Institute
and Faculty
of Actuaries

Sessional ARC Diabetes

An analysis of diabetes mortality risk

by B. Grechuk*, A. Gorban, E. Mirkes, S. Reid

Presented to the Institute & Faculty of Actuaries

18 March 2026
Edinburgh

An analysis of diabetes mortality risk

B. Grechuk* A. Gorban E.Mirkes S. Reid

18 March 2026

Abstract

Objective: This research sought to update understanding following improvements to treatment and deepen the understanding of the mortality risk associated with Type 1 or Type 2 diabetes, including relative risk in the presence of comorbidities. Specifically, a model to provide mortality predictions at a granular level for lives with and without diabetes. The model is tailored for use by the insurance industry to provide an updated source from which to appreciate the risk posed when underwriting people with diabetes. By providing an updated and deeper understanding of mortality risk, the research's aim is to improve access to insurance for those individuals living with diabetes.

Method: The model combines industry standard underwriting risk factors, such as age, gender, deprivation index, body mass index, smoker status, blood pressure, and cholesterol level, with various comorbidities related to diabetes. A comprehensive analysis of mortality risk factors, between 2010 and 2019, for people with and without diabetes is undertaken on over 1.2 million records based on Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES), and Office for National Statistics (ONS) death registrations data. Cox proportional hazards models are used to estimate the probability of death, stratified by gender across three distinct populations: Type 1 diabetes, Type 2 diabetes, and a general population sample.

Results: The model output produced are permutations of the following: gender; population split by general sample, Type 1 and Type 2 diabetes; and a time dependent exponential model and a time invariant homogeneous model. A Shiny model application allows interaction with the model outputs (https://0jv7e6-scott-reid.shinyapps.io/diabmdl/) and spreadsheets provide additional explanation. Useful insights were obtained through industry discussions on the variation of existing market practice against that implied by the results. Key rating factors were generally aligned with market practice such as age, BMI, Blood pressure and Cholesterol and years since a diabetes diagnosis. However, for a few significant mortality risks impacting co-morbidities the results did not adhere to prior expectations. Exploratory work suggested that the order and sequencing of key co-morbidities for diabetes must be included in future model development.

Keywords: Diabetes; Mortality risk; Cox proportional hazards model; Insurance

***Correspondence to:** Bogdan Grechuk, School of Computing and Mathematical Sciences, University of Leicester, University Road, LE1 7RH, UK. E-mail: bg83@leicester.ac.uk

Contents

All tables and figures referenced herein can be found in a separate 'Tables and Figures' document.

1 Foreword by the Diabetes Steering Group on behalf of the Institute and Faculty of Actuaries

The Diabetes Steering Group (DSG), on behalf of the Institute and Faculty of Actuaries (IFoA), are delighted to introduce this paper, which, to our knowledge, is the first study that provides a comprehensive analysis of various mortality risk factors for individuals diagnosed with both Type 1 and Type 2 diabetes (Reid et al., 2023). This report includes a model which provides the user with a mortality risk prediction for individuals living with and without diabetes.

This research is considered important to the primary audience, the UK insurance industry, as it furthers the DSG's key research aim as set out in our initial paper (Reid et al., 2023); that is, to widen access to insurance products for customers living with diabetes by improving the insurance industry's access to information on the specific mortality risk for those living with diabetes. This primary intended audience for this research is the actuarial community and insurance industry, specifically those insurers and reinsurers that write protection and longevity risk products in the UK. However, the IFoA has a Royal Charter under which the DSG also seeks to provide information for the wider public interest.

The Diabetes Working Party produced a sessional paper (Reid et al., 2023) published on 18th May 2023 which provided insights to diabetes mortality risk via a comprehensive literature review and a global underwriting survey. Following this, the working party sought to further the research aim by commissioning a research project. That research project is detailed herein and was commissioned by the IFoA Actuarial Research Centre (ARC) and a group of industry participants, specifically Pacific Life Re, Partner Re, Swiss Re, Legal & General, and Zurich Insurance Group. The research was carried out by world-leading experts in risk analysis, risk modelling and risk evaluation at the University of Leicester, namely Dr Bogdan Grechuk[1], Dr Evgeny Mirkes and Prof Alexander Gorban. These academic researchers are supported by the Real World Evidence Centre and the Leicester Diabetes Centre, a unique, collaborative partnership between the NHS and the University of Leicester.

Individuals with diabetes, both Type 1 and Type 2, run a greater risk of developing one or more severe health complications, including cardiovascular and cerebrovascular disease. Diabetes is also a leading cause of blindness in working-aged individuals and a common cause of kidney failure. Life expectancy following a diagnosis of diabetes has historically been lower than in those without diabetes, given that inadequate glycaemic control gives rise to several complications that cause premature death, along with increased risks of long-term disability.

In recent years, early detection and management of diabetes, both from a personal as well as a physician-led perspective, has improved such that survival with diabetes has increased. New pharmaceuticals, coupled with enhanced monitoring and modern insulin dosage systems, have transformed the lives of individuals living with diabetes. Life expectancy with optimal glycaemic management has been extended in those with diabetes; however, the long-term impact of new pharmaceuticals has yet to be fully appreciated.

The overarching aim of the research project is to develop a deeper understanding of the mortality risk associated with a diagnosis of Type 1 and Type 2 diabetes and the impact of recent improved medical treatments. More specifically, the DSG's objectives are to:

- Gain insights from recent data by considering advanced data analytic techniques to understand relative mortality risk factors and interactions.
- Produce a model that can predict mortality across a wide age range at a granular level for lives living with and without both Type 1 and Type 2 diabetes. The model should include co-morbidities such that the impact on mortality of living with and without diabetes in the presence of a wide range of co-morbidities can be understood.

The work carried out by the University of Leicester is detailed within this paper; specifically, the following areas are detailed: a background on the data sources used, e.g. Clinical Practice Research Datalink (CPRD) and linked data; the process of data ingestion and cleaning; the data analysis; and the fitting of a Cox proportional hazard model. The DSG sought a transparent model where inferences can be drawn and therefore, a regression model was selected, namely Cox proportional hazard. The DSG built an accompanying Shiny application[2] so that the insurance industry and any interested wider stakeholders, e.g. Diabetes associations and general practitioners (GPs), can interact with the model output.

[1] https://le.ac.uk/people/bogdan-grechuk
[2] https://0jv7e6-scott-reid.shinyapps.io/diabmdl/

The DSG is satisfied that the research outputs have been carried out to a high quality and in line with expectations to produce findings that are relevant to the primary intended audience (e.g. the actuarial community and insurance industry).

The interactive Shiny application will enable the user to consider the impact of different attributes on a general sample population, as well as a population of diabetes Type 1 and Type 2. The DSG would like to draw the reader's attention to the benefits and limitations Section 6.6 and to advise that care be exercised when interpreting results and utilising the Shiny application. Section 7 includes a general industry discussion followed by an applied industry discussion. The modelling was appreciated in the general discussion but there are many complications such as the quality of the data and how to handle co-morbidities correctly. Building a general model that covers Type 1 and Type 2 for the all the different co-morbidities was ambitious and we understand this is the first time that has been done. The general feedback was that creating a model based on CPRD data was worthwhile as an initial approach that can be improved in future based on the feedback from practitioners.

2 Introduction

2.1 Background

Both Type 1 and Type 2 diabetes are serious diseases and there has been a dramatic increase in prevalence over the last few decades. The number of worldwide cases has quadrupled between 1980 and 2004 (Zhou et al., 2016) and has increased for both Type 1 and Type 2 diabetes (Patterson et al., 2019). More than half a billion individuals are living with diabetes worldwide as of 2021, and this number is predicted to grow to 1.3 billions in the year 2050 (Ong et al., 2023). Over 300 million individuals have prediabetes (Atlas et al., 2015) and it is estimated that almost half of all individuals (49.7%) living with diabetes are undiagnosed (Cho et al., 2018).

In the UK, the number of individuals living with diabetes in all its forms is approaching 5 million, and this number is predicted to rise to 5.5 millions by 2030 (Cho et al., 2018). The expected lifetime of an individual with diabetes is significantly lower than for a non-diabetes individuals of similar age and other conditions (Bertoni et.al (Bertoni et al., 2002), Tancredi et.al. (Tancredi et al., 2015)). The analysis of diabetes-related mortality is complicated by the fact that individuals rarely die from diabetes directly, instead, individuals with diabetes have increased risk of death from other diseases (Defronzo et.al (DeFronzo, 2009)). In the last decade, significant advances in medical treatments for Type 1 (Boscari and Avogaro, 2021) and Type 2 (Chee and Dalan, 2024) diabetes seem to have reduced the mortality risk for many individuals. However, the effect of these treatments on mortality risk is not yet fully understood. Studies to date which provide information on mortality risk for individuals with diabetes is derived from data ten or more years old which does not reflect those more recent treatments.

2.2 Overview of this research project

This research project develops models for mortality prediction for individuals with Type 1 and Type 2 diabetes. These models are based on recent data (namely the years 2010 to 2020) and as such include the impact of recent treatments. The model output is mortality predictions for an individual based on risk factors including age, gender, body mass index (BMI), blood pressure (BP), cholesterol level (CL), smoking status, index of multiple deprivation (IMD), blood glucose (sugar) level (HbA1c), duration since diabetes diagnosis, and existence of various co-morbidities. With this information on each risk factor as an input, the models provide an output of a probability of death for the individual within the next t years for any given $t > 0$. To compute this probability, the models allow for the effect of the risk factors, listed above, separately and, for some of the factors, in combination. For example, coronary heart disease significantly increases the mortality risk for all individuals with diabetes, but the increase is particularly strong for individuals with high BMI.

It is important to note that this analysis does not attempt to draw causal epidemiological conclusions. Instead, it seeks to pinpoint risk factors that are reliably correlated with mortality in the diabetic population. Such an approach reflects the requirements of underwriting practice, which relies on consistent and interpretable risk differentiators to support equitable, evidence-driven decisions.

The analysis of data in this research project, alongside the existing literature (Huxley et al., 2015; Kautzky-Willer et al., 2023) reveal many risk factors which correlate with mortality differently depending on whether the diagnosis is of Type 1 or Type 2 diabetes, and gender. Therefore, separate mortality models were

developed by gender for (i) general population, (ii) individuals living with Type 1 diabetes, and (iii) individuals living with Type 2 diabetes. There are therefore six models produced in total.

All of the models are Cox proportional hazards models (Cox, 1972). A major advantage of this model is that the computed Cox coefficients explicitly show the effect of each factor on the mortality risk. The primary audience, the insurance industry, are required by regulation to have transparency in pricing terms and as such a transparent model where inferences can be drawn was preferred.

These models can be used to inform the insurance industry on the specific mortality risk posed by individuals with diabetes with the aim of achieving availability of insurance products for those living with diabetes and also more appropriate pricing and reserving for life and health insurance products. These models provide a better understanding of the mortality risk factors for individuals with diabetes which may also be of interest to wider stakeholders.

2.3 Comparison with the existing literature

The literature on diabetes is very rich. However, many studies investigate the effect of only one or at most several risk factors. Indeed, each risk factor considered herein can be found in the existing research as it affects the prevalence of diabetes or the associated health outcomes including: age (de Miguel-Yanes et al., 2011; Constantino et al., 2013), gender (Ohkuma et al., 2019; Mauvais-Jarvis, 2018), ethnicity (Goff, 2019), physical activity (Wahid et al., 2016), body mass index (Chatterjee et al., 2017), alcohol consumption (Baliunas et al., 2009), socio-economic status (Evans et al., 2000) and co-morbidities such as cardio-vascular disease (Leung et al., 2009; Wilmot et al., 2012; Riley and Cowan, 2014; Stamler et al., 1993), heart failure (Ohkuma et al., 2019), coronary heart disease (Peters et al., 2014), stroke (Collaboration et al., 2010), hypoglycaemia (Elwen et al., 2015), high blood pressure (Chiriacò et al., 2019), depression (Ali et al., 2006), dementia (Bello-Chavolla et al., 2019), fatty liver disease (Song et al., 2021) and COVID 19 (Hussain et al., 2020; Pal and Bhadada, 2020) (non-exhaustive list).

In contrast, this project builds a model that allows for many risk factors, separately and, for some of the factors, in combination. Such studies are far less common in literature. Jensen et.al. (Jensen et al., 2014) developed a clustering methodology which can be used to identify new risk factors related to diabetes. Golovenkin et.al. (Golovenkin et al., 2020) studied a selection of more than 100, 000 hospitalisation cases with individuals suffering from diabetes characterised by 55 attributes. The outcome of interest was hospital readmission, rather than mortality.

To the best of our knowledge, none of the existing recent literature provides a comprehensive analysis of the impact of various risk factors on the mortality of individuals living with diabetes. In addition, the literature does not provide a general model for mortality prediction for individuals with and without diabetes based on a large variety of risk factors. Herein, this is the focus of this research.

3 Data structure

3.1 Data collection

This research project requires information regarding the outcome of interest, death. It also requires information regarding possible risk factors for that outcome of interest, such as diabetes diagnosis type and date, age, gender, etc. The first was sourced using the Office for National Statistics (ONS) Death Registration Data, which enables a status of alive or deceased to be mapped to an anonymised individual record, and, in the latter case, the date of death. The second was sourced using the CPRD (Clinical Practice Research Datalink) database (Herrett et al., 2015), which contains anonymised individual data records from a network of general practices (GP) across the UK. In addition, information from the Hospital Episode Statistics (HES) database was utilised to gain further information on co-morbidity diagnosis from the National Health System (NHS) hospitals' admissions and outpatient appointments data. Deprivation is considered a significant risk factor for mortality (Evans et al., 2000). To allow for this, an Index of Multiple Deprivation (IMD) score was mapped at a GP and at an individual level.

These datasets and the linkage of these datasets were necessary to enable the influence of various risk factors on the mortality of those living with diabetes to be understood. Samples of the full database were taken over a restricted time and geographical scope so as to work with the least amount of data to achieve the research aim. This approach aligns with the principle of data minimisation, ensuring that only the minimum amount of data necessary to adequately address the research question was used, thereby reducing

computational burden and safeguarding data privacy. These datasets are fully anonymised, intended for use by researchers and their use in this project was approved by CPRD in 2021.

The time period for this research spans from 1 January 2010 (the 'study start date') to 31 December 2019 (the 'study end date'). This was the most recent 10-years period that was not impacted by the COVID-19 pandemic. Although some of the datasets are UK wide, the HES data is restricted to England and as such, the geographical scope of the research was narrowed to England as this linkage was necessary to understand the influence of co-morbidities on the mortality of diabetes patients. Data from the private healthcare system is not included; this is not considered material because the vast majority of diabetes care and associated hospital admissions in England occur within the NHS, making NHS Hospital Episode Statistics a comprehensive and representative source for the population under study.

The following samples were requested from the CPRD database:

- A diabetes sample: anonymised individual records from the database who have at least one record related to diabetes (Type 1 or Type 2), were alive at the study start date, and were at least 18 years of age by that date (i.e. a year of birth is 1991 or earlier). This sample resulted in $621,115$ individual records. Given that approximately 4 million adults in England were living with diabetes (diagnosed and undiagnosed) in 2019, about 15% of them were included in this sample.
- A general sample: a random sample of $250,000$ anonymised individual records, regardless of diabetes status (i.e. general) who were alive and were at least 18 years of age at the study start date. This sample is about 0.5% of the relevant population.
- An additional sample: a random sample of $25,000$ anonymised individual records, regardless of diabetes status, with any of the following diseases: Heart failure, Coronary heart disease, Angina, Heart attack, Stroke, Amputation, Macrovascular disease, Asthma, Atrial fibrillation, Cancer, CKD (kidney disease), COPD (Chronic obstructive pulmonary disease), Dementia and Epilepsy.

These samples are not mutually exclusive and as such, some records appear in several samples. In total, $1,205,657$ unique records were selected across the three samples. This is referred to herein as the 'total dataset'.

For each individual record in the total dataset, the following were sought:

- the entire medical record available in CPRD was requested;
- the Hospital Episode Statistics (HES) data for that record was linked;
- the ONS Death registration data for that record was linked;
- and the IMD was mapped at an individual and GP level.

3.2 CPRD data

The CPRD data arrived in the following tables: Patients, Practice, Staff, Consultation, Clinical, Additional Clinical Details, Referral, Immunisation, Test and Therapy (9 tables in total referred to herein as the 'main tables'), plus additional tables without necessary information (referred to herein as 'ignored tables'). The list of ignored tables is presented in Table 12.

Each table has its own set of columns of various lengths. The first column is Patient Identifier ('patid'), which is an encrypted unique identifier given to an individual record in the CPRD dataset. All tables and figures can be found in the separate 'Tables and Figures' document. Table 1 provides a list of columns and their description. The field patid is used to find records in all tables that correspond to a given individual record. Other forms of linkage are provided, Tables 1-10 describes all columns in the main tables, together with all forms of linkage. For convenience, links between fields are also summarised in Table 11 and Figure 1.

There are also two additional dictionaries tables, 'Medical dictionary' and 'Product dictionary'. The diseases in CPRD tables are encodes using Read version (v) 2 codes also known as 'Medcodes' e.g. Heart failure has Medcode 'G58..00'. A list of Medcodes with descriptions of the corresponding medical terms is presented in the Table 'Medical dictionary' and products (e.g. prescribed drugs) are encoded using 'prodcodes'. Table 'Product dictionary' lists all such drugs and their descriptions. We refer you to Tables 13 and 14 for columns descriptions in the dictionary tables.

3.3 Data from linked sources

The linked data from HES are organised into 11 tables: Patients, Hospitalizations, Episodes, Diagnoses by episodes, Diagnoses by hospitalisation, Primary diagnoses across a hospitalisation, Procedures, Augmented Care Periods, Critical Care, Maternity, and Health Resource Group. See Tables 15-25 for the descriptions of all table columns. For this research, only the first 6 tables were necessary as these contain information about diagnoses. All tables contain column 'patid' to assist in linkage. Individual patients may contribute data to more than one GP practice. In this case, multiple patid's may represent the same individual. However, a column 'gen hesid' within the table 'Patients' in the HES data provides a CPRD generated unique key which allows information in CPRD tables and in HES tables to be identified where it refers to the same individual. Within the total dataset, there were $15,107$ cases which had multiple patid's.

The information about diagnosis in HES is recorded using the International Classification of Diseases version 10 (ICD-10) coding frame. This differs from the CPRD's medcodes e.g. Heart failure has Medcode 'G58..00' while the corresponding ICD-10 code is 'I50.9'. For this reason, a mapping between ICD-10 codes and medcodes was required.

The linked ONS Death registration data consists of only one table. It contains the required information of patid, date and cause of death. A full list of columns is shown in Table 26.

The IMD data consists of two files - one contains a dictionary to identify the GP county, and another with the deprivation score at the individuals and GP level. A description of the columns is shown in Table 27.

Not all individuals in the total dataset are 'eligible' for the linked data: HES, ONS and IMD. There is a text document linkage eligibility *new patids.txt* which provides which individuals are 'eligible' for linked data. Individuals eligible for HES linked data have attribute 'hes e' equal to 1. Similarly, individuals eligible for ONS linking have 'death e' equal to 1, while individuals eligible for IMD linking have 'lsoa e' equal to 1.

4 Data treatment and missing data analysis

4.1 Information required for mortality risk analysis

The selection of variables for this study was informed by a combination of evidence from the academic literature (as discussed in Section 2) and established actuarial practice within the insurance industry. Many of the included factors, such as age, gender, smoking status, body mass index (BMI), and indicators of co-morbid conditions, have consistently been associated with variations in mortality risk among individuals with diabetes. These are widely recognised both in clinical studies and in underwriting guidelines as relevant indicators of long-term health outcomes. Additionally, variables such as blood pressure, cholesterol levels, and HbA1c were included based on their routine use in risk stratification and premium setting by insurers, where actuarial judgment and historical claims experience inform expectations of mortality risk, even in the absence of strong causal evidence. As mentioned in Section 2.2, the objective of this analysis is not to establish epidemiological causation, but rather to identify variables that show meaningful correlations with mortality outcomes within the diabetes population. This distinction aligns with the practical needs of insurance underwriting, where the goal is to detect consistent, interpretable differentiators in mortality risk that can support fair and evidence-based decision making.

The following data items are available across the linked data. The aim of data treatment was therefore to produce a data table, where each individual (patid) has the following information, which is appropriately cleaned and suitable for analysis:

- Outcome of interest:
  - Individual's status as alive or deceased as at study end date 31.12.2019
  - Date of death, where applicable
- Mortality risk factors of interest:
  - Static data items
    - Age at the start date
    - Gender
    - Index of multiple deprivation (IMD)
    - Smoking status
  - Regularly recorded data items

* Body mass index (BMI)
* Blood pressure (BP)
* Cholesterol level (CL)
* Blood glucose (sugar) level (HbA1c)
* Indicator for diagnosed diseases which are considered 'potentially significant' (see Table A). We emphasise that this is the original list of 'potentially significant' diseases for the mortality of individuals with diabetes. This list was formed based on the literature review (Section 2), and consultations with specialists from Leicester Diabetes Research Center. We do not claim that all these diseases are indeed significant. In fact, identification of a subset of significant diseases is a major part of this research project.
* Date of diagnosis, where applicable

The following data items, although available across the linked data and considered potential risk factors of interest, were not included for analysis for the reasons shown below.

- Marital status – high level of missing data
- Ethnicity – high level of missing data and not used by the UK insurance industry

The following sections provide insights for each item of data used including commentary on missing data analysis where relevant and data treatment. Further commentary can be found alongside the relevant tables within the 'Tables and Figures' document.

Table A: List of abbreviations used for medical conditions

DT1	Type 1 Diabetes	DT2	Type 2 Diabetes
AF	Atrial Fibrillation	AID	Acquired immune deficiency syndrome
Amp	Amputation	Ang	Angina
Ano	Anorexia	Anx	Anxiety
Ast	Asthma	Bli	Blindness
Bro	Bronchiectasis	CHD	Coronary heart disease
Can	Cancer	CLD	Chronic Liver Disease
CS	Chronic sinusitis	COP	Cryptogenic organising pneumonia
CyF	Cystic Fibrosis	DDI	Diverticular disease of intestine
Dem	Dementia	Dep	Depression
Epi	Epilepsy	Hem	Hemiplegia
HeL	Hearing loss	HF	Heart failure
Hyp	Hypertension	IBS	Irritable bowel syndrome
LD	Learning disabil.	MLD	Mild Liver Disease
MVD	Macrovascular	MSc	Multiple sclerosis
PrD	Prostate disorders	PsE	Psoriasis or eczema
PuF	Pulmonary Fibrosis	PUD	Peptic Ulcer Disease
Sch	Schizophrenia	RhA	Rheumatoid arthritis
Str	Stroke	SLD	Moderate or Severe Liver Disease
ThD	Thyroid disorders		

4.2 Data for outcome of interest

There are two potential sources of date of death: 'deathdate' field in the original CPRD data and 'ONSDeath' field in the linked ONS data. For some individuals there are discrepancies in these fields. Some individuals have different dates of death (see Table 46) and there are $86,825$ records with an ONS date of death without a 'deathdate' in the CPRD data (this is 16% out of the total of $543,668$ individuals in the total sample linked in ONS). The lack of any recorded activities (such as GP visits, hospitalizations, etc.) after the ONS date of death is considered adequate evidence that the ONS date of death is correct. Within the insurance industry, the ONS data is considered the 'gold standard' for death records and this is supported in the literature (Gallagher et al., 2019).

All individuals who died before the study start date are removed. To ensure that individuals with deaths pre-dating the ONS death register's creation in 1998 are not included within the study period, the CPRD date of death is used if it is before 1998. For deaths after 1998 the ONS death data is used.

The CPRD date of death is considered fit for the filtering purpose above but is not considered fit for the purpose of mortality analysis. Therefore, where a link between the ONS and CPRD data cannot be established, these cases are removed from the study. That is, only individuals with ONS linked data are used. Although, this reduces all sample sizes it is considered appropriate because the CPRD date of death information is considered to be unreliable for mortality analysis. In the diabetes sample, there were $621,115$ individuals initially. A link can be established to the ONS for $283,057$ (45%). Excluding those individuals who died before the study start date, the sample contains $242,461$ individuals (85%). In the general sample, there were $250,000$ individuals initially. A link can be established to the ONS for $103,597$ (41%). Excluding those individuals who died before the study start date, the sample contains $95,786$ individuals (92%). These data are summarised in the table below.

Sample Type	Initial Individuals	Linked to ONS (%)	Alive at Study Start (%)
Diabetes Sample	621,115	283,057 (45%)	242,461 (85%)
General Sample	250,000	103,597 (41%)	95,786 (92%)

- The effect of excluding the non-ONS linked individuals is compared between the initial sample (all individuals) and the linked sample (only individuals linked with ONS). Standard tests for statistical significance such as Kolmogorov-Smirnov test (KS-test) are not useful for large datasets[3]. Therefore, a manual threshold is introduced. A 1% difference in fractions for linked and non-linked sub-samples is considered a sufficient effective size. The impact of filtering needs to be considered by the main variables of interest for mortality analysis including gender, age, and comorbidities.

- Table 47 shows the contingency of gender from the initial to the linked sample. Female individuals comprise 47.8% among the initial sample and 47.6% among the linked sample. The change in the composition of the samples, i.e. the effect size is 0.2%. Because this is less than 1%, it can be assumed that linking and gender are independent.

- Figure 3 presents the contingency of age from the initial to the linked sample. The age distributions are similar and the mean age with 59.79 for the initial sample and 60.82 for the linked sample. The difference in mean age is nearly 1 year, which is considered as negligible alongside the similar age distribution.

- Table 48 presents a comparison of various diseases associated with the individuals in the initial and the linked sample. Most of the diseases (28 out of 41) show statistically significant differences in prevalence between the initial and the linked samples, indicating that linkage to the ONS is not independent of disease status for these conditions. This suggests that certain comorbidities may affect the likelihood of successful linkage, potentially due to systematic differences in healthcare usage or death registration practices associated with specific conditions. As a result, this introduces the possibility of selection bias in analyses relying solely on the linked sample. While the effective sizes for many of these conditions remain modest, this dependency needs to be acknowledged in any inference about disease-related mortality risks to avoid over- or under-estimating effects due to differential linkage rates.

A careful investigation of the database reveals the existence of some individuals of old age that are linked to ONS, have no death date records, but also have no recorded GP activity after the study start date. We suspected that these individuals are in fact deceased or immigrated and no longer residing within the UK[4], and performed a more careful investigation.

Of the $1,205,657$ records within the total sample:

- $53,045$ records have an age at the study start date of at least 80.
- According to CPRD data, $33,079$ were alive at the study start date.
- According to the ONS linked data, $8,566$ of the above were linked to ONS and alive at the study end date.
- The overlapping cases from the above two datasets indicated as 'alive' is $8,347$.

For those 8,347 'alive' individuals, we then search for test dates and diagnoses dates, and we have discovered that

[3] Such tests almost always return a statistically significant difference, due to the large data size.

[4] There may be other similar reasons, e.g. transferring from GP to a care provider.

- $1{,}353$ individuals have the last date of diagnoses before the study start date.
- $1{,}884$ individuals have the last date of test before the study start date.
- There are totally $1{,}251$ individuals without any recorded GP activity after the study start date.

Table 49 shows the age distribution for the samples before and after removing those individuals. In particular, there are 5 individuals older than 120 with no death date records, and all 5 cases have no GP recorded activity after the study start date. It is unlikely that individuals alive at these ages are no longer interacting with the registered GP. It was considered likely that these individuals were deceased, but the death was not recorded. This may occur where an individual immigrated from the UK some time ago. On this basis, all records with an age greater or equal to 80 at the study start date, which do not have any GP recorded activity after the study start date, are removed from the analysis.

The removal of the above $1{,}251$ individuals, which we call 'fake' individuals, has a crucial effect on the research objective, especially in insurance context. While the proportion of these individuals in the entire database is negligible ($1{,}251$ out of $1{,}205{,}657$, which is about 0.1%), their proportion among $8{,}347$ individuals of age 80+ linked to ONS is 15%. This percentage increases when we increase age threshold, and reaches 100% for the individuals aged 110+. Therefore, if 'fake' individuals were not excluded, they would lead to a severe underestimation of the mortality rate at old ages.

After removal these individuals, the number of individuals left in the total sample which are linked to ONS is $656{,}410$. All the statistic for the total sample described below is made based on this set of individuals.

4.3 Data for mortality risk factors of interest

The following data were processed as either static or regularly monitored.

4.3.1 Static data items

The following data items are considered static and therefore extracted from the data as at the study start date. Although some items may vary through the study period, such changes are not considered materially significant for the purposes of mortality modelling. For instance, occasional corrections to recorded date of birth or gender are rare and typically reflect data quality improvements rather than actual changes in the individual's characteristics. In the context of this study, where the objective is to evaluate broad mortality risk patterns across a large population, treating these variables as fixed at the study start date simplifies the analysis without introducing meaningful bias. This assumption aligns with industry practice in insurance risk modelling, where underwriting is typically based on characteristics known at a specific point in time.

Age and Gender Age is defined as the number of full years of age at start of the study. It has been computed as 2009 minus the year of birth. In Cox proportional hazard model, we will also have a parameter t that represents time from the study start to any given time moment. The age at any moment can then be computed as $A + t$, where A is the age at the study start. Having A constant simplifies the analysis.

Gender is recorded as binary value: 1 – Male, 2 – Female, as recorded at birth.

IMD Index of multiple deprivation (IMD) takes an integer value from 1 to 10, and is computed based on the postcode recorded at the study start. For individual records, the data item 'IMD decile' is missing for nearly 58% of total sample and nearly 61% of the general sample. In these cases, the missing value is impute with the IMD for the individuals registered GP. The remaining missing values, 291 cases (171 for which the IMD is recorded as 0 and 120 for which it is recorded as NaN) are allocated a value 5, which is a middle value for the index.

We acknowledge that the proportion of missing individual-level IMD data is substantial and that GP catchment areas are geographically broader than individual postcodes. However, excluding these records would result in a loss of over half the study population, drastically reducing statistical power and likely introducing significant selection bias, as data missingness is rarely random. Consequently, we utilised the GP-level IMD as a pragmatic proxy, operating on the reasonable assumption that patients generally register with a GP practice in close proximity to their residence, thereby sharing similar socioeconomic characteristics.

Smoking status Files *Smoke_ent_*.txt* contain information about smoker status (1 – Yes, 2 – No, 3 – Ex). From *Smoke_ent_*.txt* we received $12{,}511{,}090$ records. After removing records with value 0 'Data Not Entered' we have $12{,}509{,}309$ records. After removing duplicated records, we have $11{,}068{,}831$ records.

These files should be accompanied by *Smoke_Clin_*.txt* to identify dates. There were $10{,}714$ records with missing event dates. These dates were substituted by system date, which is the date the record was made in the database.

In addition, files *Smoke_Int_*.txt* and *Smoke_test_*.txt* present smoking information using medcodes. ICD-10 codes and medcodes for smoker status are presented in Table 35. Total number of records from files *Smoke_Int_*.txt*, *Smoke_test_*.txt* and *Smoke_diag.txt* is $16{,}248$. After removing all of records with medcodes unrelated to smoke status we have $4{,}107$ records.

In the total dataset we have $75{,}727$ individuals or 6.7% without smoker status recorded. In General sample we have $50{,}080$ or 21.6% such individuals. Results of analysis of randomness of missingness are presented in Table 36. They show that values missed not completely at random. This means that we cannot remove records with missing values, but we should impute data. The imputation method we used is set value 'Non-smoker' for records without smoker status. This approach is based on the heuristic of clinical coding practices, where positive risk factors (such as smoking) are actively recorded by clinicians, whereas the absence of a record is frequently used to implicitly denote a negative (non-smoking) status.

4.3.2 Regularly recorded data items

The following data items are considered to be regularly recorded and therefore extracted from the data on an ongoing basis between the start date and end date of study.

It is a known issue in health data analysis that observational data can be recorded and missing in a biased way. This is because it is more likely for health metrics to be recorded where health is poor or there is a diagnosis which requires more frequent interaction with the health system or management of that metric (Rusanov et al., 2014). Therefore, missing data analysis and treatment is carefully considered herein and the impacts discussed within the results and conclusions.

Body mass index (BMI) There are two sources of BMI observation data. Information was extracted from both sources, merged, cleaned, and then analysed for missing data bias.

The first source is any table containing data with medcodes specified in Table 28 and Table 29. In total, there are $489{,}756$ records with medcodes from Tables 28 and 29 which represent 4% of all BMI measurements. For the 277 observations with an interval code and no corresponding event date (0.05%), the system date is substituted. That is, the date the information was entered into CPRD which is, for those cases where both dates are known, within a few days of the event date.

The second source of BMI measurement observations is within the 'Additional Clinical details' table of the HES data alongside the date of the observation as found in the clinical information file. There are $14{,}317{,}076$ records in this table, which is 96% of all BMI measurements. Each record contains patid (ID of individual), adid (ID of measurement), and the value of the measurement. These records were filtered to remove a) those with no measurement value ($308{,}392$ measurements, 2.5%) and b) any duplicate with the same combination of (patid, adid) ($1{,}824{,}379$ duplicates, 15%).

Joining the two sources provides a file with a time series of BMI measurement observations. Individuals have from 1 to a maximum of 1,508 observations within the study period. Records with obvious data input errors such as negative, nil, very small/large values ($16{,}221$, 0.13%) are removed. After this, there are $12{,}193{,}099$ BMI measurement observations for $978{,}380$ individuals. We will call this 'BMI total sample'.

BMI is known to be recorded and missing in a biased way. This is because it is more likely for BMI to be recorded where health is poor or there is a diagnosis which requires more frequent interaction with the health system or management of weight (Nicholson et al., 2019). For the analysis of missing data, a given individual is considered to have a known value where there is at least one BMI measurement, otherwise it is missing. This disregards the lack of recording frequency which is expected for healthier lives. While this approach does not distinguish between the clinical context of individuals with differing health needs, it offers a pragmatic framework for handling large-scale, routinely collected data. It is important to acknowledge that the presence of a single BMI measurement may reflect varying levels of data quality: for instance, in individuals with chronic conditions such as diabetes, infrequent BMI recording may indicate incomplete data, whereas in healthier individuals, infrequent contact with healthcare services may reasonably result in fewer recordings. Although this introduces some bias, treating the presence of at least one BMI value as a known measurement provides a consistent and transparent definition of observed versus missing data.

Table 30 presents the results of the relation between missingness of BMI and diagnosed diseases of interest in the BMI total sample. For most diseases of interest, the missingness of BMI is not random, as would be expected based on clinical practices. This aligns with prior expectations, since conditions such as diabetes, cardiovascular disease, and obesity are known to prompt more regular weight and BMI monitoring as part of ongoing management and risk assessment. Conversely, there are only three diseases—Epilepsy, AIDS, and Multiple Sclerosis—for which the association with BMI missingness is statistically insignificant. This also seems reasonable, as these conditions may not routinely require weight monitoring as part of standard care pathways. One further condition, Learning Disability, shows a small p-value suggestive of statistical dependence but a very small effect size, indicating minimal practical impact. This may reflect heterogeneous care patterns in this group, or variability in healthcare engagement and data recording practices. Overall, the observed pattern of associations largely corresponds with clinical expectations and known drivers of BMI measurement in routine care.

For the general sample of $250,000$ individuals, there are $769,757$ BMI measurements for $162,238$ individuals (65%). Therefore, only 65% of the sample have at least one BMI measurement observation. Table 31 shows the results of missing data analysis where the χ^2 independence test shows that there are only 4 attributes for which we do not have enough evidence to reject hypothesis about independence with confidence level 99%. There is only one attribute (diagnosed stroke) which has significant p-value (0.758).

In conclusion, the missingness of BMI is highly correlated with age, sex and comorbidities. This means that removing records with missing values is not appropriate and therefore impute the missing data would be preferable. For this analysis, missing BMI measurements are input based on the 'nearest neighbour' approach with regard to characteristics age, gender and comorbidities. Specifically, we used the 1NN method, in which the imputed value corresponds to the actual BMI of the nearest neighbour, rather than an average or moving average of several nearest neighbours. This preserves the granularity of real observations while ensuring that imputations reflect biologically and clinically plausible values from similar individuals. This method is considered suitable in this context because these characteristics are strong predictors of BMI and are consistently recorded with high completeness in the dataset.

Blood pressure (BP) Blood pressure measurements are found within the additional file named *BP_medval_*.txt* which contains over 32m observations. These records were filtered to remove those with a) no measurement value ($544,923$, 1.7%) and b) any duplicate with the same patient ID ($3,875,362$, 12%). Where diastolic field was greater than the systolic field ($61,286$, 0.2%), the values were reversed as this is a usual input error. The systolic and diastolic fields were reviewed for reasonableness (systolic range [30,120], diastolic range [50, 300]). Where a record had values outside reasonable range, the record was removed ($281,831 + 201,426$, 1.5%). For the 475 cases with a missing event date, the date was substituted by system date with the same logic as per BMI.

After the above data treatment, there are over 27m BP measurement observations which correspond to individuals in the data and so leave the remaining without BP observations: 7.6% ($85,651$) individuals in the total dataset and 26% ($65,286$) individuals in the general sample. Table 31a shows that the analysis of the randomness of the missing data is not completely at random, as expected. Therefore, it is not appropriate to remove records with missing values and the preferred treatment is to impute the data. The imputation method adopted was to take the mean value from individuals with the same gender and the same (or similar, that is, closest in the Hamming distance) set of comorbidities.

Cholesterol level (CL) The majority of cholesterol measurement observation data is found in the 'Test' tables and some results are also found in the 'Referral' tables. The data can be filtered specifically for cholesterol tests using readcodes and medcodes for cholesterol tests that are presented in Table 33.

The following values are measured during a CL test with the normal levels presented in Table 32.

- HDL (high-density lipoprotein) – the higher the better
- Non-HDL (also written as LDL)– the lower the better. This group includes IDL, VLDL and lipoprotein
- Total cholesterol (TC) or serum cholesterol – this is the total amount of cholesterol in the blood and includes both HDL and non-HDL cholesterol
- Triglycerides – the lower the better
- The TC to HDL ratio (TC:HDL) – the lower the better

Under extreme conditions (e.g. homozygote with defective genes), total cholesterol can be above 1000mg/dL which corresponds to 26 mmol/L (Kattah et al., 2019). Therefore, any cases where the observation is above 40 mmol/L within the data is assumed to be an incorrect coding of units. That is, the true unit is mg/dL and as such, it is converted to mmol/L[5].

For the 256 cases with a missing event date, the date was substituted by system date with the same logic as per BMI. In total there are over 12m CL measurement observations. After removing duplicates, there are over 10m observations of 903,041 individuals, about 75% of all individuals in the total dataset. For the general sample, cholesterol measurements are observed for about 33% individuals. Therefore, 67% of individuals do not have any CL measurements observations. Results of analysis of randomness of missingness (see Table 34) show that values are missed not completely at random. Therefore, it is not appropriate to remove records with missing values and the preferred treatment is to impute the data. The imputation method used is 1NN in space of age, gender and comorbidities- this is the same method as we used for BMI.

Blood glucose (sugar) level (HbA1c) HbA1c information can be found mainly in table 'Test' within the HES data. The result of HbA1c measurements can be presented either as interval or as a scalar data. To extract measurements of HbA1c, the medcodes presented in Table 38 are used. There are 100,064 measurement observations with interval data among 13,930,560 observations from the 'Test' table.

For those records with a missing event date, the date of data registration is taken. This is considered appropriate because, based on available records with both values present, the date of measurement and the registration date typically occur within a few days of each other. Therefore, using the registration date provides a reasonable approximation of the measurement time with minimal impact on the temporal alignment of the data.

The units of measurements for all individuals and for the general sample individuals are presented in Tables 39 and 40, respectively. As we can see from these Tables, there are many units of measurements used, and the intervals of reasonable values in these units have non-empty intersections. Therefore, if the unit of measurement is not given, it is impossible to guess it based on the value. For this reason, measurements with some units such as 'No Data Entered', 'No unit', etc., were removed. When the unit is given, we convert the measurements to % (A1c), which is one of the standard representations[6].

Extreme value observations were treated as incorrect data inputs and the observations removed. The sample data contains 10 records with HbA1c less than 2%, and 190 records with values greater than 20%. These values fall outside the ranges for reasonableness, because HbA1c value of 3.2% is considered as a pathologically low (Joob and Wiwanitkit, 2018), while, on the other hand, at values greater than 10% an individual 'needs injectable therapy' (Association et al., 2017), and values exceeding 17% have been observed by medics only at rare cases (Petznick, 2011). The removal of these 200 outlier records (0.0014% of the 13.9 million total observations) is not considered to materially affect the results, given the minimal proportion and their likely erroneous nature. These exclusions help to preserve the overall quality and reliability of the dataset without introducing meaningful bias.

There are 1,522,020 measurement observations with a date recorded but a measurement value missing. In these cases, the observations are removed.

In total, HbA1c has been measured for about 61% of individuals in the total sample. That is, 39% of individuals do not have any HbA1c measurement observation data. This varies significantly across samples. In the diabetes sample, about 10% of individuals have missing HbA1c value. In the general sample, the missingness of HbA1c is 85%.

It is anticipated that this measure is missing for those without a diabetes diagnosis because HbA1c is not measured where medics do not consider an individual at risk of or showing signs of diabetes. Therefore the analysis of missing data is approached separately for the two groups.

Results of analysis of randomness of missingness show that there are values missed not completely at random, see Table 41. Therefore, it is not appropriate to remove records with missing values and the preferred treatment is to impute the data. For individuals without a diabetes diagnosis for whom HbA1c has never been measured, a value within the 'normal' range is input (say, 5%)[7].

For individuals with a diabetes diagnosis, it is anticipated that HbA1c is missing without a significant bias by age (at least for adults) and gender (Weykamp, 2013). This prior expectation is supported by our data.

[5]Conversion of units follows the following rule: mg/dL = mmol/L $\times 38.6$ for cholesterol and mg/dL = mmol/L $\times 88.5$ for triglyceride.

[6]There are several standard units of measurement for HbA1c: % (A1c), mg/dl (eAG1), mmol/l (eAG2). For these standard units, there are formulae for conversion: eAG1 = 28.7A1c -46.7, eAG2 = eAG1 $\times 0.055766$. Mmol/mol can be transformed to % using the formula Hb[%]$= 2.1 + 0.092$Hb[mmol/mol].

[7]For individuals without diabetes, a normal range for HbA1c value is 4% to 5.6% inclusive. HbA1c levels between 5.7% and 6.4% are indicative of pre-diabetes. A HbA1c value of 6.5% or higher is indicative of diabetes (ElSayed et al., 2024).

Hence, we focus the analysis based on the dependence on comorbidities. For every individual P with an unknown HbA1c value, a search is performed across all individuals P_t in the diabetes sample such that (i) HbA1c is known and (ii) the Hamming distance between comorbidities of P and P_t is less than k, where k is the minimal value such that number of nearest neighbours is not less than 10. The input HbA1c value of P as the average of HbA1c of its nearest neighbours. While we acknowledge that this method has some limitations, e.g. it does not incorporate the temporal sequence of HbA1c observations, this approach allows us to preserve comorbidity-related structure in the data, which is particularly relevant given the strong clinical association between certain conditions (e.g., cardiovascular disease, renal impairment) and glycaemic control.

Diagnosed Diseases of interest Information is required regarding diseases (co-morbidities) that are expected to have a significant effect on the mortality risk . The list of diseases has been constructed by analysing the literature, and in consultation with DSG and colleagues from Leicester Diabetes Research Center. The resulting list of diseases is presented in Section 4.1.

The objective of the research is to differentiate mortality risk for individuals with diabetes. For example, those with coronary heart disease may have a higher mortality risk when there is also a diabetes diagnosis present compared to absence. Therefore, variables are coded to flag diabetes status and if any other disease is present.

Diagnosed diseases can be found in the CPRD data and in the HES linked data.

In the CPRD data, diagnosed diseases are encoded in the form of 'medcodes' and 'readcodes'; the codes relevant to diseases listed in Section 4.1 are summarised in Table 45. If any of the non-diabetes codes are present in the CPRD data for an individual, a corresponding disease status variable is set to equal to 1.

A diabetes status variable is coded based on a list of diabetes-related medcodes/readcodes as found in Tate et.al (Tate et al., 2017). This list is used here because it was freely available, and using it as a starting point significantly sped up the analysis.

The codes are split into three groups to define diabetes diagnosis status:

- 127 codes that are clearly related to Type 1 diabetes (D1T), see Table 42.
- 124 codes that are clearly related to Type 2 diabetes (D2T), see Table 43.
- 106 'unspecified' codes from which it is not possible to identify the Type of diabetes, see Table 44.

For each individual (patID) within the data, two variables are defined for modelling purposes based on if items from the diabetes specific list of medcodes/readcodes are present or not before the end of the study period:

- $d_1 = 1$ if Type 1 diabetes have ever been diagnosed, $d_1 = 0$ otherwise.
- $d_2 = 1$ if Type 2 diabetes have ever been diagnosed, $d_2 = 0$ otherwise.
- For individuals with unspecified codes, these are classified in the following way:
  - If $d_1 = d_2 = 1$, unspecified codes are disregarded.
  - If $d_1 = 1$ and $d_2 = 0$, unspecified codes are classified as Type 1.
  - If $d_1 = 0$ and $d_2 = 1$, unspecified codes are classified as Type 2.
  - If $d_1 = d_2 = 0$, the classification is age dependent. If the first code appears below age 28, set $d_1 = 1$ and $d_2 = 0$, otherwise set $d_1 = 0$ and $d_2 = 1$.

Classification of the unspecified codes by age is considered appropriate because Type 1 Diabetes is usually diagnosed earlier in life compared to Type 2 (Thomas et al., 2018). The distribution of first diagnosis within the total sample by Type of Diabetes is presented on Figure 2. It can be seen that this also holds within the diabetes sample; for individuals with age of first diagnosis of diabetes less than 28 it is more likely to be Type 1 diabetes and for first diagnosis of diabetes at age 28 or more it is more likely to be Type 2 diabetes.

In the HES data, diagnosed diseases are encoded using International Classification of Diseases (ICD) codes instead of medcodes/redcodes. At the time of the research, it was not possible to find a clear list of ICD codes corresponding to each diseases as in Tables 42, 43, 44, and 45. To utilise the HES data, the ICD codes have been converted to medcodes/redcodes. Then we identify the diagnosis as per the process outline above and summarised in Tables 42, 43, 44, and 45.

For the conversion, a dictionary from file ICD10.DBF from the 'Clinical Terminology Browser' was used (the only version released on 19.03.2018) available online at https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9/subpack/8/releases. It was developed to translate readcodes to ICD10 and there may be several

readcodes for one ICD10 code. There are $14{,}702$ different ICD10 codes covering $116{,}374$ readcodes in the file. Unfortunately, there are 458 ICD10 codes used in the HES data within the sample that are not defined in this dictionary. In these cases, a nearest match is used to identify the diagnosis. For any given code, the ICD code with the same four symbols that could be found in the dictionary ICD code is used and then the corresponding row is used for decoding; otherwise the ICD code with the same three symbol code is used.

For every disease, the date of the first record with the corresponding ICD, readcode, or medcode is used as the date of diagnosis. In effect, all records with the same disease at a later date are ignored. Some records are missing a date of diagnosis. However, all these cases contain a date of inclusion of information into the system (known as 'system date'). Therefore, where the diagnosis date is missing the system date is used. Adopting this approach, reduces the 0.6% of records in the total sample with missing date of diagnosis to nil.

5 Direct data analysis

To undertake reasonableness checks and better appreciate the morbidity and mortality risk within the samples, data analysis was undertaken. We computed various quantities of interest, such as mortality rates or diseases probabilities, directly from data, without the use of any models. This preliminary step aimed to ensure that the underlying data was consistent with expected trends and to offer initial insight into the proximity of observed outcomes, such as mortality and morbidity, to key rating factors. These analyses form a critical bridge between raw data and the structured modeling process, helping to distinguish correlation from causation and identifying patterns that may justify the inclusion or exclusion of certain variables in later stages.

The analysis in each subsection below is performed based on one of the following samples:

- general sample (GS)
- diabetes sample (DIB)

The distribution of age for GS and DIB is presented in Table 51 and Figure 4.

5.1 Mortality rates

To check the quality of data, a direct mortality estimation was performed by age bands on the GS and compared to the National Life Tables (NLT) (Office for National Statistics, 2025). These tables are appropriate for several reasons: they provide official, high-quality mortality benchmarks derived from comprehensive national data; they cover a similar geographical region as the study sample (the UK for NLT versus England for the study sample); data for various time periods are available. Furthermore, the ONS methodology is transparent, statistically robust, and widely accepted in actuarial and demographic research, making the NLT a reliable external reference point for validating age-specific mortality rates.

Mortality is estimated in age bands because, by any separate age, the direct estimate would produce inaccurate results due to small sample sizes. The age bands employed are in 10-year increments, except for the youngest group (e.g. 16-20, 20-30, 30-40, and so on).

For each age band A, the probability of death (e.g. the mortality rate) is calculated using the formula:

$$\mu_A = \frac{N_{DA}}{N_A}$$

where: N_A is number of individuals in the age band A at the start of the study, and N_{DA} is number of observed deaths within the study period from those individuals in age band A. The resulting mortality rates are presented in Table 68.

The NLT ('Tables') present, for each age x, the probability q_x that an individual aged x will die before reaching age $x+1$. The notation $q(x,t)$ therefore denotes the probability, for an individual aged x, to die during the year t. The Tables present mortality rates by three-year periods, e.g. $q(x,2011)$ is the central estimate for the period $2010-2012$. A mortality rate by 10 year age bands over the study period is computed from the Tables in order to make a comparison to the μ_A values calculated from GS. To achieve this, the number of NLT expected deaths for each year t is calculated for a given population N_x for each age x, e.g.

$$M(x,t) = N_x \cdot q(x,t).$$

The number of death in the following year $t+1$, is calculated after reducing the population at age x for the deaths in the year t and by accounting for the change in the probability of death for the cohort now 1 year

older:

$$M(x+1,t+1) = (N_x - M(x,t)) \cdot q(x+1,t+1).$$

Therefore, the number of deaths during the year $t+k$, for the cohort aged x in year t, can be calculated as

$$M(x+k,t+k) = \left(N_x - \sum_{p=0}^{k-1} M(x+p,t+p) \right) \cdot q(x+k,t+k).$$

This can also be rewritten as:

$$M(x+k,t+k) = N_x \cdot \prod_{p=0}^{k-1} (1 - q(x+p,t+p)) \cdot q(x+k,t+k).$$

The total number of deaths, for the cohort aged x at the study start, over the study period of 2010 to 2019 is therefore:

$$M_x = \sum_{k=0}^{9} M(x+k,2010+k).$$

The total number of deaths within an age band A is the summation of the total deaths for each age x within A:

$$M_A = \sum_{x \in A} M_x.$$

To arrive at an estimation of mortality from NLT, the total deaths by age band, M_A, are taken as a proportion of the population at the start of the period, $N_A = \sum_{x \in A} N_x$,

$$\mu_A = \frac{M_A}{N_A}$$

Using the N_x values from the GS, the implied mortality rate μ_A based on the NLT estimate rates are presented in Table 69. Figure 10 compares these to the GS mortality and the observed difference is negligible which provides confidence that, at least in the aspect of general mortality, our data are consistent with the general UK population.

5.2 Diseases probabilities

Before building the model for mortality prediction, we performed some direct estimates for diseases probabilities that may be of independent interest. The objective of this exercise is to assess the plausibility and internal consistency of the morbidity information before incorporating it into a multivariate modelling framework. This step is of particular interest to the actuarial audience because the presence and prevalence of certain diseases, such as cardiovascular conditions or cancer, are known to be strong predictors of increased mortality risk of the individuals with diabetes and are frequently used as underwriting factors. By estimating these probabilities directly from the data (i.e., without modelling assumptions), we aimed to uncover potential patterns, anomalies, or data quality issues that might impact downstream risk assessment.

First, the probabilities of various diseases for individuals in the general (GS) and in diabetes (DIB) samples are computed as the ratio:

$$P(A) = \frac{N_A}{N_T},$$

where N_T is the total number of individuals in the sample, and N_A is the number of individuals in this sample for which disease A has ever been diagnosed. The resulting disease probabilities are presented in Table 50. Some of the findings are highlighted in green: for example, the most frequent disease in GS is dementia, the most frequent in DIB (that is, given a diabetes diagnosis) is hypertension. Care should be taken when interpreting such direct estimates. For example, it cannot be concluded that diabetes increases the risk of hypertension as other factors are not accounted in such analysis, such as the difference in the weighted average age of the samples. DIB is on average older than GS, see Table 51 and Figure 4, and older individuals have higher risk of hypertension.

Second, the probability of various pairs of diseases in the samples is computed. That is, for any pair A and B of diseases, the following ratio is computed:

$$P(A|B) = \frac{N_{AB}}{N_B},$$

where N_B is the number of individuals in the sample for which disease B has ever been diagnosed, and N_{AB} is the number of individuals with both diseases A and B diagnosed. The results are presented in Tables 52-54 and Figure 5. In Table 52, yellow background highlights conditional probabilities greater than 50%, blue background - conditional probabilities greater than 70%, while green background - conditional probabilities greater than 99%. The same information is also presented in Figure 5 (right). As expected, the conditional probability greater than 70% happens only in cases where one disease implies the existence of another by definition. Table 53 and Figure 5 (left) present the same information but with yellow background highlighting conditional probabilities greater than 30%, while blue background, greater than 60% (the green background is still greater than 99%). Table 54 provides the actual values of these conditional probabilities. Conditional probabilities of one disease given another disease for diabetes sample are presented in Tables 55 - 57 and Figure 6.

These analyses provide valuable insights into the structure and co-occurrence of morbidity within the general and diabetic populations. From an actuarial perspective, understanding the marginal and conditional probabilities of diseases supports several key objectives: it aids in risk stratification, informs underwriting guidelines, and provides empirical justification for the inclusion of specific health indicators in predictive models. Furthermore, patterns of disease clustering can point to latent health dependencies that may materially affect mortality or morbidity risk. Although these estimates do not account for confounding factors and should not be interpreted causally, they offer a transparent, data-driven foundation for exploring how chronic conditions interact and concentrate within higher-risk groups—insights which are critical for both model calibration and the development of fair, evidence-based rating structures.

5.3 Relative information gain

The entropy for disease A is given by the formula:

$$H(A) = H(N_A, N_T) = -\frac{N_A}{N_T}\log_2\frac{N_A}{N_T} - \frac{N_T - N_A}{N_T}\log_2\frac{N_T - N_A}{N_T}.$$

The conditional entropy of A given B is:

$$H(A|B) = \frac{N_B}{N_T}H(N_{AB}, N_B) + \frac{N_T - N_B}{N_T}H(N_A - N_{AB}, N_T - N_B).$$

The relative information gain (RIG) of A given B is then defined as:

$$RIG(A|B) = 1 - \frac{H(A|B)}{H(A)}.$$

The objective of using entropy and relative information gain (RIG) in this context is to quantify the degree of uncertainty reduction about the presence of one disease (A) given knowledge of another disease (B). Entropy provides a measure of unpredictability or information content, while conditional entropy captures the remaining uncertainty in A once B is known. RIG then expresses the proportion of uncertainty in A that is resolved by knowing B, thus offering a normalised measure of association strength between diseases.

The RIG metric is particularly useful because it accounts not just for co-occurrence frequency but for how much predictive information one disease provides about another. A higher RIG value indicates that disease B is strongly informative about the likelihood of disease A occurring. For example, a RIG of 20% means that knowing whether an individual has disease B reduces the uncertainty about their status regarding disease A by 20%, relative to not knowing B. The relative information gain of one disease given another disease for GS are presented in Table 58, Table 59, and Figure 7. In Table 58, yellow background highlights RIG greater than 5%, blue background highlights RIG greater than 10%, while green background highlights RIG greater than 20%. The same information is presented diagrammatically on Figure 7. Table 59 provides the actual values of all relative information gains. The same information for diabetes sample is presented in Table 60, Table 61, and Figure 8.

From an actuarial and risk modeling perspective, this information is highly valuable. It allows for the identification of disease pairs where one condition may act as a proxy or early indicator for another, which has implications for both underwriting and early intervention strategies. Furthermore, these insights can inform variable selection or interaction terms in predictive models by highlighting the most informative disease relationships. While RIG does not imply causality, it provides a clear and interpretable measure of association strength that can guide both data exploration and practical decision-making in morbidity risk assessment.

5.4 Diabetes morbidity rates

The frequencies of diabetic individuals among those with a certain disease in the sample is investigated. The objective is to identify which diseases are most strongly associated with the presence of diabetes, helping to uncover potential comorbidities and support risk profiling. This analysis provides insight into how frequently diabetes co-occurs with other conditions, which can inform both underwriting decisions and the design of predictive models that incorporate disease history as a factor in assessing health risk.

In Table 62, the diseases are listed in alphabetical order, while in Table 63 these are sorted by proportion of diabetic cases among individuals with the given disease. Where there are less than 500 cases in the GS, the presented estimates for these diseases are unreliable due to the lack of data and are therefore ignored in further analysis (5 diseases highlighted by yellow background in Table 63). Angina has the highest frequency with over 30% of individuals with angina are also diabetic. Irritable bowel syndrome (IBS) has the lowest frequency with about 8% of individuals being diabetic. This distribution appears reasonable when considering the known associations between diabetes and various comorbid conditions. Angina, a cardiovascular condition, is strongly linked to diabetes due to shared risk factors such as hypertension, obesity, and dyslipidemia, which may explain the high co-occurrence. In contrast, irritable bowel syndrome (IBS) is a functional gastrointestinal disorder with less established metabolic or vascular overlap with diabetes, which supports the observed lower co-prevalence. The pattern aligns with expectations based on current clinical and epidemiological understanding.

The frequencies of individuals with diabetes among individuals with a certain combination of diseases was also investigated. The objective was to identify multi-morbidity patterns that are particularly associated with a higher prevalence of diabetes, potentially revealing synergistic effects between conditions that elevate diabetes risk. Only combinations for which there are at least 500 individuals in the GS are included. The results are presented in Table 64. The findings show that certain combinations, particularly those involving cardiovascular diseases (e.g., angina and hypertension) are associated with markedly higher diabetes prevalence with values often exceeding 40%. This is consistent with known clinical risk factors and pathophysiological mechanisms that link these conditions with diabetes. On the other hand, there are combinations for which the prevalence of diabetes is around 10%, which is substantially lower than the overall proportion of adults with diabetes in UK. The use of a 500-individual threshold helps ensure that the presented estimates are statistically robust.

These data are used to build a simple linear regression model for diabetes morbidity rates. This model and its analysis are simplistic; it requires assumptions that are unlikely to be reasonable. This simplistic analysis, however, does help meet the objective of illustrating broad patterns and potential associations between disease prevalence and diabetes rates across different conditions or combinations thereof. As it provides useful contextual appreciation of the data to the reader, it serves as a starting point for understanding where comorbidities may signal higher diabetes risk. In this model, each individual is characterised by a vector $x = (x_1,\ldots,x_n)$, where n is the number of diseases (excluding diabetes) and x is an indicator function where $x_i = 1$ if and only if disease i has ever been diagnosed, otherwise $x_i = 0$. The simplistic assumptions required are: that disease i increases the probability of having diabetes by a factor γ_i, and that the influence of all diseases are independent. Then the probability $p(x)$ that an individual also has a diabetes diagnosis is:

$$p(x) = p(0) \cdot \prod_{i \in S(x)} \gamma_i,$$

where $0 = (0,0,\ldots,0)$, and $S(x) = \{i : x_i = 1\}$ is the set of diseases of this patient. Then

$$\log p(x) = \log p(0) + \sum_{i \in S(x)} \log(\gamma_i) = \alpha_0 + \sum_{i \in S(x)} \alpha_i = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i,$$

where $\alpha_0 = \log p(0)$ and $\alpha_i = \log(\gamma_i)$, $i = 1,\ldots,n$. This is a linear regression model, and the regression coefficients α_i can be found from minimising the sum-of-squares error in this approximation. If $\alpha_i \approx 0$ for a given i, it can be concluded that disease i is not significant, and so exclude it, and repeat the calculation. The results (for all individuals in GS) are presented in Figure 9 and Table 65. Figure 9 presents the process of selection of statistically significant subsets of attributes. The final regression coefficients are presented in Table 65. It can be seen that there are several attributes with negative coefficients. This means that the presence of these diseases is associated with the absence of diabetes within the sample.

The simplicity of this model is beneficial in this exploratory context because it allows for clear visualization and interpretation without the added complexity of more sophisticated methods, which may obscure the basic relationships. However, the assumptions of the model are unlikely to hold in a real-world, multi-morbid population. Moreover, confounding factors (e.g., age, sex, socioeconomic status) are not controlled for in this

model, and disease interactions are not accounted for beyond simple additive effects. Therefore, readers should view the results as illustrative rather than predictive or causal. The model can suggest areas for deeper analysis but should not be used to draw firm conclusions about risk or policy implications without further, more rigorous statistical modelling.

5.5 Diseases influence on mortality rates

The overall probability of death, e.g. the mortality rate, can be estimated as:

$$P(D) = \frac{N_D}{N_T},$$

where: N_T is the total number of individuals in the sample, and N_D is the number of individuals who died during the study period.

In GS, the overall probability of death is

$$P(D) = 0.158730.$$

For comparison, the age-averaged probability of death within the next 10 years calculated from National Life Tables in about 0.152. The absolute difference between these estimates is marginal (< 0.007), indicating that the mortality profile of the study cohort is highly representative of the general population. The slight deviation is within the expected range given that the NLT cover the entire population, whereas the CPRD sample is subject to minor variations based on the geographic distribution of participating GP practices.

To appreciate the relative increase in mortality risk for a given disease, the ratio of deaths for each disease A can be calculated as follows:

$$P(D|A) = \frac{N_{DA}}{N_A},$$

where: N_A is a number of individuals in a sample with disease A ever diagnosed, and N_{DA} is the number of such individuals who died during the study period. This simple ratio provides an intuitive measure of observed mortality among individuals with a specific diagnosis. However, it has important limitations: it does not account for differences in age, sex, comorbidities, or follow-up time, all of which can strongly influence mortality. It also does not imply causality or isolate the excess risk attributable to the disease itself, as confounding factors may bias this estimate.

For example, from the GS with A set to be diabetes, the overall probability of death is much higher than the wider GS:

$$P(D|A) = 0.408154.$$

Of course, this is partially due to the fact that individuals with diabetes are generally older.

Similarly, let

$$P(D|\overline{A}) = \frac{N_{D\overline{A}}}{N_{\overline{A}}},$$

where $N_{\overline{A}}$ is a number of individuals in a sample for whom disease A has never been diagnosed, and $N_{D\overline{A}}$ is the number of such individuals who died during the study period. Then the influence of disease A on the mortality can be estimated as the following odds ratio:

$$OR(A) = \frac{P(D|A)}{P(D|\overline{A})}.$$

Now consider two diseases, A and B. Let N_{AB} be the number of individuals in the sample with both diseases diagnosed, $N_{A\overline{B}}$ - the number of individuals with disease A diagnosed but without disease B present, $N_{\overline{A}B}$ - the number of individuals with disease B diagnosed but without disease A present, and $N_{\overline{AB}}$ - the number of individuals for which neither disease A nor B have ever been diagnosed. Let N_{DAB} be the number of individuals with both diseases diagnosed who died during the study period, and let $N_{DA\overline{B}}$, $N_{D\overline{A}B}$ and $N_{D\overline{AB}}$ be defined similarly.

Then the following probabilities are computed:

$$P(D|AB) = \frac{N_{DAB}}{N_{AB}}, \quad P(D|\overline{A}B) = \frac{N_{D\overline{A}B}}{N_{\overline{A}B}}, \quad P(D|A\overline{B}) = \frac{N_{DA\overline{B}}}{N_{A\overline{B}}} \quad \text{and} \quad P(D|\overline{AB}) = \frac{N_{D\overline{AB}}}{N_{\overline{AB}}}$$

to enable the investigation of the following odds ratios:

$$OR(A|B) = \frac{P(D|AB)}{P(D|\overline{A}B)}, \quad OR(A|\overline{B}) = \frac{P(D|A\overline{B})}{P(D|\overline{A}\overline{B})},$$

$$OR(B|A) = \frac{P(D|AB)}{P(D|A\overline{B})}, \quad \text{and} \quad OR(B|\overline{A}) = \frac{P(D|\overline{A}B)}{P(D|\overline{A}\overline{B})}.$$

For example, $OR(A|B)$ measures the influence of disease A on the mortality rate of individuals having disease B. The meaning of ratios $OR(A|\overline{B})$, $OR(B|A)$ and $OR(B|\overline{A})$ is similar.

Table 66 presents the above applied to GS with $A =$diabetes, and $B = F$ is any other disease. The disease with the highest probability of death, given diabetes, is heart failure; more than 73% of individuals with heart failure and diabetes died. For comparison, approximately 70% of individuals with heart failure (diabetic or not) died. While this suggests that diabetes adds a modest absolute risk of death once heart failure is established, this finding aligns with the understanding that heart failure itself is the dominant driver of mortality in this comorbidity cluster (Seferović et al., 2018). However, this conditional probability analysis masks the critical epidemiological link: the probability of heart failure being present is markedly higher for diabetic individuals ($> 17\%$) compared to the general sample ($< 5\%$). This is highly consistent with established literature, including data from the Framingham Heart Study, which demonstrated that diabetes independently increases the risk of developing heart failure by two- to five-fold compared to age-matched controls (Kannel et al., 1974).

As mentioned above, $P(D|A)$ for $A =$diabetes is 0.408154. It can be seen that the combination of diabetes with some other diseases is below that average rate. Those with chronic sinusitis and diabetes experience a mortality rate of less than 31%. Hence, data suggest that chronic sinusitis is associated with a lower mortality risk for diabetics. There are few other diseases with this property, see Table 66. While this suggests a statistical association with lower mortality, it should not be interpreted as a biological protective effect. Rather, this finding likely reflects the 'Healthy User Effect' or selection bias common in observational data (Shrank et al., 2011). Patients diagnosing and managing non-life-threatening conditions like sinusitis often demonstrate higher health-seeking behavior and engagement with primary care. Furthermore, the presence of such a code may act as a marker for a subgroup free from more severe, immediately life-limiting comorbidities (such as heart failure or metastatic cancer) that would otherwise dominate the clinical picture.

The influence of combination pairs of diseases B, C on diabetic individuals mortality rate is analysed by computing $P(D|ABC)$ where A is diabetes. The results are presented in a separate document 'Diseases influence on mortality rates'. As expected, the data shows that, for most pairs of diseases B, C, their influence on the mortality of diabetes individuals is not independent, and therefore pairwise interactions of the diseases should be added into the final model.

6 Model overview

6.1 Model Type

6.1.1 Aim and suitability

The aim of the research is to model the data such that, given certain input information about an individual, the resulting model output provides an estimate of the probability of death within the next t years and provides estimates of how that varies for each element of the input information. This mimics the underwriting process for insurers; given a set of input information provided at the time of underwriting by an individual, the underwriters must assess the alteration to a 'standard' mortality risk posed by that individual based on that information. The key area of interest for this research is how that varies by the disease status of diabetes.

Given this objective, a key output of the model is the probability of observing a death within a given period of time (t), given a set of information (z). The probability of observing an event within a given period of time and assessing the impact of a set of variables (z) upon that probability, lends itself to a Cox proportional hazard model.

The model type selected was a Cox proportional hazards model because it is well-suited to the task of estimating the probability of death over time while quantifying the effect of multiple covariates on that risk. Its main advantage in our context is *interpretability*. The model estimates hazard ratios for each covariate, providing a clear and interpretable measure of how each factor (e.g., presence of diabetes, age, sex, lifestyle factors) alters the risk of death relative to a baseline or 'standard' individual. This aligns with the underwriting analogy, where insurers adjust standard mortality expectations based on individual characteristics.

The main limitation of the Cox model is that it assumes that all attributes have independent influence on the mortality probability. We overcome this limitation by introducing separate attributes for pairwise interactions of the diseases.

6.1.2 Cox proportional hazard model description

This model type evaluates how the time to an observed event varies based on a set of explanatory variables. Here the event of interest is death. The set of explanatory variables is the input information encoded as the vector $z = (z_0, z_1, \ldots, z_m)$, where m is the number of attributes, and z_i is a value of attribute i, with $i = 0, 1 \ldots, m$.

When considering the time until the event of death is observed, the cumulative probability of death over a given period of time (t) from a given age (x) is considered. Let X be a random variable indicating the age (in years) of death. While in the rest of the paper we define age as the number of full years of age at start of the study, see Section 4.3.1, for the purpose of the discussion of this section, by 'age' we mean exact age as a real number – this makes the limits below well-defined. The cumulative distribution function of X, $F_X(x) = P(X \leq x)$, is the probability of death up to and including age x. And so, $S(x) = 1 - F_X(x)$ is the probability of survival up to and including age x. Then the conditional probability of death for an individual, who having attained age x, dies between age x and $x + \Delta x$ is defined as:

$$P_x(\Delta x) = P(x < X < x + \Delta x | X > x) = \frac{F_X(x + \Delta x) - F_X(x)}{1 - F_X(x)}$$

When the time observed beyond age x (Δx) is limited to be infinitesimally small, the probability of death (q_x) becomes the force of mortality (μ_x). The force of mortality is obtained by limiting Δx to 0 and is therefore defined as:

$$\mu(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x | X > x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x (1 - F_X(x))} = \frac{f_X(x)}{S(x)},$$

where $f_X(x) = \frac{d}{dx} F_X(x)$ is the density function of random variable X. Because $f_X(x) = \frac{d}{dx}(1 - S(x)) = -S'(x)$,

$$\mu(x) = \frac{f_X(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{d}{dx} \ln S(x).$$

From the above, the probability of survival from age x to age $x + t$, e.g. $S_x(t)$, is given by:

$$S_x(t) = \exp\left(-\int_x^{x+t} \mu(y) dy \right),$$

and so, the probability of death from age x to $x + t$ is $P_x(t) = 1 - S_x(t)$.

For a Cox Proportional hazard model, the instantaneous probability of the event occurring is termed the 'hazard rate'. Therefore when the event of interest is death, the hazard rate is equivalent to the force of mortality. The hazard rate at time t for an individual with the input information defined in vector z is denoted by $\lambda(t, z)$. For a Cox proportional hazard model, the hazard rate is calculated as:

$$\lambda(t, z) = \lambda_0(t) \cdot \exp\left(\sum_{i=0}^{m} \beta_i z_i \right), \tag{1}$$

where $\lambda_0(t)$ is called the baseline hazard rate, and β_i are Cox coefficients.

For each input information i within vector z, there is a corresponding β_i coefficient which adjusts the baseline hazard rate. The input information represents the risk factors affecting the hazard rate. Therefore, if $\beta_i > 0$ the input information i (or risk factor i) increases the hazard rate, and $\beta_i < 0$ decreases the hazard rate. The absolute value $|\beta_i|$ therefore measures the significance of risk factor i.

In this application of the model:

- z_0 is set to be the non-negative function of age;
- z_i, where i is length of time since diagnosis, is also set to be the non-negative function of time; and
- all other z_i are equal to either 0 or 1, as an indicator of if the risk factor i is present ($z_i = 1$) or absent ($z_i = 0$).

Given the above, (1) simplifies to[8]:

$$\lambda(t,z) = \lambda_0(t)e^{\beta_0 z_0} \cdot \exp\left(\sum_{i=1}^{m} \beta_i z_i\right) = \lambda_0(t)e^{\beta_0 z_0} \cdot \exp\left(\sum_{i\in I} \beta_i\right) = \lambda_0(t)e^{\beta_0 z_0} \prod_{i\in I} k_i, \tag{2}$$

where I is the set of indices i such that $z_i = 1$ and $k_i = e_i^{\beta}$.

The model provides estimates that are used to alter the overall hazard rate based on an individual's input information. For any input information (risk factor) $i = 1,\ldots,m$, a change in the value of z_i from 0 to 1, alters the hazard rate $\lambda(t,z)$ by a factor specific to that risk factor, e.g. k_i. If there is a change to more than one risk factor, say both z_i and z_j from $(z_i,z_j) = (0,0)$ to $(z_i,z_j) = (1,1)$, then this alters the hazard rate $\lambda(t,z)$ by a factor $k_i k_j$. This implies the factors i and j are acting independently. This is a key underlying assumption of the Cox proportional hazard rate model.

For example, where factors i and j indicate the presence or absence of certain diseases then the model provides coefficients to alter the force of mortality based on the disease states. For example, if an individual is diabetic (D) this alters the hazard rate by a factor of k_D and if the individual also has hypertension (H) this alters it by k_H. That is, an individual with both diabetes and hypertension has a hazard rate altered by factor $k_D k_H$. It is important to note that this way of modelling is simplistic and ignores the known dynamic of a 'combination of risk factors effect'. For example, hypertension may increase the probability of death to a greater extent for diabetic individuals than for non-diabetic individuals (Bell, 2008). So far in this model description, this has not been allowed for as the model is solving the coefficient independently.

To take account of such combination effects, the model must solve these factors independently and then also in combination. This requires the introduction of new risk factors to represent any pair of risk factors i and j. This new risk factor indexed as ij is such that $z_{ij} = 1$ if an individual has both risk factors, e.g. if $z_i = z_j = 1$; otherwise $z_{ij} = 0$. Continuing the example for an individual with both diabetes and hypertension, the input information is therefore $z_D = 1$, $z_H = 1$, and now, $z_{DH} = 1$. Then, for an individual with only these two risk factors, the hazard rate formulae (2) becomes:

$$\lambda(t,z) = \lambda_0(t)e^{\beta_0 z_0} k_D k_H k_{DH}. \tag{3}$$

The interpretation of k_{DH} is the coefficient that accounts for the mortality risk arising from the combination of diabetes and hypertension that remains after accounting for diabetes and hypertension individually. Therefore, the factor k_D represents the alteration to baseline hazard rate for diabetes, k_H for hypertension, k_{DH} for the alteration when both are present.

It should be noted that the model does not account for which condition pre-dates the other, which may epidemiologically cause a variation in the force of mortality. For instance, the development of hypertension prior to diabetes may have different clinical implications than the reverse order. However, for the purpose of this research objective of mimicking the information available at the time of underwriting, this simplification is considered appropriate.

In the underwriting context, risk assessment is typically based on a snapshot of available health information at the time of application, without access to detailed longitudinal records or the precise sequence of diagnoses. Therefore, modelling comorbid conditions as coexisting covariates, rather than in a temporally ordered fashion, aligns with the real-world data structure and decision-making process used by insurers.

Moreover, many of the granular subgroups defined by disease order (e.g., diabetes preceding hypertension vs. hypertension preceding diabetes) are relatively small and statistically underpowered, which limits the reliability and interpretability of stratified hazard estimates. Attempting to model these pathways separately could introduce instability or noise into the model, especially in population-based datasets not designed to capture such detail robustly.

6.2 Model build

Models were built iteratively with modelling choices, detailed below with supporting rationale, made along the way. These decisions were made in conjunction with ongoing discussions between Leicester University and DSG. The objective of this working relationship was to ensure the output was fit for purpose for the intended audience and as such, some decisions herein may vary from medical or epidemiological research. Namely, risk factors are taken as the latest observation to the study start even if the risk factor would not remain static thereafter and there is evidence that a change in that risk factor would imply a change in the mortality risk going

[8] For simplicity, we will ignore the length of time since diagnosis risk factor in this discussion

forward. Therefore, it is known that there will be drift between mortality risk classification groups after the study start date. However, this is considered appropriate for this research objective as it mimics the underwriting process; analysing mortality risk based only on the latest observation found in the input information accessible during the underwriting stage.

The model build was undertaken in stages, each with an objective, to reach a final model to fit. The stages are:

- Data partitioning;
- Risk factor transformation;
- Risk factor selection.

Each stage is described in a separate subsection below.

6.2.1 Data partitioning

Sex at birth is an important factor in understanding of mortality risk (Zarulli et al., 2021). Indeed, many risk factors including the age shape of mortality risk vary by sex (Wu et al., 2021). To build more accurate models, the dataset is divided by sex and the model fit for males and females separately. This approach offers several advantages. Stratification allows the model to capture sex-specific hazard functions and avoids the assumption of proportionality across sexes, which may be violated if key risk factors behave differently for men and women. It also enhances interpretability by enabling the direct comparison of risk profiles within each sex group, aligning with clinical evidence that diabetes presents distinct risk trajectories in males and females (Kautzky-Willer et al., 2016).

6.2.2 Risk factor transformation

The following risk factors are modelled to test the statistical significance of the factor on the mortality rate. A summary of risk factor transformation is summarised in Table 70 and the mortality statistics associated with each risk factor is presented in Table 71.

The following risk factor transformation is introduced:

Age Age is calculated as the number of whole years at the study start and assigned the risk factor z_0 in the model. This approach uses exact single-year age rather than broader age bands. This reflects standard practice in life insurance and actuarial modelling, where mortality assumptions are typically constructed on a per-year-of-age basis.

Using single-year age allows for more granular modelling of the age-related variation in mortality, which is known to change markedly even across adjacent years, particularly at older ages. Given the size and richness of the dataset, there is sufficient statistical power to estimate risk at this fine level of resolution without the need to aggregate into coarser categories. This avoids potential loss of information and smoothing over important inflection points in the age–mortality relationship that could influence model accuracy and underwriting decisions. Moreover, insurers routinely rely on single-age mortality tables (NLT), and using a consistent age definition enhances the interpretability and practical relevance of the model outputs for actuarial and underwriting audiences.

In the model, the hazard rate (e.g. force of mortality) depends on age as:

$$\lambda(t,z) \sim e^{\beta_0 z_0},$$

where z_0 is a function of age. To determine z_0, the force of mortality as it varies with age must be assessed. For example, the force of mortality by age for males is presented in Figure 11. It can be seen that the force of mortality is almost linear in logarithmic coordinates. This suggests that $z_0 =$age is already a reasonable model.

However, the fit to data can be improved by assuming that z_0 is a piecewise-linear function of age. For example, the manual splitting of age interval [17-100] into three intervals of [17-27], [28-73], and [74-100] significantly improves the fit to data, see Figure 12. To identify intervals with the best accuracy automatic segmentation is used. The results of this segmentation are presented in Table 72, Figure 13, and Figure 14. Applying the Elbow rule to Figure 13 suggests that only 2 segments should be used. That is, z_0 is given by the formula:

$$z_0 = \begin{cases} k \cdot (\text{age} - z^*) + z^*, & \text{if age} \leq z^* \\ \text{age}, & \text{if age} \geq z^*, \end{cases} \tag{4}$$

where z^* and k are fit to the partitioned data, e.g. for males and females separately. As indicated in Table 72, the optimal break point z^* is 50 for males and 65 for females. The optimal value of k is 0.6510 for males and 0.7335 for females.

Body mass index (BMI) The latest observation of the BMI measurement at the study start is transformed into covariates to improve model simplicity. The ranges employed are in line with those adopted by the NHS (NHS, 2023*b*). However, the groupings are altered to improve model accuracy: specifically, the obese range was subdivided into the 3 classes shown below, as it significantly improved model accuracy. The following risk factor transformation is introduced:

- Underweight: $z_{b1} = 1$ if BMI< 18.5, and $z_{b1} = 0$ otherwise.
- Overweight: $z_{b2} = 1$ if BMI≥ 25.0, and $z_{b2} = 0$ otherwise.
- Obese Class 1: $z_{b3} = 1$ if BMI≥ 30.0, and $z_{b3} = 0$ otherwise.
- Obese Class 2: $z_{b4} = 1$ if BMI≥ 35.0, and $z_{b4} = 0$ otherwise.
- Obese Class 3: $z_{b5} = 1$ if BMI≥ 40.0, and $z_{b5} = 0$ otherwise.

For example, an individual with BMI 32.5 has a BMI which is both ≥ 25.0 and ≥ 30.0 and so is described by a vector $(z_{b1}, z_{b2}, z_{b3}, z_{b4}, z_{b5}) = (0,1,1,0,0)$. Therefore, coordinates are not independent as, for example, if if $z_{b3} = 1$ then $z_{b2} = 1$.

This structure reflects a cumulative threshold-based encoding that mirrors the way underwriters and actuaries typically think about BMI risk categories - each level of obesity builds on the previous. This encoding captures the escalating nature of risk associated with higher BMI thresholds, which is especially useful in underwriting, where risk loading increases progressively across weight categories.

Blood pressure (BP) Measurements for Systolic blood pressure (SBP) and diastolic blood pressure (DBP) are used at the latest observation as at study start. The risk factor transformation introduced is shown below and is considered in line with NHS ranges used (NHS, 2023*a*):

- Low blood pressure: $z_{p1} = 1$ if either SBP≤ 90 or DBP≤ 60. Otherwise $z_{p1} = 0$.
- Pre-hypertension: $z_{p2} = 1$ if either SBP≥ 120 or DBP≥ 80. Otherwise $z_{p2} = 0$.
- High BP (Stage 1): $z_{p3} = 1$ if either SBP≥ 140 or DBP≥ 90. Otherwise $z_{p3} = 0$.
- High BP (Stage 2): $z_{p4} = 1$ if either SBP≥ 160 or DBP≥ 100. Otherwise $z_{p4} = 0$.

In terms of insurance underwriting, blood pressure is routinely assessed at point of application, and risk assessment typically involves identifying whether an individual exceeds particular thresholds that align with these exact categories. Many underwriting manuals apply stepwise premium loadings based on these categories, and reinsurer rating guides often treat elevated BP levels as risk multipliers, particularly in the presence of comorbid conditions like diabetes or obesity.

Cholesterol levels (CL) Measurements for Cholesterol levels (CL) are used at the latest observation as at study start. The risk factor transformation introduced is shown below and is informed/based on widely used clinical guidelines (Superdrug Online Doctor, 2023):

- High CL: $z_{c1} = 1$ if CL> 5 mmol/l. Otherwise $z_{c1} = 0$.
- Very high CL: $z_{c2} = 1$ if CL> 6.5 mmol/l. Otherwise $z_{c2} = 0$.
- Extremely high CL: $z_{c3} = 1$ if CL> 7.8 mmol/l. Otherwise $z_{c3} = 0$.

In the underwriting context, elevated cholesterol is a known risk factor and often triggers further review or a premium loading. While insurers may use more comprehensive lipid panels, total cholesterol remains a common and interpretable metric. The use of the most recent value at the time of underwriting is standard practice, as underwriters make risk assessments based on an individual's current health status, not historical averages or trends. The cumulative structure of the indicators (e.g., someone with CL = 8.0 mmol/L will satisfy all three conditions) reflects the increasing risk load and provides flexibility in the model to capture marginal effects across ranges.

Smoking status Smoking status is based on the categorisation defined at the study start. The risk factor transformation introduced is shown below and is informed by epidemiological and actuarial research demonstrating the differential mortality risks associated with current, former, and never smokers (Doll et al., 2004).

- Ever smoked: $z_{s1} = 1$ if an individual ever smoked. Otherwise $z_{s1} = 0$.
- Current smoker: $z_{s2} = 1$ if an individual is current smoker. Otherwise $z_{s2} = 0$.

That is to say, at the study start, $(z_{s1}, z_{s2}) = (0,0)$ is an individual that never smoked, $(z_{s1}, z_{s2}) = (1,0)$ an individual who is an ex-smoker, and $(z_{s1}, z_{s2}) = (1,1)$ an individual who is a smoker.

This approach is reasonable and aligns well with underwriting practices, where smoking status is a key and typically static data point collected at application. While smoking behavior can change over time, underwriting generally relies on smoking status reported or verified at the point of underwriting, which is assumed to be constant for pricing and risk classification purposes.

Index of multiple deprivation (IMD) The IMD is taken as at the study start. The risk factor is transformed using 4 variables set to either a value of 0 or 1.

- $z_{md1} = 1$ if IMD level is at least 3, and $z_{md1} = 0$ otherwise.
- $z_{md2} = 1$ if IMD level is at least 5, and $z_{md2} = 0$ otherwise.
- $z_{md3} = 1$ if IMD level is at least 7, and $z_{md3} = 0$ otherwise.
- $z_{md4} = 1$ if IMD level is at least 9, and $z_{md4} = 0$ otherwise.

In other words, IMD levels 1 and 2 are coded as $(0,0,0,0)$, IMD 3 and 4 as $(1,0,0,0)$, IMD 5 and 6 as $(1,1,0,0)$, IMD 7 and 8 as $(1,1,1,0)$, and IMD 9 and 10 as $(1,1,1,1)$.

From an underwriting perspective, the IMD is assumed to be a static data point, reflecting the individual's health or risk profile at the point of underwriting (i.e., study start). This is consistent with insurance risk assessment practices where data used for pricing or risk classification is typically fixed at application. Although individuals' health status may change over time, underwriting decisions usually cannot account for such longitudinal variations. Additionally, using this binary stepwise transformation avoids potential issues with sparse data if the IMD were used as a continuous variable or in very granular categories. It also simplifies interpretation for underwriters and actuaries, who can associate specific IMD threshold levels with corresponding risk adjustments.

Diabetes disease status The diabetes disease status is encoded based on diagnoses as at the study start. A later diagnosis is not considered relevant because this approach mimics the real-world underwriting process in insurance. In insurance underwriting, the assessment of mortality risk and premium setting is based on the information available at the time of application, which includes the individual's current health conditions and medical history up to that date. Later developments or diagnoses occurring after the policy issue date are generally not factored into the initial risk classification, as underwriting decisions and pricing are fixed at inception. The transformation of this data item is:

- $z_{D1} = 1$ if an individual has been diagnosed Type 1 diabetes, and $z_{D1} = 0$ otherwise. Similarly,
- $z_{D2} = 1$ if an individual has been diagnosed Type 2 diabetes, and $z_{D2} = 0$ otherwise.

For example, an individual diagnosed with Type 2 diabetes in June 2010 has z_{D1} and z_{D2} of $(0,0)$.

It is considered reasonable to model Type 1 and Type 2 diabetes separately when estimating mortality risk because these conditions differ significantly in their pathophysiology, typical age of onset, progression, complications, and associated mortality patterns. Numerous epidemiological studies have demonstrated statistically significant differences in mortality risk between individuals with Type 1 and Type 2 diabetes, see e.g. (Genuth et al., 2021).

Duration since Diabetes diagnosis This risk factor is only considered for individuals with diabetes. For an individual with Type i diabetes, $i = 1,2$, it measures the duration (in years) since the Type i diabetes was diagnosed. This is computed by the formula '2009 - calendar year of the diagnosis' given the study start of 01/01/2010. For example, an individual diagnosed in June 2009 with Type 2 diabetes has a duration since diagnosis of 2009-2009 =0. Duration therefore increases in whole numbers from 0.

Including duration since diagnosis as a risk factor is reasonable and important for mortality risk modeling because the length of time an individual has lived with diabetes is strongly associated with the progression of complications and overall mortality risk (Wright et al., 2020).

From an underwriting perspective, while duration since diagnosis may not always be explicitly recorded, it is often estimated or indirectly assessed through medical history or reported diagnosis dates. Incorporating this information aligns with underwriting practices that seek to assess not only the presence of a condition but also its chronicity and severity, which influence risk classification and pricing decisions. Assuming duration as a static variable at study start is consistent with underwriting timelines, where risk is assessed based on information available at application. This approach balances model complexity and practical utility, providing a meaningful metric to enhance mortality risk prediction for individuals with diabetes.

Blood glucose (sugar) level (HbA1c) Measurements for blood glucose (sugar) level (HbA1c) are used at the latest observation as at study start. The risk factor is transformed using the ranges defined by guidelines such as those from the American Diabetes Association (ADA) and the World Health Organization (WHO):

- Pre-diabetes: $z_{h1} = 1$ if HbA1c$\geq 6.0\%$. Otherwise $z_{h1} = 0$.
- Diabetes: $z_{h2} = 1$ if HbA1c$\geq 6.5\%$. Otherwise $z_{h2} = 0$.
- Diabetes high: $z_{h3} = 1$ if HbA1c$\geq 7.5\%$. Otherwise $z_{h3} = 0$.
- Diabetes very high: $z_{h4} = 1$ if HbA1c$\geq 9.0\%$. Otherwise $z_{h4} = 0$.

These variables define certain HbA1c ranges. For example, an HbA1c$< 6.0\%$ is defined by $(0,0,0,0)$, an HbA1c between 6.0 and 6.5 by $(1,0,0,0)$, an HbA1c between 6.5 and 7.5 as $(1,1,0,0)$, etc.

From an underwriting perspective, HbA1c is a key biomarker routinely assessed and recorded in medical underwriting questionnaires or clinical tests. Using the latest available value at study start reflects standard underwriting practice, where the most recent medical data available at application is used for risk classification. Although HbA1c can fluctuate, the latest measurement provides a practical snapshot of glycemic control and risk status. The cumulative binary coding captures clinically meaningful cutoffs while maintaining simplicity and interpretability for underwriting decisions. It balances granularity with data availability and aligns well with mortality risk modeling objectives focused on diabetes-related risk stratification.

Disease status For each of the diseases listed in Section 4.1, z_i is set to 1 if disease i has ever been diagnosed at the study start and $z_i = 0$ otherwise. Diagnoses do occur during the study period. To keep vector z constant for each individual, the following strategy is introduced:

- If disease i has never been diagnosed then $z_i = 0$;
- If disease i has been diagnosed before the study start, then $z_i = 1$;
- If disease i has been diagnosed x years after the study start, then we treat the given individual as two individuals: one with $z_i = 0$ for x years, and one with $z_i = 1$ for $10 - x$ years.

Pairwise interaction of risk factors For each pair (i,j) of risk factors discussed above, a new risk factor is introduced $z_{ij} = z_i z_j$. For example, if each z_i and z_j takes values in $\{0,1\}$, then $z_{ij} \in \{0,1\}$, and $z_{ij} = 1$ if and only if $z_i = z_j = 1$.

Higher order interactions are not considered because on balance the working group prioritised having coefficient value outputs which could be easily interpreted.

6.2.3 Risk factor selection process

An iterative model fitting process was employed; before fitting a final model with all risk factors and meaningful pairwise interactions, we build some simplified models with only some of the attributes included. In all models, the coefficients β_i can be computed by maximum likelihood method. An important feature of Cox proportional hazard model is that the baseline hazard rate $\lambda_0(t)$ is not needed to compute β_i, hence they can be computed first. This is used here to allow for risk factor selection.

- Model A: a model without data partitioning and with no diseases and no interaction attributes. All other risk factors described in Section 6.2.2 are included. Risk factor selection is then considered via backward and forward feature selection process. In Model A, the data is not partitioned by sex, this ensures the same set of risk factors are included for males and females. The objective of model A aligns with our desire to understand the influence of 'industry standard' underwriting risk factor (such as age, gender, BMI, etc.) on the mortality risk of individuals from GS. Another objective is to have a simple but reasonable model as a benchmark with which the more complicated models can be compared.
- Model B: a model with data partitioning, allowing for disease status independently but not in combination (all risk factors in 6.2.2 except pairwise disease interactions). As a minimum, the model must include 'industry standard' underwriting risk factor classifications. These risk factors are protected, and are not subject to the selection process. Other risk factors (diseases) are selected via backward and forward feature selection process. The objective of model B is to understand the influence of various comorbidities on the mortality risk of individuals from GS as stand-alone factors, without taking into account interaction. Another objective is to have this model as a benchmark with which our main model (with interactions) can be compared.

The objective of including pairwise interactions of attributes in the model is to capture non-additive effects and synergistic relationships between risk factors—especially among comorbidities—that may jointly influence mortality risk in ways not apparent when considered in isolation. Certain combinations of risk factors (e.g., cardiovascular disease and high BMI) may result in elevated or diminished hazard beyond what would be expected from their individual effects.

Risk factor selection: backward and forward feature selection The objective of this step is to build Model A by selecting meaningful non-desease and non-interaction risk factors, independent of sex, that have significant influence on the mortality of individuals with diabetes.

Model A includes all risk factors listed in section 6.2.2, except the disease status and interaction risk factors. As the data is not partitioned, a risk factor of sex is introduced to account for variation. The model is fit on the GS data. The resulting coefficients are presented Table 76, alongside the standard errors (SE) and p-values. Where the p-value measures the significance of the risk factor and a higher p-value indicates a lower significance. Risk factors with a p-value greater than 0.01 are considered insignificant. Table 76 therefore contains 7 insignificant attributes. Risk factor selection can be reviewed by starting with the full list of 20 risk factors (as above) and removing those considered insignificant via the backward feature selection (BFS) process. The process of backward feature selection (BFS) (or dimensionality reduction) removes variables with the greatest p-value one-by-one til there is at least one statistically insignificant variable. The results are presented in Table 77. The model resulting from this process is presented in the right part of Table 76. Assuming that the baseline hazard rate is given by the exponential model (8) described below, the computed parameters are:

- $a = 0.0552$, $b = -12.3610$ for the full model (before exclusion of insignificant risk factors);
- $a = 0.0553$, $b = -12.3236$ for the reduced model (after exclusion of insignificant risk factors).

Correlation matrix between risk factors of the full model is presented in Table 78. The removed risk factors are strongly correlated with the remaining risk factors.

Table 76 shows some risk factors with negative coefficients, that is, the presence of the corresponding risk factor decreases the baseline force of mortality. Most risk factors with negative coefficients in the model also exhibit a moderate to strong positive correlation with age. This suggests that their apparent protective effect may be confounded by age: these factors become more common in older individuals, but age itself is already a strong predictor of increased mortality. Thus, when age is controlled for in the model, these variables may appear to lower risk simply because they are positively associated with age, not necessarily because they are truly protective. However, the variable CI3 is an exception. Despite having a negligible

correlation with age (0.032), it retains a significant negative coefficient. This finding aligns with the 'Cholesterol Paradox' often observed in observational studies of chronic diseases (Wang et al., 2023). In this context, the inverse association between cholesterol and mortality is typically attributed to reverse causality, where low cholesterol levels serve as a marker for underlying frailty, malnutrition, or chronic inflammation, rather than higher cholesterol levels conferring a direct biological benefit.

To understand the meaning of a negative coefficient (see Table 76), a direct estimation of the force of mortality for age bands is calculated using the GS data (we do not have enough data for mortality estimation for each year separately). Results are presented in Tables 79-84. As you can see, in all cases negative coefficients were explained by the supporting data. In other words, the observed negative dependence is really present in the data, it is not an artefact of the model.

Alternatively, forward feature selection (FFS) can be employed; starting with one attribute (e.g. age), and adding the next 'most significant' (measured by p-value) risk factor one-by-one, until there are no more significant risk factors. The results are presented in Table 85 and Table 86.

The model resulting from BFS and FFS are both the same, which provides confidence in the robustness and stability of the selected set of risk factors. This convergence suggests that the final feature set is not overly sensitive to the choice of selection method and likely reflects genuinely important predictors within the data.

Risk factor selection: industry standard risk classification In Model B and the final output models, attributes describing age, IMD, BMI, blood pressure, smoker status, and cholesterol level are classified as 'industry standard' and therefore must be included in the model, irrespective of their significance. There are other risk factors commonly used in the industry that are not included in our models. For example, family medical history, occupation, physical activity levels are often considered in traditional underwriting are omitted due to data availability constraints.

The exclusion of these factors presents both benefits and limitations. On the one hand, the streamlined model may be easier to implement and explain, reducing complexity and potential biases introduced by subjective or self-reported data. On the other hand, it may compromise predictive accuracy, especially for edge cases where lifestyle or family medical history plays a crucial role. It is not possible, at this time, to test the impact however as the data is not recorded within the available data.

Risk factor selection: disease significance considered by sex Model B partitions the data by sex and considers all risk factors from Section 6.2.2, excluding only the disease pairwise interactions.

As explained in Section 6.2.2, diseases are diagnosed during the study period, and, to keep attribute vector z constant for each individual, we treat the given individual as two: one without this disease from the beginning of the study to the date of diagnosis, and another one with this disease from the date of diagnosis. In case of several different diagnosis, an individual record can be split into more than two records. After such split of general sample linked to ONS, we obtained a dataset with $157{,}143$ records.

Table 87 presents the resulting Cox coefficients. There are 12 negative coefficients. Comparison of Cox coefficients for models A and B (e.g. with and without diseases) is presented in Table 88. A notable change is that a cholesterol observations in the 'high' range (i.e. CI) has a significant negative coefficient value in Model B (without disease status) to a significant positive coefficient value in Model C (with diseases). This suggests that the Cholesterol paradox is diminished somewhat after allowing for disease.

Risk factors with a p-value greater than 1% can be considered 'insignificant' but may still be included where these are considered 'industry standard risk classification'. With the objective of minimising 'over-fit' of the data, other insignificant attributes can be removed one-by-one until all the remaining attributes are significant. This process is undertaken on the partitioned data, e.g. for males and females separately. The Cox coefficients for the models fit to the partitioned data are presented in Tables 89 and 90, for male and female respectively. In terms of statistical significance, some notable differences emerge due to sex partitioning and disease status. Several covariates demonstrate significance in one sex but not the other, highlighting underlying biological and epidemiological variations. Specifically, two male-specific diseases were found to be statistically significant for males but not for females. Conversely, seven female-specific diseases are significant only in the female cohort. The observed variation underscores the reasonableness and necessity of sex-specific modelling approaches. Sex-based biological differences (e.g., hormonal influences, genetic expression) and social determinants (e.g., healthcare access and behaviour) contribute to distinct patterns of morbidity and mortality. Incorporating these nuances into survival models enhances both statistical robustness and appropriate understanding of risk.

Risk factor selection: pairwise interactions considered by sex To form models with interactions we consider all individual attributes and all possible pairs of attributes. Initially, we have 59 individual attributes

and $59 \cdot 58 = 1711$ pairs. However, some 'interactions' are duplicates[9], and were therefore removed. We also removed interactions for which there are less than 100 individuals in the dataset. After this, $1{,}616$ attributes remain for males and $1{,}573$ for females.

Fitting Cox coefficients for these many attributes is not possible in one go due to insufficient memory, even on a supercomputer. Therefore, the following procedure is employed:

(i) Start with 'industry standard risk factors', that is, attributes describing age, IMD, BMI, blood pressure, smoker status, and cholesterol level.

(ii) Add one of several other attributes (e.g. diseases attributes or interaction attributes).

(iii) Compute Cox coefficients of the resulting model.

(iv) Remove insignificant attributes, that is, with p-value greater than 0.01. If an 'industry standard risk factor' attribute is insignificant, it can be removed unless it is the last attribute in the corresponding group. That is, age attribute is protected and cannot be removed. Out of 4 IMD attributes, at least one should stay, and so on.

(v) Go to step (ii) and repeat the process, until all attributes are considered.

The result of the described procedure depends on the order we add the attributes in step (ii). To make the procedure reproducible, we must fix the order. To establish the order in which the attributes are added on step (ii), we have used Lasso (least absolute shrinkage and selection operator) method. This method is based on minimizing sum of squares error. It has been originally developed for regression analysis and has been adapted for the Cox model (Tibshirani, 1997). It has been implemented in the glmnet package for R. The main advantage of this method is that it works much faster than the procedure (i)-(v) above, and produces a ranking of the attributes. The disadvantage is that the resulting rating is inaccurate (some significant interactions are ranked low by Lasso and vice versa), and therefore cannot be used alone to build the final model. However, it can be used to form the initial order of attributes before applying the main procedure (i)-(v).

For female and male models, the resulting list of attributes and their Cox coefficients and p-values are presented in Tables 93 and 94, respectively.

6.3 Model fitting

6.3.1 Computing coefficients values

The coefficients β_i in (1) can be computed by maximum likelihood method. We record the order in which we have observed deaths, write down the likelihood function that individuals die in this particular order, and then find the parameters that maximise the likelihood.

6.3.2 Computing the baseline hazard rate: Time homogenous and Exponential models

After computing the Cox coefficients, the next step is computing the base rate $\lambda_0(t)$. Time is measured in years, therefore the study period corresponds to the interval $t \in [2010, 2020)$. We first assume that $\lambda(t)$ is a piece-wise constant function, that is, $\lambda(t) = \lambda_j$ is a constant during year j, $j = 2010, 2011, \ldots, 2019$. We then estimate each λ_j separately from the data.

We next consider one year of data during a specific year j. Then $\lambda(t) = \lambda_j$ is a constant, and $\lambda(t,z)$ in (1) becomes $\lambda(z)$, and (1) can be rewritten as:

$$\lambda(z) = \lambda_j \cdot \exp\left(\sum_{i=0}^{m} \beta_i z_i\right) = \lambda_j B(z), \tag{5}$$

where $B(z) = \exp\left(\sum_{i=0}^{m} \beta_i z_i\right)$.

We will estimate λ_j using method of moments. Because we need to estimate just one parameter, we can use just the first moment – expectation. Let $A = (a_1, \ldots, a_n)$ be the set of individuals that remained alive during the whole year j and let $D = (d_1, \ldots, d_k)$ be the set of individuals who died during the year j. Let X be the random variable indicating the number of deaths in this sample. Then, the expected number of deaths is:

$$E[X] = \sum_{i \in A \cup D} E[X_i],$$

[9] For example, BMI4 attribute is equal to 1 if BMI ≥ 35.0, BMI5 attribute is equal to 1 if BMI ≥ 40.0. The 'interaction' attribute equal to 1 if and only if both BMI4 and BMI5 are 1, but this is just BMI5.

where X_i is the number of deaths for individual i, and the summation is over the whole sample $A \cup D$. As X_i can only take values 0 or 1, the formula becomes:

$$E(X_i) = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = P(X_i = 1) = 1 - \exp(-\lambda_j B(z_i)) \approx \lambda_j B(z^i),$$

where z^i is the vector of covariates for individual i, and the last approximate equality follows from the fact that $B(z^i)$ is typically small and, for small x, $1 - \exp(-x) \approx x$. The generalised formula therefore reads as:

$$E(X) = \sum_{i \in A \cup D} E[X_i] = \sum_{i \in A \cup D} (1 - \exp(-\lambda_j B(z^i))) \approx \lambda_j \sum_{i \in A \cup D} B(z^i).$$

If the observed number of deaths is k, then let us find λ_j from the assumption that the observed number is equal to the expected. Then:

$$\lambda_j \approx k \left(\sum_{i \in A \cup D} B(z^i) \right)^{-1}. \tag{6}$$

In this way we can estimate λ_j based on data from years $j = 2010, 2011, \ldots 2019$.

Time-homogeneous model We next consider models in which $\lambda_0(t)$ is a continuous function. We first consider a time-homogeneous model, which is based on the following assumption

- Consider two individuals A and B with the same set of attributes $z_1, \ldots, z_m$ but different age. Consider two different time moments t_1 and t_2. Assume that the age of A at time t_1 is equal to the age of B at time t_2, and this age is greater than z^* defined in (4). Then the force of mortality of A at t_1 is equal to the force of mortality of B at t_2.

In other words, the assumption states that the probability of death depends on risk factors and current age, but not on the calendar year. For example, if individual A was born in 1940 and individual B in 1945, and these individuals are otherwise identical, then the probability of death of A in 2010 (age 70) is the same as the probability of death of B in 2015 (also age 70).

From (2), the mortality rate of A at t_1 is:

$$\lambda_0(t_1) e^{\beta_0 z_0(A)} \cdot \exp\left(\sum_{i=1}^{m} \beta_i z_i \right),$$

while the mortality rate of B at t_2 is:

$$\lambda_0(t_2) e^{\beta_0 z_0(B)} \cdot \exp\left(\sum_{i=1}^{m} \beta_i z_i \right),$$

where $z_0(A)$ and $z_0(B)$ are the values of attribute z_0 for A and B, respectively.

Hence, the time-homogeneity assumption implies that

$$\lambda_0(t_1) e^{\beta_0 z_0(A)} = \lambda_0(t_2) e^{\beta_0 z_0(B)}$$

or equivalently,

$$\frac{\lambda_0(t_1)}{\lambda_0(t_2)} = e^{\beta_0 (z_0(B) - z_0(A))}.$$

Because the age is greater than z^*, (4) implies that $z_0(A)$ and $z_0(B)$ are the ages of A and B, respectively, at the study start. The difference in age is therefore, $z_0(B) - z_0(A)$. Because the age of A at time t_1 is equal to the age of B at time t_2, this implies that

$$z_0(B) - z_0(A) = t_1 - t_2.$$

Hence,

$$\frac{\lambda_0(t_1)}{\lambda_0(t_2)} = e^{\beta_0 (t_1 - t_2)}.$$

The only function that satisfies this equality for every pair of real numbers t_1, t_2 is the function

$$\lambda_0(t) = e^{\beta_0 t + b} \tag{7}$$

for some parameter b. This parameter can be found to best fit the data λ_j, $j = 2010, \ldots, 2019$ found from (6).

Exponential model The only disadvantage of a time-homogeneous model is that it ignores the improvement of medicine over time. In reality, a 50 year-old individual with cancer has a higher chance to survive in 2020 than in 1970. To incorporate this effect, we can adjust the model as:

$$\lambda_0(t) = e^{(\beta_0 - \epsilon)t + b},$$

where ϵ is the term responsible for medicine improvement. Denoting $a = \beta_0 - \epsilon$, we obtain:

$$\lambda_0(t) = e^{at+b}. \tag{8}$$

Parameters a and b can then be found to best fit the data λ_j, $j = 2010, \ldots, 2019$ found from (6). We will call (8) **exponential** model for baseline hazard rate.

Obviously, we may also shift numeration of years, so that t is the time (measured in years) from study start. Then t ranges in $[0, 10)$ for the study period.

In both the homogeneous model (7) and exponential model (8) the hazard rate (1) has the form:

$$\lambda(t) = A_z e^{\gamma t},$$

where A_z is some constant that depends on vector z of attributes, and γ is equal to β_0 and a in homogeneous and exponential models, respectively.

Then the probability $S(t_0, t_1)$ to survive from time t_0 to time t_1 is:

$$S(t_0, t_1) = \exp\left(-\int_{t_0}^{t_1} \lambda(t)dt\right) = \exp\left(-\int_{t_0}^{t_1} A_z e^{\gamma t} dt\right) = \exp\left(-\frac{A_z}{\gamma}(e^{\gamma t_1} - e^{\gamma t_0})\right). \tag{9}$$

Once we computed the survival probability, we can easily compute any other quantities of interest. For example, the expected survival time at time t_0 is given by:

$$E(t_0) = \int_0^\infty S(t_0, t_0 + t)\, dt = \int_0^\infty \exp\left(-\frac{A_z e^{\gamma t_0}}{\gamma}(e^{\gamma t} - 1)\right) dt. \tag{10}$$

Having fit the model, the survival or death probability for a given set of inputs can be found using formula (9). From this, an estimate of life expectancy can be found using formulae (10).

6.3.3 Models fit by sample

The process of building the final models with all attributes and interactions is undertaken upon the different data samples to ensure that the models capture population-specific risk dynamics and allow for accurate estimation of mortality across distinct subgroups. Stratifying the data in this way enables the identification of both shared and unique predictors of mortality within and between groups, accounting for differences in sex and disease pathology. Namely, the models are fit to the following:

- Mortality models for general population, for males and females separately.
- Mortality models for individuals with Type 1 diabetes, for males and females separately.
- Mortality models for individuals with Type 2 diabetes, for males and females separately.

Therefore, a time homogeneous and exponential model is run separately for male and females on three samples. The 'final models' therefore refers to this set of 12 model runs (e.g. 2 mortality estimation methods, 2 data partitions, 3 samples).

Coefficient values & hazard rate: General Sample In order to compare the risk of diabetics to a 'standard' individual, a probability of death for the general population is required. The mortality models are fit to GS.

HbA1c attributes are not included in the model fit because most individuals in the GS will have no HbA1c measurements given there is no diabetes diagnosis or prior investigations. All other risk factor, as discussed in Section 6.2.2 are included, and data is partitioned by sex. The age attribute z_0 is given by (4) with $z^* = 50$ and $k = 0.6510$ for males and $z^* = 65$ and $k = 0.7335$ for females.

First the Cox coefficient values are solved for, then to the baseline force of mortality, and finally the probability of death can be computed for a given set of input information. For female model, the resulting list of attributes and their Cox coefficients and p-values are presented in Table 93. The baseline hazard rate

$\lambda_0(t)$ is computed, as per Section 6.3.2, via the homogeneous model (7) and exponential model (8). For the homogeneous model (7), the value of parameter b that best fit the data is $4.0430 \cdot 10^{-7}$. For the exponential model, the resulting baseline hazard rate is:

$$\lambda_0(t) = \exp(0.0616t - 14.3738) = A\exp(0.0616t),$$

where $A = e^{-14.3738}$. The probability to survive from time t_0 to time t_1 is then given by (9):

$$\exp\left(-\frac{A_z}{0.0616}(e^{0.0616t_1} - e^{0.0616t_0})\right),$$

where A_z is a constant that depends on the vector z of the attribute values.

For male model, the resulting list of attributes and their Cox coefficients and p-values are presented in Table 94. For the homogeneous model with the baseline hazard rate (7), the value of parameter b that best fit the data is $3.8460 \cdot 10^{-6}$. For the exponential model, the baseline hazard rate (8) is given by:

$$\lambda_0(t) = \exp(0.0442t - 12.1609),$$

while the probability to survive from time t_0 to time t_1 is given by (9):

$$\exp\left(-\frac{A_z}{0.0442}(e^{0.0442t_1} - e^{0.0442t_0})\right).$$

Coefficient values & hazard rate: Type 1 diabetes Next, the DS data is used to study the force of mortality of individuals with Type 1 diabetes. All risk factors in Section 6.2.2 are included. Recall that the age attribute z_0 is given by (4) with $z^* = 50$ and $k = 0.6510$ for males and $z^* = 65$ and $k = 0.7335$ for females.

Models are generated in turn with the risk factor selection, as described in Section 6.2.3. The resulting Cox coefficients for males and females are presented in Table 95 for industry standard risk classifications, Table 96 for industry standard risk classifications and disease status. A comparison of Cox coefficients in these models is presented in Table 97 for female models and in Table 98 for male models.

For the final model including pairwise interactions, Table 99 and 100 provides a list of risk factors, the Cox coefficient value and p-value for females and for males, respectively.

For the homogeneous model (7) baseline hazard rate, the value of parameter b that best fit the data is $2.1956 \cdot 10^{-5}$ for females and $1.5983 \cdot 10^{-5}$ for males. For the exponential model (8), the baseline hazard rate is:

$$\lambda_0(t) = \exp(0.0447t - 8.9282)$$

for females and

$$\lambda_0(t) = \exp(0.0416t - 9.0859)$$

for males. The probability to survive from time t_0 to time t_1 can then be computed by (9).

Coefficient values & hazard rate: Type 2 diabetes Next, the DS data is used to study the force of mortality of individuals with Type 2 diabetes. All risk factors in Section 6.2.2 are included. A similar process as outlined in the section above is undertaken; fit to industry standard risk classifications, then include disease status, then pairwise interactions. The results are presented for male and female model fits in Table 101 and Table 102, and a comparison for female in Table 103 and for male in Table 104. Table 105 and 106 include the resulting list of attributes and the Cox coefficient values and p-values for females and for males, respectively.

For the homogeneous model (7) baseline hazard rate, the value of parameter b that best fit the data is $1.0148 \cdot 10^{-6}$ for females and $2.8654 \cdot 10^{-6}$ for males. For the exponential model (8), the baseline hazard rate is:

$$\lambda_0(t) = \exp(0.0783t - 11.6993)$$

for females and:

$$\lambda_0(t) = \exp(0.0657t - 10.6501)$$

for males. The probability to survive from time t_0 to time t_1 can then be computed by (9).

The coefficients for exponential and homogeneous baseline hazard rates for different models are summarised in Table 107.

6.4 Model testing

6.4.1 Model accuracy estimation: Brier score

A standard way to estimate the accuracy of a prediction model is Brier score. If the model predicts, for each individual i, the probability p_i of some event A_i, then the Brier score is given by:

$$Br = \frac{1}{N}\sum_{i=1}^{N}(p_i - I_i)^2,$$

where N is the number of individuals, and I_i is equal to 1 is the event A_i happened for i and $I_i = 0$ otherwise.

In our case, let A_i be the event that an individual i survived during the whole study period. Then p_i is given by (9) with $t_0 = 0$ and $t_1 = 10$, and

$$I_i = \begin{cases} 1, & \text{if patient } i \text{ survived} \\ 0, & \text{if patient } i \text{ died.} \end{cases} \tag{11}$$

Table 108 summarises the computed Brier score for the final models. In all cases, the time homogeneous version of the model has a slightly bit lower score, which means that it is a slightly better model. The final models fit to DS have a higher score than those fit to GS because of the higher overall mortality rate in that sample.

6.4.2 Coefficient accuracy estimation: confidence intervals

If modelling Cox coefficients $\beta_1,\ldots,\beta_m$ in (1) as normal random variables with given variances and covariances, then $S = \sum_{i=0}^{m}\beta_i z_i$ is also a random variable with normal distribution, whose variance can be computed as:

$$Var[S] = \sum_{i=0}^{m}\sum_{j=0}^{m} z_i z_j Cov(\beta_i,\beta_j).$$

Given this information, a confidence interval $[S_{min},S_{max}]$ for S at any given confidence level can be computed. Substituting values S_{min} and S_{max} instead of $S = \sum_{i=0}^{m}\beta_i z_i$ into (1), the confidence interval $[\lambda_{min}(t,z),\lambda_{max}(t,z)]$ for the mortality rate $\lambda(t,z)$ can be computed. Further, substituting $\lambda_{min}(t,z)$ and $\lambda_{max}(t,z)$ in place of $\lambda(t)$ in (9), the confidence intervals for the survival probabilities can be computed. The resulting confidence intervals depend on the input vector z, and as such there values are computed for each individual. That is to say, some individuals can have a more accurate prediction than others. The resulting confidence intervals are included for the final models described below.

6.4.3 Medical justification review

At all stages of our work (from attribute selection to outcome verification) we consulted with medical specialists from Leicester Diabetes Research Centre.

There is a risk with data partitioning that risk factors where there may not be a prior expectation of reasonable differentiation in the coefficient value by sex are included. The comparisons for Cox coefficients for both models are presented in Tables 91 and 92. As we can see from Table 92, one attribute has different signs for male and female: Angina is risk factor for male but reduce risk of death for female. This divergence, while initially counter-intuitive, may reflect real-world clinical and diagnostic patterns. In males, angina is often an indicator of significant underlying coronary artery disease and is more likely to be associated with acute coronary events, thereby justifying its positive association with mortality risk. In contrast, for females, angina symptoms are frequently less specific, often underdiagnosed or misattributed, and may lead to more conservative management or greater healthcare engagement. As a result, women with diagnosed angina may actually represent a more health-aware, treatment-compliant subgroup, potentially explaining the observed inverse association with mortality. Therefore, although this appears a counter-intuitive differentiation in the coefficient value by sex, we do not think this poses issues with the data or analysis.

To ensure the models are not 'over-fit' to the data, the sex-variation in coefficient values by each risk factor (independently and in combination) is reviewed by medical experts.

Computed coefficients do not always correspond to our intuition. Some diseases that are intuitively dangerous have negative influence on mortality, some attributes that are intuitively important turned out to be insignificant, etc. Some of seemingly 'paradoxical' coefficients, such as, for example, the u-shaped association between BMI and mortality, have been confirmed by our colleagues from Diabetes Research

Centre as meaningful and consistent with the literature. In other cases, after consulting with medical experts, the influence does not appear meaningful or consistent with the literature. For example, this is the case for the influence of Coronary heart disease. These coefficients are the optimal values from a complex multidimensional optimisation problem which is at the current time the best modelling that could be achieved with the data available. As stated in Section 4.1, it is not the objective of this analysis to establish epidemiological causation, but to identify variables that show meaningful correlations with mortality outcomes. The reader should take all results with the benefits and limitations in mind, and be satisfied, based on multiple sources of information that the differentiators in mortality risk are fair and evidence based.

6.5 Model outputs

6.5.1 Summary of model output

Each of the models developed in this project are summarised on a separate tab in the Model Output tool (an Excel document titled 'Model_Output_Tool.xlsx'. There are 12 models in total covering the combination of samples, data partitioning, and mortality estimation method, as summarised in the table below.

Sample	Data partitioning	Mortality Estimation Method
General Sample (GS)	Male	Exponential model
Type 1 Diabetes (D1T)	Female	Homogeneous model (Homo)
Type 2 Diabetes (D1T)		

For example, D2T Male Homo is a model that estimates mortality using the homogeneous model for the baseline force of mortality and is appropriate for an individual who is male with a Type 2 diabetes diagnosis. This model is built and fitted based on the DS sample dataset. Models without 'Homo' in the title use the exponential model (8) for baseline hazard rate.

Within the Model Output Tool, an individual's vector z_i can be entered to the 'Input Information'. Guidance on the format and relevant period for the data is provided in the tool and is in line with Section 6.2.2.

The Model Output Tool transforms the input information as described in Section 6.2.2 and returns the corresponding risk factor or Cox coefficients values for each attributes. The final input information required is a 'Date of application'. From this set of inputs, the Model Output Tool provides the following:

- Covariate influence;
- Initial hazard influence;
- Time independent part of risk;
- Probability of death within within t-th year after application, for $t = 1,2,3,4$;
- Cumulative probability of death within t years after application, for $t = 1,2,3,4$;
- 95% confidence intervals (CIs) for each probability of death above;
- The CI multiplier: a ratio of the upper to lower value in a confidence interval; and
- Squared standard error calculated through covariance matrix: this is a variance of a normal random variable with the given confidence interval.

The probabilities of death output by the model meets the research objective of improving the understanding of mortality risk stratification across both the general and diabetic populations in the UK, with a particular focus on how risk factors vary by sex and disease type. This output can be helpful to the intended audience of life insurers, actuaries, medical underwriters, and policy advisors, as it provides:

- a new source of information based on the UK general compared to diabetic population;
- a new source of information that allows insurers to critique existing information and understanding of the risk; and
- a data-driven basis for more tailored underwriting practices and risk stratification, supporting product innovation and pricing refinement for individuals with diabetes.

Both time homogeneous and time inhomogeneous models are included in the output rather than one final preferred model because each offers distinct insights into the nature of mortality risk, and their inclusion supports a more flexible and comprehensive analysis. Time-homogeneous models assume that the effect of risk factors remains constant over time, offering simplicity and ease of interpretation, which can be useful for baseline comparisons and traditional actuarial applications. In contrast, time-inhomogeneous models takes into account the improvement of medicine over time, capturing changes in risk dynamics such as age-related shifts in disease progression or impact of improved treatment. Commentary from the industry on the Model Outputs and the benefits and limitations is provided within the Discussion in Section 7.

6.6 Benefits and limitations of model outputs

The final models herein have the following benefits and limitations:

Regarding data: Routinely collected data suffers from data quality issues. As detailed in Section 4, care is taken however issues remain including but not limited to:

- The data are limited to England only, and are restricted to the study period.
- As expected, most of deaths happened for reasonably old individuals, say between age 70 and 90. Therefore, we had not enough data for mortality of young individuals. We also have limited data about very old individuals (say, after 90), because not many individuals survived until this age. As a result, our mortality predictions for young individuals and for very old individuals are likely to be less accurate.
- Because we find that CPRD death information is often incorrect, we used only individuals linked to ONS. This reduces the number of individuals. Moreover, linked data belongs to England only, not the whole of UK. More importantly, the analysis in Table 48 shows that most of the diseases (28 out of 41) are not independent from linkage. In conclusion, using only ONS linked individuals significantly reduces the observed death counts in the data and the missingness is not independent of disease; this could introduce both hazard rate and coefficient misestimation risk.
- As detailed in Section 4, missing data is imputed where possible and reasonable however imputed data is not as informative as actual values and can introduce the risk of over confidence by coefficients. Some risk factors such as IMD or Smoking status have a high proportion of inputted data. Some other risk factors had such an extent of missingness that these could not be included for analysis, however these are only considered a limitation where the risk factor usually forms an allowable insurance industry underwriting risk factor (e.g. marital status).
- It should be noted that, many individuals who leave the UK do not inform NHS, and are still registered. The absence of diagnoses and death records appears in the data as if these individuals are all alive and have no new diseases. This is partially corrected by removing older individuals with no activity records during the study period, but in general it is not possible to reliably decide for every individual whether they are absent from observations due to migration or due to being healthy and not attending GP. Under-observation in routinely collected data occurs for various reasons. Section 4 discusses how this is allowed for where a death observation is missing. However other reasons for under observation, such as lax de-registering of individuals, is not accounted for. This may introduce bias to the results in a direction that depends on how the unobserved group differs from the remaining population. Individuals who emigrate without notifying the NHS are disproportionately younger, healthier, and more mobile. Therefore, the bias may be stronger in subgroups associated with higher likelihood of migration—such as younger adults, students, recent migrants to the UK, or certain socioeconomic or ethnic groups—leading to differential underestimation of rates by these risk factors. Consequently, hazard ratios for risk factors correlated with migration could be biased towards the null, while comparisons involving older, less mobile groups may be less affected.

Regarding risk factor treatment and transformation:

- Diabetes classification, as described in Section 4, is not exact. As such, there is potential that the results are affected. Misclassification is more likely to occur in individuals diagnosed at ages near the threshold used for classification (e.g., late-onset Type 1 or early-onset Type 2 diabetes). If some Type 2 cases are misclassified as Type 1, the apparent incidence of Type 1 will be inflated and its association with risk factors more typical of Type 2 (such as obesity, lower physical activity, and certain ethnic backgrounds) may be exaggerated. Conversely, if some Type 1 cases are misclassified as Type 2, the incidence of

Type 2 will be overestimated in younger individuals, potentially diluting associations with classic Type 2 risk factors and biasing hazard ratios towards the null. Overall, the misclassification is likely to attenuate differences between the two Types, particularly for age and BMI.

- For dynamic risk factors such as BMI, blood pressure, or cholesterol level, we took into account only the latest observation, not the whole time series. This is a limitation because single, most recent measurements may not accurately represent an individual's long-term exposure or trajectory. For dynamic risk factors, values can change substantially over time, and the latest observation may reflect changes caused by early disease symptoms, treatment, or temporal lifestyle changes, rather than the individual's typical baseline status. This can introduce reverse causation bias if the latest measurement is taken after disease onset or in response to emerging health issues. It may also attenuate associations if variability over time is ignored, particularly for risk factors that fluctuate with age, season, or life events.
- For each disease (excluding diabetes, where duration was modelled), we utilised a binary attribute indicating whether the disease had ever been diagnosed. Consequently, this definition treats the condition as a permanent risk factor, effectively ignoring the possibility of full recovery. We justify this approach based on the chronic and progressive nature of the high-impact co-morbidities included in the model, where 'recovery' to a baseline risk state is clinically rare. For episodic or acute conditions (such as stroke or cancer), we assume that a historical diagnosis serves as a permanent marker of physiological frailty and elevated recurrence risk. The statistical consequence of this approach is conservative: if a subset of 'fully recovered' individuals is included in the disease group, it would dilute the association between the disease and mortality, potentially underestimating - but not overestimating - the true severity of the condition.
- For comorbidities, we only take into account whether a disease has been diagnosed before the start date or not, and did not take into account the date of the diagnosis. In reality, a year-old diagnosis and a 30 year-old diagnosis of the same disease may have different influence of the mortality probability.

Regarding modelling:

- Much of the existing literature investigates the effect of one specific or several attributes on diabetic mortality risk. Here, a large number of attributes are investigated simultaneously and in combination. However, not allow for temporal aspect means that we are effectively assuming they remain constant over follow-up, which is rarely the case. Many risk factors—such as BMI, smoking status, blood pressure, and treatment use—can change significantly over time, sometimes in response to early disease processes or clinical interventions. Treating them as fixed can miss-classify exposure, dilute observed associations, and introduce reverse causation if post-diagnosis changes are incorrectly interpreted as pre-existing risk factors. As a result, the simultaneous modelling of many attributes may underestimate the true effects of dynamic risk factors and overstate the stability of their relationships with diabetic mortality.
- The Cox model explicitly shows the effect of each risk factor on mortality via the Cox coefficients. Many papers compute Cox coefficients only. Here, the baseline hazard rate is also estimated in explicit form. Therefore, the models herein provide an estimate of mortality risk over the specified time period.
- Most of the risk factors analysed are not independent from each other. The Cox model allows a set of attributes which are useful for estimation of mortality risk to be selected. Risk factors removed were not significant, but this does not mean that these risk factors are irrelevant. Some risk factors removed which are not statistically significant can still be used through correlation with included attributes.
- The significance of risk factors may depend on the order in which they are removed.
- Homogeneous and non-homogeneous models were developed separately for males and females, and separately for each sample; GS, Type 1, and Type 2 diabetes. The number of model types, fit by sample, and extent of risk factors is more comprehensive than most of the literature.
- As discussed in Section 6.1.2, the Cox model assumes risk factors influence the hazard rate independently. The model herein has the benefit of including pairwise interactions. Higher order interactions were not considered to ensure the coefficient value outputs were easier to interpret. Deep learning models which represent mortality rates as complicated functions of the risk factors were not considered suitable, even though these can account for interactions of all orders and so potentially more accurately estimate mortality. This is because, for such models, the exact influence of each attribute is not transparent which was a necessary objective of the research.

Regarding model output

- Modelling stages were reviewed for justification by data and medical expertise. The Cox proportional hazards model provides coefficients that adjust a baseline hazard rate—this baseline is unspecified but represents the hazard function for an individual with all covariates set to zero. What is considered 'standard' in this context is not a predefined life, but rather a relative risk framework where individual hazard rates are expressed as a function of covariate effects layered on top of this baseline.

7 Discussion

To provide a meaningful discussion section, it is most valuable to consider the results within the context of potential application and gain the views of potential users of the research. To achieve this, the discussion section is provided as a general industry discussion followed by an applied industry discussion. Within this discussion section, the key discussion points from the general industry discussion roundtable event are summarised. Finally, the underwriting process is outlined to provide context for the key summary points of the applied industry discussion.

7.1 General Industry Discussion

The IFoA Actuarial Research Centre (ARC) held a roundtable event with industry practitioners on the 10 July 2024 to discuss the commissioned work on an analysis of diabetes. The participants were from insurers, reinsurers, and academia to discuss the outputs of the University of Leicester model ('The Model') from an underwriting and actuarial perspective ('the group'). Overall, the feedback from the group was that the research was considered important and a significant contribution to the insurance industry which provides insights into the complex condition of diabetes for the benefit of consumers. Specific discussions were held in the following areas and the key views summarised below.

7.1.1 The data and data item quality

There was a brief discussion on some of the features of the CPRD dataset that required significant adjustment. Firstly, the death records on CPRD were unreliable and so data was used only when it could be linked to the ONS death records. Secondly, the adjustment at older ages which saw individuals removed above age 80 from 1 October 2010 where the GP record had not changed for a 10-year period ('the study period'). Individuals may move abroad or move GP practice outwith those included in the CPRD network. The group were satisfied with how the data issues were handled but it highlights data quality issues at older ages.

The data field to indicate the Type of diabetes field within the CPRD dataset is not always populated and so uncertain. It was therefore necessary, where this information was not available, to introduce an age related assumption to determine the Type of diabetes. For a diabetes diagnosis below age 28 it was assumed to be Type 1 and equal to or above 28 Type 2. The feedback from group was that this was a reasonable assumption to make. As childhood obesity and consequently earlier ages of developing Type 2 are increasing, this assumption could be challenged in any future iterations of the models. It is accepted this is not a perfect proxy; however it is reasonable enough not to adversely affect The Model.

The HbA1c is a key metric used in underwriting to access the risk of the applicant applying for insurance. For example, an individual diagnosed with diabetes 10-years ago who manages their diabetes well is considered to have lower mortality risk than an individual diagnosed with diabetes 5-years ago who has not been able to retain good management. NICE guidelines do recommend HbA1c is recorded at least annually within GP records; however, the data does not reflect this frequency of recording. Based on insurance data, it was estimated that around 50% of applicants disclosed at the time of underwriting that they did not know their HbA1c but would characterise their diabetes management as 'amazing control', based on underwriter's experience in the group. The Model uses the last HbA1c measure recorded whereas an underwriter may look at the last 3 years of data recorded, if available. It is expected that a recently diagnosed individual, within the last 10 years, may have much better management of their condition, reflecting better medication and insights. Acceptable control of HbA1c for access to insurance is key (between 6% and 8%, percentage of haemogloben in the blood that is glysosylated), particularly for Type 1. The medical definition is different where optimal control is defined as 7% (53 mmol/mol) or less.

7.1.2 The variables of interest, applicability, and results

Ethnicity is an important factor for diabetes. However, it not considered within The Model as the data item is not reliable. Further, insurance companies do not use ethnicity as a factor when setting prices in UK and Europe to comply with the Equality Act 2010. Therefore, although ethnicity is an important risk factor for diabetes, its absence in the modelling does not affect the usability for the intended users of the research.

Underwriters often use family history as a guide to the risk of the person applying for life insurance. Family history was not considered as a variable within The Model. Within GP data records, family history is a free text field and is filled in selectively for individuals with a high risk of cancer. It is more likely to be recorded where there is a family history of stroke and heart disease, than for diabetes. If diagnosed with diabetes, family history would be considered irrelevant for mortality risk; however, family history of stroke and heart disease would still have a relevance when underwriting a risk.

The Group discussed the differences between Type 1 and Type 2 based on outputs from The Model:

- For a Type 2 diabetic, the number of years (duration) since diagnosis is a key consideration as the mortality rates increase each year by about 4% for every year since diagnosis. On the other hand, for Type 1 diabetes duration since diagnosis is less material as there is a much smaller trend from year to year. The Model also shows that if someone has Type 2 diabetes for a long time, the mortality rate is similar to Type 1. The Group thought this was reasonable.

- One underwriter explained that Type 1 diabetes base ratings tend to be higher than Type 2, given that the majority of applications tend to be for individuals in their mid-30s. For Type 1, complications are likely to have an impact at an earlier age given the individual is diagnosed at a much younger age. The only time when Type 2 is higher is if there are poor control and significant complications from comorbidities. Generally, a Type 2 diagnosis tends to occur over the age of 45, so the cause of death is likely to happen in the next 35 years. The risk of developing Type 2 diabetes increases with age. Below the age of 45, it is estimated to be 1.5% of population and above this, increasing from 8% to 20% as age increases[10].

The Model results indicate gender is an important predictor of mortality risk for those living with diabetes. With regard to the diabetes population, overall it is seen that females with Type 1 diabetes have higher mortality than males, in contrast to the general sample. On the other hand, in Type 2 diabetes, males and females are more aligned. Women lose their pre-menopausal cardiovascular protection with age so become closer to the males. Younger females with Type 1 tend to have poorer control.

BMI is considered to be an important variable with relation to diabetes mortality risk. The Model produced the expected U-shaped graph[11]. The exception is female with Type 1 diabetes with a BMI 20-39, which showed a relatively flat relation to expected mortality rate by BMI category. The Group considered that the impact of BMI is likely minimal when other health problems are not present, but The Model does indicate that BMI is important as a factor.

Socio-economic status (SES) is considered to have a material impact on mortality risk, often Index of Multiple Deprivation (IMD) is used as a SES proxy. Within the Model, the IMD is based on the postcode of the individual, or the GP postcode in the over 50% of cases where individual postcode is not present. For the diabetes population, there is a distinct increase in mortality for males with diabetes from affluent to less affluent. This is not as apparent for female with diabetes; the mortality risk is relatively flat by IMD; however, the confidence intervals are wide so care must be exercised in drawing any firm conclusions. The general sample population has a slight increase by IMD. Affluent males do have lower risk, possibly due to better access to health care services and being more aware of health issues. It was discussed why affluent females, on the other hand, do not benefit from affluence which could be due to females being generally better engaged with health services across socio-economic groups.

Co-morbidities are an important indicator of mortality risk for the general and the diabetes populations as can be seen in The Model results. The co-morbidity attributes in The Model are a complex area to understand given the large number of possible co-morbidities and the interactions involved. The group's key discussion points are summarised below:

- Heart failure has the largest impact on the general sample followed by Type 2 and then Type 1. This is due to the relative increase in q_x for the different populations where the general sample has the lowest q_x followed by Type 2 and then Type 1.

[10] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/riskfactorsforprediabetesandundiagnosedtype2diabetesinengland/2013to2019

[11] The U-shaped curve between BMI and mortality risk, including diabetes mortality risk, is widely recognised in medical and epidemiological research (Di Angelantonio et al., 2016).

- CHD does not have an impact on the general sample but does for females with diabetes. For males, there is no impact due to the interaction with hypertension. The model does not include the severity of the comorbidity case. As discussed, this highlights a deficiency in the model that could be improved by pre-defining CHD as a standard condition to be modelled.
- In underwriting, a member of the group at the roundtable event anecdotally stated that about $15-20\%$ of applicants for life insurance declare a second condition, which is lower than what The Model shows. Underwriters receive targeted reports for diabetes which may not pick up other complications and those applicants not declaring; however, the model does utilise other factors such as BMI.
- There was interest in the accumulative impact of co-morbidities. Traditionally, some underwriters treat BMI and diabetes separately, but data does reveal a different impact if combined. The model does allow for pair-wise interactions. But, the model is limited in its data treatment for inter-study period disease diagnoses and co-morbidities.
- Type 2 diabetes and BMI can be reversed but lifestyle changes are difficult and there is an epigenetic legacy with diabetes. Remission is rarely achieved and rare to maintain.
- There was a suggestion of further research looking at the pathway from diagnosis to heart disease event, kidney failure, etc. to inform critical illness cover.

7.1.3 The final models and practical use

In the discussion, both models homogenous and non-homogenous were discussed. The homogenous model was considered more reliable than the non-homogenous model (includes a trend factor over time). The non-homogenous model was based on only 10 years of data which was considered an insufficient period to establish a trend.

7.2 Applied Industry Discussion

To gain the views of potential users of the research in an applied sense, a sub-group of the roundtable discussion were asked to review a set of example applicants with diabetes profiles and share their underwriting outcome for comparison to the final models. Fifteen pre-constructed applicants with diabetes customer profiles ('Example Cases') were agreed with 5 profiles for Type 1 diabetes and 10 for Type 2. The Example Cases were derived based on realistic cases reinsurers underwrite. The sub-group from the roundtable asked to participate in the applied industry feedback were the reinsurers present ('the participants'). This is considered indicative of the industry as direct insurers frequently defer underwriting outcomes ratings to be based on reinsurers' manuals. Therefore, inclusion of all members of the group at the roundtable would have lead to bias; results duplication where one or more reinsurers' underwriting manuals are used.

To provide context for the applied industry discussion below, first the underwriting process is outlined in general.

- Applicants applying for insurance are subject to an underwriting process where many attributes are considered. If the applicant has diabetes, further information is requested such as: Type of diabetes; duration since diagnosis; HbA1c; BMI; and any comorbidities (Table 109).
- The term 'healthy lives' herein refers to applicants without a diabetes diagnosis and no pre-existing health issues from the list of comorbidities conditions used in The Model. These individuals also have: a BMI in the range 20-25, an IMD of 1 (most affluent group), Systolic/Diastolic Blood pressure in the range 90 to 120 and 60 to 80 respectively, and finally, total cholesterol level less than or equal to 5 mmol/L.
- Generally, and in simplified terms, 'healthy lives' would be accepted as a result of the underwriting process without the need for additional underwriting outcomes to be applied. When an applicant is accepted without additional underwriting outcomes, standard rates are considered appropriate i.e. these applicants are accepted on healthy lives expected mortality ('healthy lives q_x').
- Generally, and in simplified terms, applicants with pre-existing health issues such as a diagnosis of diabetes may be accepted as a result of the underwriting process with the additional underwriting outcomes applied. When an applicant is accepted with additional underwriting outcomes, non-standard rates may be considered appropriate i.e. these applicants are accepted with a rating above the healthy

lives expected mortality ('rated lives q_x'). For example, an applicant with diabetes may have a higher premium to compensate for the higher expected mortality.

- Alternative terms are available but as an ultimate decision, if the mortality risk is assessed to be excessive (for example, a rating of $+300\%/400\%$ above healthy lives expected mortality) then the underwriting outcome may be to decline the application for death insurance benefits.

The 'rating' applied above the 'healthy lives q_x' are the underwriting outcomes the participants were asked to share for the Examples Cases. In order to comply with competition law, the IFoA collected the information and retained confidential information:

- Only the perceptions of additional risk posed are shared. This removes the risk of sharing either premiums or health lives rates.
- The underwriting outcomes are anonymised and only shared when at least 2 participants haven't declined.
- Only the final analysis and conclusions are published.

These outcome decisions were aggregated for Example Cases to enable comparison to Model results. Where a participant's underwriting decision would be to accept with a rating applied, the average rating is shown and a range in Table 110. The number of participants who would decline is shown; where the number of declines is high then no average rating is shown (where only 1 reinsurer rating is available).

Results of the University of Leicester model ('The Model') are applied to the Example Cases. Where comorbidities are present, these are input to The Model where there is an option to select it. The output from The Model is based on the following:

- A separate average ratio for males and females of the rated lives q_x over healthy lives q_x over a 10 and 20-year period. The q_x is projected by cohort, e.g. age 27 duration 0 will be age 28 duration 1 a year later. A weighting is applied to the rating by q_x to allow for the impact of q_x increasing with age.
- Only the homogenous model rather than the non-homogeneous model is used. This is appropriate as the homogenous model has no time trend (e.g. a policyholder aged 26 and duration 0 is the same as aged 25 duration 1 which is equivalent to age 24 duration 2, etc.).
- To calculate the rating, a general sample healthy lives (baseline with no loadings by smoker status) is compared to The Model results for a Type 1 or Type 2 diabetes.

To enable a reasonable comparison, insured lives tend to be more affluent than the general population which in turn means that insured lives have on average lower levels of mortality compared to the general population. This occurs due to the well-established correlation between mortality and socio-economic determinants. This is an important factor in explaining inequalities in life expectancy; life expectancy is significantly lower in more deprived areas compared to less deprived areas in England. In order to allow for the discrepancy in expected mortality and therefore proxy the expected mortality of the insured lives sample, the Index of Multiple Deprivation (IMD) of 1 (scale is from 1 to 5 where 1 is the most affluent) is used in the Example Cases. This is not precise as the population general sample used in this research as a benchmark will have different experience compared to an insured population across different ages and genders.

7.2.1 Key summary of comparison

In general An underwriting decision is based on the health of the applicant at the time of application. The resulting rating applied to the standard rate expected mortality is applied throughout the duration of the contract which could be over 25 or more years. In comparison, The Model provides a projected q_x which develops over time[12]. Therefore, these two methods can be compared to assess how the ratio between the rated lives qx and the healthy lives qx changes over the duration of the contract. That is, how the rating varies over time. The outcome of The Model is that ratings tend to reduce over time as the policyholder ages. This is something the industry could also consider as ratings are normally based on the health of the applicant at the time of underwriting and not over the term of the life contract (up to 25 years and beyond).

[12] q_x is the probability of death aged x.

Specifically for Type 1 The average ratings produced by The Model for Type 1 diabetes are within range of the underwriting ratings for cases 1, 2 and 4 where the number of declines is 0 or 1 (Table 110). Male outcomes from The Model tend to be at the lower end of the range compared to the average rating of the participants' underwriting outcomes whereas female outcomes from The Model are much higher above the upper rating.

It is also observed that the participants ratings are highly variable with a wide range, which illustrates that each has a different view of the risk. A varying view of the risk may occur when there is a highly uncertain estimate, or different participants have different sources of information for the basis of the risk perception.

The use of the applicant's gender by participants in the risk assessment of diabetes is not used. This may reflect greater sensitivities on the use of gender created by the EU Gender Directive, even though the guidance indicating its use would be acceptable in relation to the different health outcomes. The Model rating is lower when the comparing the 20-years average to a 10-year average. The rating for q_x reduces as the applicant gets older. The 20-year average model rating is either close or below the lower rating range for males.

For cases 3 and 5, the risk is considered high and is declined by most of the participants. In these cases, we do have higher ratings generated by the model compared to cases 1, 2 and 4. In the Example cases with co-morbidities (cases 3, 4 and 5), the model does not include explicitly Chronic Kidney Disease, Retinopathy, Peripheral vascular disease, or diabetic neuropathy. The reason for this is due to high correlation of these conditions with other co-morbidities, particularly hypertension (Table 55 and 56). A learning for future model iterations would be to design the selection of comorbidities so the model includes these key comorbidities for diabetes as standard. This would provide more refinement than using the broader hypertension condition.

Specifically for Type 2 The Model has much lower ratings for Type 2 diabetes compared to the participants ratings. This is across all Type 2 diabetes Example cases from 6 through to 15.

A plausible explanation for this is due to the underwriting process. Those applicants that are successful and go through the process to be accepted will be healthier, that is with a lower q_x compared to a healthy sample from the general population sample. The healthy sample from the population will include individuals with diabetes and other conditions that have not been diagnosed yet through either a visit to their GP or hospital[13]. The insured population will be healthier as there is a selection process to categorise lives into healthy lives, rated lives and, where risks are too high, would be declined. This difference may explain why a small difference is seen between a healthy general sample and a sample of Type 2 diabetes for each case for The Model outcomes. Also, those diagnosed with diabetes will be receiving treatment and general health checkups which may reduce the mortality[14] against a general sample that includes individuals that have undiagnosed diabetes without treatment.

However, the rating for case 6 should be near a healthy life as this life has no co-morbidities, that is the only diagnosis is Type 2 and with good control of HbA1c. This example case is overweight with a BMI of 37, but this variable within The Model does not lead to a predicted higher mortality at older ages. The average participants rating is around $+100\%$.

Another feature to explain why the resulting ratings in The Model are lower for Type 2 diabetes is to consider the HbA1c metric. The HbA1c measure for all the Example cases is between 5.7 and 10 which reduces the risk of mortality, compared to levels outside this range. The Model does show a typical U-shared curve where below 5.7 and above 10 shows higher mortality. The calibration of the Model could be more refined as it categorises diabetes as follows: ≤ 5.7, 5.7 to 6.5, 6.5 to 7, 7 to 10 and ≥ 10. The category '7 to 10' could be unevenly distributed with good risks closer to 7 and poorer risk closer to 10.

Most participants declined Example Cases 10, 13, 14 and 15. This is predominantly due to the serious comorbidities associated with these cases and the HbA1c readings were toward the higher range. The model here gives a result that is close to a 0 rating. This highlights a deficiency in the model in that the comorbidities are based on a single average factor based on the data using a proportional cox model, so refinements such as the severity of the co-morbidity are not captured. The model is also based on the latest readings and not an average that can provide more information on the variability for certain measures such as HbA1c. Other factors that will have a significant influence on the results is the individual pathway for comorbidities, such as treatments and drugs. These factors are not captured by the model.

[13] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/riskfactorsforprediabetesandundiagnosedtype2diabetesinengland/2013to2019

[14] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/articles/riskfactorsforundiagnosedhighbloodpressureinengland/2015to2019

7.3 Further work since discussion

There were some significant challenges made to the findings:

- Ratings for females are very high compared to males for Type 1 diabetes; and
- Rating for Type 2 that resulted in negative loadings where there are potentially very serious co-morbidities / complications present.

Further work was carried out to address the later point. A simplified model was created, exploring a noted limitation to the model; namely, the ordering of risk factors when fitting the model. In this exploratory work, the order in which co-morbidities are added into the model was altered, and we only included important co-morbidities for individuals with diabetes. The order we proposed was as follows: CHD, Stroke, Chronic Kidney Disease, Peripheral vascular disease, Diabetic neuropathy, Retinopathy, Hyperlipidemia, Amputation and Hypertension.

The purpose of this exploratory modelling was to check the significance of this limitation, as well as the significance of these diseases for mortality prediction without other comorbidities. This additional modelling does suggest that of the co-morbidities tested, all are significant including CHD, Stroke, MVD, Blindness, Amputation and then finally Hypertension.

Therefore, when fitting the model there are significant important model design choices in terms of which co-morbidities are included and the ordering of the risk factors. We would suggest having co-morbidities retained that are most relevant for individuals with diabetes and factored into the model in the order of importance for individuals with diabetes.

This exploratory model is an initial step and further work in the model designed would be required to overcome the challenges made.

7.4 Conclusion

The general industry and applied industry discussions provided a context to the results through the Example cases based on typical cases underwritten where the applicant has diabetes.

The following insights were discussed:

- BMI exhibited a U-shape which is aligned with what underwriters would expect to see.
- IMD exhibited a different pattern for males compared to female, where male has a much stronger increase in mortality rates as IMD goes from affluent to less affluent but on the other hand the pattern for females was not well defined, however, confidence intervals were wide.
- A useful insight provided by the reinsurer loadings is the variation in ratings between the reinsurers based on a different view of the risk.
- It was also interesting that the model loadings depend on the period the rating is applied for, as the rating tends to reduce over time as the population ages.
- The model results for Type 1 were within range from the underwriting results, except female where it is much higher. This was challenged but underwriters typically use gender neutral ratings so a benchmark was not possible.
- Type 2 results were challenged as some of the more serious co-morbidities were not adhering to prior expectations. Further exploratory modelling work suggested the order and the sequence of the key co-morbidities for diabetes is important.
- Overall, the patterns between different factors were aligned with expectations.

The modelling was appreciated in the general discussion but there are many complications such as the quality of the data and how to handle co-morbidities correctly. Building a general model that covers Type 1 and Type 2, and all the different co-morbidities was ambitious. We understand this is the first time that has been done. The general feedback was that creating a model based on CPRD data was worthwhile as an initial approach that can be improved in future based on the feedback from practitioners. This research therefore provides a useful discussion point and an additional data-driven source of information for the intended audience but it is not yet a sole source of the understanding of diabetes mortality risk.

Acknowledgements

Steering Group

The project Steering Group was to ensure that the project delivered research outputs that are of a high quality and in line with expectations, producing findings that are relevant to the actuarial community, industry, and other key stakeholder groups. Ian Catchpole chaired this group (initially Nicky Draper was a chair).

The project Steering Group included representatives from partner organisations. Independent academic guidance was provided by Professor Les Mayhew from Bayes Business School, City St George's, University of London.

Technical Subgroup

Given the complex nature of this research, a Technical Sub-Group was set up to provide additional oversight and input. Members of this group included: Scott Reid, Wui Hua (John) Ng, Mei Sum Chan, Roshan Tajapra, Han Yan, Kishan Bakrania, Jon Lambert and Chris Bagnall.

Industry Discussion

The roundtable event held on the 10 July 2025 was conducted under Chatham House Rules. This was to encourage openness in the discussion. The insurance/reinsurance companies represented were as follows: Zurich, L&G, SCOR, Partner Re, RGA, Hannover Re, Gen Re, Munich Re and Pacific Life Re. Swiss Re provided written feedback post event. Academic institutions represented included: University of Leicester and Bayes Business School. Nicky Draper from Crystallise Ltd attended given her medical expertise; previously she was Co-Chair of the Diabetes working party.

The industry feedback and challenge were extremely valuable, and we would like to thank those who took part of this event. Section 7 covers details of this General Industry discussion and the Applied Industry discussion based on Example cases post roundtable event.

Leicester Diabetes Research Centre

At all stages of our work (from attribute selection to outcome verification), we consulted specialists from the Leicester Diabetes Research Centre, in particular Dr. Joseph Henson, who provided substantial guidance throughout the project.

The Shiny App

We thank Maros Bobulsky for developing (together with the fourth author) the Shiny App, to make the outputs more accessible to a wider audience.

Independent peer reviewer

Josephine Robertson provided independent peer review and edited/re-drafted many of the sections. This work was invaluable to challenge and increase the quality of the overall paper. We also thank Pacific Life Re and Gen Re, who provided extensive feedback on the paper and model.

Data used

This study is based in part on data from the Clinical Practice Research Datalink (CPRD) obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author(s) alone[15]. We also used linked Hospital Episode Statistics (HES) data and Office for National Statistics (ONS) mortality data.

References

Ali, S., Stone, M., Peters, J., Davies, M. and Khunti, K. (2006), 'The prevalence of co-morbid depression in adults with type 2 diabetes: a systematic review and meta-analysis', *Diabetic medicine* **23**(11), 1165–1173.

Association, A. D. et al. (2017), 'American diabetes association standards of medical care in diabetes–2017', *Diabetes care* **40**(Suppl. 1), S1.

Atlas, D. et al. (2015), 'International diabetes federation', *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation* .

Baliunas, D. O., Taylor, B. J., Irving, H., Roerecke, M., Patra, J., Mohapatra, S. and Rehm, J. (2009), 'Alcohol as a risk factor for type 2 diabetes: a systematic review and meta-analysis', *Diabetes care* **32**(11), 2123–2132.

Bell, D. S. (2008), 'Hypertension and diabetes—a toxic combination', *Endocrine Practice* **14**(8), 1031–1039.

Bello-Chavolla, O. Y., Antonio-Villa, N. E., Vargas-Vázquez, A., Ávila-Funes, J. A. and Aguilar-Salinas, C. A. (2019), 'Pathophysiological mechanisms linking type 2 diabetes and dementia: review of evidence from clinical, translational and epidemiological research', *Current diabetes reviews* **15**(6), 456–470.

Bertoni, A. G., Krop, J. S., Anderson, G. F. and Brancati, F. L. (2002), 'Diabetes-related morbidity and mortality in a national sample of us elders', *Diabetes care* **25**(3), 471–475.

Boscari, F. and Avogaro, A. (2021), 'Current treatment options and challenges in patients with Type 1 diabetes: Pharmacological, technical advances and future perspectives', *Reviews in Endocrine and Metabolic Disorders* **22**(2), 217–240.

Chatterjee, S., Khunti, K. and Davies, M. J. (2017), 'Type 2 diabetes', *The Lancet* **389**(10085), 2239–2251.

Chee, Y. J. and Dalan, R. (2024), 'Novel therapeutics for Type 2 diabetes mellitus - a look at the past decade and a glimpse into the future', *Biomedicines* **12**(7), 1386.

Chiriacò, M., Pateras, K., Virdis, A., Charakida, M., Kyriakopoulou, D., Nannipieri, M., Emdin, M., Tsioufis, K., Taddei, S., Masi, S. et al. (2019), 'Association between blood pressure variability, cardiovascular disease and mortality in type 2 diabetes: A systematic review and meta-analysis', *Diabetes, Obesity and Metabolism* **21**(12), 2587–2598.

Cho, N., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., Ohlrogge, A. and Malanda, B. (2018), 'Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045', *Diabetes research and clinical practice* **138**, 271–281.

Collaboration, E. R. F. et al. (2010), 'Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies', *The Lancet* **375**(9733), 2215–2222.

Constantino, M. I., Molyneaux, L., Limacher-Gisler, F., Al-Saeed, A., Luo, C., Wu, T., Twigg, S. M., Yue, D. K. and Wong, J. (2013), 'Long-term complications and mortality in young-onset diabetes: type 2 diabetes is more hazardous and lethal than type 1 diabetes', *Diabetes care* **36**(12), 3863–3869.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

de Miguel-Yanes, J. M., Shrader, P., Pencina, M. J., Fox, C. S., Manning, A. K., Grant, R. W., Dupuis, J., Florez, J. C., D'Agostino, R. B., Cupples, L. A. et al. (2011), 'Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms', *Diabetes care* **34**(1), 121–125.

DeFronzo, R. A. (2009), 'From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus', *Clinical Diabetology* **10**(3), 101–128.

Di Angelantonio, E., Bhupathiraju, S. N., Wormser, D., Gao, P., Kaptoge, S., De Gonzalez, A. B., Cairns, B. J., Huxley, R., Jackson, C. L., Joshy, G. et al. (2016), 'Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents', *The lancet* **388**(10046), 776–786.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004), 'Mortality in relation to smoking: 50 years' observations on male british doctors', *Bmj* **328**(7455), 1519.

ElSayed, N. A., Aleppo, G., Bannuru, R. R., Bruemmer, D., Collins, B. S., Ekhlaspour, L., Gaglia, J. L., Hilliard, M. E., Johnson, E. L., Khunti, K. et al. (2024), 'Diagnosis and classification of diabetes: Standards of care in diabetes—2024.', *Diabetes Care* **47**.

Elwen, F. R., Huskinson, A., Clapham, L., Bottomley, M. J., Heller, S. R., James, C., Abbas, A., Baxter, P. and Ajjan, R. A. (2015), 'An observational study of patient characteristics and mortality following hypoglycemia in the community', *BMJ Open Diabetes Research and Care* **3**(1), e000094.

Evans, J. M., Newton, R. W., Ruta, D. A., MacDonald, T. M. and Morris, A. D. (2000), 'Socio-economic status, obesity and prevalence of type 1 and type 2 diabetes mellitus', *Diabetic Medicine* **17**(6), 478–480.

Gallagher, A. M., Dedman, D., Padmanabhan, S., Leufkens, H. G. and de Vries, F. (2019), 'The accuracy of date of death recording in the c linical p ractice r esearch d atalink gold database in e ngland compared with the o ffice for n ational s tatistics death registrations', *Pharmacoepidemiology and drug safety* **28**(5), 563–569.

Genuth, S. M., Palmer, J. P. and Nathan, D. M. (2021), 'Classification and diagnosis of diabetes'.

Goff, L. M. (2019), 'Ethnicity and Type 2 diabetes in the UK', *Diabetic Medicine* **36**(8), 927–938.

Golovenkin, S. E., Bac, J., Chervov, A., Mirkes, E. M., Orlova, Y. V., Barillot, E., Gorban, A. N. and Zinovyev, A. (2020), 'Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data', *GigaScience* **9**(11), giaa128.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., Van Staa, T. and Smeeth, L. (2015), 'Data resource profile: clinical practice research datalink (cprd)', *International journal of epidemiology* **44**(3), 827–836.

Hussain, A., Bhowmik, B. and do Vale Moreira, N. C. (2020), 'Covid-19 and diabetes: Knowledge in progress', *Diabetes research and clinical practice* p. 108142.

Huxley, R. R., Peters, S. A., Mishra, G. D. and Woodward, M. (2015), 'Risk of all-cause mortality and vascular events in women versus men with type 1 diabetes: a systematic review and meta-analysis', *The lancet Diabetes & endocrinology* **3**(3), 198–206.

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J. and Brunak, S. (2014), 'Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients', *Nature communications* **5**(1), 1–10.

Joob, B. and Wiwanitkit, V. (2018), 'Very low HbA1C, is it a problem?', *Iranian journal of pathology* **13**(3), 379.

Kannel, W. B., Hjortland, M. and Castelli, W. P. (1974), 'Role of diabetes in congestive heart failure: the Framingham study', *The American journal of cardiology* **34**(1), 29–34.

Kattah, L., Gómez, A., Gutiérrez, S., Puerto, K., Moreno-Pallares, E. D., Jaramillo, A. and Mendivil, C. O. (2019), 'Hypercholesterolemia due to lipoprotein x: case report and thematic review', *Clinical Medicine Insights: Endocrinology and Diabetes* **12**, 1179551419878687.

Kautzky-Willer, A., Harreiter, J. and Pacini, G. (2016), 'Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus', *Endocrine reviews* **37**(3), 278–316.

Kautzky-Willer, A., Leutner, M. and Harreiter, J. (2023), 'Sex differences in type 2 diabetes', *Diabetologia* **66**(6), 986–1002.

Leung, A. A., Eurich, D. T., Lamb, D. A., Majumdar, S. R., Johnson, J. A., Blackburn, D. F. and McAlister, F. A. (2009), 'Risk of heart failure in patients with recent-onset type 2 diabetes: population-based cohort study', *Journal of cardiac failure* **15**(2), 152–157.

Mauvais-Jarvis, F. (2018), 'Gender differences in glucose homeostasis and diabetes', *Physiology & behavior* **187**, 20–23.

NHS (2023*a*), 'Blood pressure test', https://www.nhs.uk/tests-and-treatments/blood-pressure-test/. Accessed: 2025-08-09.

NHS (2023*b*), 'Obesity', https://www.nhs.uk/conditions/obesity/. Accessed: 2025-08-09.

Nicholson, B., Aveyard, P., Bankhead, C., Hamilton, W., Hobbs, F. and Lay-Flurrie, S. (2019), 'Determinants and extent of weight recording in uk primary care: an analysis of 5 million adults' electronic health records from 2000 to 2017', *BMC medicine* **17**(1), 222.

Office for National Statistics (2025), 'National life tables – life expectancy in the UK: 2021–2023', Statistical bulletin, Office for National Statistics.
URL: *https://www.ons.gov.uk/releases/nationallifetableslifeexpectancyintheuk20212023*

Ohkuma, T., Komorita, Y., Peters, S. A. and Woodward, M. (2019), 'Diabetes as a risk factor for heart failure in women and men: a systematic review and meta-analysis of 47 cohorts including 12 million individuals', *Diabetologia* **62**(9), 1550–1560.

Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H. et al. (2023), 'Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the global burden of disease study 2021', *The Lancet* .

Pal, R. and Bhadada, S. K. (2020), 'Should anti-diabetic medications be reconsidered amid covid-19 pandemic?', *Diabetes research and clinical practice* **163**.

Patterson, C. C., Karuranga, S., Salpea, P., Saeedi, P., Dahlquist, G., Soltesz, G. and Ogle, G. D. (2019), 'Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the international diabetes federation diabetes atlas', *Diabetes research and clinical practice* **157**, 107842.

Peters, S. A., Huxley, R. R. and Woodward, M. (2014), 'Diabetes as risk factor for incident coronary heart disease in women compared with men: a systematic review and meta-analysis of 64 cohorts including 858,507 individuals and 28,203 coronary events', *Diabetologia* **57**(8), 1542–1551.

Petznick, A. (2011), 'Insulin management of type 2 diabetes mellitus', *American Family Physician* **84**(2), 183–190.

Reid, S., Oliver, N., Bagnall, C., Catchpole, I., Chadwick, P., Lambert, J., Li, J., Schneider, M., Tajapra, R., Yan, H. et al. (2023), 'An analysis of diabetes mortality and morbidity risk'.

Riley, L. and Cowan, M. (2014), 'Noncommunicable diseases country profiles 2014', *Geneva: World Health Organization* .

Rusanov, A., Weiskopf, N. G., Wang, S. and Weng, C. (2014), 'Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research', *BMC medical informatics and decision making* **14**(1), 51.

Seferović, P. M., Petrie, M. C., Filippatos, G. S., Anker, S. D., Rosano, G., Bauersachs, J., Paulus, W. J., Komajda, M., Cosentino, F., De Boer, R. A. et al. (2018), 'Type 2 diabetes mellitus and heart failure: a position statement from the heart failure association of the european society of cardiology', *European journal of heart failure* **20**(5), 853–872.

Shrank, W. H., Patrick, A. R. and Alan Brookhart, M. (2011), 'Healthy user and related biases in observational studies of preventive interventions: a primer for physicians', *Journal of general internal medicine* **26**(5), 546–550.

Song, T., Jia, Y., Li, Z., Wang, F., Ren, L. and Chen, S. (2021), 'Effects of liraglutide on nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus: A systematic review and meta-analysis', *Diabetes Therapy* pp. 1–15.

Stamler, J., Vaccaro, O., Neaton, J. D., Wentworth, D., Group, M. R. F. I. T. R. et al. (1993), 'Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial', *Diabetes care* **16**(2), 434–444.

Superdrug Online Doctor (2023), 'Cholesterol levels', https://onlinedoctor.superdrug.com/cholesterol-levels.html. Accessed: 2025-08-09.

Tancredi, M., Rosengren, A., Svensson, A.-M., Kosiborod, M., Pivodic, A., Gudbjörnsdottir, S., Wedel, H., Clements, M., Dahlqvist, S. and Lind, M. (2015), 'Excess mortality among persons with type 2 diabetes', *New England Journal of Medicine* **373**(18), 1720–1732.

Tate, A. R., Dungey, S., Glew, S., Beloff, N., Williams, R. and Williams, T. (2017), 'Quality of recording of diabetes in the uk: how does the gp's method of coding clinical data affect incidence estimates? cross-sectional study using the cprd database', *BMJ open* **7**(1), e012905.

Thomas, N. J., Jones, S. E., Weedon, M. N., Shields, B. M., Oram, R. A. and Hattersley, A. T. (2018), 'Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional, genetically stratified survival analysis from uk biobank', *The lancet Diabetes & endocrinology* **6**(2), 122–129.

Tibshirani, R. (1997), 'The lasso method for variable selection in the Cox model', *Statistics in medicine* **16**(4), 385–395.

Wahid, A., Manek, N., Nichols, M., Kelly, P., Foster, C., Webster, P., Kaur, A., Friedemann Smith, C., Wilkins, E., Rayner, M. et al. (2016), 'Quantifying the association between physical activity and cardiovascular disease and diabetes: a systematic review and meta-analysis', *Journal of the American Heart Association* **5**(9), e002495.

Wang, T.-Y., Chang, W.-L., Wei, C.-Y., Liu, C.-H., Tzeng, R.-C. and Chiu, P.-Y. (2023), 'Cholesterol paradox in older people with type 2 diabetes mellitus regardless of lipid-lowering drug use: a cross-sectional cohort study', *Nutrients* **15**(14), 3270.

Weykamp, C. (2013), 'Hba1c: a review of analytical and clinical aspects', *Annals of laboratory medicine* **33**(6), 393.

Wilmot, E. G., Edwardson, C. L., Achana, F. A., Davies, M. J., Gorely, T., Gray, L. J., Khunti, K., Yates, T. and Biddle, S. J. (2012), 'Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis', *Diabetologia* **55**(11), 2895–2905.

Wright, A. K., Suarez-Ortegon, M. F., Read, S. H., Kontopantelis, E., Buchan, I., Emsley, R., Sattar, N., Ashcroft, D. M., Wild, S. H. and Rutter, M. K. (2020), 'Risk factor control and cardiovascular event risk in people with type 2 diabetes in primary and secondary prevention settings', *Circulation* **142**(20), 1925–1936.

Wu, Y.-T., Niubo, A. S., Daskalopoulou, C., Moreno-Agostino, D., Stefler, D., Bobak, M., Oram, S., Prince, M. and Prina, M. (2021), 'Sex differences in mortality: results from a population-based study of 12 longitudinal cohorts', *Cmaj* **193**(11), E361–E370.

Zarulli, V., Kashnitsky, I. and Vaupel, J. W. (2021), 'Death rates at specific life stages mold the sex gap in life expectancy', *Proceedings of the National Academy of Sciences* **118**(20), e2010588118.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., Bixby, H., Cowan, M. J., Ali, M. K., Taddei, C. et al. (2016), 'Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants', *The Lancet* **387**(10027), 1513–1530.

Institute and Faculty of Actuaries

London

1-3 Staple Inn Hall · High Holborn · London · WC1V 7QJ
Tel: +44 (0) 20 7632 2100 · Fax: +44 (0) 20 7632 2111

Edinburgh

Space · 1 Lochrin Square · 92-94 Fountainbridge · Edinburgh · EH3 9QA
Tel: +44 (0) 20 7632 2100

Oxford

1st Floor · Park Central · 40/41 Park End Street · Oxford · OX1 1JD
Tel: +44 (0) 1865 268 200 · Fax: +44 (0) 1865 268 211

Beijing

Level 14 · China World Office · No.1 Jianguomenwai Avenue · Chaoyang District · Beijing, China 100004
Tel: + +86 (10) 6535 0248

Hong Kong

1803 Tower One · Lippo Centre · 89 Queensway · Hong Kong
Tel: +11 (0) 852 2147 9418

Singapore

5 Shenton Way · UIC Building · #10-01 · Singapore · 068808
Tel: +65 8778 1784

www.actuaries.org.uk