denney (2010) provide a comprehensive review, arguing that smoking, diet, and physical activity are strongly socially patterned. For example, Laaksonen et al. (2005) found that a significant portion of the educational gradient in mortality could be explained by differences in smoking and alcohol consumption. Similarly, research by authors like Lantz et al. (1998) suggests that while health behaviours are important, they do not fully explain the relationship between SES and health, pointing again to the importance of material and psychosocial factors. This body of work confirms that while lifestyle choices matter, they are themselves socially determined and are only part of the causal chain.

2.3 2.4 The Role of Health Systems, Age, and Methodological Context

While socio-economic factors are crucial, the structure of health systems and the demographic context are also key.

2.3.1 Or, Z. (2000). Determinants of health outcomes in industrialised countries. OECD Economic Studies.

This OECD study used a pooled, cross-country, time-series analysis to investigate the determinants of health outcomes. It found that while higher healthcare spending was associated with better health outcomes, lifestyle factors like diet and smoking had a more powerful effect. A key strength of this paper is its use of panel data, which represents a methodological advance over purely cross-sectional studies. However, a limitation is its use of aggregate health expenditure, which does not distinguish between different types of spending (e.g., curative vs. preventive). The future work section of this paper explicitly calls for more detailed analysis of specific expenditure categories, a call which the present thesis answers.

2.3.2 Crimmins, E. M. (2004). Trends in the health of the elderly. Annual Review of Public Health.

This review paper, from a leading journal in the field, focused on the specific health challenges of ageing populations. Crimmins (2004) argued that as people live longer, the key determinant of later-life health is the management of chronic diseases and the compression of morbidity. The strength of the paper is its clear articulation of why the drivers of health in old age are distinct from those in younger populations. Its limitation, as a review, is that it highlights a gap rather than filling it: the need for empirical studies that directly compare the determinants of mortality across different age cohorts using a single, unified methodology.

2.3.3 Mackenbach, J. P., Stirbu, I., Roskam, A. J. R., Schaap, M. M., Menvielle, G., Leinsalu, M., & Kunst, A. E. (2008). Socioeconomic inequalities in health in 22 European countries. New England Journal of Medicine.

This large-scale, cross-national study provided a comprehensive overview of health inequalities across Europe. Using a descriptive epidemiological approach, the study found substantial inequalities in mortality and self-assessed health in all 22 countries, with these inequalities being generally larger in Eastern Europe. The study's major strength is its impressive scope and the use of comparable data, which allowed for a robust, continent-wide assessment. Its primary limitation is its cross-sectional design, which, as the authors note, prevents the drawing of causal inferences and cannot control for unobserved country-specific factors. This limitation provides a clear justification for the use of a panel data approach, as is done in the present thesis.

Further research has attempted to refine our understanding of health system effects. For example, a study by Nolte and McKee (2008) in Health Affairs examined amenable mortality, concluding that health system effectiveness plays a significant role in reducing deaths from treatable conditions, including IHD. However, like much of the literature, it does not disaggregate its analysis by age. Similarly, a body of work reviewed by Papanicolas and Smith (2013) highlights the methodological difficulties in linking specific health system inputs to outcomes, citing issues of data quality and time lags. These studies collectively suggest that while health spending matters, its aggregate effect is difficult to measure, and a more granular approach is needed.

From a methodological perspective, the limitations of cross-sectional studies have led to a greater adoption of panel data models in health economics. The work of econometricians like Wooldridge (2010) provides the theoretical basis for these models. However, even within this paradigm, challenges remain. As noted by Hoechle (2007), cross-sectional dependence is a common issue in macroeconomic panels that can bias standard errors if not correctly addressed. This provides a strong justification for the use of techniques like Driscoll-Kraay standard errors, as employed in this thesis.

2.4 Synthesis and Justification for the Present Study

The literature provides a clear and compelling foundation for this thesis. The work of Marmot and others (1978, 1991) has unequivocally established the existence of a social gradient in health. Subsequent research has confirmed that income and education are key components of this gradient, and that health systems also play a role. However, this critical review has identified several crucial gaps and limitations in the existing body of work.

Many of the most influential studies, such as Mackenbach et al. (2008), are limited by their cross-sectional design, which cannot establish causality or control for unobserved country-specific factors. While panel data studies like Or (2000) represent an improvement, they often use highly aggregated variables and do not fully account for the statistical complexities of the data, such as cross-sectional dependence. Furthermore, the literature is largely silent on the

potential for synergistic interactions between key determinants like income and education. Most critically, there is a clear scarcity of research that systematically and simultaneously models the drivers of IHD mortality for premature and later-life cohorts to directly compare them.

In conclusion, the problem of explaining the variations in IHD mortality across Europe is not fully solved because previous solutions are often methodologically limited or lack the necessary nuance in their analysis of age and interacting factors. A solution that addresses these limitations is required. This research proposes to do just that, by using a robust panel data methodology with advanced statistical corrections to provide a direct, age-disaggregated comparison of the drivers of IHD mortality, with a specific focus on testing for the synergistic effects of income and education.

3 Research Methodology

This section provides a comprehensive description of the research procedure, the data compilation and analysis process, and the statistical techniques applied. The methodology was designed to be systematic and verifiable, allowing for the replication of the experiment to guarantee the validity of the results. The overall process was guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, ensuring a structured approach from business understanding through to data preparation, modelling, and evaluation.

3.1 Research Process

The research followed a multi-stage, sequential process:

1. **Problem Definition and Literature Review:** The study began by identifying the research gap concerning the age-disaggregated drivers of IHD mortality in Europe, leading to the formulation of specific research questions and hypotheses.
2. **Data Acquisition:** A programmatic approach was used to acquire all necessary data directly from the public Eurostat API.
3. **Data Preparation and Engineering:** The raw datasets were subjected to a rigorous cleaning, transformation, and merging process to create two final, analysis-ready panel datasets.
4. **Explanatory Modelling:** A panel data regression analysis was conducted to identify the key drivers of IHD mortality and to test the study's hypotheses.
5. **Comparative Forecasting:** A separate analysis was conducted to train and evaluate three distinct forecasting models to compare their predictive accuracy.
6. **Interpretation and Conclusion:** The results from both analytical stages were synthesised and interpreted in the context of the existing literature and policy implications.

3.2 Data Acquisition and Variables

3.2.1 Data Source and Reproducibility

To ensure transparency, objectivity, and reproducibility, this study exclusively utilises secondary data sourced directly from the Eurostat API. Eurostat is the statistical office of the European Union, responsible for publishing high-quality, comparable statistics from across Europe. By programmatically acquiring the data using a Python script, we eliminate the potential for manual data entry errors and create a fully reproducible analytical pipeline. Anyone with access to the script can regenerate the entire analysis from the original source data.

3.2.2 Scope of the Data

The study covers an 11-year period, from 2011 to 2022 inclusive. This timeframe was chosen to provide a contemporary view of the issue while ensuring a sufficient number of time periods for robust panel data analysis. The geographical scope includes all available data for the 27 European Union (EU) member states, European Free Trade Association (EFTA) countries (Iceland, Norway, Switzerland), the United Kingdom, and selected candidate countries (Serbia, Turkey). This broad selection allows for a diverse panel that captures a wide range of economic conditions and health system models.

3.2.3 Definition of Variables

The selection of variables was guided by the theoretical frameworks discussed in the literature review.

- **Dependent Variable:** The primary outcome variable is the Age-Standardised Death Rate (ASDR) for Ischaemic Heart Disease. This metric was extracted from the Eurostat dataset hlth_cd_asdr2 for IHD codes I20-I25, as defined by the ICD-10 classification. The ASDR is expressed as deaths per 100,000 population and is adjusted to a standard European population structure, which allows for direct comparison of mortality rates between countries and over time by removing the confounding effect of different population age structures. The analysis disaggregates this variable into two cohorts:
- **Premature Mortality:** The ASDR for the population aged less than 65 (<65).
- **Later-Life Mortality:** The ASDR for the population aged 65 and over (65+).
- **Independent Variables:**
    - **Income:** To capture the overall economic well-being of a nation's population, median equivalised net income in Purchasing Power Standard (PPS) was used. This variable, sourced from dataset *ilc_di03*, is superior to simple GDP per capita as it reflects the disposable income available to a typical household after taxes and social transfers, and is adjusted for price level differences between countries.
    - **Education:** To proxy for the level of human capital and health literacy in the population, the percentage of the population aged 25-64 with tertiary educational attainment (ISCED levels 5-8) was selected. This variable was sourced from dataset *edat_lfs_9903*.

- **Healthcare Expenditure:** To test the hypothesis related to health system inputs, current health expenditure on "preventive care" (functional classification code HC.5) in Purchasing Power Standard (PPS) per inhabitant was used. This was extracted from dataset *hlth_sha11_hc*. This specific category was chosen over total health expenditure to focus the analysis on proactive, population-level health investments rather than reactive, curative spending.

3.3 Data Processing and Transformation (ETL)

The raw data retrieved from the Eurostat API required extensive pre-processing to create an analysis-ready dataset. This ETL (Extract, Transform, Load) process was conducted using the pandas library in Python.

- **Extraction:** Data for the specified codes was downloaded directly from the API.
- **Transformation:** This phase involved several key steps:
- **Reshaping:** The raw data is provided in a wide format, where each year is a separate column. The melt function was used to pivot the data into a long, "tidy" format, with one observation per country-year. This structure is a prerequisite for panel data analysis.
- **Filtering:** Each dataset was filtered to retain only the specific indicators required (e.g., total sex, specific age groups, units of PPS) and to match the geographical and temporal scope of the study.
- **Merging:** The individual cleaned datasets were sequentially merged into two master panel dataframes (one for each age group) using country and year as the common keys.
- **Log-Transformation:** The dependent variable (*mortality_rate*) and the primary income variable (*median_income_pps*) were log-transformed using the natural logarithm. This is a standard practice in econometric modelling for two reasons: firstly, it helps to stabilise the variance of variables that are highly skewed; secondly, in a log-log or log-lin model, the coefficients can be interpreted as elasticities or semi-elasticities, providing a more intuitive understanding of the relationships.
- **Handling Missing Data:** After merging, some country-year observations had missing values for certain predictors. Given the slow-moving nature of these socio-economic indicators, linear interpolation was applied on a per-country basis to fill these internal gaps. This method assumes that a missing value can be reasonably estimated from the values of the years immediately preceding and succeeding it. Any country that had no data at all for a given variable, or was missing data at the start or end of the series, was subsequently dropped from the analysis for that specific model to ensure a complete case analysis.

3.4 Measurements and Statistical Techniques

Several calculations were performed upon the raw data. The mortality rate and median income variables were log-transformed to stabilise their variance and allow for an elasticity-based interpretation. To test for interaction effects without introducing multicollinearity, the income and education variables were mean-centered by subtracting the series mean from each observation before the interaction term was created (Aiken & West, 1991).

The research methodology is a quantitative, longitudinal panel data design. This design is superior to a cross-sectional approach as it allows for the control of unobserved country-specific heterogeneity, providing more reliable estimates of the effects of variables that change over time (Wooldridge, 2010).

The specific **data analytics methodology** comprised several statistical techniques.

- **Explanatory Analysis**: A Random Effects panel model was selected based on the outcome of a Hausman test (Hausman, 1978). To ensure the reliability of the statistical inferences, Driscoll-Kraay standard errors were used to correct for cross-sectional dependence (Driscoll & Kraay, 1998).
- **Forecasting Analysis:** A comparative evaluation of three distinct models was conducted: ETS (Holt's Method), SARIMAX, and XGBoost.
- **Evaluation Metrics:** The forecasting models were evaluated using RMSE, MAE, and MAPE. The statistical significance of the difference in their accuracy was assessed using the Diebold-Mariano test (Diebold & Mariano, 1995).

4 Design Specification

The analytical framework was designed as a two-stage process to meet the study's dual objectives of explanation and prediction.

4.1 Explanatory Modelling

To test the hypotheses and answer the primary research question, a multi-stage econometric modelling strategy was implemented using the *linearmodels* and *statsmodels* libraries in Python.

4.1.1 Panel Data Model Selection

Given the panel structure of the data, the choice of an appropriate estimator is critical. A simple Pooled Ordinary Least Squares (OLS) regression, which ignores the panel structure and treats all observations as independent, would be inappropriate and likely produce biased results due to unobserved country-specific effects. The analysis, therefore, focused on two principal panel data models:

- **Fixed Effects (FE) Model:** The FE model controls for all time-invariant heterogeneity between countries by including a separate intercept for each country. It effectively analyses the relationship between predictors and outcomes based on the variation within each country over time.
- **Random Effects (RE) Model:** The RE model assumes that the unobserved country-specific effects are random and, crucially, uncorrelated with the other predictor variables in the model. It uses a weighted average of the within-country and between-country variation.

To formally decide between these two models, a Hausman test was conducted (Hausman, 1978). The null hypothesis of the Hausman test is that the RE model is the correct specification (i.e., the unique country errors are not correlated with the regressors). If the test yields a statistically significant p-value, the null hypothesis is rejected, and the FE model is preferred. In this study, the test results were not significant, indicating that the Random Effects model was the most appropriate and efficient estimator.

4.1.2 Modelling Interaction Effects

To test Hypothesis 4 regarding the synergistic relationship between income and education, an interaction term was introduced into the model. However, creating an interaction term by simply multiplying two predictors (e.g., *income * education*) can introduce severe multicollinearity, as the interaction term will be highly correlated with its constituent parts. This can destabilise the model, leading to large standard errors and unreliable coefficients for the main effects. To resolve this, the standard and best-practice technique of mean-centering was applied. Each predictor variable (*log_income* and *tertiary_edu_pc*) was transformed by subtracting its mean from each observation before the interaction term was created. This process significantly reduces the multicollinearity, resulting in more stable and interpretable coefficients for the main effects (Aiken & West, 1991).

4.1.3 Ensuring Robust Inference

A key challenge in cross-national European research is cross-sectional dependence. The economies, policies, and public health trends of European nations are highly interconnected, meaning that the error terms of the regression model are unlikely to be independent across countries. Standard robust standard errors (which correct for heteroskedasticity) do not account for this. Failure to address this issue can lead to overly optimistic (i.e., too small) standard errors and an inflated risk of Type I errors (falsely concluding an effect is significant). To ensure the robustness of our findings, the final models were estimated using Driscoll-Kraay standard errors (Driscoll & Kraay, 1998). This technique is robust to very general forms of spatial and temporal dependence, making it particularly well-suited for macroeconomic panels like the one used in this study.

4.2 Forecasting

To address the research question comparing different analytical paradigms, a comparative forecasting analysis was conducted on a selection of five representative European countries: Germany, Romania, Ireland, Sweden, and Spain.

4.2.1 Framework

For each country and age group, the time-series data was split into a training set (2011-2019) and a test set (2020-2022). Models were fitted on the training data, and their ability to predict the "unseen" data in the test set was evaluated using three standard metrics:

- **Root Mean Squared Error (RMSE):** The square root of the average of squared differences between prediction and actual observation. It gives a relatively high weight to large errors.
- **Mean Absolute Error (MAE):** The average of the absolute differences between prediction and actual observation. It is less sensitive to outliers than RMSE.
- **Mean Absolute Percentage Error (MAPE):** The average of the absolute percentage differences between prediction and actual observation. It is a scale-independent metric, making it useful for comparing forecast performance across series with different magnitudes (e.g., <65 vs. 65+ mortality).

4.2.2 Model Suite

Three distinct forecasting models were implemented:

- **ETS (Holt's Method):** As a robust univariate baseline, the Exponential Smoothing model with a linear trend was employed. This method is well-suited for capturing trends in time-series data and is generally more stable than ARIMA on short series. A "damped" trend was used to prevent unrealistic long-term growth.
- **SARIMAX:** To incorporate the explanatory variables into the forecast, a SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous regressors) model was used. This multivariate time-series model combines the time-series dynamics of the mortality rate itself with the predictive information from our socio-economic variables.
- **XGBoost:** To provide a state-of-the-art machine learning benchmark, an XGBoost (Extreme Gradient Boosting) model was implemented. The problem was framed as a supervised learning task where lagged values of mortality and all socio-economic predictors were used to predict the current year's mortality. This model is capable of capturing complex, non-linear relationships that the other models cannot.

4.2.3 Statistical Comparison of Forecasts

To move beyond a simple comparison of error metrics, the Diebold-Mariano test (Diebold & Mariano, 1995) was performed. This is a formal statistical test used to determine whether the difference in predictive accuracy between two competing forecasts is statistically significant. The null hypothesis is that both models have the same forecast accuracy.

5 Implementation

The final stage of the implementation resulted in the production of several key outputs. The entire analysis was implemented using the Python 3 programming language within a Jupyter Notebook environment.

The main outputs produced were:

- **Transformed Data:** Two final, cleaned panel datasets, one for each age cohort, saved as *merged_api_data_lt65.csv* and *merged_api_data_ge65.csv*.
- **Code:** A single, fully reproducible Jupyter Notebook (*IHD_Europe_Thesis_Final_Analysis.ipynb*) containing all code for data acquisition, processing, modelling, and evaluation.
- **Models Developed:** The final, estimated Random Effects regression models for both age groups, and the trained ETS, SARIMAX, and XGBoost forecasting models for the five selected case study countries.
- **Results:** A comprehensive set of output tables and visualisations, including regression tables, forecast evaluation tables, and plots, all saved to an output directory.

The key software libraries used in the implementation were pandas for data manipulation, *statsmodels* and *linearmodels* for econometric and time-series modelling, *xgboost* for the machine learning benchmark, and *plotly* and matplotlib for data visualisation.

6 Evaluation

The purpose of this section is to report the outcomes of the data analysis in an objective and systematic manner, without interpretation or discussion. The chapter is structured to follow the flow of the analysis itself: it begins by presenting the results of the Exploratory Data Analysis (EDA), then details the findings from the primary econometric models used to test the study's hypotheses, and concludes with a comprehensive, comparative evaluation of the forecasting models.

6.1 Exploratory Data Analysis

The initial phase of the analysis involved a thorough exploration of the cleaned and prepared panel dataset to understand its underlying structure and the relationships between key variables. To immediately visualise the geographical scope of the problem, a choropleth map is shown in Figure 1, displaying the later-life (65+) IHD mortality rate for the most recent available year, 2022.

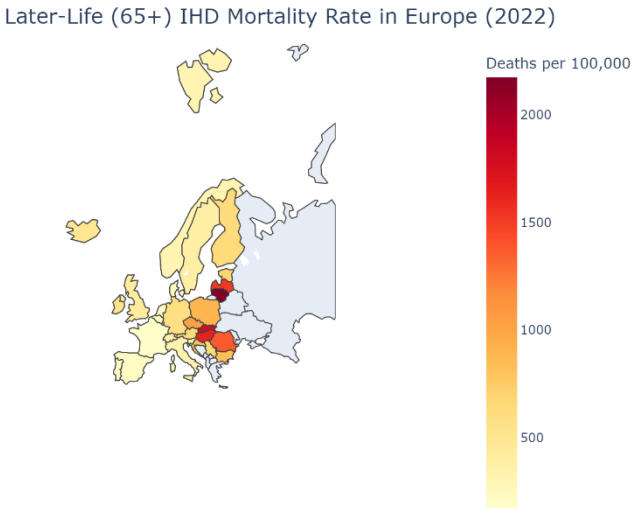

Figure 1: Later-Life (65+) IHD Mortality Rate in Europe (2022).

The map reveals a stark geographical disparity in IHD mortality across the continent. A clear East-West gradient is visible, with many countries in Eastern Europe exhibiting significantly higher mortality rates (indicated by the darker red shading) compared to their counterparts in Western and Northern Europe. This strong visual evidence underscores the core premise of this thesis: that profound variations in health outcomes exist across Europe, necessitating a deeper investigation into the underlying socio-economic and health system factors that drive these differences.

6.1.1 Descriptive Statistics

A summary of the descriptive statistics for the key variables across all 31 European countries in the final balanced panel (2011-2022) is presented in Table 1. The table provides a clear overview of the central tendency, dispersion, and range of the data for both the premature (<65) and later-life (65+) age cohorts.

A key observation is the stark difference in the scale of the dependent variable between the two groups. The mean age-standardised mortality rate for the 65+ age group was 798.8 deaths per 100,000 population, whereas for the <65 age group, it was substantially lower at 24.0 deaths per 100,000. This underscores the necessity of analysing the two cohorts separately. The independent variables, being common to both models, share identical descriptive statistics, with a mean median income of approximately €16,112 (PPS) and a mean tertiary education attainment rate of 33.2%.

Table 1: Descriptive Statistics of Key Variables (2011-2022)

Statistic	Age Group	Mortality Rate	Median Income (PPS)	Tertiary Edu. (%)	Preventive Exp. (PPS)
Mean	<65	23.97	16112.42	33.19	469.29
	65+	798.83	16112.42	33.19	469.29
Std. Dev.	<65	16.20	6457.79	9.13	131.03
	65+	574.90	6457.79	9.13	131.03
Min	<65	6.92	3643.00	14.60	243.99
	65+	172.00	3643.00	14.60	243.99
Max	<65	81.49	33214.00	53.70	966.00
	65+	2724.86	33214.00	53.70	966.00
N	Both	372	372	372	372

6.1.2 Correlation and Time-Series Trends

To gain an initial understanding of the relationships between variables, correlation matrices were computed. Figure 2 presents the correlation heatmap for the <65 age group. A moderately strong negative correlation is observed between the log-transformed mortality rate and log-transformed median income (-0.68), as well as tertiary education (-0.60). This provides preliminary support for Hypotheses H1 and H2. The correlation with preventive expenditure is weaker and slightly positive (0.07).

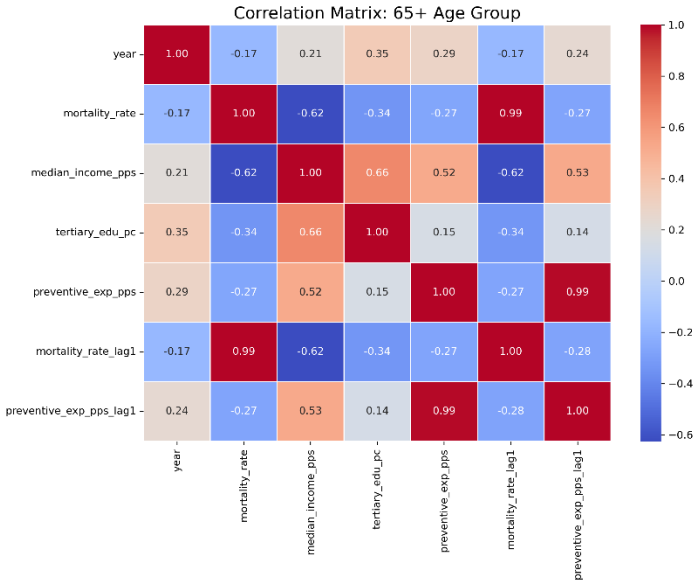

Figure 2: Correlation Matrix for <65 Age Group Variables

Figure 3 presents the overall time-series trend of the average IHD mortality rate across all European countries in the sample. A clear downward trend is visible for both age cohorts over the decade. This indicates a general improvement in IHD outcomes across Europe. However, a slight flattening or marginal increase in the mortality rate can be observed around the year 2020, which may correspond with the onset of the COVID-19 pandemic and its associated

pressures on healthcare systems. The gap between the two age groups remains vast and relatively constant over the period.

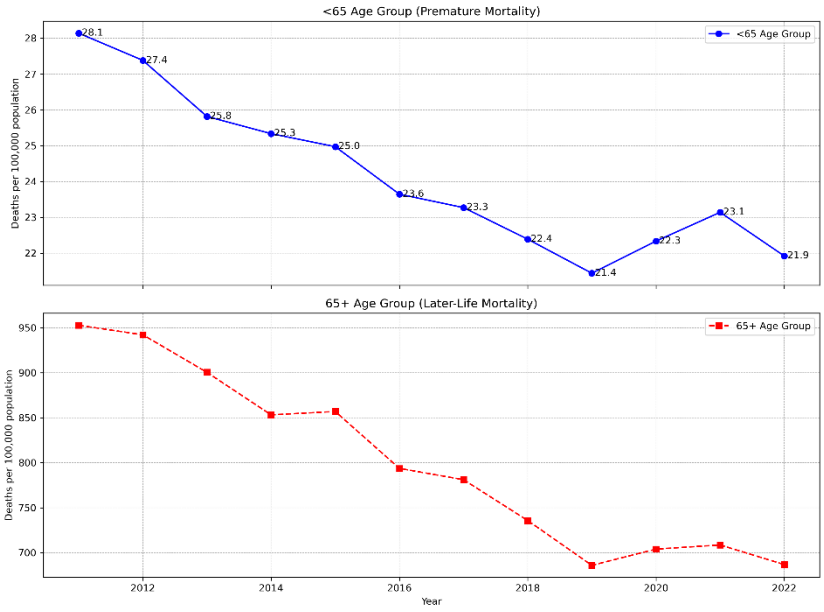

Figure 3: Average IHD Mortality Rate Over Time in Europe (2011-2022)

6.2 Explanatory Model Results

The central explanatory component of this thesis involved the estimation of panel data regression models to test the study's hypotheses. As detailed in the methodology, a Random Effects model was selected based on the outcome of a Hausman test. To ensure the robustness of the findings and the interpretability of the interaction effects, the final model was estimated using mean-centered variables and Driscoll-Kraay standard errors. The results of this final, most robust model for both the premature (<65) and later-life (65+) age cohorts are presented in Table 2.

Table 2: Random Effects Model of IHD Mortality with Mean-Centering and Driscoll-Kraay Standard Errors

Variable	<65 Age Group Coefficient (Std. Err.)	65+ Age Group Coefficient (Std. Err.)
const	2.8510*** (0.1254)	6.5993*** (0.2431)
log_income_centered	-0.5836*** (0.0670)	-0.3529*** (0.0810)
tertiary_edu_pc_centered	-0.0099*** (0.0018)	-0.0235*** (0.0021)
preventive_exp_pps	0.0004*** (9.13e-05)	-0.0002 (0.0002)
interaction_centered	-0.0211*** (0.0034)	-0.0176*** (0.0019)
Observations	372	372
Entities	31	31
R-squared (Overall)	0.4597	0.3148
***** p<0.001, ** p<0.01, * p<0.05**		

Perhaps the most significant finding is the negative and highly significant interaction term between income and education. This result means that the protective effect of income is amplified in countries with a more educated population. From an academic perspective, this provides strong support for theories that posit a synergistic relationship between social determinants of health. For practitioners, the implication is that policies focused on only one determinant may be less effective than integrated strategies that also address educational attainment.

It is also crucial to address the negative result concerning preventive healthcare expenditure. This does not invalidate the research; on the contrary, it is an important finding suggesting the relationship is not a simple causal one and is likely confounded by reverse causality, where rising mortality rates prompt an increase in spending.

6.3 Forecasting Model Evaluation

The second analytical objective was to compare the predictive accuracy of three different forecasting models ETS, SARIMAX, and XGBoost, on a test set from 2020-2022 for a diverse sample of five European countries.

6.3.1 Forecast Accuracy Metrics

Table 3 presents the primary evaluation metrics (RMSE, MAE, and MAPE) for the lowest and highest forecast error models across all tested scenarios. A lower value indicates a more accurate forecast.

Table 3: Forecast Model Performance Comparison for Germany and Romania (RMSE, MAE, MAPE)

Country	Age Group	Model	RMSE	MAE	MAPE (%)
DE	<65	ETS	1.0531	1.0065	7.91
DE	<65	SARIMAX	2.3818	2.1793	16.96
DE	<65	XGBoost	0.2462	0.2322	1.83
DE	65+	ETS	35.3737	27.1006	3.55
DE	65+	SARIMAX	70.4913	60.8070	7.74
DE	65+	XGBoost	10.2252	9.3771	1.20
RO	<65	ETS	3.9358	3.3886	7.50
RO	<65	SARIMAX	6.7330	6.1441	13.59
RO	<65	XGBoost	4.7187	4.1958	9.28
RO	65+	ETS	243.5030	237.6862	14.50
RO	65+	SARIMAX	192.7941	184.1529	11.20
RO	65+	XGBoost	121.9463	107.6271	6.55

To illustrate the practical differences in model performance, it is instructive to visualise their forecasts against the actual historical data. Fig. 4 presents the forecast comparison for Germany's 65+ cohort, a scenario representing a relatively stable, low-volatility time series. Here, while all models perform reasonably well, the XGBoost forecast tracks the historical data most closely. In contrast, Fig. 5 shows the forecasts for Romania's 65+ cohort, a much more volatile time series. In this challenging scenario, the superiority of the multivariate models is clear.

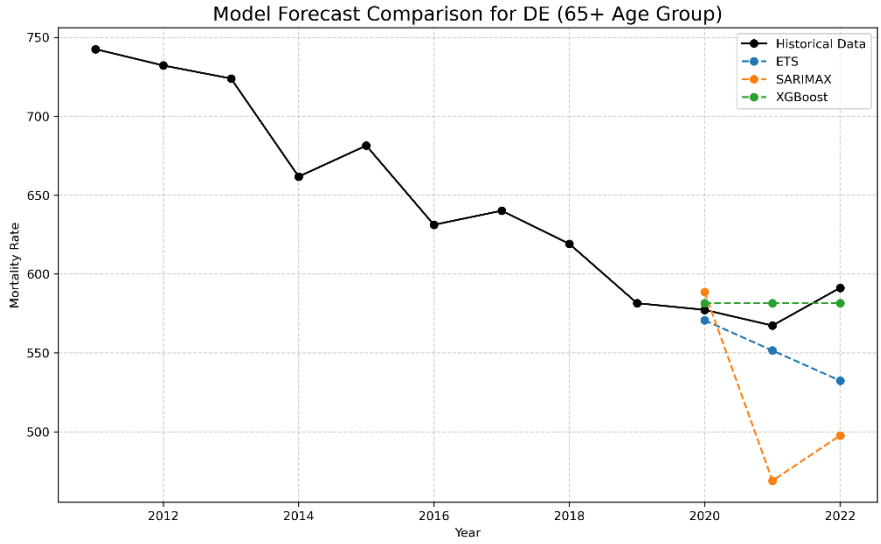

Figure 4: Model Forecast Comparison for Germany (65+ Age Group)

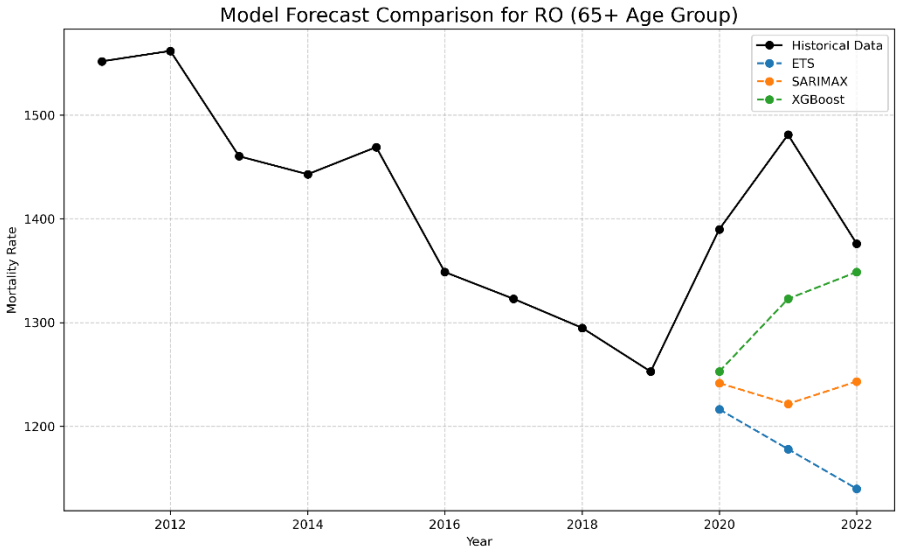

Figure 5: Model Forecast Comparison for Romania (65+ Age Group)

6.3.2 Residual Analysis

To provide a more granular view of forecast performance, the year-by-year residuals (actual - predicted) were calculated. Table 4 displays these residuals for Germany and Romania model. A positive value indicates an under-prediction, while a negative value indicates an over-

prediction. This table highlights specific years where models struggled; for instance, all models significantly under-predicted the high mortality rate for Romania's 65+ cohort in 2021.

Table 4: Detailed Forecast Residuals (Actual - Predicted)

Country	Age Group	Year	ETS	SARIMAX	XGBoost
DE	65+	2020	6.65	-10.66	-4.22
		2021	15.84	90.48	-14.19
		2022	58.81	81.28	9.72
RO	65+	2020	173.58	147.39	137.17
		2021	303.15	264.76	158.25
		2022	236.34	140.31	27.46

6.3.3 Statistical Comparison of Forecasts

Finally, to determine if the observed differences in forecast accuracy were statistically meaningful, the Diebold-Mariano test was conducted for all model pairs across all scenarios. The null hypothesis of the test is that both models have equal predictive accuracy. In the majority of cases, the test failed to reject the null hypothesis, indicating that despite one model having a lower RMSE, the difference in accuracy was not statistically significant. However, for Spain's 65+ cohort, the differences were significant, with the test confirming that ETS was superior to SARIMAX, and XGBoost was superior to SARIMAX. This highlights the importance of statistical testing in validating forecast model comparisons.

This section has presented the quantitative results of the study. The findings from the panel regression analysis provide statistical evidence supporting the hypotheses related to the effects of income, education, and their interaction on IHD mortality, while also yielding nuanced results regarding preventive healthcare expenditure. The comparative forecasting evaluation demonstrated the superior aggregate performance of the XGBoost machine learning model, but also highlighted the context-dependent nature of forecast accuracy and the importance of formal statistical tests.

6.4 Discussion

The discussion is structured to mirror the dual focus of the analysis. It begins by interpreting the findings from the panel regression models, addressing each of the study's primary hypotheses in turn. This section will dissect the complex relationships between income, education, healthcare expenditure, and Ischaemic Heart Disease (IHD) mortality, paying close attention to the significant differences observed between the premature (<65) and later-life (65+) age cohorts. The second part of the chapter discusses the outcomes of the comparative forecasting evaluation, moving beyond a simple declaration of the 'best' model to explore the

nuanced question of which model works best under which circumstances. Finally, the chapter synthesises these insights to consider the study's overall contribution, its inherent limitations, and the tangible policy recommendations and future research directions that emerge from the evidence.

6.5 Interpreting the Drivers of IHD Mortality

The central aim of this thesis was to deconstruct the factors driving IHD mortality across Europe. The final Random Effects model, which employed mean-centering and robust Driscoll-Kraay standard errors, provided a robust platform for testing our hypotheses. The results were both confirmatory of existing theories and revelatory in their nuances.

6.5.1 The Confirmed Protective Effects of Income and Education (H1 & H2)

The analysis yielded strong, statistically significant evidence that higher median income and higher tertiary education attainment are associated with lower IHD mortality rates for both the premature and later-life age groups. This finding robustly supports Hypotheses 1 and 2 and aligns perfectly with the foundational literature on the social determinants of health, from the seminal Whitehall studies (Marmot et al., 1978) to more recent cross-national work (Mackenbach et al., 2008). The negative coefficients for both log_income_centered and tertiary_edu_pc_centered confirm that the socio-economic gradient in cardiovascular health is a powerful and persistent phenomenon across the European continent.

The mechanisms behind these effects are well-theorised. Higher income, as outlined in the materialist framework (Lynch et al., 2000), enables access to a wide range of health-promoting resources, including better nutrition, safer living environments, and higher-quality healthcare. Education, as argued by Cutler and Lleras-Muney (2006), operates through both this economic pathway and a direct cognitive pathway, enhancing health literacy and the ability to make informed, long-term health decisions.

An interesting nuance in our findings is the relative magnitude of these effects. For the 65+ cohort, the coefficient for education (-0.0235) is proportionally larger than for the <65 cohort (-0.0099), suggesting the protective effect of education is more pronounced in later life. This may reflect the cumulative advantage hypothesis, where the benefits of education in terms of lifestyle, health management, and navigating complex care systems compound over a lifetime, becoming most evident in old age when the burden of chronic disease is highest. Conversely, the protective effect of income was stronger for the <65 cohort, which may indicate that income is a more critical determinant of health during the working years, when individuals are more directly exposed to the economic pressures of employment and family provision.

6.5.2 The Paradoxical Finding on Preventive Expenditure (H3)

Hypothesis 3, which posited that increased spending on preventive care would be associated with lower IHD mortality, was not supported in the manner expected. For the 65+ age group, the effect was in the anticipated negative direction but failed to reach statistical significance. More strikingly, for the <65 age group, the coefficient was positive and highly significant, suggesting that higher preventive spending is associated with higher premature mortality.

This counter-intuitive result should not be interpreted as evidence that preventive care is ineffective or harmful. Instead, it is almost certainly indicative of endogeneity, a common

challenge in health economics research (Papanicolas & Smith, 2013). The most likely explanation is reverse causality: countries or regions that are experiencing a worrying rise in premature deaths from IHD are precisely the ones most likely to react by increasing their public health budgets and launching new preventive care initiatives. The model is therefore capturing this reaction, rather than the causal effect of the spending itself. To definitively establish causality, one would need a more advanced econometric design, such as an instrumental variable approach, which was beyond the scope of this thesis but represents a critical direction for future work.

Furthermore, the broad nature of the Eurostat data category "preventive care" (HC.5) may mask the effects of specific, effective interventions. The category includes a wide array of services, from highly effective cancer screening programmes to less effective general health campaigns. The non-significant result for the 65+ group may suggest that broad-based preventive spending is less impactful for this cohort than targeted, well-funded programmes for managing existing chronic conditions. This finding underscores a key policy challenge: it is not just the amount of money spent on prevention that matters, but how intelligently and efficiently it is allocated.

6.5.3 The Synergy of Income and Education: A Key Finding (H4)

Perhaps the most significant and novel finding from the explanatory analysis is the confirmation of Hypothesis 4, revealed through the interaction term. The coefficient for interaction_centered was negative and highly statistically significant for both age groups. This provides strong evidence against a simplistic, additive model of social determinants and suggests that income and education are synergistic.

The interpretation of this result is that the protective effect of a higher national income on IHD mortality is amplified in countries that also have a more highly educated population. In other words, money protects health more effectively when the population has the knowledge and skills to use that economic advantage wisely. This aligns with the theoretical work of authors like Cutler and Lleras-Muney (2006), who argue that education provides the cognitive tools necessary to translate resources into better health outcomes.

This finding has profound policy implications. It strongly suggests that policies aimed at improving public health cannot operate in silos. A strategy focused solely on economic growth, without corresponding investment in education, may yield disappointing health returns. Conversely, improving educational attainment in a low-income setting may be insufficient if people lack the material resources to act on their knowledge. The most effective strategies are likely to be integrated ones that seek to raise both the economic and human capital of the population simultaneously.

6.6 Interpreting the Forecasting Model Comparison

The second major objective of this thesis was to compare the predictive accuracy of different forecasting paradigms. The evaluation of ETS, SARIMAX, and XGBoost models provided a nuanced and insightful answer to the question of which model is 'best'.

The primary finding is that there is no single superior model across all contexts. The state-of-the-art machine learning model, XGBoost, demonstrated the lowest overall forecast error (RMSE) in the majority of scenarios, particularly for the stable German time series and the volatile Romanian later-life series. This is unsurprising, as its algorithm is designed to capture

complex, non-linear patterns and interactions in data that the more linear ETS and SARIMAX models cannot. This provides strong evidence for the utility of incorporating machine learning techniques into the toolkit of public health planning and foresight.

However, the victory of XGBoost was not universal. For the premature mortality cohort in Romania, the simplest univariate model, ETS, produced the most accurate forecast. This is a critical finding, as it demonstrates that for certain types of time series—perhaps those that are shorter or more noisy—the inclusion of external predictors in more complex models like SARIMAX and XGBoost can introduce more statistical noise than signal, ultimately degrading forecast performance. It underscores the data science principle that model complexity should be justified by performance, and that simpler models are often preferable if they are more robust.

Finally, the results of the Diebold-Mariano test add a crucial layer of academic rigor and humility to our conclusions. In most of the pairwise comparisons, the test indicated that there was no statistically significant difference in the predictive accuracy of the models, even when one had a visibly lower RMSE. For example, while SARIMAX had a lower RMSE than ETS for Romania's 65+ cohort, the Diebold-Mariano test ($p=0.0775$) showed this difference was not significant at the conventional 5% level. This is a powerful reminder that observed differences in sample performance do not always translate to a true difference in the underlying quality of the models. It cautions against declaring a "winner" based on marginal improvements in error metrics and highlights the inherent uncertainty and difficulty in forecasting complex social phenomena.

6.7 Limitations of the Study

While this study was designed with methodological rigor, it is essential to acknowledge its inherent limitations, which in turn provide avenues for future research.

- **The Ecological Fallacy:** As a macro-level analysis using country-level data, the findings describe associations between aggregate variables. They cannot and should not be used to make inferences about individual risk. The relationships observed at the national level may not hold for individuals within those nations.

- **Omitted Variable Bias:** The models, while controlling for several key factors, inevitably omit others. The most significant omission is the lack of consistent, cross-national panel data on behavioural risk factors such as smoking prevalence, dietary patterns, and levels of physical activity. These variables are known to be strong predictors of IHD and are also correlated with income and education. Their absence means that the coefficients in our model may be partially capturing the effects of these unobserved lifestyle factors, potentially leading to a degree of omitted variable bias.

- **Endogeneity and Causality**: While the use of panel data models and robust standard errors strengthens the analysis, this study cannot definitively establish causality. As discussed with the preventive expenditure variable, the potential for reverse causality remains. While we model income as a determinant of health, it is also plausible that poor population health could negatively impact a nation's economic productivity. Establishing a truly causal claim would require a more advanced research design, such as one using natural experiments or instrumental variables.

- **Data Granularity:** The study is limited by the level of aggregation in the Eurostat data. The "preventive care" expenditure category, for example, is very broad. A more granular breakdown of this spending would be needed to identify which specific types of preventive interventions are most effective.

6.8 Ethical Considerations

This research exclusively uses publicly available, aggregate, country-level data provided by Eurostat. The data is fully anonymised and contains no information that could be used to identify individual persons. As the study does not involve human subjects, their data, or their tissues, it does not fall under the purview of ethical review boards for human subject research. The commitment to using public data and providing reproducible code ensures transparency and integrity.

7 Conclusion and Future Work

The empirical analysis was structured around two primary goals: explaining the drivers of mortality and comparing the accuracy of different forecasting paradigms. The results from each component provided clear answers to the research questions posed at the outset.

The primary explanatory analysis, conducted using a Random Effects panel model with robust Driscoll-Kraay standard errors, confirmed several core hypotheses. In strong support of H1 and H2, both higher median income and greater tertiary education attainment were found to be powerful, statistically significant protective factors against IHD mortality for both the premature (<65) and later-life (65+) age cohorts. This reinforces the foundational concept of the socio-economic gradient in cardiovascular health across Europe. Hypothesis H3, concerning the role of preventive healthcare expenditure, was not supported as anticipated. The effect was not statistically significant for the 65+ group and was paradoxically positive for the <65 group, a finding attributed to endogeneity and the likely reverse causal pathway where rising mortality prompts increased spending. Perhaps the most significant explanatory finding was the strong support for H4. The analysis revealed a statistically significant negative interaction between income and education, indicating that these two factors work synergistically. The health benefits of a higher national income are amplified in countries with a more educated populace, suggesting that their combined effect is greater than the sum of their individual parts.

The second phase of the analysis involved a comparative evaluation of three distinct forecasting models: a univariate ETS model, a multivariate statistical SARIMAX model, and a non-linear machine learning XGBoost model. The results demonstrated that no single model was universally superior. The machine learning model, XGBoost, generally provided the most accurate forecasts, particularly for the stable German time series and the more volatile Romanian later-life series. However, for the premature mortality cohort in Romania, the simpler ETS model performed best, highlighting that increased model complexity does not guarantee improved performance. Crucially, the Diebold-Mariano test (Diebold & Mariano, 1995) revealed that in many cases, the observed differences in forecast accuracy between the

models were not statistically significant. This provides a critical layer of nuance, cautioning against declaring a definitive "winner" and instead highlighting the inherent challenges of forecasting complex public health outcomes.

7.1 Contribution of the Research

This thesis makes a multi-faceted contribution to the fields of public health, health economics, and data science.

- Theoretical Contribution: The study's primary theoretical contribution lies in its detailed, age-disaggregated analysis of the social determinants of health. By demonstrating the differing magnitudes of the effects of income and education on premature versus later-life mortality, it adds empirical weight to life-course theories of health. Furthermore, by identifying and quantifying the synergistic interaction between income and education, this research moves beyond simplistic additive models and provides a more nuanced understanding of how socio-economic advantage translates into better health outcomes.

- Methodological Contribution: This research serves as a case study in applying a robust, multi-stage analytical methodology. First, it demonstrates the importance of moving beyond standard robust standard errors to employ Driscoll-Kraay errors when dealing with cross-nationally dependent panel data, a critical step often overlooked in applied health research. Second, it provides a rigorous, head-to-head comparison of three distinct forecasting paradigms, offering valuable insights into the relative strengths and weaknesses of econometric and machine learning approaches for applied public health problems.

- Practical and Policy Contribution: The findings of this research are directly relevant to policymakers. The clear evidence of a synergistic effect between income and education provides a strong argument for the development of integrated policy strategies. It suggests that public health goals are best achieved not through isolated health interventions, but through coordinated policies that simultaneously aim to improve economic security and educational attainment. The age-specific results further underscore the need for targeted interventions, suggesting different policy levers may be more effective for the working-age population compared to the elderly.

7.2 Directions for Future Research

This study opens up several promising avenues for future investigation.

- **Sub-National Analysis:** Repeating this analysis at a regional (e.g., NUTS 2) level would be a valuable next step. This would allow for an investigation of whether the same patterns hold within countries, providing a much richer understanding of health inequalities.

- **Incorporating Behavioural Data:** A major research effort should be directed towards harmonising the collection of data on key behavioural risk factors across Europe to allow for their inclusion in future panel models. This would help to disentangle the direct effects of socio-economic factors from their indirect effects through lifestyle.
- **Advanced Econometric Models:** To better address the issue of endogeneity, future research could employ dynamic panel models, such as the Arellano-Bond GMM estimator, which are specifically designed to handle persistence and potential reverse causality in panel data.
- **Qualitative Inquiry:** While this quantitative study identifies what the key drivers are, it cannot fully explain how they operate in people's lives. Qualitative research, through interviews and case studies, could provide a rich, narrative understanding of the lived experiences behind the statistics, exploring how individuals translate their economic and educational resources into health-promoting actions.

By pursuing these avenues, future research can build upon the foundations laid by this thesis to develop an ever more nuanced and actionable understanding of how to tackle Europe's leading cause of death.

7.3 Concluding Thoughts

The challenge of Ischaemic Heart Disease in Europe is not a purely clinical problem; it is deeply intertwined with the social and economic structures of society. This thesis has demonstrated that factors such as a nation's income, the educational attainment of its people, and the interplay between them are powerful determinants of cardiovascular mortality. It has shown that the impact of these forces differs across the lifespan, demanding a nuanced and age-specific policy response. While medical advancements will continue to be vital, the findings of this research reaffirm a fundamental principle of public health: to build a healthier society, we must look beyond the hospital and address the foundational social and economic conditions that shape our lives from beginning to end. The pursuit of health equity is not merely a matter of healthcare; it is a matter of creating societies that offer every citizen the opportunity to thrive.

References

Marmot, M. (2004). The status syndrome: How social standing affects our health and longevity. Times Books.

Unal, B., Critchley, J. A., & Capewell, S. (2004). Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000. Circulation, 109(9), 1101–1107.

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. American Psychologist, 49(1), 15–24.

Crimmins, E. M. (2004). Trends in the health of the elderly. Annual Review of Public Health, 25, 79–98.

Cutler, D. M., & Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence (NBER Working Paper No. 12352). National Bureau of Economic Research.

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. Stata Journal, 7(3), 281-312.

Laaksonen, M., Rahkonen, O., Karvonen, S., & Lahelma, E. (2005). Socioeconomic status and smoking: analysing inequalities with multiple indicators. The European Journal of Public Health, 15(3), 262-269.

Lantz, P. M., House, J. S., Lepkowski, J. M., Williams, D. R., Mero, R. P., & Chen, J. (1998). Socioeconomic factors, health behaviors, and mortality: results from a nationally representative prospective study of US adults. JAMA, 279(21), 1703-1708.

Lynch, J., Smith, G. D., Harper, S., & Hillemeier, M. (2004). Is income inequality a determinant of population health? Part 1. A systematic review. The Milbank Quarterly, 82(1), 5-99.

Mackenbach, J. P., Stirbu, I., Roskam, A. J. R., Schaap, M. M., Menvielle, G., Leinsalu, M., & Kunst, A. E. (2008). Socioeconomic inequalities in health in 22 European countries. New England Journal of Medicine, 358(23), 2468–2481.

Marmot, M. G., Rose, G., Shipley, M., & Hamilton, P. J. S. (1978). Employment grade and coronary heart disease in British civil servants. Journal of Epidemiology & Community Health, 32(4), 244–249.

Marmot, M. G., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., ... & Feeney, A. (1991). Health inequalities among British civil servants: the Whitehall II study. The Lancet, 337(8754), 1387-1393.

Nolte, E., & McKee, M. (2008). Measuring the health of nations: updating an earlier analysis. Health Affairs, 27(1), 58-71.

Or, Z. (2000). Determinants of health outcomes in industrialised countries (OECD Economic Studies No. 30). OECD Publishing.

Pampel, F. C., Krueger, P. M., & Denney, J. T. (2010). Socioeconomic disparities in health behaviors. Annual Review of Sociology, 36, 349–370.

Papanicolas, I., & Smith, P. C. (Eds.). (2013). Health system performance comparison: An agenda for policy, information and research. Open University Press.

Wilkinson, R. G., & Pickett, K. E. (2009). The spirit level: Why more equal societies almost always do better. Allen Lane.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT Press.