The Relationship between Occupation Characteristics and Mortality in England and Wales, 2003–2017

Christopher McDonald

Kellogg College, University of Oxford

September 2021

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science

MSc in Statistical Science
DECLARATION OF AUTHORSHIP

Please submit the completed form with your dissertation.

Name (in capitals): CHRISTOPHER MCDONALD Candidate number: 1023069

College (in capitals): KELLOGG Supervisor: Prof. Robin Evans

Title of dissertation (in capitals): THE RELATIONSHIP BETWEEN OCCUPATION CHARACTERISTICS AND MORTALITY IN ENGLAND AND WALES, 2003-2017

Word count: 11431

Please tick to confirm the following:

I have read and understood the University's disciplinary regulations concerning conduct in examinations and, in particular, of the regulations on plagiarism (The University Student Handbook Section 8.8; available at https://www.ox.ac.uk/students/academic/student-handbook).	☒
I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at http://www.ox.ac.uk/students/academic/guidance/skills?wssl=1	☒
The dissertation I am submitting is entirely my own work except where otherwise indicated.	☒
It has not been submitted, either partially or in full, for another qualification of this University (except where the Special Regulations for the subject permit this), or for a qualification at any other institution.	☒
I have clearly signaled the presence of all material I have quoted from other sources, including any diagrams, charts, tables or graphs.	☒
I have clearly indicated the presence of all paraphrased material with appropriate references.	☒
I have acknowledged appropriately any assistance I have received in addition to that provided by my supervisor.	☒
I have not copied from the work of any other candidate.	☒
I have not used the services of any agency providing specimen, model or ghostwritten work in the preparation of this dissertation. (See also section 2.4 of Statute XI on University Discipline under which members of the University are prohibited from providing material of this nature for candidates in examinations at this University or elsewhere: http://www.admin.ox.ac.uk/statutes/352-051a.shtml).	☒
I agree to retain an electronic version of the work until the publication my final examination result. I agree to make any such electronic copy available to the examiners should it be necessary to confirm my word count or to check for plagiarism.	☒

Candidate's signature: cmcdonald Date: 19/09/2021

Acknowledgements

First, I would like to express my gratitude to my supervisor Prof. Robin Evans, for his generous support and guidance throughout this project. The project would not have been possible without him. I would also like to thank Dr Neil Laws, my course supervisor, for his helpful advice during the year.

Third, I am extremely grateful to Alison Sizer of UCL and the Centre for Longitudinal Study Information & User Support (CeLSIUS). Her ideas and suggestions were invaluable, and the project would not have been possible without her tireless patience and support, which covered all aspects of the project.

Finally, I am eternally grateful to partner Francisca Vasconcelos, and to my parents, whose consistent support and encouragement were the reason I was able to study in Oxford this year.

The permission of the Office for National Statistics to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is funded by the ESRC under project ES/V003488/1. The author alone is responsible for the interpretation of the data.

This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Abstract

Jobs are an important determinant of mortality, influencing health through both the direct effects of physical exposure and psychological conditions, and the indirect effects of income, social status and lifestyle. This project investigates the relationship between job characteristics and incidence of mortality. 51 interpretable characteristics were extracted by factor analysis from the US Occupational Information Network (O*NET) database, and linked at the occupation level to the UK Office for National Statistics Longitudinal Study. The sample consisted of $n = 279,368$ individuals from England and Wales whose occupations were recorded at the 2001 UK census. Cox proportional hazards regression models were used to examine the relationship between the occupation features and mortality over 15 years of follow-up (2003–2017). Models were fit which adjusted for socioeconomic status—as measured by education, marital status, housing status and vehicle ownership—and for self-reported health in 2001.

Repetitive work contexts—characterised by repetitive tasks, repetitive physical motions, high levels of automation and time pressure—predict the greatest mortality risk, with a hazard ratio of 1.39 (1.24, 1.56) under the model adjusted for socioeconomic status. Other results support previous findings, such as greater job control (here measured as "self-directedness") and more technical occupations predicting lower mortality (Bosma *et al.* 1997; Lee 2011). These results are likely due to a mixture of the direct and indirect effects of occupations, as well as potential residual confounding. Further evaluation of the results—including model goodness-of-fit and sensitivity to alternative analysis choices—would be desirable.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

For working individuals, jobs consume a large fraction of daily activities. They are a substantial determinant of individuals' resources and social status, and are closely related to lifestyle factors such as diet, social life, sleep, and exercise.

The relationships between the workplace, the individual and the environment in determining mortality are complex (N. J. Johnson *et al.* 1999). Differences in exposure to physical hazards, such as disease, radiation, dust, toxic substances, hazardous work conditions or equipment, are one pathway by which jobs influence mortality (Baxter *et al.* 2010). More recently, interest has grown in the relationship between mortality and 'psychosocial' job conditions, for example self-perceived social status, job demands and job control, or social support (J. V. Johnson and Hall 1988; Bosma *et al.* 1997). The relationship between income—a key feature of jobs—and mortality is long established (Backlund *et al.* 1996; M. Marmot 2005).

This project investigates the relationship between a variety of job characteristics and incidence of mortality. Occupation-level features are taken from the US Occupational Information Network (O*NET) database, and linked to the UK Office for National Statistics Longitudinal Study (ONS LS). The relationship between the occupation features and mortality is assessed for 279,368 individuals of working age, 25-64, over a period of 15 years. The results identify interpretable characteristics of jobs which predict mortality across the working population of England and Wales.

1.2 Previous work

In England, the relationship between occupation and mortality has been studied since at least the mid-19th century, when the British Registrar General produced estimates of mortality rates by occupation based on the 1851 UK census (Registrar General 1855). Similar statistics were produced over the next 100 years based on the 10-yearly censuses carried out in the UK (Fox and Adelstein 1978). Following these trends, Katikireddi *et al.* (2017) report mortality rates by occupation in England and Wales over the period 2001-2011. They find that health professionals, teachers and managers experienced the lowest mortality rates, while construction and factory workers experienced the highest mortality rates.

A weakness of this approach is that it does not account for confounding—that is, differences between occupational groups (e.g. in income, status, or health behaviour) which might otherwise explain differences in mortality. This issue can be framed as the choice of a suitable reference mortality rate. For example, to isolate the direct effects of occupation, past studies compare mortality of employed indi-

viduals to their spouses' mortality, or to individuals in the same social class. Fox and Adelstein (1978) compares these two methods, and using a χ^2 statistic cautiously concludes that around 80% of variation in mortality between occupational groups is explained by social class, with specific occupation accounting for only 20%. In contrast, N. J. Johnson *et al.* (1999) suggests that, within the US during the 1980s, specific occupation predicted mortality risk more strongly than social class.

Another potential issue with this approach comes from health-related selection into occupations, known in this context as the "healthy-worker effect" (Fox *et al.* 1982). It is long-established that employed individuals tend to be in better health than comparable non-employed individuals (McMichael 1976); it is also reasonable that occupations vary in the degree to which they select individuals with better or worse health. Fox and Collier (1976) find that this "healthy-worker" tends to diminish over time; this may be explained either as worsening health over time due to employment, or more simply as regression of individuals' health to the mean.

Many causal mechanisms or pathways have been proposed to explain the observed relationship between occupation and mortality. These can broadly be categorised as direct, through occupation-related exposure, or indirect, for example through income, prestige, or health behaviour. The most obvious direct mechanism is via physical job conditions, such as exposure to hazardous conditions (Baxter *et al.* 2010) or physical stresses (Fletcher *et al.* 2011). Analysis of work-related mortality by Coggon *et al.* (2010) indicates that mortality directly attributable to work in England and Wales has declined substantially since 1980. Other approaches look at 'psychosocial' job conditions, such as self-perceived social status, job demands and job control, or social support. M. G. Marmot *et al.* (1991) suggests that jobs which combine high demands with low levels of control may increase the risk of coronary heart disease. Self-perceived social status is closely related to occupation and also predicts greater mortality (Tang *et al.* 2016). Lee (2011) looks at a variety of psychosocial occupation characteristics and concludes that "Job IQ"—the degree to which the occupation requires originality, reasoning and other cognitive abilities—may be the factor which most consistently and strongly relates to mortality.

Occupation may also affect health through indirect mechanisms. These include the income, prestige and social ties gained from an occupation, as well as occupation-influenced preferences and constraints on housing, nutrition, sleep, exercise, drug use, health care, neighbourhood characteristics (e.g. air quality) and other determinants of health.

In summary, studies linking job characteristics and mortality vary according to: (1) analysed causes of mortality, (2) considered effect types (physical, psychosocial or indirect), (3) examined groups (nationally representative samples vs individual

workplaces), (4) analysis of job characteristics at the job- versus occupation-level, (5) consideration of occupation as a measure of social status versus an exposure variable, and (6) use of statistical techniques (e.g. standardisation vs statistical models). Studying mortality rates by occupation is useful for identifying particular occupations which strongly predict mortality; however this type of study does not necessarily say much about the mechanisms by which higher or lower mortality occurs. Studying mortality directly attributable to work more accurately identifies causal mechanisms, but is limited to the most direct effects of occupation on mortality. Studying the relationship of job characteristics to mortality may give the most detailed picture of their relationship.

1.3 Approach

This work investigates the relationship between a variety of job characteristics and likelihood of mortality. More specifically, the project quantifies the strength of relationships between job characteristics and mortality across the population of England and Wales. The studied characteristics are relatively simple, interpretable and easy to measure.

The methods used to achieve these objectives are summarised in figure 1; the approach followed is similar to Lee (2011). Individual-level data were taken from the Office for National Statistics Longitudinal Study (ONS LS), the largest Longitudinal data resource in England and Wales. The selected sample consists of the $n = 279,646$ members of the LS who reported an occupation at the 2001 UK census. A set of interpretable occupation features—extracted from the US Occupational Information Network (O*NET) database and converted from US to UK occupations—were linked to this sample at the occupation level. Regression models were used to examine relationships between each occupation feature and mortality during 15 years of follow-up (2003–2017). Model goodness of fit was examined, and the reliability of the results was briefly explored by varying the sample selection criteria. Further evaluation of the reliability of the results would be desirable.

Various design choices were made during the above analysis: for example the choice of data sources, data preprocessing methods, sample selection criteria, statistical model, and model predictor variables. These choices can be considered "parameters" of the analysis; they determine both the questions the analysis attempts to answer and the confidence justified in these answers. Parameter choices are the main focus of sections 2 and 4. Where possible, the choices made are justified, and compared to possible alternative options; given the number of choices made it was not possible to discuss them all in detail.

1.4 Outline

The remainder of this report is structured as follows. Section 2 describes the approach of the project, the two main data sources—the O*NET database and the ONS LS—and how they were preprocessed in preparation for modeling. Section 3 provides a graphical summary of the key features of each data source. Further background on the data sources can be found in appendix B. Following this, section 4 describes the statistical methods used to prepare the occupation features and relate them to mortality; the results of this are presented in section 5. Finally, section 6 evaluates the findings with respect to prior expectations, and suggests weaknesses and potential improvements to the methods used.

2 Data Preparation

The first stage of the analysis was the construction of a suitable data set. Specifically, the aim was to produce a data set linking individuals with their job characteristics and mortality. Section 2.1 discusses the approach taken to this problem, and the choice of data sources. Following this, sections 2.2 and 2.3, discuss the methods used and key design choices made when preparing the data for modeling.

2.1 Approach and Choice of Data Sources

The approach taken in this project was to (1) construct a data set of *occupation* features, and (2) link these features to a separate data set of individuals, their measured occupations, and their mortality. Here, an occupation is defined as a cluster of similar jobs, for example nurse or chef. While only one individual can have a given job, many individuals share the same occupation. "Features" of an occupation can be measured as averages or aggregates over its constituent jobs.

The US Occupational Information Network (O*NET) database (National Center for O*NET Development 2021) was used to construct the data set of occupation features. The 2010 version of the O*NET 'Standard Occupational Classification" (O*NET 2010 SOC) distinguishes 1,110 occupations, though data are only collected for 974 'data-level' occupations. For each of these 974 data-level occupations, the O*NET database defines over 400 occupation variables. These occupation variables are organised into high-level categories known as "domains", as shown in figure 2. There is no comparable source of detailed occupational information in the UK (Dickerson and Morris 2019). Further information on the O*NET database can be found in appendix B.1.

The Office for National Statistics Longitudinal Study (ONS LS) was used to construct the data set of individuals, their occupation, and their mortality. It consists of linked census and life events data for a 1% sample of the population

of England and Wales, representing largest longitudinal data resource in these countries (Office for National Statistics 2021). Further information on the LS can be found in appendix B.2. The use of the ONS LS for this project was approved by the UK Statistics Authority's Data Ethics team.

The main advantage of this approach—of linking job characteristics to individuals via their occupations—is the large sample size and substantial scope and detail of the occupation features it allows. This makes it possible to precisely estimate relatively complex models, and examine a number of job features simultaneously, with relatively robust results. Also, using data from across England and Wales ensures that the results will apply generally across the population, rather than just to a specific workplace, occupation or geographic region.

The main weaknesses of the approach are two-fold. First, three types of error are introduced when using occupation features. These are (1) error in the initial measurement of the occupation features; (2) error due to the difference between the occupations for which the features were measured and the occupations to be described; and (3) error due to differences between individuals' job features and the (aggregated) features of their measured occupations. Second, the LS measures a limited number of variables—for example, individuals' incomes are not measured—and has a low frequency of updates—once per 10 years for census data. This increases the level of noise in the results, and makes it more difficult to isolate the relationship between occupation features and mortality which is independent from other factors. More practically, a further weakness of the LS data is its strict confidentiality rules, which make it challenging to access, analyse, and report results.

2.2 O*NET

The task of preparing the occupation feature data set was broken into three stages. First, an initial set of O*NET variables was extracted (section 2.2.1). $p = 246$ variables from eight different domains were selected, using version 17.0 of the O*NET database. Second, dimensionality reduction was carried out, with the aim of creating a smaller and more interpretable set of occupation features (section 2.2.2). The choice of dimensionality reduction technique is discussed below; the theory and implementation of the chosen method—factor analysis—is discussed later in section 4. This resulted in a reduced set of $q = 51$ occupation features. Finally, the occupation features were converted from O*NET 2010 SOC to the UK occupation classification (section 2.2.3). The result of this process was a data set of occupation features which can be linked to individuals via their measured occupations.

The methods used and key design choices made during these stages are discussed in more detail below.

2.2.1 Variable Selection

More than 40 versions of the O*NET database have been published since the first release in 1998[1]. A number of considerations were relevant when deciding which of these versions to use. First, it was required that the O*NET version use O*NET 2010 SOC (rather than the 2000 or 2020 versions), since this version is most easily and accurately mapped to the UK 2000 SOC (see section 2.2.3). This restricted the possible versions to those published between 2010 and 2020 (versions 15.1–25.0). Second, there was a trade-off between the quality and completeness of the data in the O*NET version, and its validity for the purposes of this project. The quality and completeness of data increases in later O*NET versions, as more data were collected and refinements to methods were made. On the other hand, the extent to which the variables accurately describe individuals' occupations over the full follow-up period 2003–2017 likely decreases with later O*NET versions, since occupation characteristics change over time. Version 17.0 of the O*NET database (published in July 2012) was selected as a compromise between these two considerations.

Each version of the O*NET database contains over 400 variables arranged into 6 broad categories and further broken down into "domains". Each domain describes a different aspect of occupations. As shown in figure 2, these domains vary according to whether their corresponding variables are comparable across occupations ("Cross Occupation") or specific to individual occupations ("Occupation Specific"), and whether they mainly describe the "character" of occupations ("Job-Oriented") or the individuals who fill them ("Worker-Oriented").

Since the aim of the project was to explore the relationship between a variety of occupation characteristics and mortality outcomes, as many relevant domains were included as possible. To be included, domains were simply required to: be comparable across occupations ("Cross Occupation" rather than "Occupation Specific"), be measured as numeric variables, have a plausible relationship with mortality / health, and apply meaningfully to UK (as well as US) occupations.

For example, the "Intermediate / Detailed Work Activities" domains were not included, as they are not comparable across occupations; descriptors in these domains only apply to certain groups of occupations. The "Education" domain ("Prior educational experience required to perform in a job") was not included since education methods, certification and occupational requirements generally differ substantially between the US and the UK.

The resulting data consisted of eight domains, containing a total of 407 variables. For four of these domains—knowledge, skills, abilities and work activities—both the "importance" and "level" of each occupation characteristic is measured.

[1] `https://www.onetcenter.org/db_releases.html`

However, a high degree of correlation (typically > 0.9) is observed between these two measurements (Handel 2016). To simplify the interpretation of the results, variables measuring the "level" of occupation characteristics were dropped. The final variable set consisted of 246 variables, across eight domains. They are summarised in table 1.

2.2.2 Dimensionality Reduction

Dimensionality reduction can be defined as the transformation of high-dimensional data into meaningful low-dimensional representations (Van Der Maaten *et al.* 2009). Techniques for dimensionality reduction are used for a variety of purposes including classification, visualisation, and data compression. In this case, the aim was to reduce the initially selected $p = 246$ O*NET variables to a smaller and more interpretable feature set, for use as an input to regression models.

Many techniques have been proposed for dimensionality reduction: Espadoto *et al.* (2021) lists 44 techniques discussed in recent reviews of dimensionality reduction in machine learning and information visualisation. They suggest that techniques vary along eight main "traits": for example input type, ease of use, linearity and computational complexity. Among the techniques with traits most appropriate for this project are principal component analysis (PCA), probabilistic PCA (PPCA), sparse PCA (SPCA), factor analysis (FA), independent component analysis (ICA), and locality preserving projection (LPP). The technique selected was factor analysis. This choice was based on the interpretability of its results (Gorsuch 1983), and its empirically-measured performance on similar tasks (Espadoto *et al.* 2021). Here interpretability can be defined as the ease with which meaningful labels can be applied to the dimensions of the resulting latent variables.

More specifically, six of the eight variable domains were reduced separately using a robust variant of the normal linear factor model (Bartholomew *et al.* 2011). The remaining two domains—Interests and Work Values—were left in their original form. The methods used and results found from these factor analyses are presented in sections 4 and 5. The resulting data set consists of $q = 51$ occupation features, organised into eight domains.

2.2.3 Mapping

The occupation features produced by dimensionality reduction were initially coded to the O*NET 2010 SOC. By contrast, in the LS data set (section 2.3) occupation was measured according to the UK 2000 SOC, the occupation coding scheme used for the 2001 census. Therefore it was necessary to map the occupation features from the O*NET 2010 SOC to the UK 2000 SOC. This task was achieved by mapping the occupation feature data set: (1) from the O*NET 2010 SOC to the

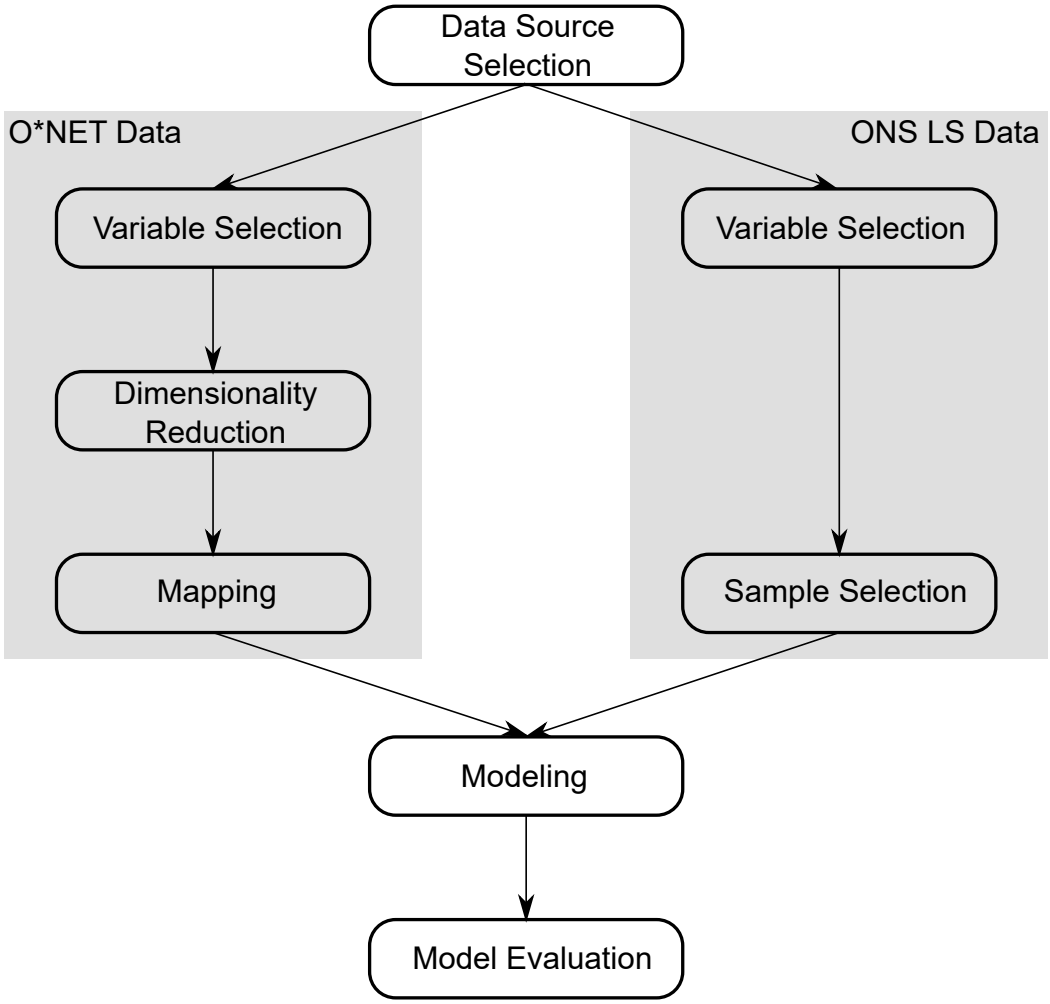


Figure 1: Summary of analysis stages.

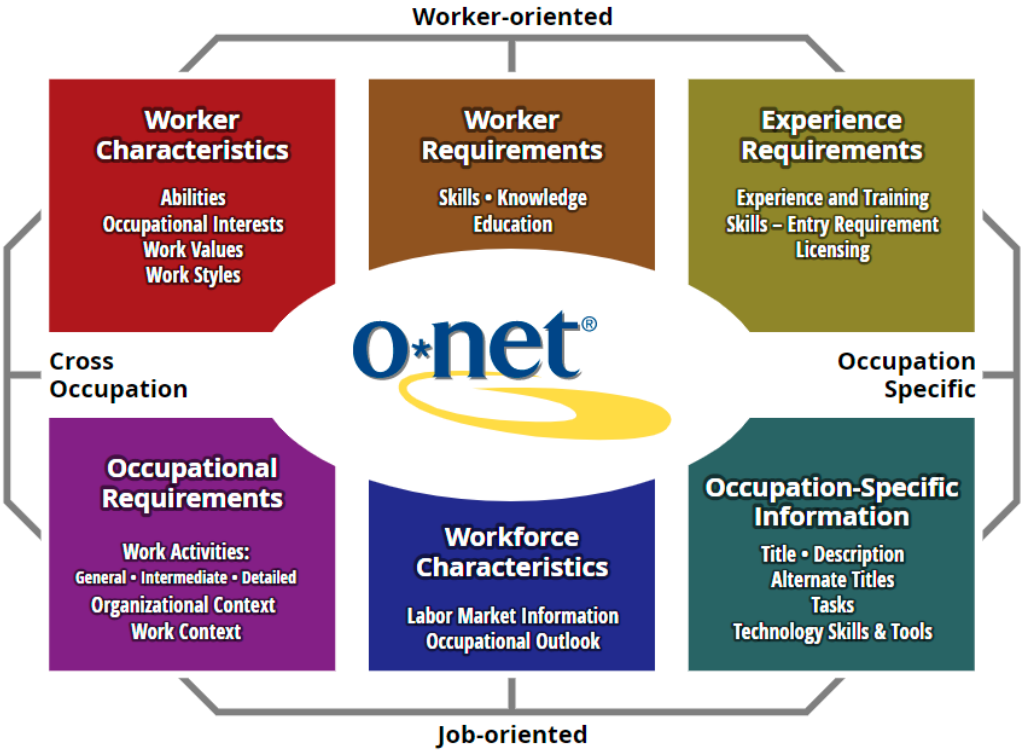


Figure 2: The O*NET content model. From "The O*NET®Content Model" (by the National Center for O*NET Development, used under the CC BY 4.0 license).

Table 1: Summary of O*NET domains used in this study.

Category	Domain	Description	Rater	Range	Variables	
					Original	Reduced
Worker Characteristics	Abilities	(Importance of) Enduring attributes of the individual that influence performance	Analyst	1-5	52	7
Worker Characteristics	Interests	Preferences for work environments and outcomes	Analyst	1-7	6	-
Worker Requirements	Knowledge	(Importance of) Organized sets of principles and facts applying in general domains	Job Incumbent	1-5	33	8
Worker Requirements	Skills	(Importance of) Developed capacities of the individual that influence performance	Analyst	1-5	35	6
Occupational Requirements	Work Activities	(Importance of) Work activities that are common across a very large number of occupations	Job Incumbent	1-5	41	7
Occupational Requirements	Work Context	Physical and social factors that influence the nature of work	Job Incumbent	1-3 / 1-5	57	7
Worker Characteristics	Work Styles	(Importance of) Personal characteristics that can affect how well someone performs a job	Job Incumbent	1-5	16	4
Worker Characteristics	Work Values	Work values and needs that are reinforced or satisfied by the job	Analyst	1-7	6	-
Total					246	51

UK 2010 SOC (the 2010 version of the UK occupation classification system); and then (2) from the UK 2010 SOC to the UK 2000 SOC.

The same basic method was used for both of these two mappings. First, a matrix of (normalised) weights W was constructed. The elements of W describe the weight of each 'input' occupation as a component of each 'output' occupation. Specifically, the weight w_{jk} can be thought of as the fraction of individuals classified into occupation j in the 'output' SOC system, who were classified into occupation k in the 'input' SOC system. Next, let x_{ki} denote the value of the occupation feature i for input occupation k, and y_{ji} denote the value of the same feature for output occupation j. The formula used to convert between occupation sets was a simple weighted average[2]

$$y_{ji} = \sum_k w_{jk} x_{ki} \tag{2.1}$$

or in matrix form

$$Y = WX \tag{2.2}$$

where the rows of X correspond to the occupation features for each 'input' occupation, and the rows of Y are the occupation features for each 'output' occupation.

Following the methods of Dickerson and Morris (2019), two data sources were used to construct the weight matrix for the O*NET 2010 SOC to UK 2010 SOC mapping. These were (1) an O*NET SOC – UK SOC matching matrix, produced in October 2020 by the "Labour Market Information (LMI) for All" project[3] (UK Department for Education 2021); and (2) the May 2011 version of the US Bureau of Labor Statistics' (BLS) Occupational Employment Statistics (OES) (US Bureau of Labor Statistics 2021). The weight matrix for the UK 2010 SOC to UK 2000 SOC mapping was derived directly from ONS correspondence tables between these two SOCs (Office for National Statistics 2012). Further details of the mapping procedure can be found in appendix D.

Finally, each mapped occupation feature y_{ji} was standardised to take values in the interval [0, 1] by the linear transformation

$$\hat{y}_{ji} = \frac{y_{ji} - \min_k(y_{ki})}{\max_k(y_{ki}) - \min_k(y_{ki})}. \tag{2.3}$$

2.3 ONS LS

The task of preparing the LS data was broken into two stages. First, a set of LS variables was constructed (section 2.3.1). 20 variables from the 2001 census

[2]While other formulas are possible, for example $y_{ji} = x_{ki}\mathbb{1}(k = \text{argmax}_l\, w_{li})$, a simple weighted average is arguably the most reasonable choice.

[3]`https://www.lmiforall.org.uk/`

were selected initially, from which further variables were derived. A total of 11 LS variables were used as predictors in the final regression models. Second, a suitable sample was derived (section 2.3.2). Individuals were included in the sample if they were aged 25-64 and reported a valid occupation at the 2001 census, and if they were still alive in 2003. These criteria resulted in a sample of $n = 279,646$ individuals, who experienced $21,822$ mortality events over the follow-up period of 2003–2017.

The methods used and key design choices made during these stages are described in more detail below.

2.3.1 Variable Selection

The choice of census version involved a trade-off between two factors.

One the one hand, a longer follow-up period (and so earlier census) is desirable as it allows model parameters to be more accurately and robustly estimated. For example, using data from the 1971 census would provide more than forty years of follow-up data and of the order of 100,000 mortality events.

On the other hand, the validity of the analysis is likely greater when a more recent census and a shorter follow-up time is used. For example, the O*NET job characteristic data are only available beginning in 2001, with the quality and completeness of data increasing over time (see section 2.2). Also, individuals' occupations (and other characteristics) are only sampled at a single point in time (at the census)[4], after which their life circumstances are likely to change. These considerations favour using a more recent census—for which the O*NET data better match the true characteristics of the occupations they claim to describe—and a shorter follow-up period—to reduce the amount that individuals' characteristics drift during this time. The 2001 census was chosen as a compromise between these two competing factors.

Data from the 2001 census comprise over 100 variables in the LS, and contain a variety of information on individuals and their households. The overall aim of the variable selection process was to isolate as best as possible the association of occupation and mortality, independent of other factors. A total of 20 variables were initially selected; these are listed in appendix C.1. The resulting variables include information on individuals' biology (age, sex and ethnicity), social and physical environment (marital status, education level, occupation, geographic region), and self-reported health status.

A smaller set of 11 variables was then derived with the aim of summarising the information contained in the original 20 variables. Data reduction was per-

[4]This could be resolved by using individuals' occupations recorded at each census, to obtain a data point every 10 years. However, for feasibility reasons, and the other problems mentioned, only a single census version was used.

formed both to increase the performance of the models fit—by reducing collinearity between variables, and reducing model size—and to increase the interpretability of the results—by reducing the total number of relationships to consider, and having each variable represent distinct constructs. This process consisted of (1) summarising groups of related variables, and (2) grouping (or binning) similar values of individual variables. Groups of variables were summarised by identifying the sets of original variables which measure similar constructs, and then deriving a variable which summarises the information contained in these sets. Values of individual variables were binned where the information they contained was too detailed, or where differences between different levels of variables were not of interest. The full set of original and derived variables is shown in appendix C.1.

2.3.2 Sample selection

The initially selected sample consisted of individuals who (1) answered "Yes" to the 2001 census question "Have you ever worked?", (2) were aged 25-64 in 2001, and (3) were alive on 1st January 2003. A fourth criterion—individuals who (4) reported being in work the previous week—was treated as a parameter of the analysis, and the sensitivity of the results to the inclusion of this criterion was tested. The follow-up period was chosen as 2003–2017 inclusive, giving a total of 15 years of follow-up.

These criteria were chosen based on two main considerations. First, the sample should represent the group under study: individuals who were working (or previously working) at the 2001 census. This is achieved in the first criterion. Second, the sample selection criteria should aim to maximise the validity of the analysis. This is achieved by criteria (2)-(4) above. For example, individuals younger than 25 were excluded, as these individuals tend to move jobs more frequently (Syed *et al.* 2019) and so their recorded occupation is less representative of their prior and subsequent job-related exposure. Individuals who died prior to 2003 were removed from the sample to reduce potential biases from individuals who were near death in 2001.

The results of the sample selection process are as follows. Of the approximately $1,200,000$ total LS members, around $540,000$ have data entries corresponding to the 2001 census. Subsequent death is recorded for around $80,000$ of these individuals. After applying the sample criteria (1)–(3) above, around $281,000$ individuals remained. Individuals with missing data in one or more variables were then dropped: this group consisted of (1) individuals who not present in their household on the census night ($< 0.01\%$ of the full sample), (2) individuals whose census records have not been linked to their NHS records on the NHSCR ($\sim 0.05\%$), and (3) students living away from home, who were not required to answer certain census questions ($\sim 0.02\%$). This resulted in the final sample

of 279,646 individuals, of whom 21,822 died during the follow-up period. The characteristics of the sample are described in section 3.2.

3 Exploratory Data Analysis

3.1 O*NET

The basic structure of the O*NET data is summarised in figure 3; the "Work Styles" domain and the three occupations defined in table 7 are used as examples. Data from this domain were gathered by surveys to job incumbents; the 95% confidence intervals for the mean indicate the degree to which the survey respondents agreed on the relative importance of each "Work Style" to the occupation. Each variable is rated on a 1–5 scale; ratings in this case are heavily skewed towards the high end. "Chief Executives" score highly on the importance of all work styles, while "Freight Inspectors" and "Metal Fabricators" score less highly.

Figure 4 summarises how different versions of the O*NET database vary. Versions 15.1–25.0 are shown, corresponding to the versions which use the O*NET 2010 SOC system. The fraction of O*NET 2010 SOC occupations with complete data is more than 92% for every domain for version 17.0; by version 21.0 (August 2016) this is above 97%.

The relative standard deviation of a measurement x_{ij} of variable j for occupation i is defined here as

$$r_{ij} = \frac{s_{ij}}{b_j - a_j} \tag{3.1}$$

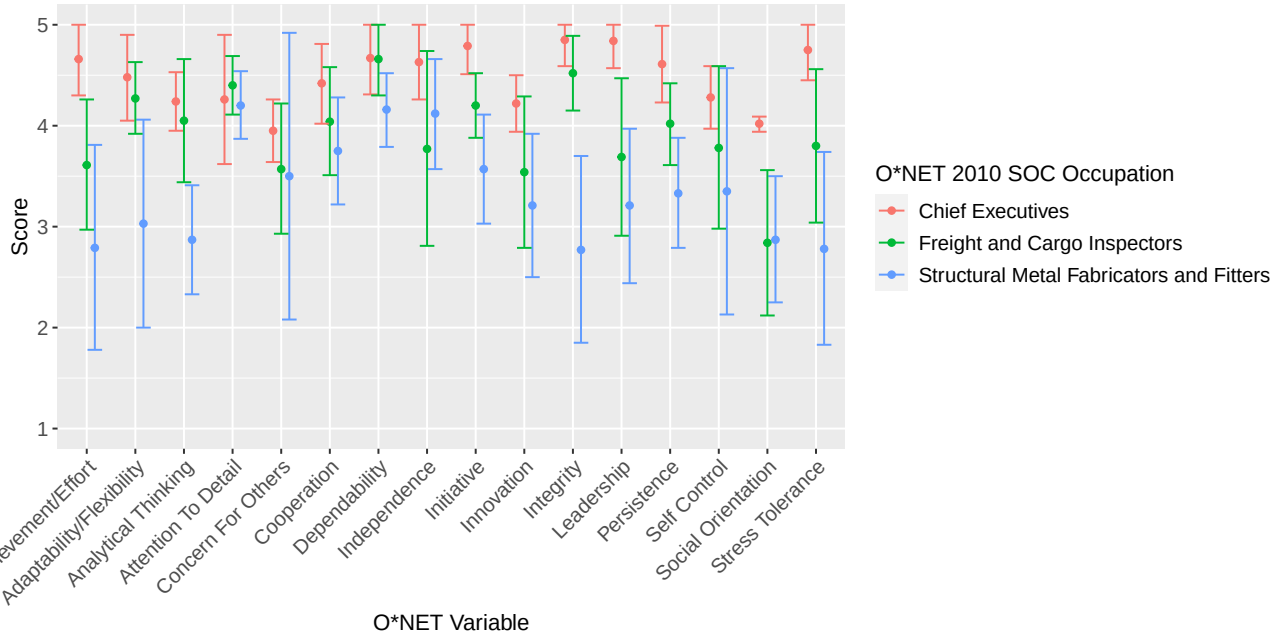
where s_{ij} is the estimated standard error of the measurement, and $x_{ij} \in [a_j, b_j]$ is the range of values which variable j can take. Figure 4b shows the relative standard deviation averaged over occupations and variables, for the six domains where standard errors were derived. In all cases, the average standard deviations are less than 10% of the full range of the variables. The abilities and skills domains appear to be more precisely estimated than the other domains.

Version 17.0 (July 2012) of the O*NET database was used in this report. Figure 5 shows when measurements were last taken for each variable and occupation, across each of the eight domains, for this version of the O*NET database. The majority of occupation data were updated between 2006–2012, with a small minority updated prior to this.

3.2 ONS LS

The extracted data set consists of 279,646 observations and 21,822 mortality events. In this section, its characteristics are summarised using exploratory data analysis plots. Underlying counts for all figures can be found in appendix G.

(a) Ratings and 95% confidence intervals for the "Work Styles" domain, over 3 example occupations.



(b) Distribution of ratings for each of the 16 O*NET variables (not labelled) in the "Work Styles" domain.

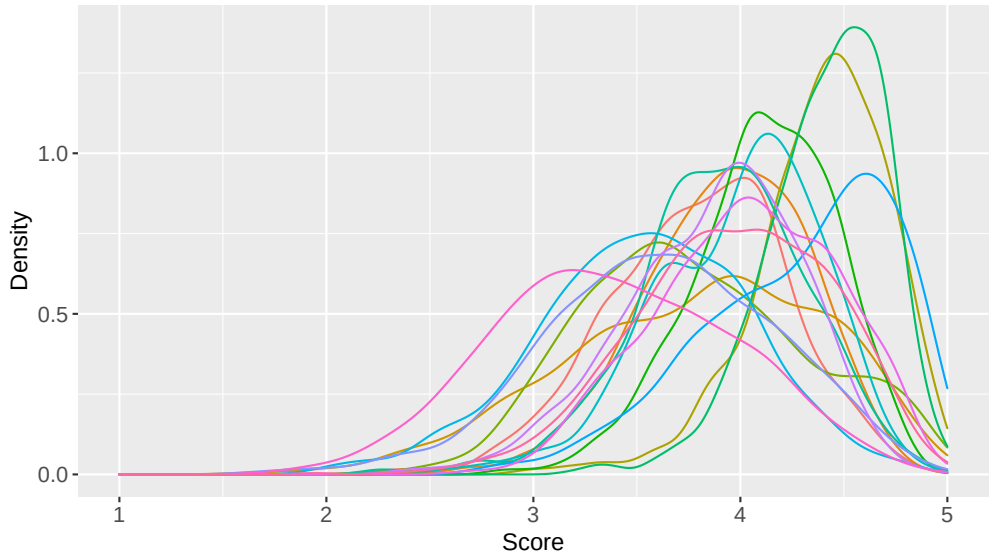
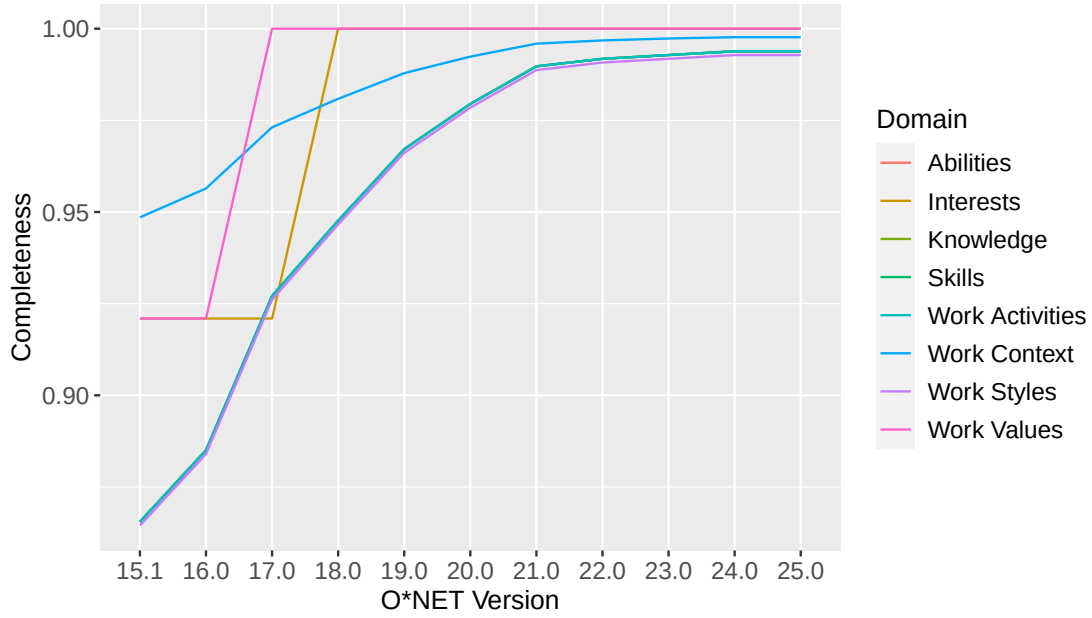


Figure 3: Example O*NET data from the "Work Styles" domain (O*NET version 17.0). The variables describe the importance of different work styles.

(a) Fraction of O*NET 2010 SOC occupations with complete data, by domain and O*NET version.



(b) Relative standard deviation averaged over occupations and variables, by domain and O*NET version.

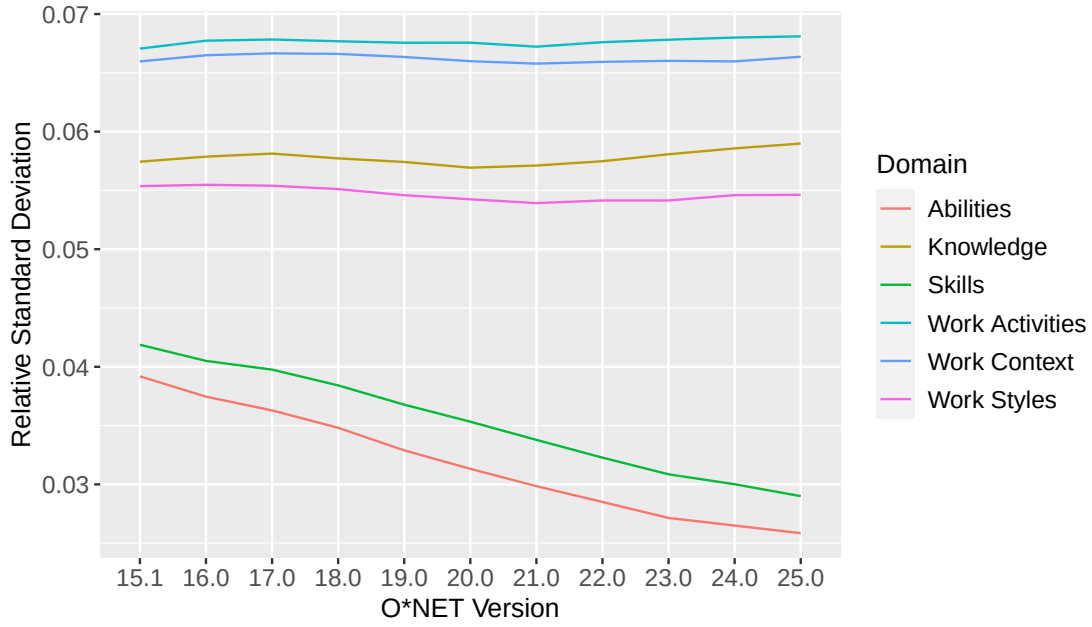


Figure 4: Summary of variation between different O*NET versions.

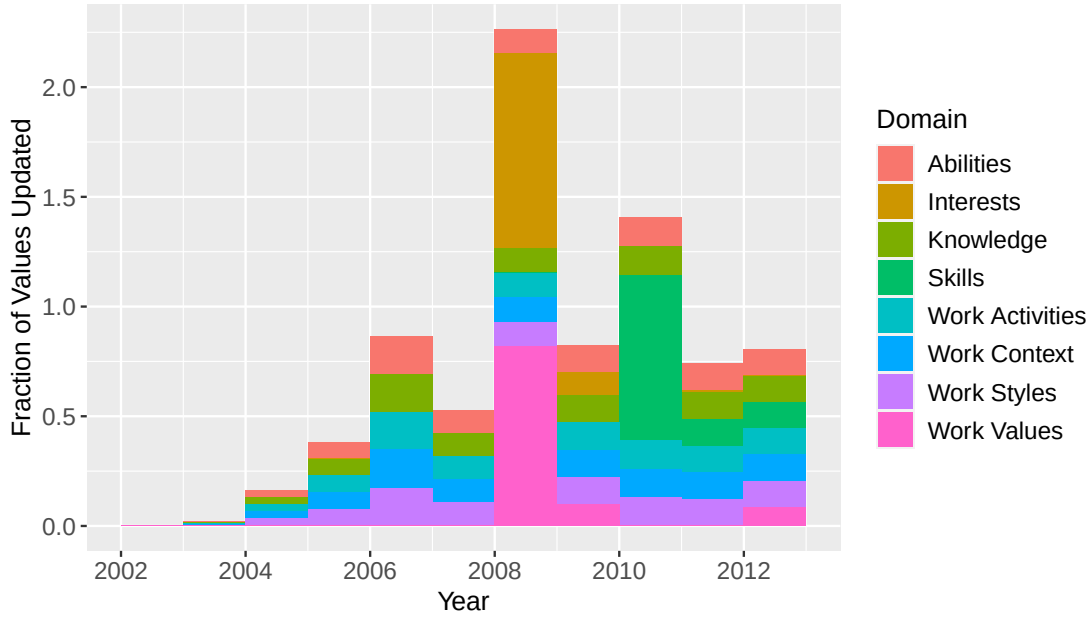
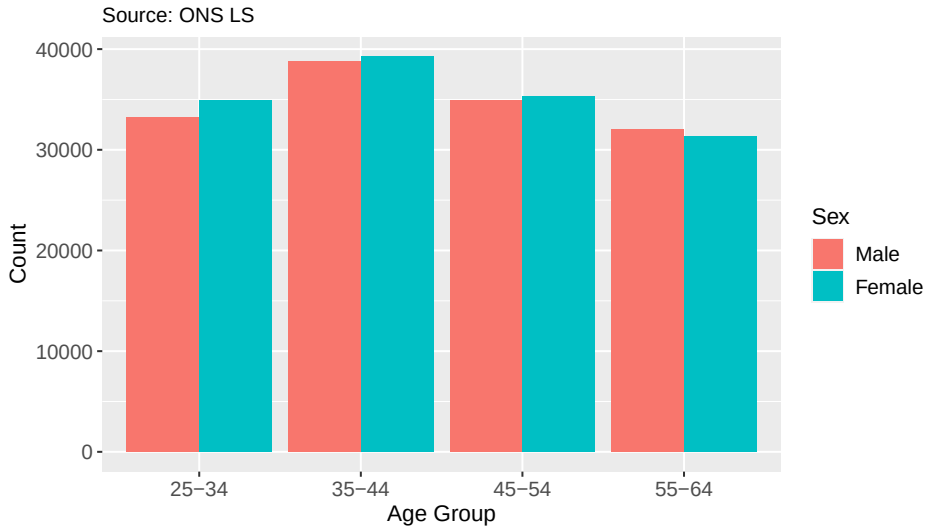
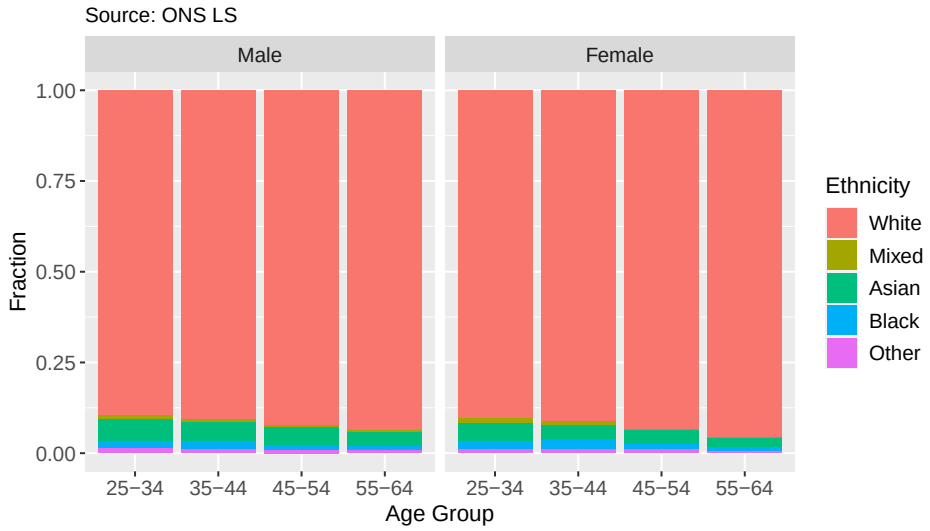


Figure 5: Distribution of last update date for measurements x_{ij} from version 17.0 (July 2012) of the O*NET database.

(a) Sample Size by Age Group and Sex



(b) Sample Fractions by Ethnicity, Age Group and Sex



(c) Sample Size by Government Office Region (GOR) and Age Group

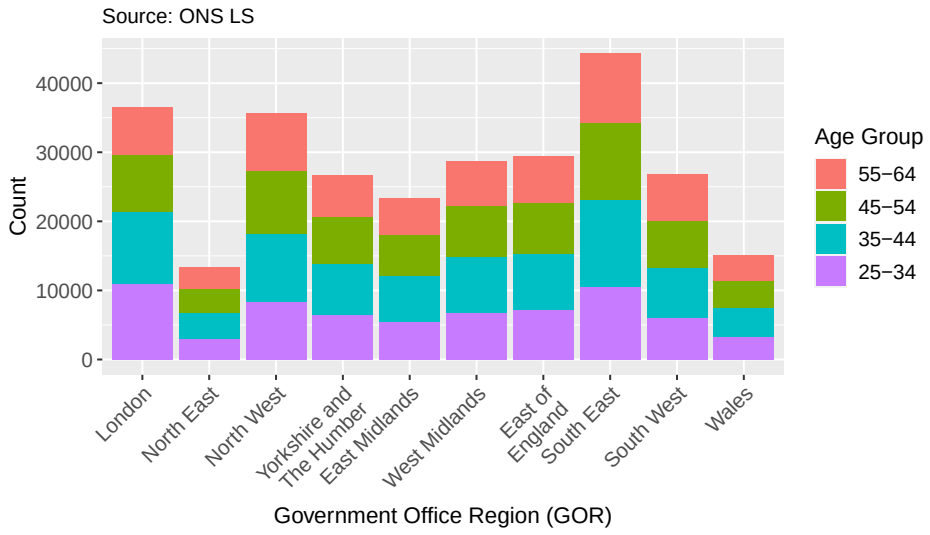
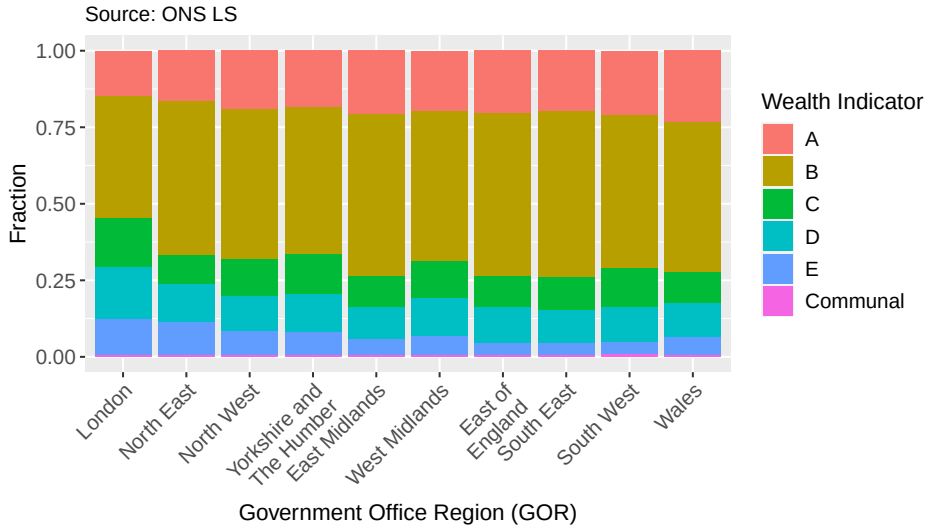
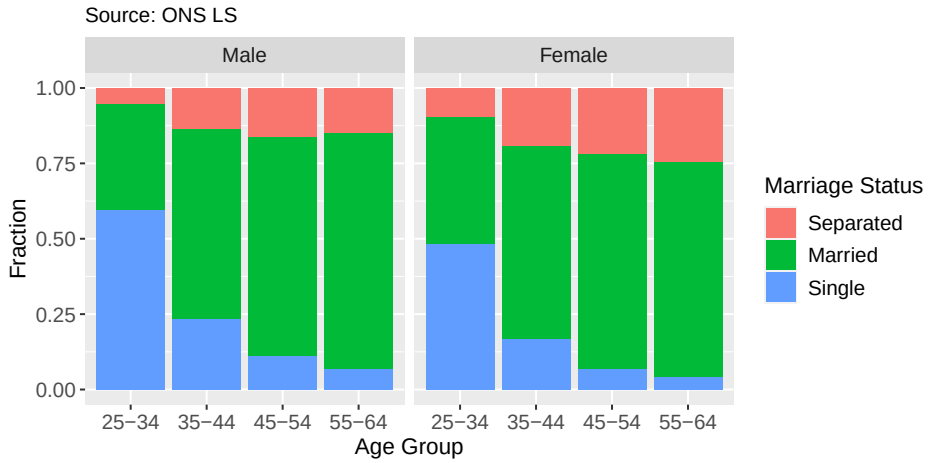


Figure 6: Summary of demographic variables. Source: ONS LS.

(a) Sample Fractions by Government Office Region (GOR) and Wealth



(b) Sample Fractions by Marital Status, Age Group and Sex



(c) Sample Fractions by Education Level, Age Group and Sex

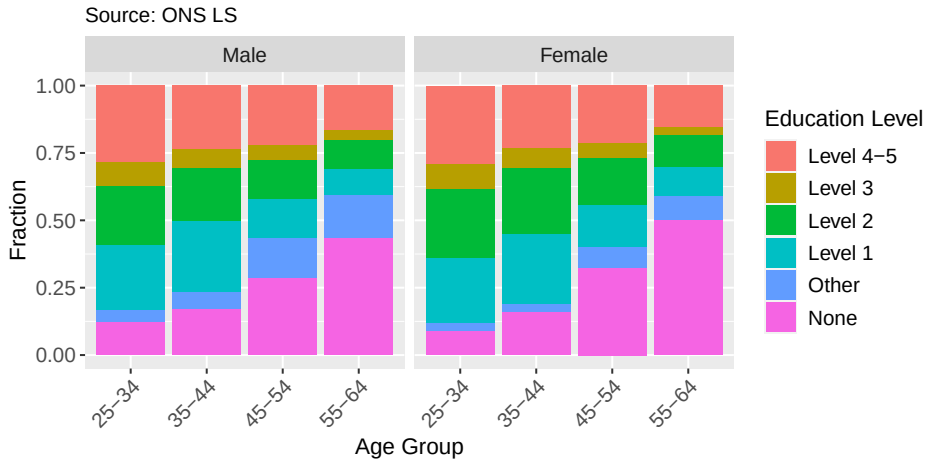
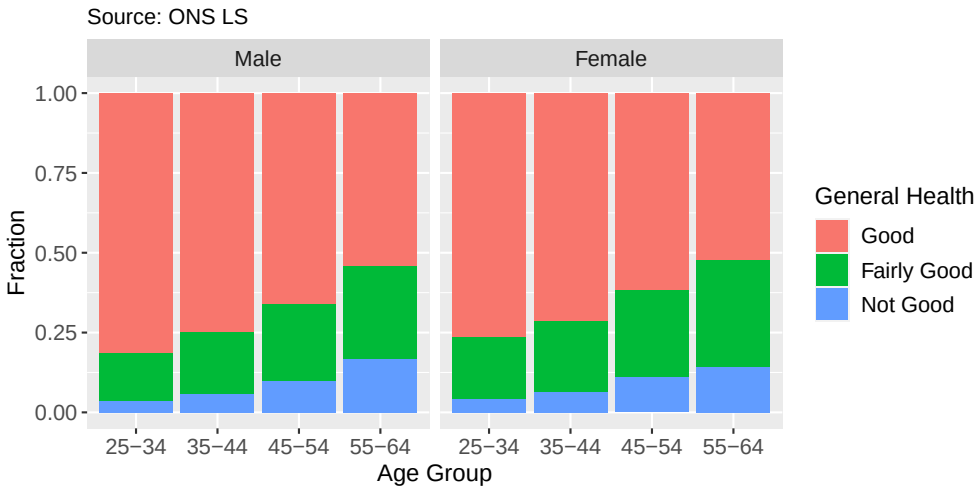
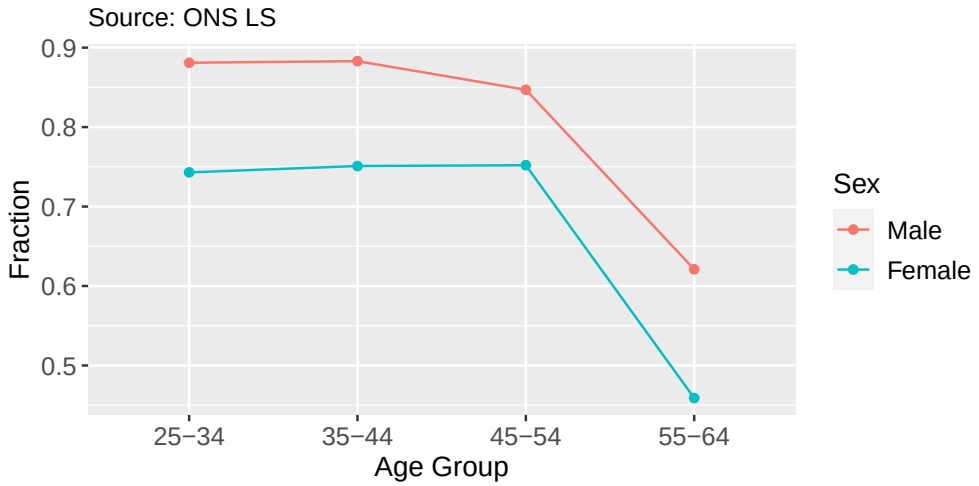


Figure 7: Summary of socioeconomic variables. In (a), the derived wealth indicator is based on property ownership and vehicle access. In (c), level 1 corresponds to 1-4 GCSEs or equivalent, level 2 to 5+ GCSEs or equivalent, level 3 to 2+ A levels or equivalent, and level 4-5 to university undergraduate or equivalent. Source: ONS LS.

(a) Sample Fractions by General Health, Age Group and Sex



(b) Sample Fractions in Employment, by Age Group and Sex



(c) Sample Fractions in Employment, by General Health and Sex

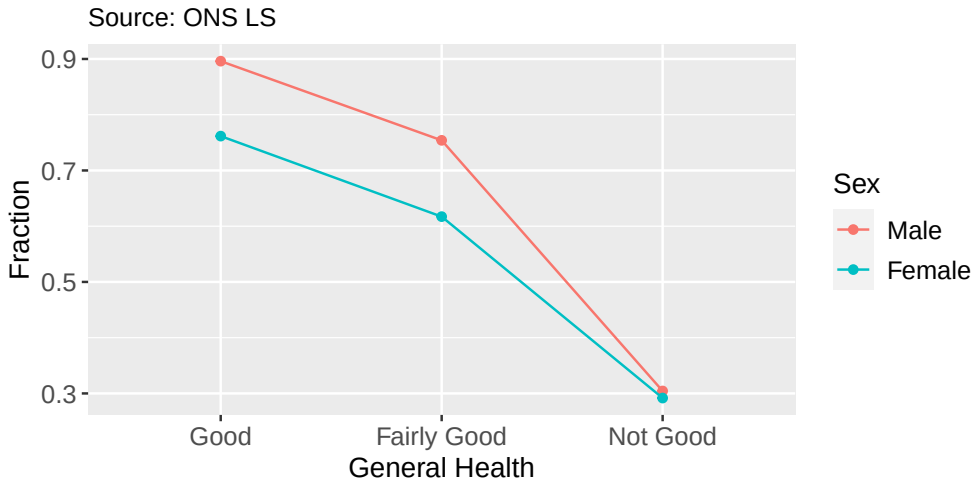


Figure 8: Relationship between self-reported general health, employment status and sex. Source: ONS LS.

Figure 6 summarises the basic characteristics of the sample. Figure 6a shows that the data are relatively balanced across age groups, with ages 35–44 comprising the largest age group and 55–64 the smallest. The sample is 50.3% female and 49.7% male; given that females comprised around 51.2% of the population in mid-2001 (ONS 2021a), they are very slightly underrepresented in the sample. Figure 6b shows that white ethnicities are a majority, comprising over 90% of the sample. The ethnic diversity of the sample is greater in younger than older age groups, with the greatest difference observed in the "mixed" ethnic group, which comprises 1.2% of the sample in the 25–34 age group compared to only 0.3% in the 55–64 age group. Figure 6c shows that the South East is the region most represented in the sample, while the North East is the least represented. The distribution of age is similar across regions with the exception of London, where there were a higher fraction of young individuals: 30% of the sample in London were aged 25–34, compared to only 22–24% of the sample in other regions.

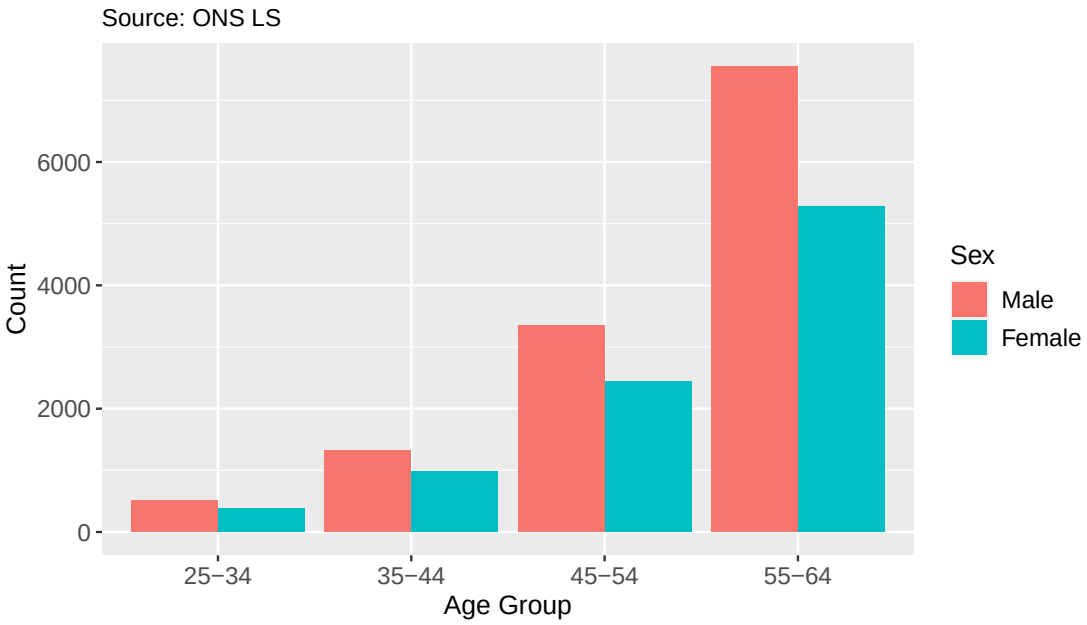
Figure 7 summarises the distribution of socioeconomic variables across the sample. Figure 7a shows that the most common wealth indicator variable in every region is wealth band B. The wealth indicator was derived by combining information on individuals' housing status and vehicle ownership. Differences in this indicator reflect not only individuals' levels of wealth, but also the affordability and desirability of alternative types of accommodation and transport options to those individuals. Wales has the highest sample fraction owning both a house and a car (band A) at 23%, while London has the lowest fraction in this category at 15%. In addition to wealth levels, this difference may reflect differing transport options, age distributions, land / property prices, and general culture between these two regions. Around 0.5% of individuals in each region are in communal establishments, such as hospitals, care homes, prisons, and student halls of residence. Figure 7b shows that the fraction of individuals who are either married or separated is greater in older age groups compared to younger age groups. A higher fraction of females than males are separated in every age group of the sample. Figure 7c shows that in general, the level of education qualifications is higher in younger than older individuals. This follows the general trend of increasing education qualifications observed in England and Wales during this time: for example, participation in higher education increased from 8.4% in 1970 to 19.3% in 1990 (Bolton 2012). The group containing the highest fraction of individuals with advanced qualifications (level 4-5) is 25–34 year-old females at 29%, while the group containing the highest fraction of individuals with no recorded qualifications is 55–64 year old females at 50%.

Figure 8 shows the relationship between self-reported health, sex and employment status. The fraction of individuals reporting "Fairly Good" or "Not Good" health is greater in older age groups, and is consistently 2-5% greater in females

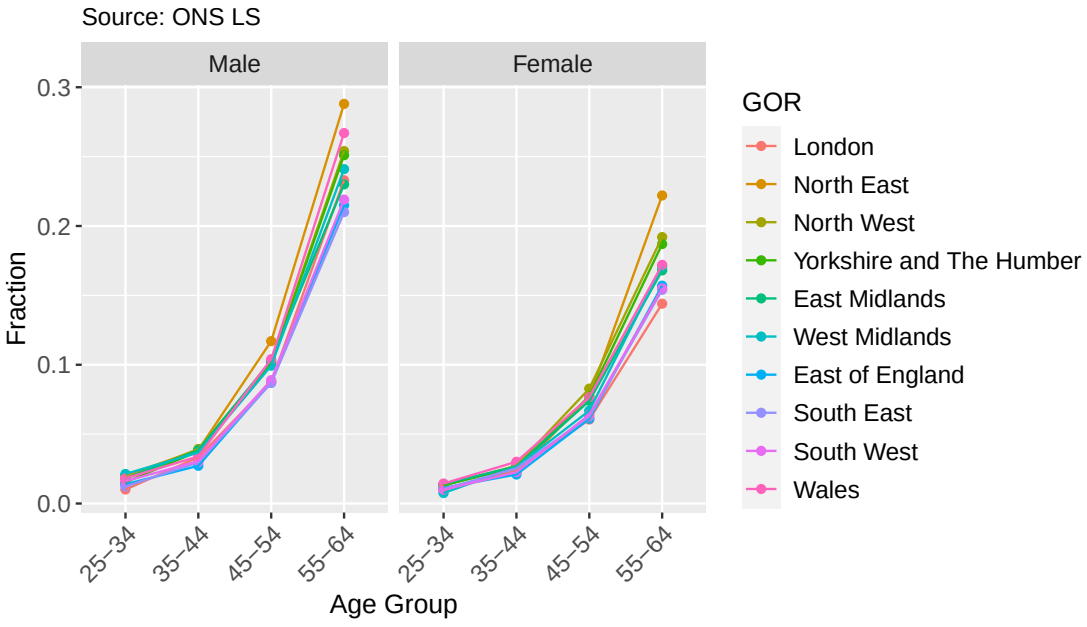
compared to males. Results for reported long-term limiting illness (not shown) are similar. According to figure 8b, a higher fraction of males than females were currently employed in every age group of the sample. Since the sample consists of individuals who recorded having an occupation in the past, this indicates that females in the sample were more likely than males to stop working prior to the census. The fraction of employed individuals appears similar across ages 25-54, but is substantially lower in the 55–64 age group; in this group, only 62% of males and 46% of females reported being in employment at the time of the census. This drop is likely due to multiple factors, including a higher likelihood of retirement or poor health in older individuals, and a longer history of opportunities to start and stop employment. Figure 8c shows that reported general health is strongly correlated with employment status in the sample; individuals reporting worse health were substantially less likely to be employed. This relationship appears to be similarly strong in males and females. This is also likely due to multiple factors, including health-related selection into employment, increasing levels of both poor health and non-employment with age (as discussed above), and the direct impact of non-employment on health.

Figure 9 summarises the distribution of mortality events over the sample. According to figure 9a, around 60% of recorded deaths occurred in the 55–64 age group, while 25%, 10% and 5% of deaths occurred in the three younger age groups respectively. As shown in figure 9b, the fraction of individuals who died during follow-up varies from around 2% of the 25–34 age group to 17% and 24% of the 55–64 age group, for females and males respectively. Wales and northern regions of England tend to have higher mortality rates, while London and southern regions of England tend to have lower mortality rates. The probability of mortality can be seen more clearly in figure 9c, which shows Kaplan-Meier estimators (Kaplan and Meier 1958) for survival probability over the study period. Mortality was substantially higher in older age groups compared to younger age groups, and in males compared to females; despite reporting poorer health (figure 8a), females were less likely than males to die during follow-up. This observation agrees with the well-established "paradoxical" finding that females tend to be diagnosed with illnesses at a higher rate, but have consistently greater life expectancy (Wingard 1984; Oksuzyan *et al.* 2008). According to figure 9d, individuals who reported worse general health were more likely to die during follow-up, with this probability approaching 30% for male individuals reporting "Not Good" health. Similar results were observed for reported long-term limiting illness (not shown). This shows that self-reported health measures have substantial correlation with subsequent mortality.

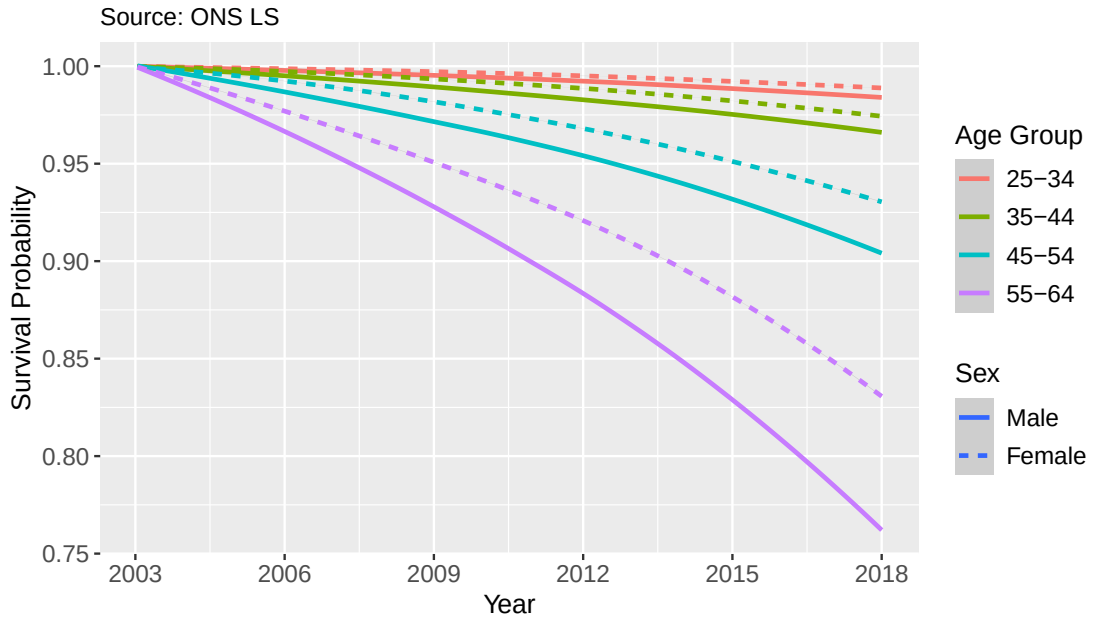
(a) Sample mortality events by Age Group and Sex



(b) Sample mortality fraction by Government Office Region (GOR), Age Group and Sex



(c) Smoothed Kaplan-Meier Estimate for Survival Probability, by Age Group and Sex



(d) Smoothed Kaplan-Meier Estimate for Survival Probability, by Sex and General Health

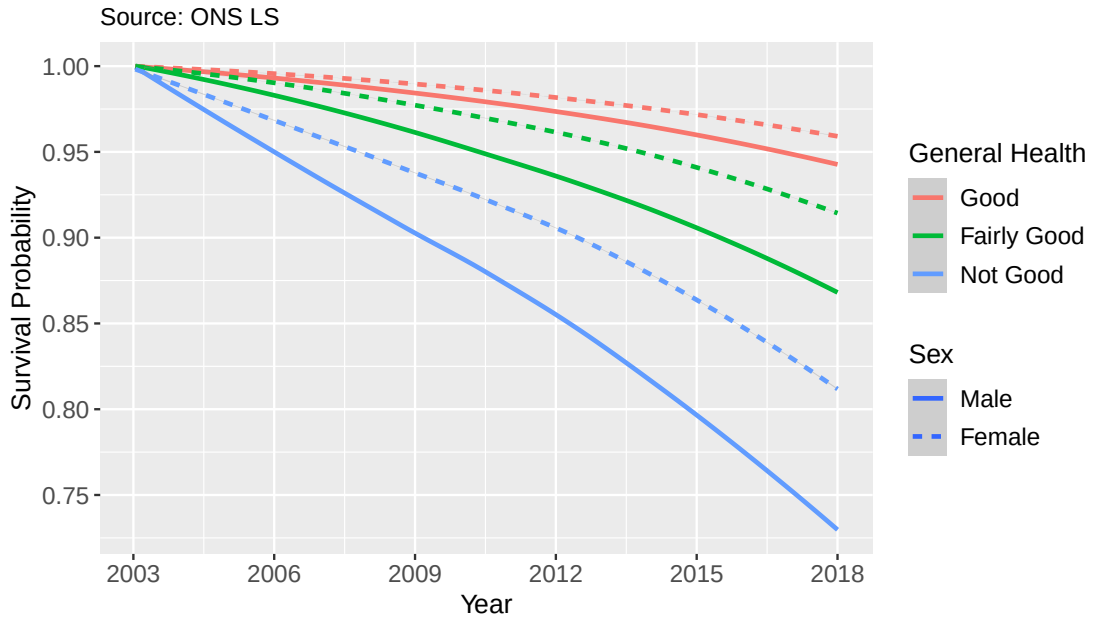


Figure 9: Distribution of mortality events over the sample, during the follow-up period 2003–2017. Source: ONS LS.

4 Methods

4.1 Factor Analysis

As discussed in section 2.2.2, the aim of this stage was to reduce the $p = 246$ initially selected O*NET variables to a smaller and more interpretable feature set. To achieve this, a latent variable model based on the normal linear factor model was fit to each domain of O*NET variables separately. These models assume that the observed occupation variables $\boldsymbol{x}$ were generated by some unobserved latent variables $\boldsymbol{z}$. The dimensions of $\boldsymbol{z}$ represent the key underlying dimensions along which occupations vary; they were assigned labels based on their relationship to the original variables. The posterior mean $\boldsymbol{m} = \mathbb{E}(\boldsymbol{z}|\boldsymbol{x})$ represents an estimate of the underlying occupation features; this quantity was estimated for each occupation, and subsequently used as an input to the regression models of section 4.2.

4.1.1 Theory

The fundamental idea of factor analysis is that observed data $\boldsymbol{x}$ are generated (or explained) by a set of underlying latent variables ("factors") $\boldsymbol{z}$. Specifically, the factor analysis model assumes that the observed data are linearly related to these latent variables[5].

Many variants of the factor analysis model are possible; the most commonly used is known as the *normal linear factor model* (Bartholomew *et al.* 2011). Given a data set $(\boldsymbol{x}_i, \boldsymbol{z}_i)_{i=1}^n$ of observed data $\boldsymbol{x}_i \in \mathbb{R}^p$ and unobserved latent variables $\boldsymbol{z}_i \in \mathbb{R}^q$, the normal linear factor model can be expressed as

$$\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbb{1}) \tag{4.1}$$

$$\boldsymbol{x}_i|\boldsymbol{z}_i \sim \mathcal{N}(\Lambda \boldsymbol{z}_i + \boldsymbol{\mu}, \Psi) \tag{4.2}$$

where Λ is a $p \times q$ matrix and Ψ is a diagonal $p \times p$ covariance matrix. The columns of Λ capture the correlations between observed variables, and are called the *factor loadings*; the diagonal components ψ_{ii} are known as *specific variances*, and represent the independent noise variances for each variable (Bishop 2006).

Given the probabilistic model in equation 4.2, the marginal density over the observed data can be derived as

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Lambda\Lambda^T + \Psi). \tag{4.3}$$

[5]The factor analysis model can be considered as a special case of a broader family of models known as General Linear Latent Variable Models (GLLVMs) (Bartholomew *et al.* 2011). Analogously to Generalised Linear Models (GLMs), these are constructed from a single-parameter exponential family along with a link function.

Whereas an arbitrary multivariate normal distribution over $\boldsymbol{x}_i$ uses $O(p^2)$ parameters, the multivariate normal defined by equation 4.3 uses only $O(pq)$ parameters. The normal linear factor model can therefore be thought of as a method of specifying a low-rank multivariate normal distribution over the data $\boldsymbol{x}_i$ (Murphy 2012).

Defining $\Sigma = \Lambda\Lambda^T + \Psi$, the posterior over the latent variables $\boldsymbol{z}_i$ is

$$\boldsymbol{z}_i|\boldsymbol{x}_i \sim \mathcal{N}(\Lambda^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}), (\Lambda^T\Psi^{-1}\Lambda + \mathbb{1})^{-1}) \tag{4.4}$$

via Bayes rule. The posterior means $\boldsymbol{m}_i = \Lambda^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$ are known as latent scores (Murphy 2012), and provide a low-dimensional summary of the original data $\boldsymbol{x}_i$.

The parameters $\boldsymbol{\mu}$, Λ and Ψ of the normal linear factor model (equation 4.2) can be estimated by maximum likelihood:

$$\begin{aligned}(\boldsymbol{\mu}, \Lambda, \Psi)_{\text{MLE}} &= \underset{(\boldsymbol{\mu},\Lambda,\Psi)}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(\boldsymbol{x}_i; \boldsymbol{\mu}, \Lambda, \Psi) \\ &= \underset{(\boldsymbol{\mu},\Lambda,\Psi)}{\operatorname{argmax}} \sum_{i=1}^{n} \log \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}, \Lambda\Lambda^T + \Psi)\end{aligned} \tag{4.5}$$

The maximum likelihood mean $\boldsymbol{\mu}$ is the sample mean; $\boldsymbol{\mu}_{\text{MLE}} = \bar{\boldsymbol{x}}$. Substituting this into equation 4.5 gives

$$(\Lambda, \Psi)_{\text{MLE}} = \underset{(\Lambda,\Psi)}{\operatorname{argmax}} \left(\frac{n}{2} \log \Sigma^{-1} - \operatorname{trace}\left(\Sigma^{-1}S\right)\right) \tag{4.6}$$

where $\Sigma = \Lambda\Lambda^T + \Psi$ as before, and

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \tag{4.7}$$

is the sample variance of the data.

To find maximum likelihood solutions for Λ or Ψ, the quantity on the right hand side of equation 4.6 is optimised iteratively with respect to Λ and Ψ. Alternatively, equation 4.5 can be optimised via the Expectation-Maximisation (EM) algorithm.

The normal linear factor model is closely related to Principal Component Analysis (PCA); in particular, probabilistic PCA (PPCA) can be obtained from equation 4.2 by setting $\Psi = \sigma^2\mathbb{1}$. However, in contrast to PCA (and PPCA), the normal linear factor model is not uniquely identified. Specifically, since $(\Lambda R)(\Lambda R)^T = \Lambda\Lambda^T$ for any orthogonal (rotation) matrix R, the likelihood (equation 4.3) is invariant under any rotation of the factor loading matrix $\Lambda \to \Lambda R$. In other words, for a given $\boldsymbol{\mu}$ and Ψ, there are a family of equivalent models corresponding to rotations of the factor loading matrix Λ. In practice, there are multiple methods for selecting a unique solution for Ψ (Murphy 2012). The most common approach is to use a heuristic method to select a rotation matrix once an initial solution for Λ has been found; this was the approach followed in this project.

4.1.2 Implementation

As discussed in section 2.2.2, a separate factor analysis was performed for each of the six domains which were reduced; the Interests and Work Values domains were left in their original forms. The stages involved in each of these analyses are described below. As above let $\boldsymbol{x}_i \in \mathbb{R}^p$ represent the observed data and $\boldsymbol{z}_i \in \mathbb{R}^q$ the unobserved latent variables.

First, the number of latent dimensions q was selected. This choice was made with the aim of maximising both the interpretability of the resulting latent dimensions and the statistical goodness-of-fit of the model. An approximate number of dimensions to use q_0 was found by applying a standard criterion, known as parallel analysis (Horn 1965), which involves comparing the eigenvalues of the true data to those of simulated replications. $q = q_0$ dimensions were used unless the interpretability of the results was significantly improved by for $q = q_0 \pm 1$, in which case this was used instead.

Second, the parameters of the normal linear factor model μ, Λ, and Ψ were estimated. The O*NET data are not typically normally distributed (figure 3b), and so the maximum likelihood estimators for these parameters may not be accurate. Instead, a robust variant on the maximum likelihood method, known as the minimum residual ("MinRes") method, was used (Harman and Jones 1966). This method attempts to minimise the sum-of-squared residuals between the off-diagonal components of the model covariance $\Sigma = \Lambda\Lambda^T + \Psi$ and the observed covariance matrix S.

Third, the estimate for Λ was rotated in order to achieve an interpretable solution. Heuristic rotation methods typically attempt to achieve "simple structure" in the loading matrix Λ (Bartholomew *et al.* 2011)—a situation where each variable loads strongly onto one latent dimension, and much more weakly onto all other dimensions. "Oblimin" rotation is the standard oblique (non-orthogonal) rotation method for this purpose; it allows the latent dimensions to be correlated, and generally gives fairly simple structure. This was the method used here.

Fourth, the resulting parameter estimates $\widehat{\mu}$, $\widehat{\Lambda}$, and $\widehat{\Psi}$ were used to compute estimates for the latent scores $\boldsymbol{m}_i = \mathbb{E}(\boldsymbol{z}_i|\boldsymbol{x}_i)$. Rather than the posterior mean under the normal linear factor model (equation 4.4), the estimator

$$\widehat{\boldsymbol{m}}_i(\boldsymbol{x}_i) = \left(\Lambda^T\Psi^{-1}\Lambda\right)^{-1}\Lambda^T\Psi^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \tag{4.8}$$

was used. This estimator, due to Bartlett (1937), can be derived as a least-squares estimator for the posterior mean, and is unbiased.

Finally, the latent dimensions Z_j were named and defined. The dimensions were named using the highest-loading variables and highest-scoring occupations. Names were based on the original set of variables, and aimed to be as simple and accurate as possible. Section 5.1 presents the results of the factor analyses.

4.2 Survival Analysis

Statistical modeling of "time-to-event" data—such as the "time-to-mortality" data used here—is known as survival analysis. A theoretical overview of the standard approach to survival analysis is given in section 4.2.1. More specifically, the aim here was to estimate the relationship between a set of predictors $\boldsymbol{x}_i$ and time until mortality $\mathcal{T}$. The Cox proportional hazards model is the regression model most commonly applied to this task, both in previous work (N. J. Johnson *et al.* 1999; Lee 2011) and in medical applications in general (Klein and Moeschberger 2003). This model was therefore initially selected for use here; it is discussed in section 4.2.2. A brief overview of other regression models for survival analysis is given in appendix E; these other models may provide a comparable or better fit, and so this may be a useful direction for future work. Following this, methods for evaluating the goodness-of-fit of the Cox model are discussed in section 4.2.3. The overall approach to modeling—including the sample selection criteria and predictor variables used in each model—is discussed in section 4.2.4.

4.2.1 Background

Survival analysis[6] refers to a set of statistical techniques used to analyse "time-to-event" data (Steinsaltz 2019). This type of data occurs in any evolving system where "events" can be clearly defined: examples include births, death and disease in biological systems; component failure in mechanical systems; and many types of events in social systems such as life events (e.g. marriage), consumer behaviour, and unemployment. The problem of analysing time-to-event data arises in a number of fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography (Klein and Moeschberger 2003).

Mathematically, the time until an event occurs t can be modeled as the realisation of a random variable $\mathcal{T}$. The distribution of $\mathcal{T}$ is typically specified using either (1) the survival function $S(t) = \mathbb{P}(\mathcal{T} > t)$, (2) the event probability density function $f(t)$, or (3) the *hazard* function $\lambda(t)$, which is defined as the conditional probability density

$$\lambda(t) = f(t|\mathcal{T} \geq t) \tag{4.9}$$

which in the case of continuous $\mathcal{T}$ can be written

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{4.10}$$

Intuitively, the hazard $\lambda(t)$ represents the "risk" of an event at time t; the event

[6] Also known as reliability theory, duration analysis, event history analysis, or lifetime modeling.

probability (per unit time) $f(t) = \lambda(t)S(t)$ is the risk experienced at time t multiplied by the probability of survival up to that time.

The one-to-one correspondence between $\lambda(t)$ and $S(t)$ can be observed by noting that

$$\begin{aligned} \lambda(t) &= -\frac{S'(t)}{S(t)} \\ &= -\frac{d}{dt}\log S(t). \end{aligned} \tag{4.11}$$

according to equation 4.10. Inverting this equation gives

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(t')dt'\right) \\ &= \exp\left(-\Lambda(t)\right) \end{aligned} \tag{4.12}$$

where $\Lambda(t) = \int_0^t \lambda(t')dt'$ is known as the cumulative hazard function. $\Lambda(t)$ can be interpreted as the total risk accumulated up until a given time t, or as the expected number of events up to time t if multiple events were allowed to occur. Intuitively, equation 4.12 generalises the constant-hazard exponential distribution $S(t) = \exp(-\lambda t)$ to arbitrary time-varying hazards $\lambda(t)$.

A distinguishing feature of time-to-event data (as opposed to other types) is *incomplete observation*. This can be due to either *censoring*, where the history of a given individual is not fully observed, or *truncation*, where only events occurring in a certain interval are observed at all. Mathematically, censoring corresponds to a situation in which the event time $\mathcal{T}$ is only known to be in a certain interval; observations are of the form $\mathcal{T} \in (t_1, t_2)$ rather than $\mathcal{T} = t$. By contrast, truncation corresponds to a situation in which the event time is observed *conditional* on falling in some interval: the observations in this case are of the form $\mathcal{T}^* = t$ with $\mathcal{T}^* = \mathcal{T}|\mathcal{T} \in (t_1, t_2)$. The different types and causes of incomplete observations are described in detail in Klein and Moeschberger (2003) and Steinsaltz (2019).

The most common forms of incomplete observation for time-to-event data are left-truncation and right-censoring. In this case, individuals are only observed if their events occur after t_l, the left-truncation time, and their event times are only known precisely if they occur before t_r, the right-censoring time. Typically the left-truncation time can be set to zero without loss of generality[7]. Given event times $\mathcal{T}_i$ for some individuals $i = 1, \ldots, n$, the situation in which both left-truncation

[7]In general, the truncation time is uniform—for example, the entry time to a medical study—while the censoring time varies between individuals—for example due to drop-outs from the study. In this case, times $\mathcal{T}$ can be measured from the left-truncation time ($t_l = 0$), and it is taken as implicit that the model describes event times conditional on survival up to t_l.

and right-censoring occur can be described using a pair of random variables

$$T_i = \min(\mathcal{T}_i, c_i) \tag{4.13}$$

$$\Delta_i = \mathbb{1}(\mathcal{T}_i \leq c_i) \tag{4.14}$$

where c_i are the censoring times for individuals $i = 1, \ldots, n$. Here T_i is an *exit time*, which indicates the time past which the individual i was no longer observed—either due to an event or due to censoring. The indicator variable Δ_i has $\Delta_i = 1$ if an event was observed and $\Delta_i = 0$ if it was censored.

Given some model $S_i(t)$ over observed data $\{(t_i, \delta_i)\}_{i=1}^n$, the likelihood is then

$$\begin{aligned} L &= \prod_{i=1}^{n} f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda_i(t_i)^{\delta_i} S_i(t_i) \end{aligned} \tag{4.15}$$

by equation 4.10. In other words, it is a product of the event probability densities $f_i(t_i)$ for uncensored observations t_i, and the discrete survival probability $S_j(t_j)$ for censored observations t_j.

Individuals are said to be "at risk" at time t if they are still under study just prior to this time—i.e. if $t \leq T_i$. The risk set $R(t)$ is defined as the set of all individuals who were at risk at time t.

4.2.2 Cox Proportional Hazards Model

Given a set of right-censored data $\{(T_i, \Delta_i, \boldsymbol{x}_i)\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a set of time-independent predictors, the Cox model is defined by the equation

$$\lambda_i(t) = \exp(\boldsymbol{x}_i^T \beta)\lambda_0(t) \tag{4.16}$$

where $\lambda_i(t)$ and $\lambda_0(t)$ are the individual and baseline hazards, and $\beta \in \mathbb{R}^p$ is a vector of parameters. The parameter β_j represents the strength of the relationship between the predictor x_j and the hazard. Typically the baseline hazard $\lambda_0(t)$ is estimated nonparametrically; in this case equation 4.16 defines a *semiparametric* model. The defining assumption of the Cox model is that differences in predictors correspond to constant multiplicative factors in the risk; this assumption is often plausible in practice.

When the T_i are continuously distributed, the observed event times will be distinct. In this case, denote the observed data $\{(t_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^n$ with $\sum_i \delta_i = m$ the observed number of events, and the event times $t_1 < t_2 < \cdots < t_m$ without loss of generality. Using equation 4.15, the likelihood of the Cox model can be written

$$L(\beta, \lambda_0(t)) = \prod_{i=1}^{n} \left[\lambda_0(t_i) \exp(\boldsymbol{x}_i^T \beta)\right]^{\delta_i} \exp\left(-\Lambda_0(t_i) \exp(\boldsymbol{x}_i^T \beta)\right). \tag{4.17}$$

The maximum likelihood estimators for β and $\lambda_0(t)$ can be computed in two steps (Klein and Moeschberger 2003). First, the nonparametric MLE for $\lambda_0(t)$ (Breslow 1972) occurs when

$$\widehat{\lambda}_0(t) = \begin{cases} \dfrac{1}{\sum_{j \in R(t)} \exp(\boldsymbol{x}_j^T \beta)} & \text{if} \quad t \in \{t_1, \ldots, t_m\} \\ \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

or in terms of the cumulative hazard

$$\widehat{\Lambda}_0(t) = \int_0^t \widehat{\lambda}_0(t')dt' \tag{4.19}$$

$$= \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \beta)}. \tag{4.20}$$

This yields a profile likelihood for β proportional to

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\boldsymbol{x}_i^T \beta)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \beta)}. \tag{4.21}$$

This likelihood can also be derived as a partial likelihood; specifically, equation 4.21 gives the probability that the individuals experience events in the observed order, conditional on the event times t_i. It is easily verified that equation 4.21 is convex in β; the maximum likelihood estimator $\widehat{\beta}$ can be computed using the Newton-Raphson algorithm.

Since event times must be recorded to a finite resolution in practice, data frequently contain ties—that is, multiple events are recorded at one discrete time. In this case, the partial likelihood in equation 4.21 must be adjusted to take account of this fact. Mathematically, let $t_1 < t_2 < \cdots < t_l$ denote the l distinct event times, with $\sum_i \delta_i = m$ $(m > l)$ the total number of events as before. Define also $D(t_i)$ as the set of individuals experiencing an event at event time t_i. The solution to the problem of ties can be found by assuming that events observed to occur at the same time actually occur in some unknown order. Given d_i events with observed times at t_i, there are $d_i!$ possible orders in which the events could have occurred, each of which correspond to a term in the partial likelihood of the form of equation 4.21. The exact partial likelihood can be computed by weighting each of these $d_i!$ terms proportional to their probabilities, $1/d_i!$.

In practice, the exact method is computationally infeasible. Instead, an approximation due to Efron (1977) is used, in which the contribution of event time t_i to the partial likelihood is

$$L_i(\beta) = \frac{\prod_{j \in D(t)} \exp(\boldsymbol{x}_j^T \beta)}{\prod_{k=0}^{d_i - 1} \left(\sum_{j \in R(t_i)} \exp(\boldsymbol{x}_j^T \beta) - \frac{k}{d_i} \sum_{j \in D(t_i)} \exp(\boldsymbol{x}_j^T \beta) \right)} \tag{4.22}$$

with the full likelihood

$$L(\beta) = \prod_{i=1}^{l} L_i(\beta). \tag{4.23}$$

The approximation in equation 4.22 is accurate provided the deaths at time t_i are not a large proportion of the risk set $R(t_i)$.

4.2.3 Model Evaluation

The extent to which the results of a fitted model can be meaningfully interpreted depends strongly on the degree to which the data satisfy the model assumptions. The assumptions of the Cox model, defined by equation 4.16, can be broken down into three main parts (Lin *et al.* 1993): (1) the time-invariance of the hazard ratio $\lambda_i(t)/\lambda_0(t)$, i.e. proportional hazards; (2) the correctness of the functional form used for the predictors in the exponential; and (3) the correctness of the exponential link function. The two diagnostic methods used in this project were Cox-Snell residuals and Martingale residuals; these were used to provide a (limited) evaluation of the above assumptions.

Cox-Snell residuals (Cox and Snell 1968) are motivated by the identity

$$\Lambda(T) \sim \text{Exp}(1) \tag{4.24}$$

where $\Lambda(t)$ is the cumulative hazard for a time-to-event random variable T. Given an arbitrary time-to-event model $\widehat{\Lambda}_i(t) = \widehat{\Lambda}(t; \boldsymbol{x}_i)$ for observed data $\{(t_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^n$, the Cox-Snell residuals are defined as

$$\widehat{e}_i = \widehat{\Lambda}_i(t_i). \tag{4.25}$$

In other words, they are the model cumulative hazards evaluated at the observed exit times. According to equation 4.24, the goodness-of-fit of the model $\widehat{\Lambda}(t; \boldsymbol{x})$ can be evaluated by comparing the distribution of the $\widehat{e}_i$ to a censored Exp(1) distribution. Specifically, goodness-of-fit can be evaluated by (1) estimating the cumulative hazard $\widehat{\Lambda}_{\text{CS}}(e)$ of the pair $\{(\widehat{e}_i, \delta_i)\}_{i=1}^n$—typically using the Nelson-Aalen estimator (Nelson 1969; O. Aalen 1978)—and (2) comparing this function to its expected value under a censored Exp(1) distribution, the line $y = x$. For the special case of the Cox model equation 4.16, equation 4.25 becomes

$$\widehat{e}_i = \widehat{\Lambda}_0(t_i) \exp(\boldsymbol{x}_i^T \beta) \tag{4.26}$$

where $\Lambda_0(t)$ is the Breslow (1972) estimator for the baseline hazard given in equation 4.20.

In order to define Martingale residuals, it is useful to define

$$N_i(t) = \Delta_i \mathbb{1}(T_i < t) \tag{4.27}$$

as the number of events experienced by individual i before time t, and

$$Y_i(t) = \mathbb{1}(T_i \geq t) \tag{4.28}$$

as the indicator for individual i being "at-risk" at time t. Given an arbitrary time-to-event model $\widehat{\Lambda}_i(t)$, the Martingale residual (Barlow and Prentice 1988) for individual i at time t is then defined as

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(t')d\widehat{\Lambda}_i(t') \tag{4.29}$$

and $\widehat{M}_i(\infty)$ is abbreviated as $\widehat{M}_i$. In other words, they represent the excess of observed events over those expected under the model, for each individual i at time t. Under the Cox model, equation 4.29 gives

$$\widehat{M}_i = \Delta_i - \widehat{\Lambda}_0(T_i)\exp(\boldsymbol{x}_i^T\beta) \tag{4.30}$$

As the name suggests, the quantities in equation 4.30 are martingales when the model is correctly specified (Klein and Moeschberger 2003). Martingale residuals can be plotted against individual predictors, to determine the "best" functional form for the predictor, or against the linear predictor $\boldsymbol{x}_i^T\beta$, to provide an overall measure of goodness-of-fit.

4.2.4 Implementation

The modeling approach taken was as follows. Four sets of models were fit, corresponding to different sets of LS predictors and different sample selection criteria. In each set, a separate Cox model was estimated for each domain of occupation features (table 1). Therefore 32 models were fit in total. This somewhat simplifies model interpretation, compared to a model including all the occupation features.

The LS predictors used are summarised in table 2, while a summary of the fitted models is given in table 3. The baseline model describes the relationship between occupation features and mortality after adjusting for only age, sex, ethnicity and geographic region. By contrast, the socioeconomic-status-adjusted (SES-adjusted) model further adjusts for marital status, wealth and education. The relationships between these socioeconomic variables are complex: occupation is a substantial determinant of wealth, and likely influences marital status; all three influence health and mortality, and are influenced by education. Therefore this "adjustment" is only approximate. Nevertheless, this model likely more accurately reflects the relationship between the occupation features and mortality which is independent of socioeconomic status—of which education, marital status and wealth are important components. The health-adjusted model further adjusts for self-reported general health and limiting long-term illness in 2001. This may help to isolate

Predictor	Group	Description	Parameters
agec	Baseline	Age in 2001 (centered)	1
age2c	Baseline	Age-squared in 2001 (centered)	1
sex	Baseline	Sex	1
agec_sexFemale	Baseline	Age-Sex interaction	1
ethnicity	Baseline	Ethnicity	4
gors0	Baseline	Geographic region	9
marriage	SES	Marital status	2
wealth	SES	Wealth indicator	5
hlqp0	SES	Education	5
illp0	Health	Limiting long term illness	1
heap0	Health	General health	2
Total			32

Table 2: Summary of LS variables used as predictors. "Parameters" indicates the corresponding number of model parameters.

the effect of health-selection into occupations, although this adjustment is limited by the unknown accuracy of the self-reported health measures. Finally, the employed subset model imposes an additional sample selection criterion: only individuals who reported being in work the previous week are included. This gives an indication of the sensitivity of the model results to the choice of sample.

The data (section 2.3) display both left truncation and right censoring. Left truncation occurs as only individuals who were alive in 2003 are included in the final sample; right censoring occurs since the time of death was not observed for individuals who died after 2018. Event times were only known up to the nearest month; maximum likelihood solutions were therefore computed using the likelihood in equation 4.23. Each model was evaluated using plots of both the Cox-Snell residuals, and of the Martingale residuals against the linear predictor. This represents a fairly limited evaluation of the model assumptions. All models were fit using the `survival` R package (Therneau 2020). The results are presented in section 5.2.

5 Results

5.1 Factor Analysis

In this section, results are presented for the factor analysis of the "Work Styles" domain. Similar results were obtained for the other five domains which were

Model Name	LS Predictors	Sample Criteria	Parameters
Baseline	Baseline	Ever worked	17
SES-Adjusted	Baseline, SES	Ever worked	29
Health-Adjusted	Baseline, SES, Health	Ever worked	32
Employed Subset	Baseline, SES	Working last week	29

Table 3: Summary of LS predictors used in each model. "Parameters" indicates the number of model parameters corresponding to the LS predictors. In addition to the sample criteria shown, individuals were required to be aged 25-64 in 2001, and alive in 2003.

reduced. The number of latent dimensions extracted within each domain is given in table 1; these dimensions are listed and defined in appendix C.2.

The parallel analysis criterion suggests that $q = 4$ latent dimensions are sufficient to explain variation in the $p = 16$ variables of the work styles domain. The fitted model is found to account for 73% of the variance in the original variables; figure 10 shows the loading matrix Λ corresponding to the model with $q = 4$ latent dimensions. As described in section 4.1.2, the model was fit using the MinRes method, and the loading matrix rotated using the oblimin method. The loading matrix achieves the criteria of "simple structure" relatively well—most variables have substantial loadings onto one latent dimension, and only small loadings onto other dimensions—and the underlying latent dimensions are relatively interpretable. As a result, the $q = 4$ solution was selected, and the latent dimensions were named as described in section 4.1.2. For example, "Ambitious" latent dimension loads most highly onto the initiative, persistence, and achievement/effort variables of the original O*NET data.

Following this, latent scores $\boldsymbol{m}_i$ were estimated using Bartlett's method. Tables 4 and 5 show the occupations with the highest and lowest scores on each latent dimension, in both the O*NET 2010 SOC space (prior to mapping), and in the UK 2000 SOC space. From this crude evaluation, the results appear broadly reasonable: for example, managers tend to score highly on the "Ambitious" dimension, while social workers and therapists score highly on the "Social" dimension. A more thorough evaluation of these results would be desirable, though difficult without occupational expertise or data sets for empirical comparison.

5.2 Survival Analysis

Occupation feature data was missing for around 300 of the $n = 279,646$ individuals in the original sample. As a result, the baseline, SES-adjusted and health-adjusted models included $279,368$ observations and $21,805$ mortality events. The employed

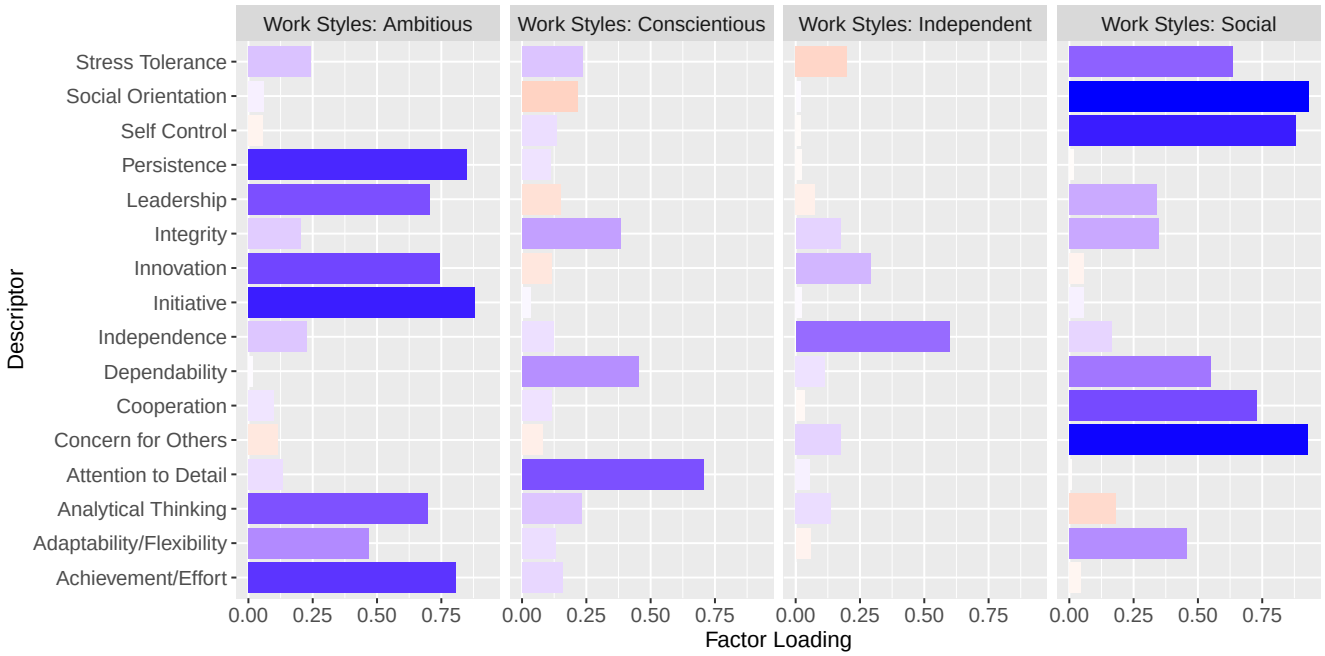


Figure 10: Estimated loading matrix $\widehat{\Lambda}$ for the "Work Styles" domain. The loading matrix relates the $p = 16$ observed variables to the $q = 4$ latent dimensions.

subset models included $209,032$ observations and $11,233$ mortality events.

5.2.1 LS Predictors

Parameters for each of the baseline, socioeconomic and health predictors were estimated between eight and 32 times, corresponding to the four sets of LS predictors and the eight sets of occupation features. It was found that parameter estimates for the LS predictors did not vary substantially over the eight sets of occupation features. For example, figure 11 shows that hazard ratio estimates for geographic region vary as different LS predictors are used, but not substantially when different occupation features are used. Therefore parameter estimates for the LS predictors are summarised by their mean (and the mean of their standard errors) over the eight sets of occupation features.

Table 4: Highest and lowest scoring O*NET 2010 SOC occupations on the four "Work Styles" dimensions.

Occupation Feature	Highest Scoring Occupations	Lowest Scoring Occupations
Work Styles: Ambitious	Choreographers; Urologists; Chief Executives; Aerospace Engineering and Operations Technicians; Industrial-Organizational Psychologists	Ushers, Lobby Attendants, and Ticket Takers; Driver/Sales Workers; Graders and Sorters, Agricultural Products; Postal Service Mail Carriers; Crossing Guards
Work Styles: Conscientious	Insurance Underwriters; Cytotechnologists; Network and Computer Systems Administrators; Court Reporters; Regulatory Affairs Specialists	Forest Fire Inspectors and Prevention Specialists; Models; Hosts and Hostesses, Restaurant, Lounge, and Coffee Shop; Dredge Operators; Mine Shuttle Car Operators
Work Styles: Independent	Hunters and Trappers; Cooks, Private Household; Massage Therapists; Physics Teachers, Postsecondary; Geographic Information Systems Technicians	Forest Fire Fighting and Prevention Supervisors; Forest Fire Inspectors and Prevention Specialists; Farmworkers and Laborers, Crop; Office Machine Operators, Except Computer; Musicians, Instrumental
Work Styles: Social	Psychiatric Aides; Psychiatric Technicians; Directors, Religious Activities and Education; Recreational Therapists; Skincare Specialists	Astronomers; Painting, Coating, and Decorating Workers; Craft Artists; Environmental Economists; Economists

Table 5: Highest and lowest scoring UK 2000 SOC occupations on the four "Work Styles" dimensions.

Occupation Feature	Highest Scoring Occupations	Lowest Scoring Occupations
Work Styles: Ambitious	Healthcare practice managers; Hospital and health service managers; Education officers, school inspectors; Registrars and senior administrators of educational establishments; Solicitors and lawyers, judges and coroners	Roundsmen/women and van salespersons; Taxi, cab drivers and chauffeurs; School crossing patrol attendants; School midday assistants; Postal workers, mail sorters, messengers, couriers
Work Styles: Conscientious	Aircraft pilots and flight engineers; Quantity surveyors; Taxation experts; Solicitors and lawyers, judges and coroners; Legal professionals n.e.c.	Farm workers; School crossing patrol attendants; School midday assistants; Market and street traders and assistants; Security guards and related occupations
Work Styles: Independent	Speech and language therapists; Fitness instructors; Driving instructors; Market and street traders and assistants; Higher education teaching professionals	Sports players; Traffic wardens; Car park attendants; Farm workers; Telephone salespersons
Work Styles: Social	Air travel assistants; Social workers; Public service associate professionals; Speech and language therapists; Occupational therapists	Roundsmen/women and van salespersons; School crossing patrol attendants; School midday assistants; Leather and related trades; Moulders, core makers, die casters

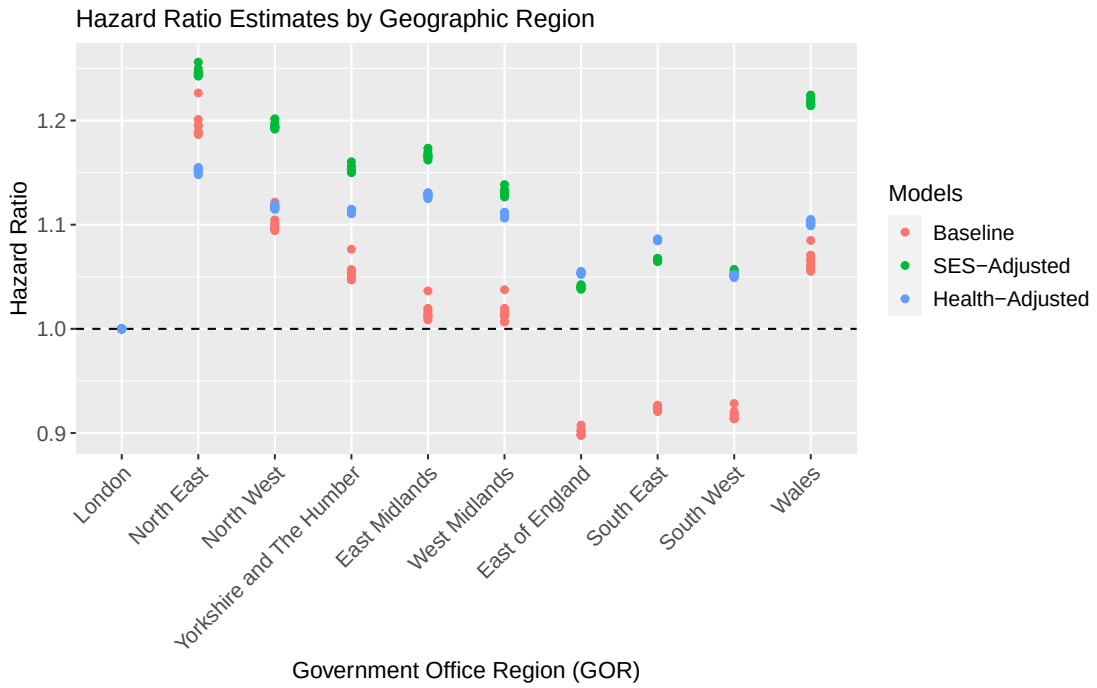


Figure 11: Hazard ratios for geographic region, under the baseline, SES-adjusted and health-adjusted models. Individual data points correspond to sets occupation features. Estimates over each of the eight sets of occupation features are similar. Source: ONS LS.

Table 6: Hazard ratios and 95% confidence intervals for the LS predictors over the baseline, SES-adjusted and health-adjusted models. Source: ONS LS.

	Model		
Predictor	Baseline	SES-Adjusted	Health-Adjusted
agec	1.09 (1.07, 1.10)	1.12 (1.10, 1.14)	1.11 (1.09, 1.13)
age2c	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
sex: Male	1.00	1.00	1.00
sex: Female	0.73 (0.70, 0.76)	0.72 (0.69, 0.75)	0.70 (0.67, 0.74)
agec_sex: Female	0.997 (0.994, 1.000)	0.996 (0.993, 0.999)	0.998 (0.995, 1.001)
ethnicity: White	1.00	1.00	1.00
ethnicity: Mixed	0.80 (0.64, 1.01)	0.70 (0.56, 0.88)	0.67 (0.54, 0.85)
ethnicity: Asian	0.72 (0.67, 0.78)	0.79 (0.73, 0.85)	0.71 (0.66, 0.77)
ethnicity: Black	0.75 (0.66, 0.85)	0.63 (0.55, 0.71)	0.62 (0.55, 0.71)
ethnicity: Other	0.52 (0.42, 0.64)	0.53 (0.42, 0.65)	0.52 (0.42, 0.65)
gors0: London	1.00	1.00	1.00
gors0: North East	1.20 (1.12, 1.28)	1.25 (1.16, 1.34)	1.15 (1.07, 1.23)
gors0: North West	1.10 (1.04, 1.16)	1.20 (1.13, 1.26)	1.12 (1.06, 1.18)
gors0: Yorkshire & Humber	1.05 (0.99, 1.12)	1.15 (1.09, 1.22)	1.11 (1.05, 1.18)
gors0: East Midlands	1.02 (0.96, 1.08)	1.17 (1.10, 1.24)	1.13 (1.06, 1.20)
gors0: West Midlands	1.02 (0.96, 1.08)	1.13 (1.07, 1.20)	1.11 (1.05, 1.18)
gors0: East England	0.90 (0.85, 0.96)	1.04 (0.98, 1.11)	1.05 (0.99, 1.12)
gors0: South East	0.92 (0.87, 0.98)	1.07 (1.01, 1.13)	1.09 (1.03, 1.15)
gors0: South West	0.92 (0.86, 0.98)	1.05 (0.99, 1.12)	1.05 (0.99, 1.12)
gors0: Wales	1.07 (1.00, 1.14)	1.22 (1.14, 1.30)	1.10 (1.03, 1.18)
marriage: Single		1.00	1.00
marriage: Married		0.70 (0.67, 0.73)	0.72 (0.69, 0.75)
marriage: Separated		0.88 (0.84, 0.92)	0.87 (0.83, 0.91)
wealth: A		1.00	1.00
wealth: B		1.07 (1.04, 1.11)	1.07 (1.03, 1.11)
wealth: C		1.36 (1.29, 1.43)	1.28 (1.22, 1.35)
wealth: D		1.76 (1.68, 1.84)	1.53 (1.46, 1.60)
wealth: E		2.50 (2.38, 2.62)	2.00 (1.91, 2.11)
wealth: Communal		3.73 (3.24, 4.30)	3.08 (2.68, 3.55)
hlqp0: None		1.00	1.00
hlqp0: Other		0.88 (0.84, 0.92)	0.92 (0.88, 0.96)
hlqp0: Level 1		0.88 (0.84, 0.92)	0.94 (0.90, 0.99)
hlqp0: Level 2		0.83 (0.80, 0.87)	0.89 (0.85, 0.94)
hlqp0: Level 3		0.81 (0.75, 0.87)	0.87 (0.81, 0.93)
hlqp0: Level 4-5		0.73 (0.69, 0.77)	0.80 (0.76, 0.84)
illp0: No			1.00
illp0: Yes			1.40 (1.35, 1.46)
heap0: Good			1.00
heap0: Fairly Good			1.34 (1.29, 1.38)
heap0: Not Good			1.93 (1.84, 2.02)

Only the baseline, SES-adjusted and health-adjusted models are considered in this section. Parameter estimates and 95% confidence intervals for these models are shown in table 6. Parameter estimates for the terms involving sex appear relatively similar across the three sets of models, while the estimates for age, ethnicity, region, marital status, wealth indicator, and education depend more on which other predictors are included. All predictors appear to predict mortality to some extent. Age is the strongest predictor of mortality; following this, the derived wealth indicator also appears to have a strong relationship with mortality. These results are discussed in more detail below.

Table 6 shows that under the models fit, individuals differing by one year of age have hazards differing by a factor of around 10%, holding the baseline, socioeconomic and health predictors constant. Consequently, the oldest sample members (age 64) experience hazards 40-50 times greater than the youngest members (age 25) under these models. The age-squared predictor, which was added to account for the possibility of a non-linear relationship between age and mortality, was not significant in any of the models fit. The hazard experienced by females was around 70% of that experienced by males, under the model, holding all other predictors constant. The interaction between age and sex was small and negative, and corresponds to an additional 10% difference in hazard in males at the oldest age (64) compared to females at the youngest age (25). These results are consistent with those found in section 3.2. The results for age, sex and their interactions were stable to the addition of marital status, the wealth indicator, education level and self-reported health in 2001 as additional predictors.

Ethnicity appears strongly related to mortality, with white individuals experiencing hazards 1.2–1.5 times higher than Mixed individuals, 1.3–1.4 times higher than Asian individuals, 1.3–1.6 times higher than Black individuals, and 1.9 times higher than individuals with other ethnicities, under the models fit, with all other variables held constant.

Hazard ratio estimates for different geographical regions vary between 0.9 and 1.3, relative to London, across the three sets of models. They are greatest in northern regions—the North East, the North West and Yorkshire and The Humber—and smallest in southern regions—London, the South East, the South West and the East of England. The estimates are dependent on the inclusion of other socioeconomic and health predictors. For example, when only age, sex and ethnicity are held constant, hazards are observed to be around 8% lower in the South East than London. However when marital status, wealth levels and education levels are additionally held constant, hazards are observed to be 7% higher in the South East relative to London; this increases to 9% when self-reported health indicators are included. This provides evidence that the reduced hazard rate observed in this group is explained by their marital status, and higher levels of education and

wealth. Under the model which adjusts for socioeconomic predictors, London is the region with the lowest observed hazard rate, while the North East is the region with the highest rate.

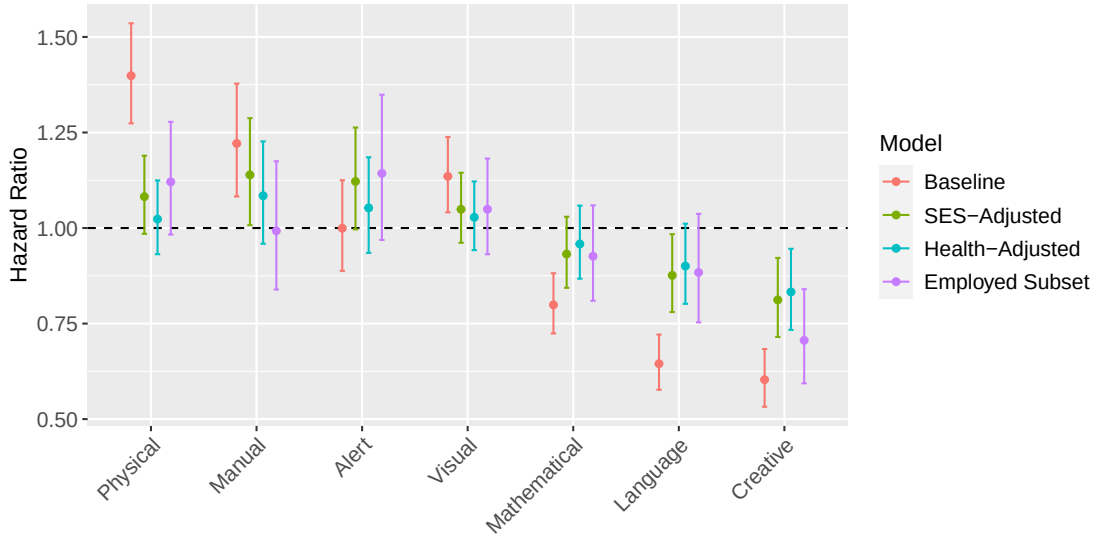
Adjusting for other socioeconomic predictors, marriage, wealth and education all predict reduced hazard rates. Married and separated individuals experience around 71% and 88% of the hazard rate of single individuals; further adjustment for self-reported health does not change this relationship substantially. Relative to wealth indicator band A, corresponding to house and car ownership, individuals in each of bands B–E were observed to experience greater hazards than those in higher bands. Individuals in communal establishments experienced the highest hazards, with a hazard ratio equal to that from a difference of fourteen years of age. Individuals reporting higher levels of education are observed to have lower hazards, with the greatest differences observed between individuals reporting some vs no qualifications (Level 1 vs no recorded qualifications), and between individuals reporting university-level vs pre-university qualifications (Level 4-5 vs Level 3). The relationship between education and mortality is weaker than its relationship with the wealth indicator, and is comparable to differences observed between different regions.

Finally, the relationship between self-reported health and mortality is strong and negative. Individuals reporting either a long-term limiting illness or "Not Good" general health have hazards 1.40 and 1.93 times greater than individuals without either of these conditions, under the model which adjusts for socioeconomic predictors. This confirms that the self-reported health measures have a strong correlation with mortality.

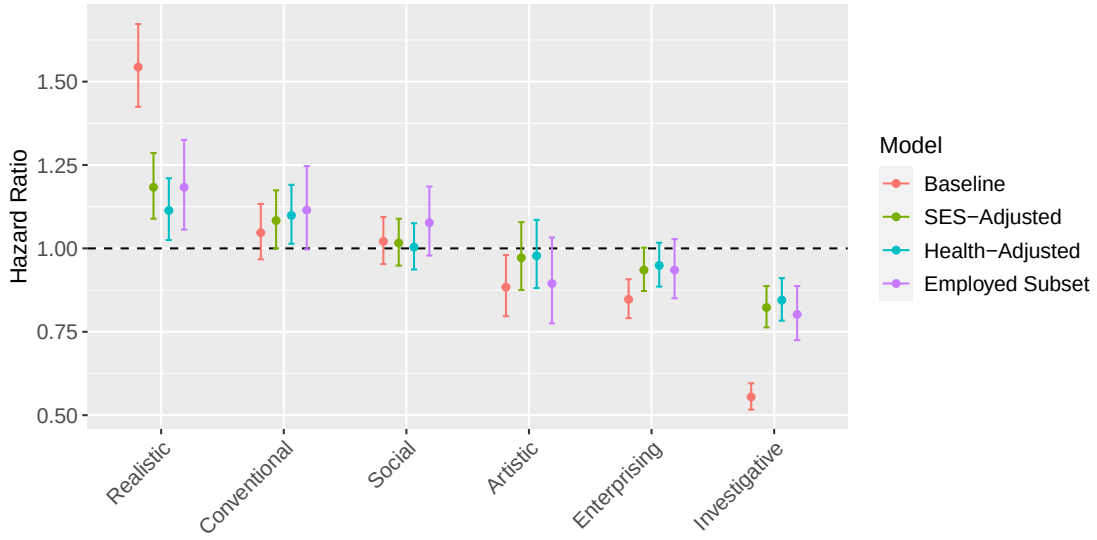
5.2.2 Occupation Features

Parameter estimates over each of the eight sets of occupation features are shown in figure 12 for the baseline, SES-adjusted, health-adjusted and employed subset models. The estimates correspond to the hazard ratio between the highest and lowest scores observed for the corresponding predictors; this provides a natural scale for evaluating and comparing the strength of the relationship of mortality with the occupation features. For example, the UK 2000 SOC occupation which scores highest on the "Abilities: Physical" predictor is *Fitness Instructors*, while the occupation which scores lowest is *Market Research Interviewers*. If the two occupations were identical over all other "Abilities" predictors, then a fitness instructor would experience a hazard 40% (27%, 54%) greater than a market research interviewer of the same baseline characteristics, under the baseline model.

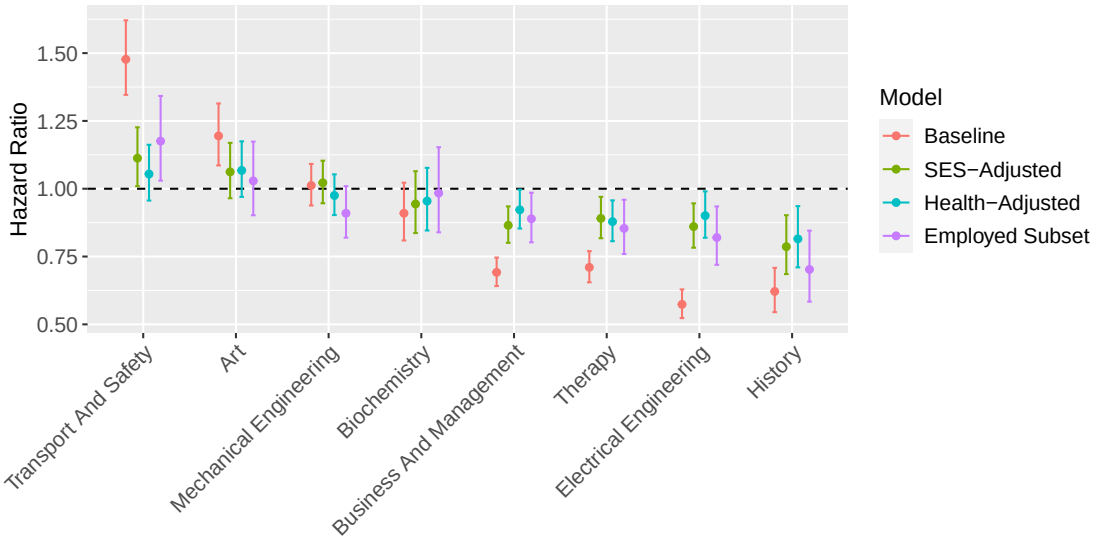
(a) Hazard Ratios: Abilities



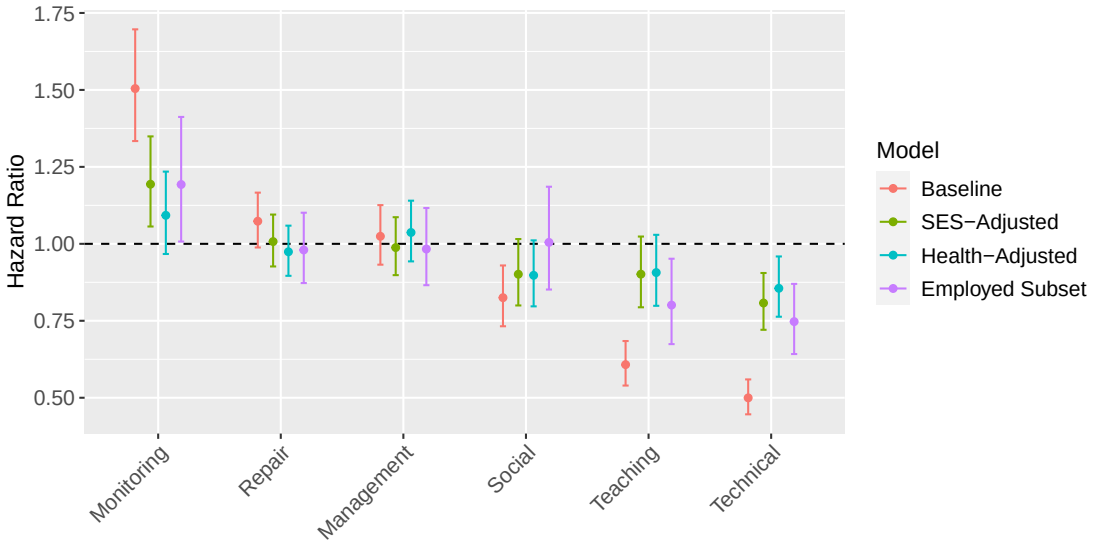
(b) Hazard Ratios: Interests



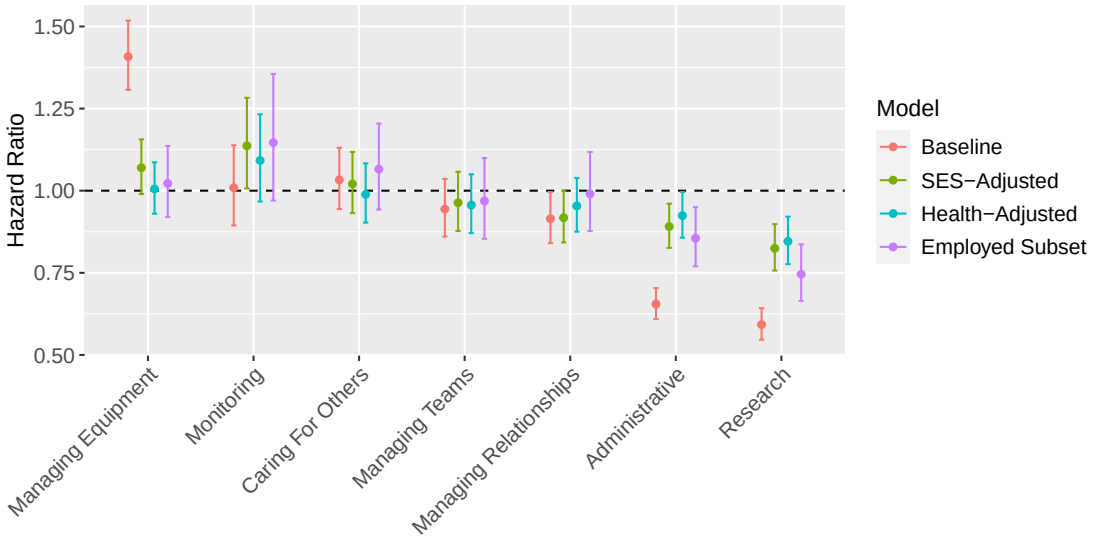
(c) Hazard Ratios: Knowledge



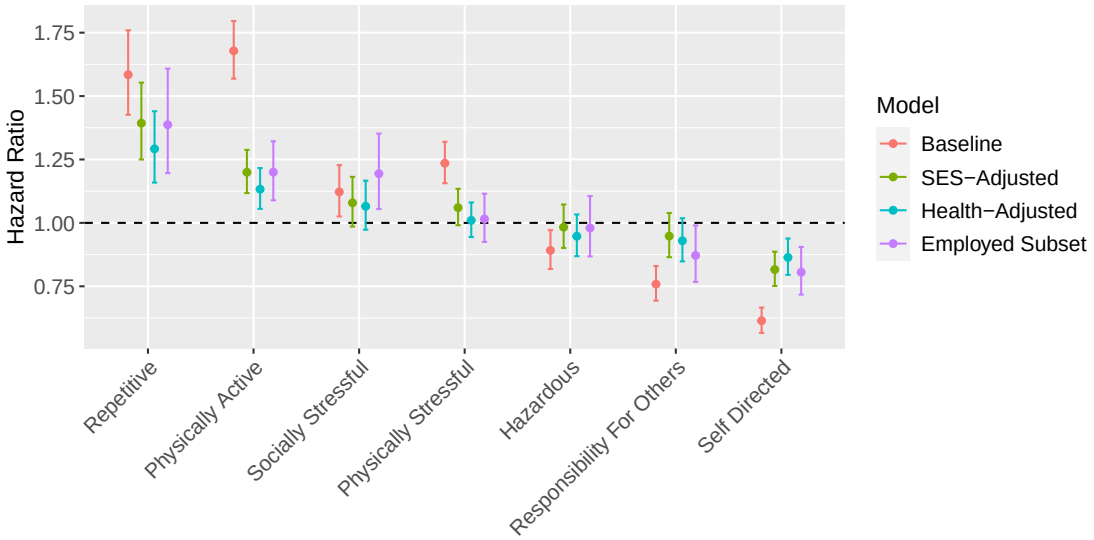
(d) Hazard Ratios: Skills



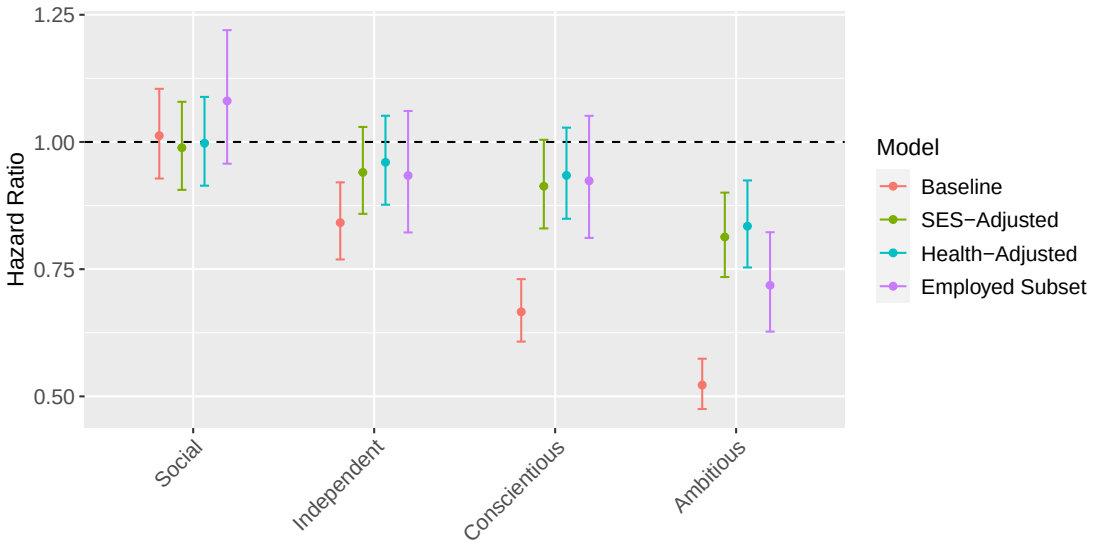
(e) Hazard Ratios: Work Activities



(f) Hazard Ratios: Work Context



(g) Hazard Ratios: Work Styles



(h) Hazard Ratios: Work Values

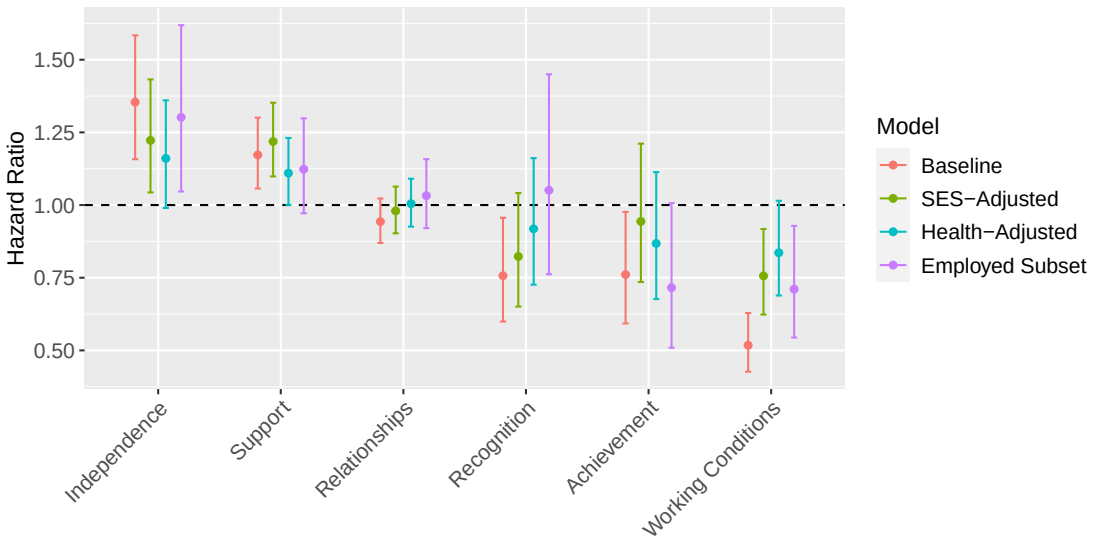


Figure 12: Hazard ratios and 95% confidence intervals for the occupation features. Results are shown for the baseline, SES-adjusted, health-adjusted, and employed subset models. Source: ONS LS.

Under the baseline model, the hazard ratios for the majority of the occupation features are significant at the 5% level, varying between 0.50 and 1.68. After adjusting for socioeconomic predictors, hazard ratios vary between 0.76 and 1.39; in general, the strength of the relationship between the occupation features and mortality is substantially reduced when socioeconomic predictors are taken into account. Adjustment for self-reported health further weakens this relationship; under this model, hazard ratios vary from 0.82 to 1.29. Models based on only the subset of the LS members who report being in employment in 2001 give results which are generally to the SES-adjusted model. This set of models was fit on approximately half the events of the previous three models, and the confidence intervals are correspondingly wider. Below, each of the occupation feature domains are examined in more detail.

The importance of physical, manual, visual and 'alert' abilities in the occupation appear to predict slightly greater hazards, while mathematical, language and creative skills predict slightly smaller hazards. Of these, creative abilities appear to have the strongest relationship, with a hazard ratio of 0.81 (0.71, 0.92) under the SES-adjusted model. Physical abilities have a strong relationship with mortality under the baseline model, but this relationship is weakened when socioeconomic predictors are included in the model.

In the occupational interests models, "Realistic" and "Conventional" work environments predict greater hazards, while "Enterprising" and "Investigative" work environments predict smaller hazards.

Greater knowledge in most domains—biochemistry, business and management, therapy, electrical engineering, and history—predicts smaller hazards; the exception to this rule is knowledge of transport and safety. Knowledge of art and mechanical engineering does not appear to have a substantial relationship with mortality.

The importance of social, teaching and technical skills in the occupation also predicts decreased hazard, while the importance of monitoring skills predicts greater hazards. Repair and management skills do not appear to be strongly related to mortality.

The work activity which most strongly predicts hazard is monitoring. The importance of managing equipment and caring for others also predicts slightly greater hazard; managing teams, managing relationships, administrative and research activities all predict smaller hazard. Occupations scoring highest on the importance of research activities experience 82% (76%, 90%) of the hazards of occupations scoring lowest on this predictor, under the SES-adjusted model.

Repetitive, physically active, socially stressful and physically stressful work contexts all predict greater mortality, while responsibility for others and self-directed work environments predict smaller mortality. Hazardous work contexts

do not appear to be related to mortality hazard.

Ambitious work styles appear to predict substantially smaller hazard; independent and conscientious work styles also appear to predict smaller hazard.

On the other hand, independence as a work value predicts greater hazard. Support as a work value also predicts greater hazards; the work values of recognition, achievement and working conditions predict smaller hazards. The "Relationships" work value does not appear to be strongly related to hazard.

A full table of results can be found in appendix F.

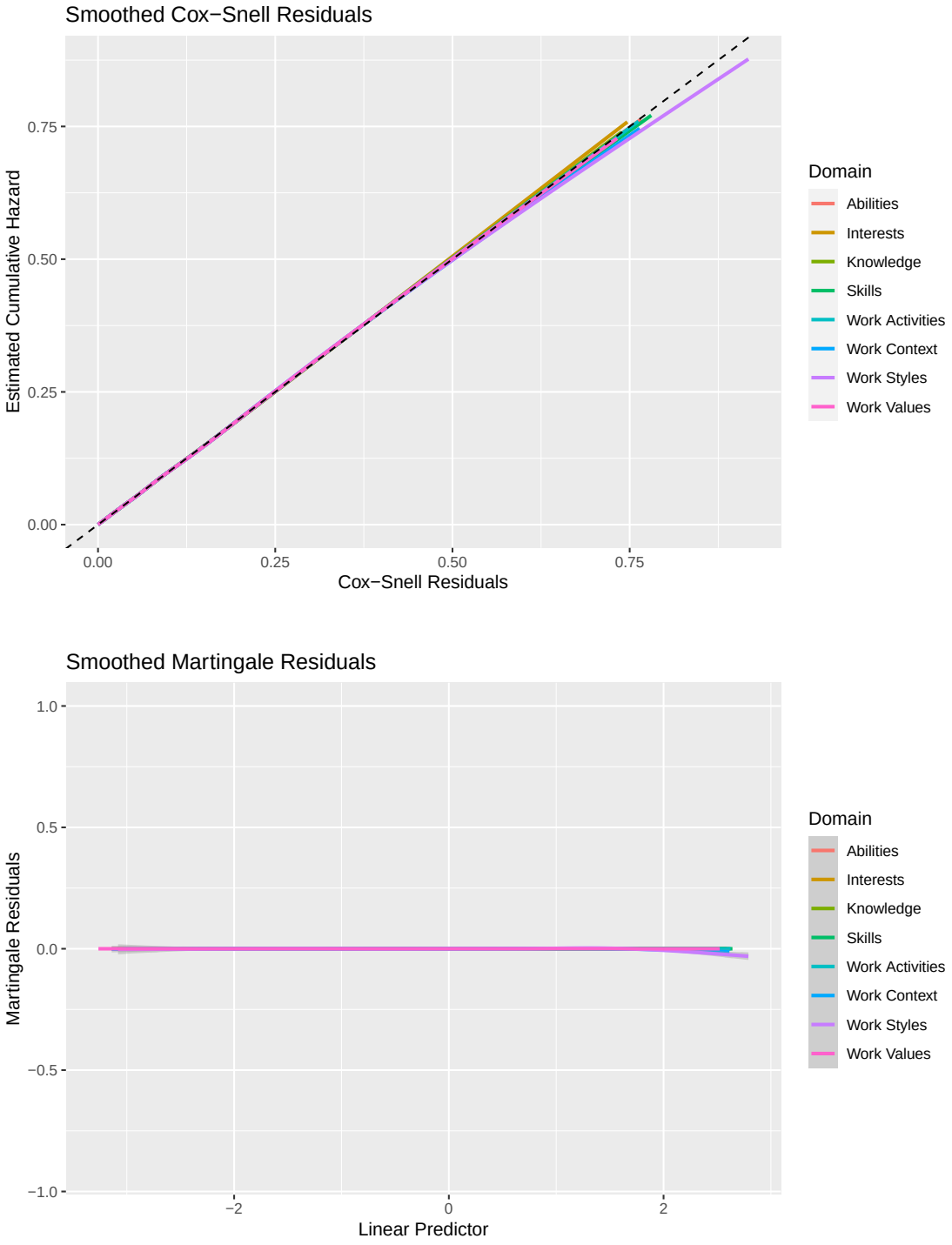
5.2.3 Model Evaluation

Plots of the Cox-Snell and Martingale residuals for the baseline, SES-adjusted and health-adjusted models are shown in figure 13. To avoid disclosure risk, individual residuals are not reported. Instead, a smooth curve—specifically a generalised additive model (GAM)—was fit to each set of residuals, in order to show general trends their values.

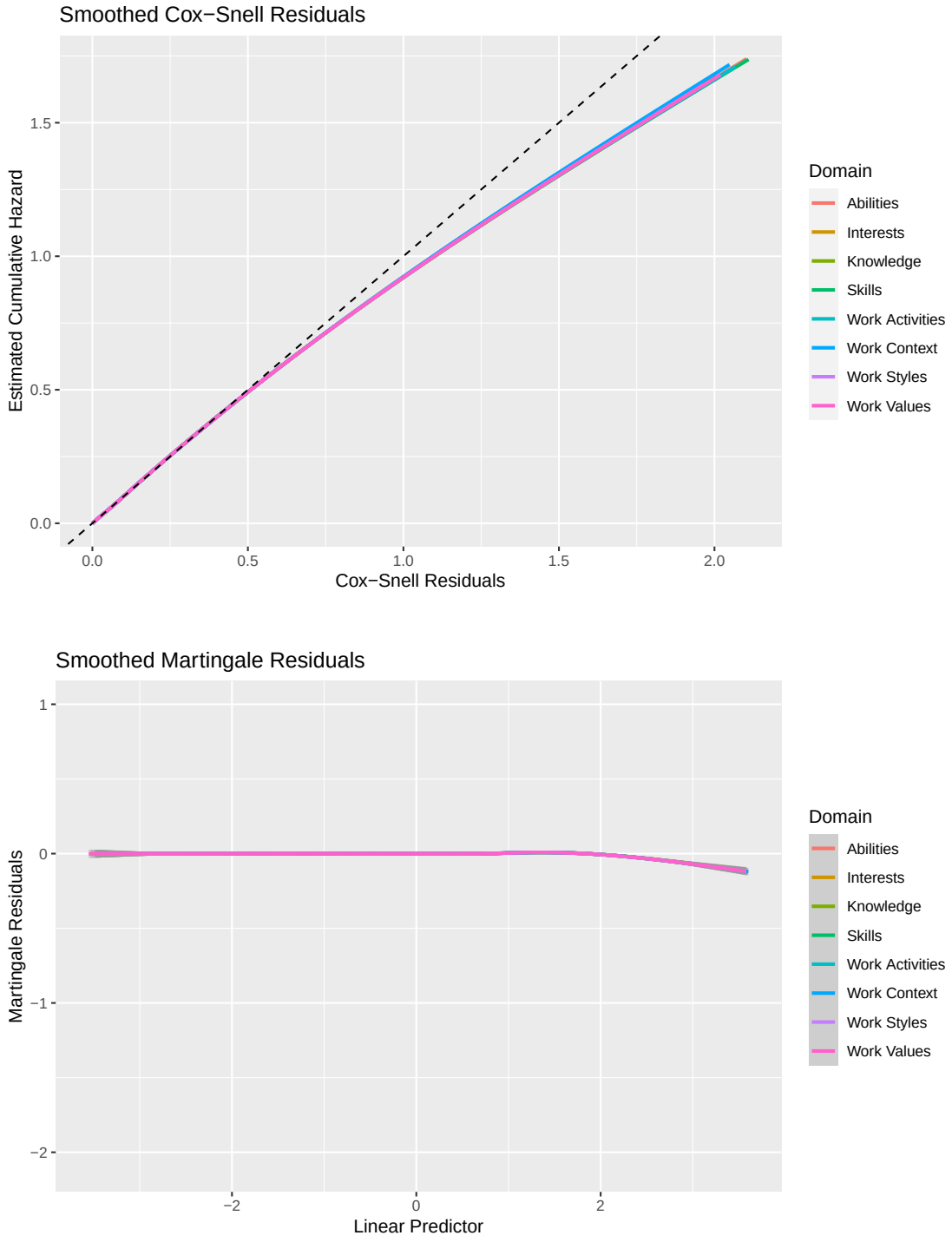
The Cox-Snell residuals show a consistent pattern across each set of LS and occupation features. For each of the baseline, SES-adjusted and health-adjusted models, the smoothed curves are similar across different sets of occupation features. The fit is best for the baseline model, which has minimal deviation from the line $y = x$, and somewhat worse in the SES-adjusted and health-adjusted models. The concave curves observed for the SES- and health-adjusted models indicate that some individuals survived longer than expected under the model.

The martingale residuals show a similar story: the smoothed curves are closest to the $y = 0$ line for the baseline model, with deviations from $y = 0$ observed under the SES-adjusted and health-adjusted models. The deviations occur in the high tail of the linear predictor, at around $\boldsymbol{x}^T\beta > 2$, and are in the negative direction. This indicates that, in general, individuals with high linear predictors were less likely to die during follow-up than predicted under the model. The smoothed curves are similar across models using the same set of LS predictors. These findings are discussed further in section 6.

(a) Baseline Model



(b) SES-Adjusted Model



(c) Health-Adjusted Model

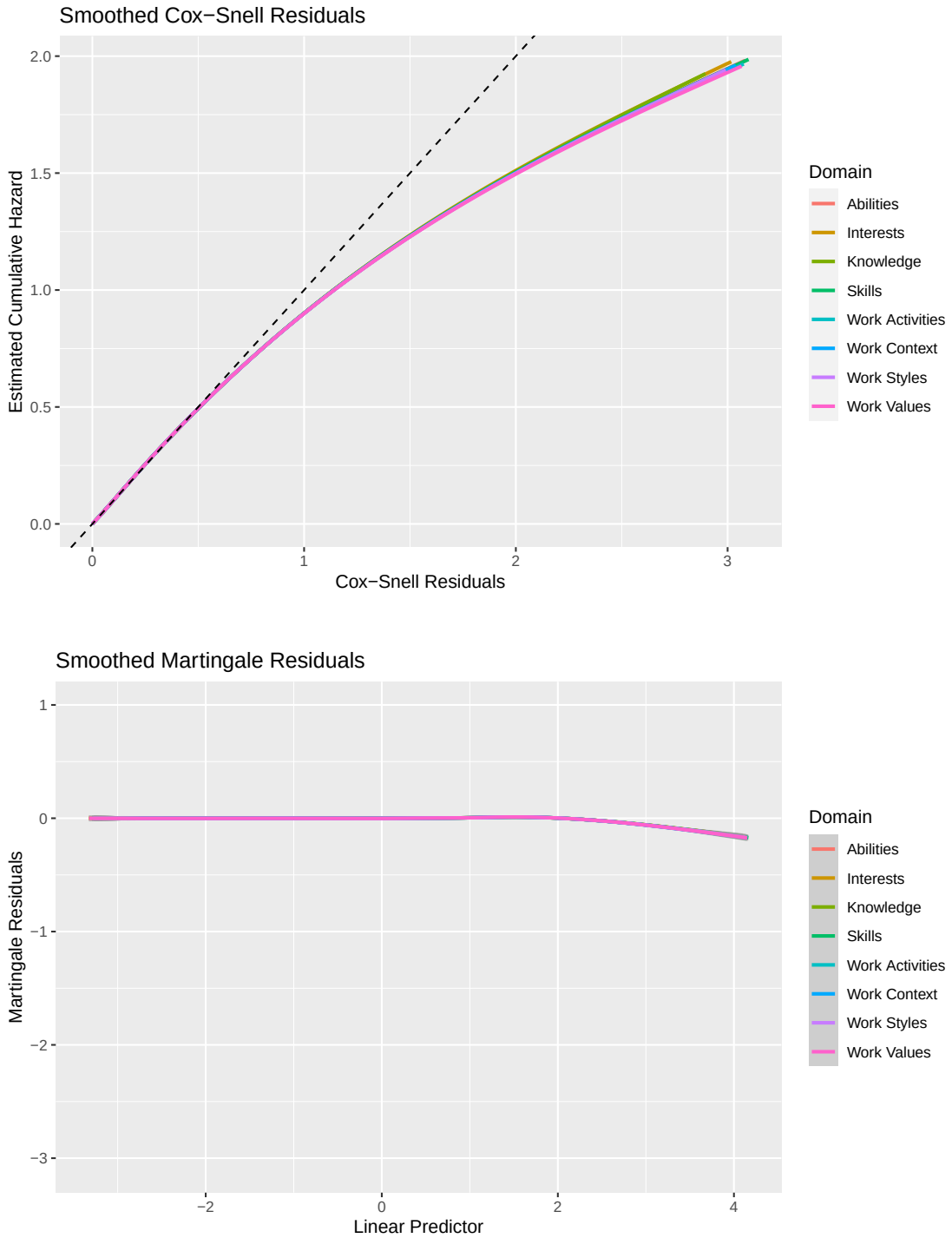


Figure 13: Smoothed Cox-Snell and Martingale residuals under the baseline, SES-adjusted and health-adjusted models. Source: ONS LS.

6 Discussion

6.1 LS Predictors

The results for the LS predictors (section 5.2.1) are generally consistent with previous findings. These include the findings that males, white and mixed ethnicities, single individuals, and individuals with the fewest assets or the fewest educational qualifications experience the greatest mortality risk, while females, black and 'other' ethnicities, married individuals, and individuals with more assets or educational qualifications experience the lowest risks (Oksuzyan *et al.* 2008; ONS 2021b; Rendall *et al.* 2011; Braveman *et al.* 2011). For example, the hazard ratio observed for white vs black ethnicity individuals is greater than the male-female ratio, and is roughly equal to equivalent to five years of age. This is consistent with findings from England and Wales[8] based on 2011 census data (ONS 2021b), which record a 5.8 year difference in life expectancy between these two groups in the years 2011–2014.

The hazard ratio corresponding to 1 year of age is increased by 25% from 1.09 (1.07, 1.10) to 1.12 (1.10, 1.14) when marital status, wealth and education are taken into account, in addition to the baseline LS predictors. This indicates that greater marriage rates, wealth or education in older individuals moderates their greater rate of mortality compared to younger ages. The age-squared term was not significant in any of the models fit. This may indicate that the relationship between age and log-hazard is purely linear, or simply that the age-squared term did not well-capture any non-linear relationship that may exist.

6.2 Occupation Features

The relationships found between the occupation features and mortality could be due to a number of effects. As discussed in section 1.2, the direct effects of physical and psychosocial job exposure, and the indirect effects of income, prestige, lifestyle and other factors represent one possible explanation. Selection effects—whereby healthier individuals select are selected into different types of occupations than less healthy individuals—represent another possible explanation. A final possibility is confounding, where an unobserved variable causally influences both the exposure and outcome variables.

The occupation features most strongly related to mortality under the SES-adjusted model are repetitive work contexts and the work value of working conditions. Repetitive work contexts—characterised by repetitive tasks, repetitive physical motions, high levels of automation and time pressure—predict the greatest mortality risk, with a hazard ratio of 1.39 (1.24, 1.56). Occupations scoring highly

[8]Results from other countries differ: see e.g. (Woolf and Schoomaker 2019).

on this feature include communications operators and production occupations. Both health-related selection into these occupations, and the direct and indirect effects of the occupations (particularly income) appear plausible as explanations for this relationship. The hazard ratio is reduced to 1.29 (1.15, 1.45) when initial health status is adjusted for, suggesting that health selection may explain some but not all of this relationship. Working conditions—which reflects the occupation's working conditions and job security—predict the smallest mortality risk, with a hazard ratio of 0.76 (0.63, 0.91). Occupations scoring highly on this feature include managers and lawyers. Again, both the direct and indirect effects of occupation appear plausible as explanations.

Other relationships are surprisingly small. For example, hazardous work contexts—characterised by exposure to disease, radiation, and hazardous conditions or equipment—did not appear to predict mortality substantially. The occupations scoring highest on this feature were medical occupations. This might suggest that the hazards experienced in these occupations are not substantial enough to affect mortality, or that the increased mortality risk they produce is balanced by the relatively high income and prestige of these occupations (Fletcher *et al.* 2011).

Lee (2011) suggests that the importance of cognitive abilities and skills—such as originality, reasoning or language skills—may be the characteristic of occupations most consistently and strongly related to mortality. The results found here generally support this conclusion, with cognitive abilities (mathematical, language, and creativity), cognitive skills (such as technical, mathematical and research skills), and investigative interests ("people who like to work with data") all predicting reduced hazards. By contrast, occupations scoring highly physical and manual abilities—for example construction and production occupations—appear to predict greater hazards.

Bosma *et al.* (1997) found that among British civil servants, low job control (or decision latitude) was associated with increased risk of coronary heart disease, even after adjusting for employment grade, negative affectivity and coronary risk factors. While all-cause mortality was the focus of this project, the results here also support this finding: self-directed work contexts—defined by high decision latitude, unstructured work, high-impact decisions—were among the strongest predictors of hazards, with a hazard ratio of 0.82 (0.76, 0.88). Occupations scoring highly on the repetitive work contexts feature may also correspond to jobs with particularly low control.

In general, the strength of the results is less than that found previously: here hazard ratios under the SES-adjusted model vary between 0.7–1.4, whereas in Lee (2011) they vary over a wider range of 0.5–1.7.

6.3 Limitations and Future Work

One weakness of the general approach taken in this project is the error introduced when describing job characteristics using occupation-level features. The magnitude of this error could be partly estimated by (1) further evaluating how well the mapped occupation features describe UK occupations, (2) evaluating how this accuracy varies over the follow-up period, and (3) evaluating the extent to which the features vary over a single occupation. As mentioned in section 5.1, these evaluations may be difficult without without occupational expertise or data sets for empirical comparison.

Another substantial weakness of the approach is that occupation (along with other variables) was measured at a single point in time, and so the results do not reflect the whole histories of individuals' working lives. For example, Moore and Hayward (1990) show that mortality risks based on current occupation differ substantially from those based on longest-held occupation. This may pose a particular problem here due to the relatively long follow-up period used—by the end of the follow-up period in 2017, individuals' characteristics may have drifted substantially from their values measured in 2001. This magnitude of this error can partly be evaluated by measuring the sensitivity of the results to varying follow-up periods.

While the model goodness-of-fit appears reasonable (section 5.2.3), further checks would be desirable to identify the reasons for the slight discrepancy between the model and the data, and ensure that the resulting bias in the estimated parameters is not substantial. This can be achieved by more detailed analysis of outlying and influential data points (or groups), and more detailed evaluation of the accuracy of the proportional hazards assumption for each predictor. For example, (Arjas 1988) plots are particularly effective for evaluating the proportional hazards assumption for a given predictor (Persson and Khamis 2007).

Given the importance of the existing LS predictors for predicting mortality, including interactions between these terms—for example between age, sex and ethnicity—may further improve model fit. Alternatively, a stratified analysis could be used. The use of geographic predictors is also limited. In addition to the high level of aggregation of government office regions (GORs), this variable describes geographies only in terms of their location, rather than their health-relevant dimensions (e.g. population density, air quality, etc.). In future work it would be desirable to include finer-grained and more health-relevant geographic predictors.

Finally, a more detailed picture of the relationship between occupation and mortality could be obtained by including additional individual- and occupation-level occupation features. For example, employment type, hours worked and organisation size are all measured by the LS, and could be included as individual-level variables; and occupational income could be included at the occupation level.

References

[1] H. Bosma, M. G. Marmot, H. Hemingway, A. C. Nicholson, E. Brunner and S. A. Stansfeld, 'Low job control and risk of coronary heart disease in Whitehall II (prospective cohort) study,' *BMJ (Clinical research ed.)*, vol. 314, no. 7080, pp. 558–565, 1997. DOI: 10.1136/bmj.314.7080.558.

[2] K. Lee, 'Essays in health economics: Empirical studies on determinants of health.,' Ph.D. dissertation, George Mason University, 2011.

[3] N. J. Johnson, P. D. Sorlie and E. Backlund, 'The Impact of Specific Occupation on Mortality in the U.S. National Longitudinal Mortality Study,' *Demography*, vol. 36, no. 3, pp. 355–367, 1999. DOI: 10.2307/2648058.

[4] P. J. Baxter, T.-C. Aw, A. Cockcroft, P. Durrington and J. M. Harrington, Eds., *Hunter's Diseases of Occupations*, 10th edition. London: CRC Press, 2010, ISBN: 978-0-340-94166-9.

[5] J. V. Johnson and E. M. Hall, 'Job strain, work place social support, and cardiovascular disease: A cross-sectional study of a random sample of the Swedish working population.,' *American Journal of Public Health*, vol. 78, no. 10, pp. 1336–1342, 1988.

[6] E. Backlund, P. D. Sorlie and N. J. Johnson, 'The shape of the relationship between income and mortality in the United States: Evidence from the National Longitudinal Mortality Study,' *Annals of Epidemiology*, vol. 6, no. 1, pp. 12–20, 1996. DOI: 10.1016/1047-2797(95)00090-9.

[7] M. Marmot, 'Social determinants of health inequalities,' *The Lancet*, vol. 365, no. 9464, pp. 1099–1104, 2005. DOI: 10.1016/S0140-6736(05)71146-6.

[8] Registrar General, 'Fourteenth Annual Report of the Registrar General of Births, Deaths, and Marriages in England,' HMSO, London, Tech. Rep., 1855.

[9] A. J. Fox and A. M. Adelstein, 'Occupational mortality: Work or way of life?' *Journal of Epidemiology & Community Health*, vol. 32, no. 2, pp. 73–78, 1978. DOI: 10.1136/jech.32.2.73.

[10] S. V. Katikireddi, A. H. Leyland, M. McKee, K. Ralston and D. Stuckler, 'Patterns of mortality by occupation in the UK, 1991–2011: A comparative analysis of linked census and mortality records,' *The Lancet Public Health*, vol. 2, no. 11, e501–e512, 2017. DOI: 10.1016/S2468-2667(17)30193-7.

[11] A. J. Fox, P. O. Goldblatt and A. M. Adelstein, 'Selection and Mortality Differentials,' *Journal of Epidemiology and Community Health*, vol. 36, no. 2, pp. 69–79, 1982. DOI: 10.1136/jech.36.2.69.

[12] A. J. McMichael, 'Standardized mortality ratios and the "healthy worker effect": Scratching beneath the surface,' *Journal of occupational medicine.*, vol. 18, no. 3, pp. 165–168, 1976. DOI: 10.1097/00043764-197603000-00009.

[13] A. J. Fox and P. F. Collier, 'Low Mortality Rates in Industrial Cohort Studies Due to Selection for Work and Survival in the Industry,' *Journal of Epidemiology & Community Health*, vol. 30, no. 4, pp. 225–230, 1976. DOI: 10.1136/jech.30.4.225.

[14] J. M. Fletcher, J. L. Sindelar and S. Yamaguchi, 'Cumulative effects of job characteristics on health,' *Health Economics*, vol. 20, no. 5, pp. 553–570, 2011. DOI: 10.1002/hec.1616.

[15] D. Coggon, E. C. Harris, T. Brown, S. Rice and K. T. Palmer, 'Work-related mortality in England and Wales, 1979–2000,' *Occupational and Environmental Medicine*, vol. 67, no. 12, pp. 816–822, 2010. DOI: 10.1136/oem.2009.052670.

[16] M. G. Marmot, S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, A. Feeney, M. G. Marmot and G. D. Smith, 'Health inequalities among British civil servants: The Whitehall II study,' *The Lancet*, vol. 337, no. 8754, pp. 1387–1393, 1991. DOI: 10.1016/0140-6736(91)93068-K.

[17] K. L. Tang, R. Rashid, J. Godley and W. A. Ghali, 'Association between subjective social status and cardiovascular disease and cardiovascular risk factors: A systematic review and meta-analysis,' *BMJ Open*, vol. 6, no. 3, e010137, 2016. DOI: 10.1136/bmjopen-2015-010137.

[18] National Center for O*NET Development, *About O*NET*, 2021. [Online]. Available: https://www.onetcenter.org/overview.html (visited on 10/07/2021).

[19] A. Dickerson and D. Morris, 'The Changing Demand for Skills in the UK,' Centre for Vocational Education Research, Research Paper 20, 2019.

[20] Office for National Statistics, *ONS Longitudinal Study*, 2021. [Online]. Available: https://www.ons.gov.uk/aboutus/whatwedo/paidservices/longitudinalstudyls (visited on 03/07/2021).

[21] M. J. Handel, 'The O*NET content model: Strengths and limitations,' *Journal for Labour Market Research*, vol. 49, no. 2, pp. 157–176, 2016. DOI: 10.1007/s12651-016-0199-8.

[22] L. Van Der Maaten, E. Postma and J. Van Den Herik, 'Dimensionality Reduction: A Comparative Review,' Tilburg University, Technical Report 2009-005, 2009.

[23] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata and A. C. Telea, 'Toward a Quantitative Survey of Dimension Reduction Techniques,' *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 2153–2173, 2021. DOI: 10.1109/TVCG.2019.2944182.

[24] R. L. Gorsuch, *Factor Analysis*. L. Erlbaum Associates, 1983, Google-Books-ID: Y5kfvgAACAAJ, ISBN: 978-0-89859-202-3.

[25] D. J. Bartholomew, M. Knott and I. Moustaki, *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley & Sons, 2011, ISBN: 978-0-470-97192-5.

[26] UK Department for Education, *Labour Market Information (LMI) for All*, 2021. [Online]. Available: https://www.lmiforall.org.uk/about-lmi-for-all/ (visited on 11/08/2021).

[27] US Bureau of Labor Statistics, *Occupational Employment Statistics (OES) Survey*, 2021. [Online]. Available: https://www.bls.gov/oes/ (visited on 11/08/2021).

[28] Office for National Statistics, *The Relationship between SOC2010 and SOC2000*, 2012. [Online]. Available: https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010 (visited on 11/07/2021).

[29] A. Syed, M. Eddolls and P. Pawelek, 'Analysis of job changers and stayers,' in *Economic review: April 2019 - Office for National Statistics*, Apr. 2019. [Online]. Available: https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/economicreview/april2019/analysisofjobchangersandstayers (visited on 10/08/2021).

[30] ONS, 'Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland,' Tech. Rep. Release Number MYE14, Jun. 2021. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland (visited on 30/08/2021).

[31] P. Bolton, 'Education: Historical Statistics,' *House of Commons Library*, vol. Standard Note SN/SG/4252, 2012.

[32] E. L. Kaplan and P. Meier, 'Nonparametric Estimation from Incomplete Observations,' *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. DOI: 10.2307/2281868.

[33] D. L. Wingard, 'The sex differential in morbidity, mortality, and lifestyle,' *Annual Review of Public Health*, vol. 5, pp. 433–458, 1984. DOI: 10.1146/annurev.pu.05.050184.002245.

[34] A. Oksuzyan, K. Juel, J. W. Vaupel and K. Christensen, 'Men: Good health and high mortality. Sex differences in health and aging,' *Aging Clinical and Experimental Research*, vol. 20, no. 2, pp. 91–102, 2008. DOI: 10.1007/BF03324754.

[35] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. New York: Springer-Verlag, 2006, ISBN: 978-0-387-31073-2.

[36] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012, ISBN: 978-0-262-30432-0.

[37] J. L. Horn, 'A rationale and test for the number of factors in factor analysis,' *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965. DOI: 10.1007/BF02289447.

[38] H. H. Harman and W. H. Jones, 'Factor analysis by minimizing residuals (minres),' *Psychometrika*, vol. 31, no. 3, pp. 351–368, 1966. DOI: 10.1007/BF02289468.

[39] M. S. Bartlett, 'The Statistical Conception of Mental Factors,' *British Journal of Psychology. General Section*, vol. 28, no. 1, pp. 97–104, 1937. DOI: 10.1111/j.2044-8295.1937.tb00863.x.

[40] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed., ser. Statistics for Biology and Health. New York: Springer-Verlag, 2003, ISBN: 978-0-387-95399-1.

[41] D. Steinsaltz, 'SB3.2 Statistical Lifetime-Models (Lecture Notes),' 2019. [Online]. Available: https://www.steinsaltz.me.uk/teaching (visited on 17/08/2021).

[42] N. E. Breslow, 'Discussion of the paper by D. R. Cox,' *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[43] B. Efron, 'The Efficiency of Cox's Likelihood Function for Censored Data,' *Journal of the American Statistical Association*, vol. 72, no. 359, pp. 557–565, 1977. DOI: 10.2307/2286217.

[44] D. Y. Lin, L. J. Wei and Z. Ying, 'Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals,' *Biometrika*, vol. 80, no. 3, pp. 557–572, 1993. DOI: 10.2307/2337177.

[45] D. R. Cox and E. J. Snell, 'A General Definition of Residuals,' *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no. 2, pp. 248–275, 1968.

[46] W. Nelson, 'Hazard Plotting for Incomplete Failure Data,' *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969. DOI: 10.1080/00224065.1969.11980344.

[47] O. Aalen, 'Nonparametric Inference for a Family of Counting Processes,' *The Annals of Statistics*, vol. 6, no. 4, pp. 701–726, 1978. DOI: 10.1214/aos/1176344247.

[48] W. E. Barlow and R. L. Prentice, 'Residuals for Relative Risk Regression,' *Biometrika*, vol. 75, no. 1, pp. 65–74, 1988. DOI: 10.2307/2336435.

[49] T. M. Therneau, 'Survival: A package for survival analysis in R,' manual, 2020, R package version 3.2-13. [Online]. Available: https://CRAN.R-project.org/package=survival.

[50] ONS, *Ethnic differences in life expectancy and mortality from selected causes in England and Wales: 2011 to 2014*, Jul. 2021. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/articles/ethnicdifferencesinlifeexpectancyandmortalityfromselectedcausesinenglandandwales/2011to2014#data-sources-and-quality (visited on 28/08/2021).

[51] M. S. Rendall, M. M. Weden, M. M. Favreault and H. Waldron, 'The Protective Effect of Marriage for Survival: A Review and Update,' *Demography*, vol. 48, no. 2, pp. 481–506, 2011. DOI: 10.1007/s13524-011-0032-5.

[52] P. Braveman, S. Egerter and D. R. Williams, 'The Social Determinants of Health: Coming of Age,' *Annual Review of Public Health*, vol. 32, no. 1, pp. 381–398, 2011. DOI: 10.1146/annurev-publhealth-031210-101218.

[53] S. H. Woolf and H. Schoomaker, 'Life Expectancy and Mortality Rates in the United States, 1959-2017,' *JAMA*, vol. 322, no. 20, pp. 1996–2016, 2019. DOI: 10.1001/jama.2019.16932.

[54] D. Moore and M. Hayward, 'Occupational Careers and Mortality of Elderly Men,' *Demography*, vol. 27, no. 1, pp. 31–53, 1990, ISSN: 1533-7790.

[55] E. Arjas, 'A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model,' *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 204–212, 1988. DOI: 10.2307/2288942.

[56] I. Persson and H. J. Khamis, 'A Comparison of Graphical Methods for Assessing the Proportional Hazards Assumptions in the Cox Model,' en, *Journal of Statistics and Applications*, vol. 2, no. 1-4, pp. 1–32, 2007.

[57] US National Research Council, *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET)*, M. L. Hilton and N. T. Tippins, Eds. Washington, D.C., UNITED STATES: National Academies Press, 2010.

[58] U. D. of Labor, *The New DOT: A Database of Occupational Titles for the Twenty-First Century*. Washington, DC: Author, 1993.

[59] M. Cifuentes, J. Boyer, D. A. Lombardi and L. Punnett, 'Use of O*NET as a job exposure matrix: A literature review,' *American Journal of Industrial Medicine*, vol. 53, no. 9, pp. 898–914, 2010. DOI: 10.1002/ajim.20846.

[60] L. Hattersley and R. Creeser, *The Longitudinal Study, 1971-1991: History, Organisation and Quality of data*. London: HMSO, 1995.

[61] N. Shelton, C. E. Marshall, R. Stuchbury, E. Grundy, A. Dennett, J. Tomlinson, O. Duke-Williams and W. Xun, 'Cohort Profile: The Office for National Statistics Longitudinal Study (The LS),' *International Journal of Epidemiology*, vol. 48, no. 2, 383–384g, 2019.

[62] J. L. Holland, *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*. Psychological Assessment Resources, 1997, ISBN: 978-0-911907-27-8.

[63] R. V. Dawis and L. H. Lofquist, *A Psychological Theory of Work Adjustment: An Individual-differences Model and Its Applications*. University of Minnesota Press, 1984, ISBN: 978-0-8357-7665-3.

[64] D. R. Cox, 'Regression Models and Life-Tables,' *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[65] S. Bennett, 'Analysis of survival data by the proportional odds model,' *Statistics in Medicine*, vol. 2, no. 2, pp. 273–277, 1983. DOI: 10.1002/sim.4780020223.

[66] O. O. Aalen, 'A linear regression model for the analysis of life times,' *Statistics in Medicine*, vol. 8, no. 8, pp. 907–925, 1989. DOI: 10.1002/sim.4780080803.

[67] L. J. Wei, 'The accelerated failure time model: A useful alternative to the cox regression model in survival analysis,' *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992. DOI: 10.1002/sim.4780111409.

A Sample R Code

```
#####################################################
## Author:      Christopher McDonald
##
## ~ Description ~
## * Selected R code produced during the project
## * Full code available on request
##
#####################################################

#####################################################
## Data Preparation - O*NET
#####################################################

# Function to read, preprocess, reduce and output occupation feature
↪  data
produce_onet <- function(
  n_factors = get_n_factors(),
  factor_names = get_factor_names(),
  fit_method = "ULS",
  rotate_method = "oblimin",
  score_method = "Bartlett",
  output_name = NULL,
  output_dir = "data/output/",
  onet_version_path = "data/raw/onet_17_0/",
  lmi_path = "data/raw/SOCtoMultipleONETCodes_FINAL.xlsm",
  oes_path = "data/raw/national_M2011_dl.xls",
  ukuk_path = "data/raw/Table 1a.xls"
){
  # Load
  onet <- read_onet_version(onet_version_path) %>% tidy_onet() %>%
  ↪  preprocess_onet()

  # Factor analysis
  onet <- reduce_onet_version(
    onet_version = onet,
    n_factors = n_factors,
    factor_names = factor_names,
    fit_method = fit_method,
    rotate_method = rotate_method,
    score_method = score_method
```

```
  )

  # Conversion from O*NET2010SOC -> UK2000SOC
  onet <- convert_onet(
    onet = onet,
    lmi_path = lmi_path,
    oes_path = oes_path,
    ukuk_path = ukuk_path
  )

  # Write to file
  if (is.null(output_name) | is.null(output_dir)) return(onet) else
    write_onet(
      onet = onet,
      output_name = output_name,
      output_dir = output_dir
    )
}

#####################################################
## Data Preparation - LS
#####################################################

## ls_analysis/scripts/preprocessing_script.R

# File paths
path_to_code <- "R/"
ls_file_name <- "ls.csv"
path_to_ls_data <- "data/"
path_to_data <- "data/"

# Load packages
require(tidyverse)
require(lubridate)
require(survival)
require(boot)

# Source R files for preprocessing and modeling functions
list.files(path_to_code, pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Read data
```

```r
full <- read_ls(input_name = ls_file_name, input_dir = path_to_ls_data)

# Get `codelist'
codelist <- get_codelist_ls(input_dir = path_to_data)

# Tidy data
full <- tidy_ls(full)

# Select sample
full <- select_sample_ls(full, drop_nonemployed = FALSE)

# Link O*NET data
full <- link_ls(full, path_to_data = path_to_data)

# Deal with missingness
full <- impute_ls(full, codelist)

# Derive variables
full <- derive_variables_ls(full, codelist, censoring_year = 2018L)

#####################################################
## Exploratory Data Analysis - O*NET
#####################################################

## mapping/scripts/eda_original.R

# Input settings
path_to_code <- "R/"

# Libraries
require(tidyverse)
require(lubridate)

# Scripts
list.files(path_to_code, pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Selected example occupations
selected_occupations <- tibble(
  onet2010 = c("11-1011.00", "51-2041.00", "53-6051.08"),
  title = c("Chief Executives", "Structural Metal Fabricators and
  ↪  Fitters", "Freight and Cargo Inspectors")
```

```
)

# Select a domain and scale, and get their possible values
selected_domain <- "work_styles"
selected_scale <- "importance"
selected_scale_range <- get_onet_scales() %>%
  filter(scale %in% selected_scale) %>%
  select(minimum, maximum) %>%
  as.numeric()

# Work Styles data from O*NET version 17.0
onet17_subset <- read_onet_version() %>%
  tidy_onet() %>%
  filter(domain %in% selected_domain, scale %in% selected_scale) %>%
  prettify_onet()

# Example O*NET 2010 SOC occupation scores (Work Styles)
example_occ_plot <- onet17_subset %>%
  right_join(selected_occupations, by = "onet2010") %>%
  ggplot(aes(x = descriptor, y = score, col = title)) +
  geom_point(position = position_dodge(width = 0.5)) +
  geom_errorbar(aes(ymin = lower_ci_bound, ymax = upper_ci_bound),
  ↪  position = position_dodge(width = 0.5)) +
  ylim(selected_scale_range) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),
        plot.margin = margin(10, 10, 10, 25)) +
  labs(x = "O*NET Variable", y = "Score", col = "O*NET 2010 SOC
  ↪  Occupation")

# What is distribution of data? Over all occupations
distr_plot <- onet17_subset %>%
  ggplot(aes(score, col = descriptor)) +
  geom_density(show.legend = FALSE) +
  xlim(selected_scale_range) +
  labs(x = "Score", y = "Density")

# Load data for all O*NET versions 15.1-25.0 (with explicit missing
↪  data)
onet <- read_onet_all() %>%
  tidy_onet() %>%
  complete(onet_version, onet2010, nesting(domain, scale, descriptor))
```

```r
# Fraction of occupations with complete data, as function of version
↪  and domain
completeness <- onet %>%
  group_by(onet_version, domain) %>%
  summarise(complete_fraction = 1 - sum(is.na(score)) / n(), .groups =
  ↪  "drop") %>%
  prettify_onet()
completeness_plot <- ggplot(completeness, aes(onet_version,
↪  complete_fraction, col = domain, group = domain)) +
  geom_line() +
  labs(x = "O*NET Version", y = "Completeness", col = "Domain")

# Mean relative standard deviation, by domain and version
precision <- onet %>%
  left_join(get_onet_scales(), by = "scale") %>%
  mutate(relative_sd = standard_error / (maximum - minimum)) %>%
  group_by(onet_version, domain) %>%
  summarise(relative_sd = mean(relative_sd, na.rm = TRUE), .groups =
  ↪  "drop") %>%
  filter(!is.na(relative_sd)) %>%
  prettify_onet()
precision_plot <- ggplot(precision, aes(onet_version, relative_sd, col
↪  = domain, group = domain)) +
  geom_line() +
  labs(x = "O*NET Version", y = "Relative Standard Deviation", col =
  ↪  "Domain")

# Last update date for O*NET version 17.0, by domain
onet17 <- read_onet_version() %>% tidy_onet() %>% prettify_onet() %>%
  mutate(year = year(date) + yday(date) / 365)
last_update_plot <- ggplot(onet17, aes(year, after_stat(density), fill
↪  = domain)) +
  geom_histogram(breaks = 2002:2013, position = "stack") +
  scale_x_continuous(breaks = seq(2002, 2012, by = 2)) +
  labs(x = "Year", y = "Fraction of Values Updated", fill = "Domain")

##################################################
## Exploratory Data Analysis - LS
##################################################

## ls_analysis/scripts/eda_script.R
```

```
# File path settings
path_to_code <- "R/"
ls_file_name <- "ls.csv"
path_to_ls_data <- "data/"
path_to_data <- "data/"

# Load packages
require(tidyverse)
require(lubridate)
library(survival)
library(broom)

# Source R files for preprocessing and modeling functions
list.files(path_to_code, pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Load data
data <- pipeline_load(
  ls_file_name = ls_file_name,
  path_to_ls_data = path_to_ls_data,
  path_to_data = path_to_data,
  check_values = TRUE,
  drop_nonemployed = FALSE,
  censoring_year = 2018L,
  verbose = FALSE
)

# Create all plots and save to list
eda_list <-
  get_eda(
    data = data,
    input_path = str_c(path_to_code, "produce_eda.R")
  )

## ls_analysis/R/write_eda.R

# Get all LS exploratory data analysis plots
get_eda <- function(data, input_path = "R/produce_eda.R"){
  function_names <- get_function_names(input_path) %>%
    str_subset("^eda_plot_")

  plot_names <- str_remove(function_names, "eda_plot_")
```

```
  map(function_names, ~ do.call(.x, args = list(data = data))) %>%
    set_names(plot_names)
}

## ls_analysis/R/produce_eda.R

# Sample size by age group and sex
# Function returns the plot and underlying data
eda_plot_age_sex <- function(data){
  plot_title <- "Number of individuals by Age Group and Sex"
  plot_subtitle <- "Source: ONS LS"

  data <- data %>%
    group_by(agegrp, sex) %>%
    summarise(count = n(), .groups = "drop")

  plot <- ggplot(data = data, mapping = aes(x = agegrp, y = count, fill
  ↪  = sex)) +
    geom_col(position = "dodge") +
    geom_text(aes(label = count),
              position = position_dodge(0.9),
              vjust = -0.5) +
    labs(x = "Age Group", y = "Count", title = plot_title, subtitle =
    ↪  plot_subtitle) +
    guides(fill = guide_legend(title = "Sex"))

  list(
    data = data,
    plot = plot
  )
}

# .... many other similar plots

#####################################################
## Factor Analysis - O*NET
#####################################################

## mapping/scripts/results_factor_analysis.R

# File path settings
```

```r
path_to_code <- "R/"

# Model parameters
fit_method <- "ULS"
rotate_method <- "oblimin"
score_method <- "Bartlett"

# Domain and factor names
selected_domain <- "work_styles"
n_factors <- 4
factor_names <-
  paste(
    selected_domain,
    c(
      "social",
      "ambitious",
      "conscientious",
      "independent"
    ),
    sep = "."
  )

# Libraries
require(tidyverse)
require(EFAtools)
require(readxl)
require(pdftools)
require(lubridate)

# Scripts
list.files(path_to_code, pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Load data - O*NET version 17.0 and O*NET & UK SOCs
onet <- read_onet_version() %>% tidy_onet() %>% preprocess_onet()
onet2010 <- load_onet2010() %>% mutate(title = str_trunc(title, width =
↪  60))
uk2000 <- load_uk2000()

# Select domain
onet_domain <- select_descriptors(onet, selected_domain)
```

```r
# Compute optimal number of factors via parallel analysis
parallel_analysis <- compute_n_pa(onet_domain, type = "EFA")
parallel_analysis

# Fit and rotate factor model
factor_model <- fit_fa(onet_domain, n_factors, method = fit_method,
↪  rotation = rotate_method)

# What are the loadings to each variable?
loadings <- get_loadings(factor_model, factor_names) %>%
  mutate(descriptor = str_trunc(descriptor, 40))

# View loadings
loadings_plot <- plot_loadings(loadings)
loadings_plot

# Latent scores in O*NET 2010 SOC
scores_onet2010 <- compute_scores(onet_domain, factor_model,
↪  factor_names, method = score_method)

# Highest & lowest scores in O*NET 2010 SOC
scores_onet2010_summary <- tabulate_scores(scores_onet2010, onet2010)

# Convert to UK2000 SOC
scores_uk2000 <- convert_onet(scores_onet2010)

# Highest & lowest scores in UK 2000 SOC
scores_uk2000_summary <- tabulate_scores(scores_uk2000, uk2000)

#####################################################
## Mapping - O*NET
#####################################################

## mapping/R/convert_onet.R

# Function to map features from O*NET 2010 SOC to UK 2000 SOC
convert_onet <-
  function(
    onet,
    lmi_path = "data/raw/SOCtoMultipleONETCodes_FINAL.xlsm",
    oes_path = "data/raw/national_M2011_dl.xls",
    ukuk_path = "data/raw/Table 1a.xls"
```

```
  ){
    # Build O*NET2010SOC -> UK2010SOC map
    onetuk <- build_onetuk(lmi_path = lmi_path, oes_path = oes_path)

    # Load UK2010SOC -> UK2000SOC map
    ukuk <- load_ukuk(input_path = ukuk_path)

    # Convert O*NET2010SOC -> UK2000SOC
    onet %>%
      convert_onetuk(onetuk) %>%
      convert_ukuk(ukuk)
  }

# Convert O*NET data from UK2010SOC to UK2000SOC
convert_ukuk <-
  function(
    onet,
    ukuk
  ){
    # Convert
    ukuk %>%
      left_join(onet, by = "uk2010") %>%
      group_by(across(any_of(c("onet_version", "uk2000", "sex")))) %>%
      summarise(
        across(-c(uk2010, weight), ~ weighted.mean(.x, w = weight,
        ↪  na.rm = TRUE)),
        .groups = "drop"
      )
  }

# Convert O*NET data from O*NET2010SOC to UK2010SOC
convert_onetuk <-
  function(
    onet,
    onetuk
  ){
    # Drop title column
    onet <- select(onet, -any_of("title"))

    # Convert O*NET2010SOC -> UK2010SOC
    onetuk %>%
      left_join(onet, by = "onet2010") %>%
```

```
      group_by(across(any_of(c("onet_version", "uk2010")))) %>%
      summarise(
        across(-c(onet2010, weight), ~ weighted.mean(.x, w = weight,
        ↪  na.rm = TRUE)),
        .groups = "drop"
      )
  }

#####################################################
## Full modeling pipeline - LS data
#####################################################

## ls_analysis/scripts/full_analysis_script.R

path_to_code = "R/"

default_settings <- list(
  # File input settings
  path_to_ls_data = "data/",
  path_to_data = "data/",
  ls_file_name = "ls.csv",

  # File output settings
  write_path = "output/",
  write_folder = NULL,

  # Print output?
  verbose = FALSE,

  # Preprocessing settings
  check_values = TRUE,
  drop_nonemployed = FALSE,
  censoring_year = 2018L,

  # Model fit settings / variables
  predictors = NULL,
  onet_domains = NULL,
  entry_time = NULL,
  exit_time =  "survyrs",
  event_indicator =  "deind",
  ties_method =  "efron",
```

```
  # Bootstrap settings
  n_bootstrap = 0,
  bootstrap_ci_level = 0.95,

  # Graphics plot settings
  residual_plots = c("coxsnell", "martingale"),
  smooth_hazards =  TRUE,
  residual_plot_type = "smooth",
  jitter_noise_level = 0.05,
  show_original_residuals = FALSE,
  arjas_covariates = NULL,

  # Graphics output settings
  graphics_width = 8,
  graphics_height = 5,
  graphics_device = "pdf"
)

# Load packages
require(tidyverse)
require(lubridate)
require(survival)
require(boot)

# R scripts for preprocessing and modeling functions
list.files(path_to_code, pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Predictor sets
baseline_predictors <- c("agec", "age2c", "sex", "agec_sexFemale",
↪  "ethnicity", "gors0")
ses_predictors <- c("marriage", "wealth", "hlqp0")
health_predictors <- c("illp0", "heap0")
onet_domains <- c("abilities", "interests", "knowledge", "skills",
                  "work_activities", "work_context", "work_values",
                  ↪  "work_styles")

# Settings for Baseline, SES-Adjusted, Health-Adjusted and Employed
↪  Subset models
model_list <- list(
  baseline = c(baseline_predictors),
  ses_adjusted = c(baseline_predictors, ses_predictors),
```

```
  health_adjusted = c(baseline_predictors, ses_predictors,
  ↪  health_predictors),
  employed_subset = c(baseline_predictors, ses_predictors)
) %>%
  imap(~ list_modify(
    default_settings,
    write_folder = .y,
    predictors = .x,
    onet_domains = onet_domains,
    residual_plots = c("coxsnell", "martingale", "deviance")
  ))
model_list[["employed_subset"]]$drop_nonemployed <- TRUE

# Fit all models, and write their results to file
walk(model_list, ~ do.call(full_pipeline_ls, .x))

## ls_analysis/R/pipeline_ls.R

# Function to read and prepare data set
pipeline_load <-
  function(
    ls_file_name = "ls.csv",
    path_to_ls_data = "data/",
    path_to_data = "data/",
    check_values = TRUE,
    drop_nonemployed = FALSE,
    censoring_year = 2018L,
    verbose = TRUE,
    ...
  ){
    # List of expected values
    codelist <- get_codelist_ls(input_dir = path_to_data)

    # Read data
    data <- read_ls(input_name = ls_file_name, input_dir =
    ↪  path_to_ls_data, verbose = verbose)

    # Check values
    if (check_values){
      unexpected <- check_values_ls(data = data, codelist = codelist)
      if (nrow(unexpected) > 0){
        warning(
```

```
          "Unexpected values found in LS data: variables ",
          str_c(unique(unexpected$term), collapse = ", ")
        )
        print(unexpected, n = 25)
      } else {
        message("LS values checked, and no unexpected values found.")
      }
    }

    # Preprocessing
    data <- data %>%
      tidy_ls(verbose = verbose) %>%
      select_sample_ls(drop_nonemployed = drop_nonemployed, verbose =
      ↪  verbose) %>%
      link_ls(path_to_data = path_to_data, verbose = verbose) %>%
      impute_ls(codelist = codelist, verbose = verbose) %>%
      derive_variables_ls(codelist = codelist, censoring_year =
      ↪  censoring_year, verbose = verbose)

    data
  }

# Function to fit Cox proportional hazards models
pipeline_model <-
  function(
    data,
    predictors = NULL,
    onet_domains = NULL,
    entry_time = NULL,
    exit_time = "survyrs",
    event_indicator = "deind",
    ties_method = "efron",
    n_bootstrap = 0,
    bootstrap_ci_level = 0.95,
    residual_plots = c("coxsnell", "martingale", "deviance"),
    smooth_hazards =  TRUE,
    residual_plot_type = "jitter",
    jitter_noise_level = 0.05,
    show_original_residuals = FALSE,
    ...
  ){
    # Model fitting function
```

```
  fit_coxph_onet <- function(onet_predictors){
    fit_coxph_ls(
      data = data,
      predictors = c(predictors, onet_predictors),
      entry_time = entry_time,
      exit_time = exit_time,
      event_indicator = event_indicator,
      ties = ties_method,
      n_bootstrap = n_bootstrap
    )
  }

  # Fit single model or multiple domains
  if (is.null(onet_domains)){
    coxph_model <- fit_coxph_onet(NULL)
  } else {
    coxph_model <-
      onet_domains %>%
      set_names() %>%
      map(~ str_subset(names(data), .x)) %>% # get O*NET variable
      ↪  names
      map(fit_coxph_onet)
  }

  # List of results
  list(
    model_summary = produce_model_summary(coxph_model),
    parameter_table = produce_parameter_table(coxph_model),
    graphics_list =
      produce_graphics_ls(
        coxph_model = coxph_model,
        data = data,
        residual_plots = residual_plots,
        smooth_hazards = smooth_hazards,
        residual_plot_type = residual_plot_type,
        jitter_noise_level = jitter_noise_level,
        show_original_residuals = show_original_residuals,
        arjas_covariates = arjas_covariates
      )
  )
}
```

```r
# Function to write results of models to file
pipeline_write <-
  function(
    model_output,
    write_folder,
    write_path = "output/",
    graphics_width = 8,
    graphics_height = 5,
    graphics_device = "pdf",
    ...
  ){
    write_ls(
      model_summary = model_output$model_summary,
      parameter_table = model_output$parameter_table,
      graphics_list = model_output$graphics_list,
      write_folder = write_folder,
      write_path = write_path,
      graphics_width = graphics_width,
      graphics_height = graphics_height,
      graphics_device = graphics_device
    )
  }

####################################################
## Model diagnostics - LS
####################################################

## ls_analysis/R/produce_graphics.ls

# Function to create all selected diagnostic plots
produce_graphics_ls <-
  function(
    coxph_model,
    data,
    residual_plots = c("coxsnell", "martingale", "deviance", "none"),
    smooth_hazards = TRUE,
    residual_plot_type = c("both", "jitter", "smooth", "none"),
    jitter_noise_level = 0.05,
    show_original_residuals = FALSE,
    arjas_covariates = NULL
  ){
    # Remove "none" from the plot list
```

```
    residual_plots <- str_subset(residual_plots, "none", negate = TRUE)

    # Extract plot settings
    arg_list <- as.list(environment())

    # Residual plots
    residual_plot_list <- if (!is_empty(residual_plots))
      residual_plots %>%
      set_names() %>%
      map(~ do.call(str_c("plot_", .x), arg_list)) else NULL

    # Arjas plots
    arjas_plot_list <- if (!is_empty(arjas_covariates))
      arjas_covariates %>%
      set_names(str_c("arjas_", arjas_covariates)) %>%
      map(~ do.call(plot_arjas, c(arg_list, covariate = .x))) else NULL

    c(residual_plot_list, arjas_plot_list)
  }

# Plot Cox-Snell residuals
plot_coxsnell <- function(
  coxph_model,
  smooth_hazards = TRUE,
  ...
){
  plot_title <- (if (smooth_hazards) "Smoothed" else NULL) %>%
    str_c("Cox-Snell Residuals", sep = " ")

  multi_plot <- (class(coxph_model) == "list")
  if (!multi_plot) coxph_model <- list(coxph_model)

  coxph_model %>%
    map(get_coxsnell) %>%
    imap(~ mutate(.x, domain = str_prettify(.y), .before = 1)) %>%
    bind_rows() %>%
    ggplot(aes(time, cumulative_hazard, col = domain)) +
    (if (smooth_hazards) geom_smooth(se = FALSE) else geom_step()) +
    geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
    (if (!multi_plot) theme(legend.position = "none") else guides(col =
    ↪  guide_legend(title = "Domain"))) +
    labs(
```

```
      x = "Cox-Snell Residuals",
      y = "Estimated Cumulative Hazard",
      title = plot_title
    )
}

# ... more similar plots

#####################################################
## Results - Occupation Features
#####################################################

## output/scripts/plot_onet_results.R

# Libraries
library(tidyverse)

# Scripts
list.files("R/", pattern = "\\.R$", full.names = TRUE) %>%
  walk(source)

# Load data
data <- read_all_models() %>%
  postprocess_onet(onet = read_onet_data())

# Create occupation feature parameter estimate plots
onet_domains <- unique(data$domain)
onet_domains %>%
  set_names() %>%
  map(~ plot_onet(data, domains = .x))

## output/R/plot_model.R

# Prepare occupation feature model results for plotting
prepare_onet <- function(data, domains = "all"){
  data %>%
    filter(predictor_type == "onet", domains == "all" | domain %in%
    ↪  domains) %>%
    prettify_model() %>%
    prettify_strings()
}
```

```r
# Plot occupation feature model results
plot_onet <- function(data, domains = "all"){
  multi_domain <- (domains == "all") | (length(domains) > 1)
  xlab <- if (multi_domain) "O*NET predictor" else NULL
  plot_title <- if (multi_domain) "Parameter Estimates" else NULL

  data %>%
    prepare_onet(domains = domains) %>%
    ggplot(aes(reorder(term, desc(estimate)), estimate, col = model,
    ↪  group = model)) +
    geom_hline(yintercept = 1, linetype = "dashed") +
    geom_point(position = position_dodge(width = 0.5)) +
    # geom_line(position = position_dodge(width = 0.5)) +
    geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.2,
    ↪  position = position_dodge(width = 0.5)) +
    (if (multi_domain) facet_wrap(~domain, ncol = 1, scales = "free")
    ↪  else NULL) +
    guides(col = guide_legend(title = "Model")) +
    scale_y_continuous(breaks = c(0.5, 0.75, 1, 1.25, 1.5, 1.75)) +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
    labs(x = xlab, y = "Hazard Ratio", title = plot_title)
}
```

B Further Information on Data Sources

B.1 O*NET

The Occupational Information Network (O*NET) program is the US' "primary source of occupational information" (National Center for O*NET Development 2021). The program collects, maintains and publishes data on the characteristics of jobs and the individuals who fill them (US National Research Council 2010). Its output is a database (the O*NET database) containing highly standardised information about occupations, as well as a set of tools and interfaces for interacting with this database (`https://www.onetonline.org/`). Since the first version of the database was published in 1998, it has been in continual development, with major updates occurring approximately yearly.

The database is constructed from two main components: a "Standard Occupational Classification" (SOC) system—a standardised method of classifying / clustering jobs into occupations; and a "content model"—a framework for organising information about individual occupations. The O*NET SOC can be thought of as providing a "taxonomy" of different occupations, while the content model provides the "anatomy" of a single occupation.

The 2010 version of the O*NET SOC distinguishes 1,110 occupations, though data are only collected for 974 'data-level' occupations. Each occupation is specified using an 8-digit code, which allows the occupation to be described at various levels of precision. For example, the "detailed O*NET SOC occupation" 29-1069.02 (Dermatologists) is contained within the "broad occupation" 29-1060.00 (Physicians and Surgeons), and at a higher-level within the "major group" 29-0000.00 (Healthcare Practitioners and Technical Occupations). Three example occupations along with their job descriptions are shown in table 7

Each of the 974 'data-level' occupations are measured along over 400 descriptive variables containing a variety of occupational information. The O*NET content model organises these variables into a structured, hierarchical taxonomy; the two highest levels of this taxonomy are shown in figure 2. The highest level of the taxonomy consists of 6 categories (known as domains), which are "designed to provide multiple windows into the world of work" (US National Research Council 2010). Each of these high-level domains contain a number of sub-categories (also known as domains), each of which consists of a number of "descriptors". Descriptors define the individual 'units' of information in the O*NET database. Finally, each descriptor is measured on one or more "scales", which both specify the precise meaning of the descriptor and define the range of values it can take. An individual variable in the O*NET database consists of a descriptor combined with a scale.

For example, the "Oral Comprehension" descriptor in the "Abilities" domain is defined as "The ability to listen to and understand information and ideas presen-

Table 7: Three example occupations from the O*NET 2010 SOC. Source: National Center for O*NET Development, used under the CC BY 4.0 license.

O*NET 2010 SOC code	Title	Description
11-1011.00	Chief Executives	Determine and formulate policies and provide overall direction of companies or private and public sector organizations within guidelines set up by a board of directors or similar governing body. Plan, direct, or coordinate operational activities at the highest level of management with the help of subordinate executives and staff managers.
51-2041.00	Structural Metal Fabricators and Fitters	Fabricate, position, align, and fit parts of structural metal products.
53-6051.08	Freight and Cargo Inspectors	Inspect the handling, storage, and stowing of freight and cargoes.

ted through spoken words and sentences". This descriptor is measured on two scales: *importance*, which indicates the importance of the oral comprehension to the occupation; and *level*, which indicates the degree to which oral comprehension is required or needed to perform the occupation. Importance is measured as a value ranging from 1 ("Not Important") to 5 ("Extremely Important"), while level is measured on a 1–7 scale.

Variables in most domains are measured through surveys to job incumbents. However, data in the "Abilities" and "Skills" domains are produced by professional occupational analysts, as job incumbents were found to generally over-rate themselves along these variables (US National Research Council 2010) The public database data represent population averages of descriptor scores for each occupation.

The O*NET SOC is based closely on the US SOC, the official US occupation classification scheme maintained and periodically updated by the US Office of Management and Budget (OMB). By contrast, the components of the content model (figure 2) were developed specifically for the O*NET program by multiple teams of researchers (contracted by the US Employment and Training Administration) during the 1990s (US National Research Council 2010). The domains were selected, defined and measured based on thorough reviews of prior research on characterising jobs. Data collection (as well as the rest of the program) is carried out by the National Center for O*NET development under a grant from the US Department of Labor / Employment and Training Administration (USDOL/ETA).

One the program's main aims is to "promote the effective education, training, counseling, and employment of the American workforce" (Labor 1993). Research uses of the O*NET database have included human resources and organizational behaviour, economic and labour market research, and occupational health (US National Research Council 2010; Cifuentes *et al.* 2010).

Its main strength for research is its detail and scope, with each of the 974 'data-level' occupations being measured along over 400 descriptive variables. There is no comparable source of detailed occupational information in the UK (Dickerson and Morris 2019). A second key strength is its frequent updates: between 2003 and 2020, around 70% of the occupations were updated[9] each year, on average. This allows for evaluation of how job characteristics change over time.

One weakness of the O*NET data for research is the substantial redundancy between occupation descriptors. For example, "problem-solving" appears as a descriptor four times in different domains (Abilities, Skills, Work Styles, and Work Activities). There is also substantial redundancy within domains. For example, the redundancy between the level and importance scales—used in the "Knowledge", "Skills", "Abilities", and "Work Activities" domains—is also substantial,

[9] `https://www.onetcenter.org/dataUpdates.html`

with correlations between the different scales generally 0.90 or greater (Handel 2016). Other weaknesses of the O*NET data include potential sampling bias, and difficulty ensuring the reliability and accuracy of survey responses (US National Research Council 2010; Handel 2016).

B.2 ONS LS

The Office for National Statistics Longitudinal Study (ONS LS) is the largest longitudinal data resource in England and Wales (Office for National Statistics 2021). It consists of linked census and life events data for a 1% sample of the population of England and Wales. The LS was set up by the Office for National Statistics[10] (ONS) in 1974 (Hattersley and Creeser 1995).

The initial study population was sampled randomly from the 1971 census on the basis of birthday; anyone born on one of four (secretly) selected birthdays was entered into the study. The same process was used to update the study after the 1981, 1991, 2001 and 2011 censuses. Life events (such as deaths) are linked from the NHS central register. New entries into the study result from births, immigration or re-entries of existing members; individuals exit the study when they die, emigrate or enlist into the military. Data for approximately 550,000 LS members are recorded at each census; the total sample size is around 1.2 million individuals.

The LS contains three types of information about study members: (1) responses to census questions, (2) life event data and (3) other variables derived from census or event data (e.g. social class). The censuses include questions on individuals (e.g. their sex, age, occupation) and their households (e.g. geographical location, accommodation type, household vehicle availability). Life event data includes entry and exit events (e.g. births and deaths of LS participants), as well as other major life events stored on the NHS central register (e.g. births to LS mothers, deaths of spouses, cancer registrations).

The LS has been used to study a variety of topics, including mortality differences between occupations; fertility patterns; inequalities in health, employment, education and geography; housing and geographical mobility; and social mobility, among others (Shelton *et al.* 2019).

[10]Formerly known as the Office for Population Censuses and Surveys (OPCS).

C Variable Definitions

C.1 LS Variables

Table	Variable	Description	Level
All	coreno	Unique identifier	LS member
CORE1	trace	NHSCR linkage indicator	LS member
CORE1	hiscen01	Present at 2001 Census indicator	LS member
CORE1	sex	Sex	LS member
CORE1	dobyr	Year of birth	LS member
ME01	gors0	Government Office Region (GOR)	Household
ME01	cavh0	Number of cars or vans available	Household
ME01	tenh0	Housing tenure	Household
ME01	dephsh0	Housing deprivation indicator	Household
ME01	occp0	Occupation	LS member
ME01	actlw0	Employment status in previous week	LS member
ME01	mstp0	Marital status	LS member
ME01	hlqp0	Highest qualification	LS member
ME01	ethgrp0	Ethnic group	LS member
ME01	illp0	Self-reported limiting long term illness	LS member
ME01	heap0	Self-reported general health	LS member
DETH	demtbde	Month of death	LS member
DETH	deyrbde	Year of death	LS member
DETH	ic10ude	WHO ICD10 cause of death code	LS member
DETH	ic10ufde	WHO ICD10 final cause of death code	LS member

Table 8: Summary of original LS variables. The "Table" column indicates the name of the original data file, while the "Level" column indicates whether the variable was measured at the individual or household level.

C.2 O*NET / Occupation Variables

Domain	Predictor	Description
Abilities	Physical	Strength, endurance, flexibility, balance and coordination
Abilities	Visual	Visual perception
Abilities	Language	Writing, speaking and listening abilities

Abilities	Manual	Manipulation and control of objects
Abilities	Alert	Concentration and perceptual speed
Abilities	Mathematical	Mathematical problem-solving abilities
Abilities	Creative	Originality, ideas and other creative abilities
Interests	Realistic	Work-related interests. See Holland (1997).
Interests	Investigative	
Interests	Artistic	
Interests	Social	
Interests	Enterprising	
Interests	Conventional	
Knowledge	Therapy	Knowledge of counseling, psychology, and social skills
Knowledge	Mechanical Engineering	Knowledge of design, technology, physics and construction
Knowledge	Business and Management	Knowledge of economics, management and other business domains
Knowledge	Electrical Engineering	Knowledge of computers, electronics and communications systems
Knowledge	Biochemistry	Knowledge of chemistry, biological systems and medicine
Knowledge	History	Knowledge of history, archeaology and geography
Knowledge	Transport and Safety	Knowledge of transport, health and safety and telecommunications
Knowledge	Art	Knowledge of fine arts, marketing and media
Skills	Social	Persuasion, negotation and social perceptiveness
Skills	Teaching	Learning, instructing and monitoring people
Skills	Technical	Programming, design, maths and other analytical skills
Skills	Repair	Installing, repairing and maintaining equipment and machines
Skills	Monitoring	Monitoring or controlling equipment and machines
Skills	Management	Management of financial, material and personnel resources

Work Activities	Managing Teams	Guiding, directing and motivating subordinates
Work Activities	Managing Equipment	Inspecting, repairing and controlling equipment
Work Activities	Administrative	Using computers, processing and recording information
Work Activities	Monitoring	Monitoring processes and surroundings
Work Activities	Managing Relationships	Public relations, sales, negotiating and resolving conflicts
Work Activities	Research	Creative thinking, self-directed work, understanding information
Work Activities	Caring for Others	Assisting, caring, training or coaching others
Work Context	Physically Stressful	Exposure to weather, enclosed spaces, temperature, sound or light
Work Context	Physically Active	Stand or move around frequently
Work Context	Socially Stressful	Frequency of social conflict situations
Work Context	Hazardous	Exposure to disease, radiation, and hazardous conditions and equipment
Work Context	Responsibility for Others	Coordinate or lead groups or teams
Work Context	Self Directed	High decision latitude, unstructured work, high-impact decisions
Work Context	Repetitive	Task repetition, repetitive motions, degree of automation, time pressure
Work Styles	Social	Social orientation, concern for others, self-control and cooperation
Work Styles	Ambitious	Initiative, persistence and achievement-orientation
Work Styles	Conscientious	Attention to detail, dependability and stress tolerance
Work Styles	Independent	Independent, innovative

Work Values	Achievement	
Work Values	Working Conditions	
Work Values	Recognition	
Work Values	Relationships	Work-related values. See Dawis and Lofquist (1984).
Work Values	Support	
Work Values	Independence	

Table 11: Definition of occupation features resulting from the factor analysis described in section 4.1. The "Occupational Interests" and "Work Values" domains were not reduced.

Variable	Description
deind	Death indicator
survyrs	Exit time, in years since 2003

Table 9: Summary of (derived) outcome variables.

Predictor	Group	Description	Parameters
agec	Baseline	Age in 2001 (centered)	1
age2c	Baseline	Age-squared in 2001 (centered)	1
sex	Baseline	Sex	1
agec_sexFemale	Baseline	Age-Sex interaction	1
ethnicity	Baseline	Ethnicity	4
gors0	Baseline	Geographic region	9
marriage	SES	Marital status	2
wealth	SES	Wealth indicator	5
hlqp0	SES	Education	5
illp0	Health	Limiting long term illness	1
heap0	Health	General health	2
Total			32

Table 10: Summary of LS predictors used in section 5. "Parameters" indicates the corresponding number of model parameters.

D Mapping of Occupation Features

Two data sources were used to construct the weight matrix for the O*NET 2010 SOC to UK 2010 SOC mapping.

The first of these is an O*NET SOC – UK SOC matching matrix, produced in October 2020 by the "Labour Market Information (LMI) for All" project[11] (UK Department for Education 2021). This data set consists of lists of the O*NET 2010 SOC occupations corresponding to each UK 2010 SOC occupation, and was produced by computer-assisted matching the O*NET SOC and UK SOC occupations (Dickerson and Morris 2019).

The second data source is the May 2011 version of the US Bureau of Labor Statistics' (BLS) Occupational Employment Statistics (OES). These data consist of the estimated number of individuals employed in each US 2010 SOC occupation, and are produced by "semi-annual mail surveys of non-farm establishments" carried out by the BLS (US Bureau of Labor Statistics 2021).

Following the method of Dickerson and Morris (2019), each match (of an O*NET 2010 SOC occupation to a UK 2010 SOC occupation) is assigned a weight proportional to the number of US individuals working in that occupation. These weights are then normalised over each UK 2010 SOC occupation (corresponding to the rows of W) to give the final weight matrix.

Mathematically, this can be described as follows. Let A_j be the set of O*NET 2010 SOC occupations matched to a UK 2010 SOC occupation j in the LMI for All matching matrix, and e_k the number of individuals employed in O*NET 2010 SOC occupation k according to the OES data. Then the weight matrix W for the O*NET 2010 SOC to UK 2010 SOC mapping is defined by

$$w_{jk} = \frac{e_k \mathbb{1}_{k \in A_j}}{\sum_{m \in A_j} e_m} \tag{D.1}$$

The mapping between the two SOCs can then be performed using equation 2.2.

The weight matrix for the UK 2010 SOC to UK 2000 SOC mapping was derived directly from ONS correspondence tables between these two SOCs (Office for National Statistics 2012). This data source specifies the fraction of individuals in each UK 2000 SOC occupation j who are classified as having UK 2010 SOC occupation k (for all values j and k); this quantity is used directly as the weight matrix element w_{jk}. The mapping from UK 2010 SOC to UK 2000 SOC can again be performed using equation 2.2.

[11] `https://www.lmiforall.org.uk/`

E Overview of Regression Models for Survival Analysis

Frequently, the aim of a time-to-event model is to examine relationships between characteristics of individuals, encoded in a set of predictors $\boldsymbol{x}_i$, and their event times. In general the predictors $\boldsymbol{x}_i = \boldsymbol{x}_i(t)$ may vary with time. In contrast to classical regression, where the object of modeling is usually the expectation of random variables, regression models for time-to-event data are typically expressed in terms of the hazard function $\lambda(t)$. Specifically, the hazards $\lambda_i(t)$ for each individual are expressed as a joint function of a *baseline hazard* $\lambda_0(t)$—common across all individuals—and the predictors $\boldsymbol{x}_i$, which modify the baseline hazard appropriately. The baseline hazard $\lambda_0(t)$ can be thought of as the risk experienced by a "baseline" individual. The four approaches most commonly used to relate the individual hazards to the baseline hazard and predictors are: (1) multiplicative hazard models, more commonly known as proportional hazards models (Cox 1972), (2) proportional odds models (Bennett 1983) (3) additive hazard models (O. O. Aalen 1989), and (4) accelerated failure time (AFT) models (Wei 1992). Each of these approaches are described briefly below.

The most common class of models used to perform regression on time-to-event data are the multiplicative hazard models. Under these models, the baseline and individual hazards are related by

$$\lambda_i(t) = \eta_i \lambda_0(t) \tag{E.1}$$

where $\eta_i = \eta_i(\boldsymbol{x}_i)$ is some non-negative "link" function, which encodes the dependence of the hazard on the predictors. The baseline hazard $\lambda_0(t)$ may have a specified parametric form, or it may be left as an arbitrary non-negative function. The most popular of the multiplicative hazard models is the Cox model, which uses link function

$$\eta_i = \exp(\boldsymbol{x}_i^T \beta) \tag{E.2}$$

for some parameter vector β. This link function is chosen for its simplicity, and the fact it is non-negative for any $\boldsymbol{x}_i^T \beta$. Under the model defined by equation E.1, hazards are directly proportional:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\eta_i}{\eta_j}, \tag{E.3}$$

and a similar relation holds for the survival function $S(t)$. As a result these models are also known as proportional hazards models. The Cox model is discussed further in section 4.2.2.

In contrast with multiplicative hazard models, additive hazard models assume that hazard functions are linearly related to predictors:

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{x}_i(t)^T \beta(t) \tag{E.4}$$

where $\beta(t)$ is a vector of time-dependent parameters. Equation E.4 is known as the Aalen additive hazards models. Unlike the Cox model, it allows the relationship between the hazard and the predictors $\boldsymbol{x}_i$ to change over time. As a result of being linear rather than multiplicative, it is also more robust to noisily measured or missing predictors than the Cox model (O. O. Aalen 1989). Its main disadvantange is that it may estimate a negative hazard under some circumstances; however this is usually not a major problem.

The relationship between the baseline and individual hazards under the proportional odds model is

$$\lambda_i(t) = \left(\frac{\eta_i}{\eta_i + (1 - \eta_i) \exp(-\Lambda_0(t))} \right) \lambda_0(t). \tag{E.5}$$

This expression results from the assumption that the odds of survival $O(t) = (1 - S(t))/S(t)$ for each individual are constant in time; equivalently, $O_i(t) = \eta_i O_0(t)$ analogously to equation E.1. The difference between the proportional odds and proportional hazards assumptions can be seen directly by comparing equations E.1 and E.5: while they give similar hazards at small times t (when $e^{-\Lambda_0(t)} \approx 1$), as $\Lambda_0(t)$ increases the hazard functions in equation E.5 tend to converge, while those in equation E.1 remain separated by a constant factor. This behaviour is useful when the effect of the predictors $\boldsymbol{x}_i$ is expected to decrease to zero over time.

Lastly, accelerated failure time (AFT) models assume that the baseline and individual hazards are related by

$$\lambda_i(t) = \eta_i \lambda_0(\eta_i t), \tag{E.6}$$

which correspond to survival functions $S_i(t) = S_0(\eta_i t)$. Intuitively, this model follows from the assumption that each individual experiences the same survival function $S_0(t)$, but "runs through" this survival function at a different rate determined by their acceleration parameter η_i.

F Further Results—Occupation Features

Table 12: Hazard ratios for the occupation features. Results are shown for the baseline, SES-adjusted, health-adjusted, and employed subset models. Exponentiated standard errors $\exp(\mathrm{se}(\widehat{\beta}))$ are given in brackets. Source: ONS LS.

	Model			
Predictor	Baseline	SES-Adjusted	Health-Adjusted	Employed Subset
Abilities				
Physical	1.40 (1.05)	1.08 (1.05)	1.02 (1.05)	1.12 (1.07)
Visual	1.14 (1.05)	1.05 (1.05)	1.03 (1.05)	1.05 (1.06)
Language	0.65 (1.06)	0.88 (1.06)	0.90 (1.06)	0.88 (1.08)
Manual	1.22 (1.06)	1.14 (1.06)	1.08 (1.06)	0.99 (1.09)
Alert	1.00 (1.06)	1.12 (1.06)	1.05 (1.06)	1.14 (1.09)
Mathematical	0.80 (1.05)	0.93 (1.05)	0.96 (1.05)	0.93 (1.07)
Creative	0.60 (1.07)	0.81 (1.07)	0.83 (1.07)	0.71 (1.09)
Interests				
Realistic	1.54 (1.04)	1.18 (1.04)	1.11 (1.04)	1.18 (1.06)
Investigative	0.56 (1.04)	0.82 (1.04)	0.85 (1.04)	0.80 (1.05)
Artistic	0.88 (1.05)	0.97 (1.05)	0.98 (1.05)	0.90 (1.08)
Social	1.02 (1.04)	1.02 (1.04)	1.00 (1.04)	1.08 (1.05)
Enterprising	0.85 (1.04)	0.94 (1.04)	0.95 (1.04)	0.94 (1.05)
Conventional	1.05 (1.04)	1.08 (1.04)	1.10 (1.04)	1.11 (1.06)
Knowledge				
Therapy	0.71 (1.04)	0.89 (1.04)	0.88 (1.04)	0.85 (1.06)
Mechanical Engineering	1.01 (1.04)	1.02 (1.04)	0.98 (1.04)	0.91 (1.05)
Business And Management	0.70 (1.04)	0.87 (1.04)	0.92 (1.04)	0.89 (1.05)
Electrical Engineering	0.57 (1.05)	0.86 (1.05)	0.90 (1.05)	0.82 (1.07)
Biochemistry	0.91 (1.06)	0.94 (1.06)	0.96 (1.06)	0.98 (1.08)
History	0.62 (1.07)	0.79 (1.07)	0.82 (1.07)	0.70 (1.10)
Transport And Safety	1.48 (1.05)	1.11 (1.05)	1.05 (1.05)	1.18 (1.07)
Art	1.19 (1.05)	1.06 (1.05)	1.07 (1.05)	1.03 (1.07)
Skills				
Social	0.83 (1.06)	0.90 (1.06)	0.90 (1.06)	1.00 (1.09)
Teaching	0.61 (1.06)	0.90 (1.07)	0.91 (1.07)	0.80 (1.09)
Technical	0.50 (1.06)	0.81 (1.06)	0.86 (1.06)	0.75 (1.08)
Repair	1.07 (1.04)	1.01 (1.04)	0.97 (1.04)	0.98 (1.06)
Monitoring	1.50 (1.06)	1.19 (1.06)	1.09 (1.06)	1.19 (1.09)
Management	1.02 (1.05)	0.99 (1.05)	1.04 (1.05)	0.98 (1.07)

Work Activities				
Managing Teams	0.94 (1.05)	0.96 (1.05)	0.96 (1.05)	0.97 (1.07)
Managing Equipment	1.41 (1.04)	1.07 (1.04)	1.01 (1.04)	1.02 (1.06)
Administrative	0.66 (1.04)	0.89 (1.04)	0.92 (1.04)	0.86 (1.06)
Monitoring	1.01 (1.06)	1.14 (1.06)	1.09 (1.06)	1.15 (1.09)
Managing Relationships	0.92 (1.04)	0.92 (1.04)	0.95 (1.04)	0.99 (1.06)
Research	0.59 (1.04)	0.83 (1.04)	0.85 (1.04)	0.75 (1.06)
Caring For Others	1.03 (1.05)	1.02 (1.05)	0.99 (1.05)	1.07 (1.06)
Work Context				
Physically Stressful	1.24 (1.03)	1.06 (1.04)	1.01 (1.04)	1.02 (1.05)
Physically Active	1.68 (1.04)	1.20 (1.04)	1.13 (1.04)	1.20 (1.05)
Socially Stressful	1.12 (1.05)	1.08 (1.05)	1.07 (1.05)	1.19 (1.07)
Hazardous	0.89 (1.04)	0.98 (1.05)	0.95 (1.05)	0.98 (1.06)
Responsibility For Others	0.76 (1.05)	0.95 (1.05)	0.93 (1.05)	0.87 (1.07)
Self Directed	0.61 (1.04)	0.82 (1.04)	0.86 (1.04)	0.81 (1.06)
Repetitive	1.58 (1.06)	1.39 (1.06)	1.29 (1.06)	1.39 (1.08)
Work Styles				
Social	1.01 (1.05)	0.99 (1.05)	1.00 (1.05)	1.08 (1.06)
Ambitious	0.52 (1.05)	0.81 (1.05)	0.83 (1.05)	0.72 (1.07)
Conscientious	0.67 (1.05)	0.91 (1.05)	0.93 (1.05)	0.92 (1.07)
Independent	0.84 (1.05)	0.94 (1.05)	0.96 (1.05)	0.93 (1.07)
Work Values				
Achievement	0.76 (1.14)	0.94 (1.14)	0.87 (1.14)	0.72 (1.19)
Working Conditions	0.52 (1.10)	0.76 (1.10)	0.84 (1.10)	0.71 (1.15)
Recognition	0.76 (1.13)	0.82 (1.13)	0.92 (1.13)	1.05 (1.18)
Relationships	0.94 (1.04)	0.98 (1.04)	1.00 (1.04)	1.03 (1.06)
Support	1.17 (1.05)	1.22 (1.05)	1.11 (1.05)	1.12 (1.08)
Independence	1.35 (1.08)	1.22 (1.08)	1.16 (1.08)	1.30 (1.12)

G Tables used in Section 2

Table 13: Sample Size by Age Group and Sex. Contains underlying counts for figures 6a and 9c. Source: ONS LS.

	Sex		
Age Group	Male	Female	Total
25–34	33238	34865	68103
35–44	38783	39275	78058
45–54	34916	35274	70190
55–64	31996	31299	63295
Total	138933	140713	279646

Table 14: Sample Size by Ethnicity, Age Group and Sex. Contains underlying counts for figure 6b. Source: ONS LS.

		Sex		
Ethnicity	Age Group	Male	Female	Total
White				
	25–34	29753	31562	61315
	35–44	35169	35885	71054
	45–54	32243	32883	65126
	55–64	30036	29924	59960
Mixed				
	25–34	367	412	779
	35–44	367	341	708
	45–54	134	179	313
	55–64	92	88	180
Asian				
	25–34	2045	1731	3776
	35–44	1958	1540	3498
	45–54	1785	1384	3169
	55–64	1278	725	2003
Black				
	25–34	630	801	1431
	35–44	885	1098	1983
	45–54	437	500	937
	55–64	379	422	801
Other				
	25–34	443	359	802
	35–44	404	411	815
	45–54	317	328	645
	55–64	211	140	351
Total		138933	140713	279646

Table 15: Sample Size by Government Office Region (GOR) and Age Group. Contains underlying counts for figure 6c. Source: ONS LS.

	Age Group				
Region	25–34	35–44	45–54	55–64	Total
London	10939	10499	8211	6842	36491
North East	2981	3727	3536	3156	13400
North West	8317	9812	9207	8311	35647
Yorkshire and The Humber	6465	7415	6759	6008	26647
East Midlands	5484	6583	5903	5354	23324
West Midlands	6786	8019	7374	6492	28671
East of England	7172	8146	7421	6667	29406
South East	10533	12542	11128	10108	44311
South West	6042	7246	6765	6660	26713
Wales	3384	4069	3886	3697	15036
Total	68103	78058	70190	63295	279646

Table 16: Sample Size by Government Office Region (GOR) and Wealth. Contains underlying counts for figure 7a. Source: ONS LS.

	Wealth Indicator						
Region	A	B	C	D	E	Communal	Total
London	5462	14508	5830	6259	4215	217	36491
North East	2207	6729	1289	1666	1461	48	13400
North West	6793	17430	4396	4069	2767	192	35647
Yorkshire and The Humber	4919	12802	3495	3301	2016	114	26647
East Midlands	4808	12380	2357	2436	1224	119	23324
West Midlands	5698	14024	3441	3563	1849	96	28671
East of England	5966	15704	2926	3530	1120	160	29406
South East	8748	23938	4920	4829	1601	275	44311
South West	5639	13335	3397	3069	1094	179	26713
Wales	3497	7390	1498	1706	893	52	15036
Total	53737	138240	33549	34428	18240	1452	279646

Table 17: Sample Size by Marital Status, Age Group and Sex. Contains underlying counts for figure 7b. Source: ONS LS.

		Sex		
Age Group	Marital Status	Male	Female	Total
25–34				
	Single	19855	16847	36702
	Married	11725	14725	26450
	Separated	1658	3293	4951
35–44				
	Single	9074	6622	15696
	Married	24447	25124	49571
	Separated	5262	7529	12791
45–54				
	Single	3928	2380	6308
	Married	25323	25206	50529
	Separated	5665	7688	13353
55–64				
	Single	2232	1351	3583
	Married	25034	22359	47393
	Separated	4730	7589	12319
Total		138933	140713	279646

Table 18: Sample Size by Education Level, Age Group and Sex. Contains underlying counts for figure 7c. Source: ONS LS.

		Sex		
Age Group	Education	Male	Female	Total
25–34				
	None	4066	3173	7239
	Other	1538	962	2500
	Level 1	8054	8381	16435
	Level 2	7229	9033	16262
	Level 3	2943	3232	6175
	Level 4-5	9408	10084	19492
35–44				
	None	6697	6252	12949
	Other	2454	1157	3611
	Level 1	10118	10283	20401
	Level 2	7769	9618	17387
	Level 3	2629	2883	5512
	Level 4-5	9116	9082	18198
45–54				
	None	9997	11401	21398
	Other	5162	2714	7876
	Level 1	5150	5570	10720
	Level 2	4949	6064	11013
	Level 3	2003	1963	3966
	Level 4-5	7655	7562	15217
55–64				
	None	13879	15665	29544
	Other	5163	2894	8057
	Level 1	3055	3317	6372
	Level 2	3418	3719	7137
	Level 3	1236	916	2152
	Level 4-5	5245	4788	10033
Total		138933	140713	279646

Table 19: Sample Size by General Health, Age Group and Sex. Contains underlying counts for figures 8a and 9d. Source: ONS LS.

		Sex		
Age Group	General Health	Male	Female	Total
25–34				
	Good	27111	26720	53831
	Fairly Good	4934	6686	11620
	Not Good	1193	1459	2652
35–44				
	Good	28997	28110	57107
	Fairly Good	7529	8612	16141
	Not Good	2257	2553	4810
45–54				
	Good	23052	21753	44805
	Fairly Good	8448	9671	18119
	Not Good	3416	3850	7266
55–64				
	Good	17351	16414	33765
	Fairly Good	9328	10440	19768
	Not Good	5317	4445	9762
Total		138933	140713	279646

Table 20: Sample Size in Employment, by Age Group and Sex. Contains underlying counts for figure 8b. Source: ONS LS.

		Employment Status		
Age Group	Sex	Employed	Non-Employed	Total
25–34				
	Male	29289	3949	33238
	Female	25892	8973	34865
35–44				
	Male	34248	4535	38783
	Female	29482	9793	39275
45–54				
	Male	29577	5339	34916
	Female	26517	8757	35274
55–64				
	Male	19876	12120	31996
	Female	14381	16918	31299
Total		209262	70384	279646

Table 21: Sample Size in Employment, by General Health and Sex. Contains underlying counts for figure 8c. Source: ONS LS.

		Employment Status		
Sex	General Health	Employed	Non-Employed	Total
Male				
	Good	86481	10030	96511
	Fairly Good	22801	7438	30239
	Not Good	3708	8475	12183
Female				
	Good	70829	22168	92997
	Fairly Good	21853	13556	35409
	Not Good	3590	8717	12307
Total		209262	70384	279646

Table 22: Sample mortality events by Age Group and Sex. Contains underlying counts for figure 9a. Source: ONS LS.

	Sex		
Age Group	Male	Female	Total
25–34	515	388	903
35–44	1321	987	2308
45–54	3342	2443	5785
55–64	7544	5282	12826
Total	12722	9100	21822

Table 23: Sample mortality events by Government Office Region (GOR), Age Group and Sex. Contains underlying counts for figure 9b. Source: ONS LS.

		Male		Female		
Age Group	Region	Died	Survived	Died	Survived	Total
25–34						
	London	54	5299	53	5533	10939
	North East	30	1448	16	1487	2981
	North West	74	3938	56	4249	8317
	Yorkshire and The Humber	48	3086	44	3287	6465
	East Midlands	49	2667	39	2729	5484
	West Midlands	71	3261	26	3428	6786
	East of England	49	3488	41	3594	7172
	South East	63	5053	57	5360	10533
	South West	47	2864	31	3100	6042
	Wales	30	1619	25	1710	3384
35–44						
	London	173	5029	129	5168	10499
	North East	71	1766	43	1847	3727
	North West	192	4692	134	4794	9812
	Yorkshire and The Humber	139	3548	97	3631	7415
	East Midlands	126	3157	88	3212	6583
	West Midlands	148	3888	105	3878	8019
	East of England	110	3927	86	4023	8146
	South East	185	6045	160	6152	12542
	South West	109	3470	83	3584	7246
	Wales	68	1940	62	1999	4069
45–54						
	London	354	3690	252	3915	8211
	North East	210	1582	134	1610	3536
	North West	470	4150	380	4207	9207
	Yorkshire and The Humber	345	3081	258	3075	6759
	East Midlands	299	2604	222	2778	5903
	West Midlands	367	3328	246	3433	7374
	East of England	320	3336	230	3535	7421
	South East	480	5048	351	5249	11128
	South West	293	3002	222	3248	6765
	Wales	204	1753	148	1781	3886
55–64						
	London	804	2643	490	2905	6842

	North East	456	1126	349	1225	3156
	North West	1069	3135	787	3320	8311
	Yorkshire and The Humber	765	2284	554	2405	6008
	East Midlands	632	2111	438	2173	5354
	West Midlands	809	2548	536	2599	6492
	East of England	718	2623	522	2804	6667
	South East	1076	4051	766	4215	10108
	South West	717	2562	525	2856	6660
	Wales	498	1369	315	1515	3697
Total		12722	126211	9100	131613	279646