The impact of type 2 diabetes mellitus on longevity using the United Kingdom primary care electronic health records

Njabulo Ncube

Doctor of Philosophy

March 2022

United Kingdom (UK)

The impact of type 2 diabetes mellitus on longevity using the United Kingdom primary care electronic health records

Njabulo Ncube

Doctor of Philosophy

School of Computing Sciences

University of East Anglia

United Kingdom (UK)

March 2022

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

Previous researchers have mainly focused on pharmacosurveillance outcomes on type 2 diabetes mellitus (T2DM) patients and there is limited literature on studies that the impact of T2DM on longevity compared to non-diabetics in the United Kingdom (UK). Several of these pharmacosurveillance studies were clinical trials which with fewer selected recruited patients. Data collected from routine electronic health records (EHR) could provide insights in longevity in the general population as opposed to selected people.

The primary aims of this study were to investigate how incidence of T2DM affects longevity in the residents of the UK adjusting for several risk factors at entry into the study. This was followed by translating the resultant model into a life expectancy model for comparisons of survival prospect between people with and without T2DM.

Medical records from 2000 to 2016, inclusive, from general practice (GP) contributing to The Health Improvement Network (THIN) database were used to develop three specific models: two performed to estimate the hazard of all-cause mortality associated with T2DM taking into account age at diagnosis (grouped and continuous age) using time-to-event as time-scale and the third to calculate the life expectancy with age as time-scale. The models were multilevel Gompertz-double-Cox with frailty regressions adjusting for socio-demographics, comorbidities and lifestyle factors. Accurately estimated life expectancies could inform future medical management by clinicians and financial planning by individuals, actuaries, insurance stakeholders and government on social security, such as retirement and life insurance.

The research found that the hazards associated with T2DM were reduced than findings in previous studies. The years of life lost to a person in medium deprived area due to T2DM after adjusted for birth cohort and age at diagnosis was between 0.1 and 6 years.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate ‘take down’ action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Contents

List of Tables

List of Figures

List of Publications

- Ncube, N., Kulinskaya, E., Steel, N. and Pchejetski, D. (2022). On the survival of individuals diagnosed with type 2 diabetes mellitus in the United Kingdom: a retrospective matched cohort study. *Diabetes Epidemiology and Management.* https://doi.org/10.1016/j.deman.2022.100065.

- Begun, A., Kulinskaya, E. and Ncube, N. A double-Cox model for non-proportional hazards survival analysis with frailty. *Statistics in Medicine. Resubmitted after minor revision.*

Acknowledgement

I would like to thank my supervisors Professor Elena Kulinskaya, Professor Nicholas Steel, Nigel Wright and Dr. Dmitry Pshezhetskiy for their time and support. In particular, I would like to specifically thank Elena for her unwavering guidance in building me up to become an independent academic researcher and original peer-reviewed article author through her vast knowledge and experience she was willing to share at all times. I also would like to thank Nicholas and Dmitry for their medical expertise and guidance in presenting my work in a way that can be easily understood by medical professionals and guidance in original peer-reviewed article writing. I also would like to extend my appreciation to Mr Nigel Wright in supporting my work in conferences and workshops.

I would love to thank my wife Primrose Kafanyota and my siblings in a special way for the prayers and support they gave me throughout my studies. Finally, I would like to thank my late parents for their love, especially my late father Jotham Ncube for helping me to make this study possible and his ever loving encouragement for me to pursue this research despite his failing health.

Chapter 1

Introduction

1.1 Background

The population's longevity prospects and associated risk have been of interest to numerous businesses and individuals who have increasingly become aware of the related exposure and the need to devise mitigation strategies (CIPR-NAIC, 2018). However, in recent years, an increase in number of deaths caused by non-communicable diseases has posed another challenge to viability of policy funds and state expenditure (WHO, 2017c). According to The Centre for Insurance Policy and Research of the National Association of Insurance Commissioners (CIPR-NAIC), longevity risk is defined as the risk that expectations or premium assumptions are exceeded by the actual survival rates and life expectancy which will result in more retirement cash flow needs than anticipated. Survival models can be used to estimate longevity/life expectancy taking into account all available relevant risk factors.

Life expectancy in several countries has been on the increase until 2010. Thereafter, there has been a deceleration in its growth rate. This may be due to the increase in the number of deaths caused by non-communicable diseases in addition to other causes. In total, 7 out of the ten most causes of death in 2019, were non-communicable diseases (WHO, 2017c). The type 2 diabetes mellitus (T2DM) was ranked 9th and 6th cause of

death in 2010 and 2015, respectively. In 2019, DM was again ranked 9th among the top ten causes of deaths, globally. About 90% of DM patients have T2DM).

ACCORD Study Group (2008) defined T2DM as a metabolic disease that is diagnosed on the basis of sustained hyperglycaemia. T2DM patients are at risk of several further health problems which include cardiovascular disease (CVD), chronic kidney disease (CKD), retinopathy and peripheral vascular disease (PVD). According to World Health Organisation (WHO) statistics, the number of T2DM deaths has been on the increase. It is therefore important to assess the impact of T2DM on both longevity and morbidity risks.

The European Commission (EC), in insurance sense, defined morbidity risk as a change of value caused by the actual disability and illness rates of the persons insured deviating from the ones expected (EC, 2007). Disability and illness both affect the individuals, societies and economies. For example, an individual may involuntarily lose a regular income and, possibly, privacy. The carers also get affected as the increase in morbidity, disability and illness of the cared for person may lead to poor work and personal life balance of a carer. Which in turn may pose a health risk to the health to carers. In addition, increases in tax revenues may be imposed as a reform to cater for social protection expenses.

In order to assess the impact of chronic diseases on longevity and morbidity risk, this research studied T2DM patients in the United Kingdom (UK)'s The Health Improvement Network (THIN) primary care database who got diagnosed with T2DM from 1984 to 2016, and met entry qualifications specified in Section 5.2.

1.2 Aims and Objectives

1.2.1 Aims

1. To analyse the impact of T2DM on longevity risk using recent robust survival modelling techniques and devise an actuarial application of the model.
2. To analyse the impact of T2DM on selected morbidity risks.
3. To develop a robust, reliable and relevant R-package that implements the resultant actuarial model.

1.2.2 Objectives

1. To carry out a literature review on the T2DM, its causes, diagnosis, treatment, management, complications, statistics and survival.
2. To perform a survival analysis on the impact of T2DM using THIN database and extended Cox model adjusting for all available relevant covariates.
3. To perform unadjusted HRs of selected morbidity risks and produce Kaplan-Meier (KM) plots associated with T2DM.
4. To determine the covariates associated with mortality and morbidity among T2DM patients through backward elimination.
5. To impute missing values using multiple imputation technique and further analysing the impact of T2DM on longevity and morbidity.
6. To translate the statistical models into actuarial models and applications.
7. To develop an R-package for actuarial and statistical use.
8. To test and debug the R-package for better functionality and to document the package

9. To estimate the impact of T2DM on the incidence of selected co-morbidities.

1.2.3 Rationale

Life expectancy has been on the increase until 2010. Thereafter, it has been near constant, especially for the elderly according to WHO. In the same period, the number of non-communicable disease deaths in both developed and developing countries has been increasing. Of particular interest is Diabetes mellitus DM which became one of the top ten causes-of-death in the world. Understanding the change in life expectancy or longevity risk and morbidity risk as a result of T2DM through robust survival techniques and models which take into account all available risk factors, will help not just insurance companies in pricing their products and better hedging against the risks but as a driving tool in improving socio-economic policies by governments and all stakeholders.

In addition, there has been limited literature on survival prospects of T2DM patients compared to non-diabetics. Most existing literature were pharmacosurveillance studies and the relevant few comparable studies only adjusted for a limited number of confounding variables. Developing a web-based R-package would help all stakeholders including non computer programmers. Actuaries and other related professions would use the package to better understand and improve their products' pricing, reserves projections, statutory and technical profits among other related insurance business calculations. Individuals may also use the application for personal financial planning.

1.3 Conclusion

The study estimates the longevity prospects of T2DM patients compared to non-diabetics. It improves and adds on to the limited available literature, by performing a matched retrospective observational study and adjusting for numerous relevant risk factors as most existing literature is pharmacosurveillance based. This thesis includes

has 9 chapters with Chapters 2-3 discussing the literature review on longevity and T2DM, Chapter 4 discusses the statistical methods applied and Chapter 5 presents the study methodology. Exploratory analysis and results from the survival models and discussed in Chapter 6 and 7. Life expectancies derived from the translated models and the developed package are discussed in Chapter 8. Finally, Chapter 9 discuses the main findings, strengths, limitations and implications of the study.

Chapter 2

Overview of Diabetes Mellitus and Life Expectancy

2.1 Life Expectancy Historical Experience

In the past decades, life expectancy has been on the increase in both developed and developing countries. Actuaries, statisticians, insurers and governments have been drawn to the challenges it has imposed. This led to various studies on longevity risk, especially in the actuarial and insurance business, Barrieu et al. (2012). However, since 2010, a significant decline in the growth of life expectancy has been experienced in several countries, including the United Kingdom (UK) (IFoA, 2011, 2017). Health challenges that include antibiotic resistance and possibly austerity measures have been the major causes (IFoA, 2016, 2017).

Recent levelling-off of death rates in older ages implied that important decisions have to be made in longevity forecasts used in pension funds and annuities. This slow-down or levelling off in the reduction of mortality rates, particularly for the old age population, has been reported for life expectancy at age 65 in the UK, United States (US) and Canada since 2011. The UK's average annual rise in life expectancy at age 65 was 2.1 months over 2000 - 2011, falling to 0.4 months after 2011, (Lu, 2017).

According to the Office of National Statistics (ONS) produced National Life Tables (UK:2013-2015 and UK:2016-2018), life expectancy at birth declined from 2013-2015 to 2016-2018 by 3.6 and 2.4 months for males and females, respectively, whereas life expectancy at 65 years ($\mathring{e}_{65}$) improved by 1.2 months for both males and females (ONS, 2016, 2020). The most common age at death in the UK in 2018-2020 was 86.7 years and 89.3 years for males and females, respectively (ONS, 2020).

In 2016, ONS reported an increase in the number of deaths in 2015 in the UK which slowed down the increase in life expectancy at birth ($\mathring{e}_0$). In the period 2013-2015, the highest number of deaths was recorded in the 80 and above age group, constituting 57.4% and 68.8% of all deaths in males and females, respectively. It is clear from the ONS report that despite the improvements in life expectancy over the last 33 years (between 1980-1982 and 2013-2015), there was a dip between 2012-2014 and 2013-2015 for older ages as depicted in Figures C.1 to C.2 (ONS, 2016).

The total mortality rates increased by 4.2% between 2014 and 2015 (by 5% for females and 3.1% for males). However, the age-standardised mortality rate (ASMR) [1] was significantly higher in males than females (1 156 and 863 per 100 000 population, respectively). Two-thirds of deaths occurred in the 75+ years age group, (ONS, 2016). Hiam et al. (2017) suggested four possible causes of significant spikes in population mortality: 1. data artefact; 2. environmental shock; 3. major epidemic and; 4. widespread failure of health and social care. However, they considered that (1) and (2) were unlikely causes of this reported increase in the UK deaths in January 2015.

Infectious diseases may have been the causative factor with reported higher than expected deaths, mainly among the elderly in the past winter seasons. However, seasonal flu was refuted as a significant contributing factor in the UK because the higher numbers of deaths reported in 2014/2015 winter occurred when mean monthly temperatures were above average (Falkous, 2017). Falkous (2017) asked,"What could be causing this, and also causing broadly similar slowdown in life expectancy trends over the same time

[1] ASMR is the weighted average of age - specific mortality rates per 100 000 people

period in some other European countries?", a question which this study aims to address.

A similar decline in $\mathring{e}_0$[2] was noted among European Union (EU) - 28 (EU-28) countries in 2015 (Eurostat, 2017), with a total of 5.2 million deaths, the highest number observed over the past five decades reported. This translated to a crude death rate of 10.2 per 1000 and the decline in $\mathring{e}_0$ in 2015 was 0.3 years compared to 2014. Eurostat (2017) added that, it was impossible to conclude whether the decline in $\mathring{e}_0$ was temporary or will continue. A further significant drop both in $\mathring{e}_0$ and $\mathring{e}_{65}$ in 2020 in EU Member States, ranging up to -1.6 years (Eurostat, 2021) remains a cause for concern.

2.2 Overview of Major Causes of Death

According to the World Health Organisation (WHO), above 54% of the worldwide deaths between 2016 and 2019 were due to the top ten causes of death which included T2DM, ischaemic heart disease (IHD) and stroke. The latter two causes of deaths are possible T2DM complications (WHO, 2017a,c). A significant increase in the number of those who died due to diabetes mellitus (DM) in 2019 compared to 2000 has been reported particularly for lower-middle-income to high-income countries. Figures C.3 to C.6 show a comparative change in the top ten causes of death between 2000 and 2019. However, it is important to note that diabetes mellitus (DM) has constantly been among the top ten causes of death since 2010. DM was reported to be the sixth and the ninth cause of death worldwide in 2015 and 2019, respectively. The number of non-communicable diseases among the top ten causes of death increased from 4 in 2000 to 7 in 2015 and 2019.

According to Eurostat (2017), diseases of the circulatory system, including those related to hypertension, cholesterol, DM and smoking, IHD and cerebrovascular disease (CeVD) were the most common causes of death. In addition, two of UK's constituencies, England and Wales, showed an exponential increase in the total number of deaths caused

[2] Life expectancy at birth

by DM between 2012 and 2014, Figure 2.1 (ONS, 2017). This exponential growth in DM mortality rates needs to be inferentially studied and mitigation solutions recommended.

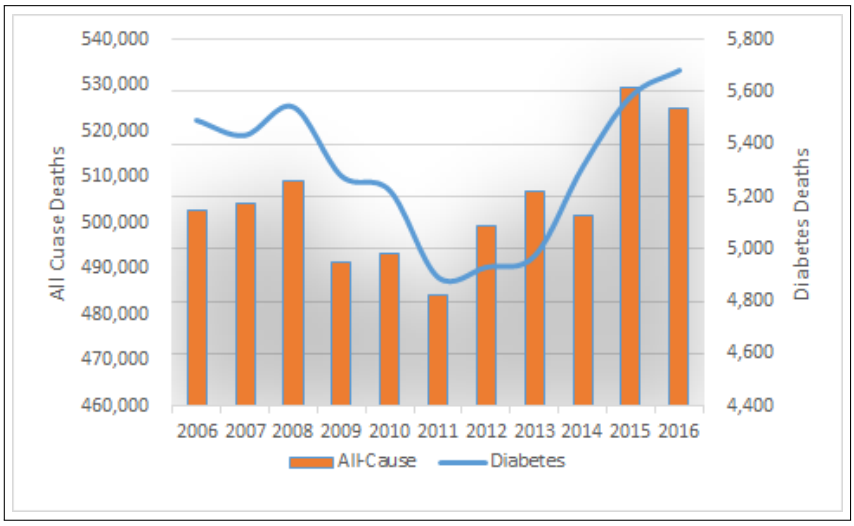


Figure 2.1: Number of deaths in England and Wales: All Cause and DM, 2006-2016. Source: ONS (2017)

2.3 Diabetes Mellitus

DM is a chronic disease that arises when either the pancreas does not produce insulin or the body fails to effectively make use of the insulin it produces (WHO, 2017b). Uncontrolled DM causes hyperglycaemia which can lead to serious damage to the body systems, especially the nerves and blood vessels. WHO estimated 8.5% of people aged 18 years and above have DM. In 2015 alone, DM directly caused 1.6 million deaths worldwide (WHO, 2017a).

Types of Diabetes Mellitus

The five main types of DM are type 1 diabetes mellitus (T1DM)[3], type 2 diabetes mellitus (T2DM), gestational diabetes mellitus (GDM), impaired glucose tolerance (IGT)

[3]T1DM can simply be defined as a serious medical condition where your blood glucose level is too high because of the body's failure to produce insulin.

and impaired fasting glycaemia (IFG). GDM is hyperglycaemia with blood sugar level higher than normal but below diagnostic level, occurring during pregnancy. Women with GDM are at higher risk of having pregnancy complications and/or developing T2DM. IFG and IGT, are intermediate conditions usually preceding T2DM.

Pre-diabetes can be broadly used as a generic term that includes IFG, IGT and glycated haemoglobin (HbA_{1c}) in the 'at risk' range ($6.0 \leq HbA_{1c} \leq 6.4$). An individual can develop any or all of the pre-diabetic conditions (Roberts et al., 2017; NICE, 2012b). IGT individuals have 60% risk of developing DM and 50% risk of having coronary heart disease (CHD) within 10 years (Yudkin and Montori, 2014; WHO, 2011).

Of much interest is T2DM, a medical condition in which the body fails to fully utilize its insulin. Individuals with pre-diabetes have annual T2DM incidence rates ranging from 3.6% to 7% depending on the pre-diabetes condition (Roberts et al., 2017). A comparison between T1DM and T2DM is provided in Table D.1. T2DM has been reported to be strongly genetic and related to life-style factors. Its insidious nature can cause vulnerable body organs such as eyes, heart, kidney and feet to be damaged before it is diagnosed (Matthews et al., 2008; Fox and Kilvert, 2007). While incidence rates of T1DM are increasing in many countries, T2DM incidence rates are currently near-epidemic and until recently, it was a disease of the adults, but is now occurring increasingly frequently in children, especially in some ethnic groups (Scobie and Katherine, 2014). T2DM has constituted between 70% and 90+% of diabetic patients worldwide (Barnett, 2006; Roberts et al., 2017). In the UK about 3.8 million people were reported to be diabetic, constituting 8.6% of the population, in 2015 in England with T2DM constituting 91.1% of diabetics (PHE, 2016c; Currie et al., 2010a). Due to the high prevalence of T2DM, this study focused on T2DM patients.

Poor diet, insulin production or utilisation and being inactive have been known to increase blood sugar levels. Furthermore, any form of stress, such as a flu, can increase blood sugar levels in diabetic patients (Scobie and Katherine, 2014).

2.3.1 The Aetiology of T2DM

The aetiology of T2DM has been known to be complex and is associated with modifiable and non-modifiable risk factors. These have been documented by Fox and Kilvert (2007) and Barnett (2006) among others.

Non-modifiable risk factors

These include age, genetics and ethnicity. The risk of T2DM increases with age and is dependent on ethnicity. Some ethnic groups are at a higher risk of T2DM than others, though the reason is still unknown. For example, BAME are at a higher risk of T2DM than white people. Furthermore, individuals with diabetic parents or siblings also have an increased risk. About 40% of people with one T2DM parent develop T2DM and 70% when both parents are affected (Ali, 2013).

Modifiable risk factors

Sedimentary lifestyles coupled with poor diet (such as excessive carbohydrates, fats and sugars intake) may increase insulin resistance (IR), a precursor to T2DM. Hyperinsulinaemia may develop as a result of IR and can continue for years until islet β-cells can no longer cope with the increased demand. As a result, they diminish, causing hyperglycaemia and other symptoms leading to possible complications. Smoking, apart from being known to cause cancer, heart and lung disease, increases the risk of T2DM by 30 to 40% compared to non-smokers according to U.S. Food and Drug Administration (FDA) and can cause challenges in the management of T2DM because increase of nicotine levels lessens the effectiveness of insulin.

Poorly managed body weight and obesity can exacerbate the risk. An unmanaged weight gain of about 13.5 kgs increases the risk of T2DM by 4-5 times compared to managed weight. However, though the incidence of T2DM increases with body mass

index (BMI), BMI does not reliably predict the risk of T2DM. Only 50% of patients with BMI$> 40\,kg/m^2$ develop T2DM (Bilous and Donnely, 2010). T2DM has also been known to develop in lean individuals (Bilous and Donnely, 2010). Unforeseen life shocking events such as injuries from traffic accidents may increase T2DM risk because the hormones produced in response to stress tend to oppose insulin's effect which increases the blood glucose levels. In addition, some medications, contraception drugs, abnormal hormones, other diseases and medical procedures may worsen the risk of T2DM and its management.

2.3.2 Diagnosis and Treatment of T2DM

Diagnosis

The diagnosis of T2DM requires the identification of a glycaemic cut-off that discriminates normal from diabetic individuals. The current cut-offs are based on the blood glucose level above which microvascular complications have been shown to increase. According to NICE, T2DM should be suspected if a patient has persistent hyperglycaemia accompanied with polydipsia, polyuria, blurred vision, unexplained weight loss, recurrent infections, tiredness, acanthosis nigricans and presence of risk factors such as family history, obesity, sedentary lifestyle, poor diet, history of gestational diabetes mellitus (GDM) and use of statins, corticosteroids, and combined treatment with a thiazide diuretic plus a beta-blocker (NICE, 2021).

Furthermore, NICE recommended a single abnormal HbA_{1c} or plasma glucose result can be used though a repeat is advised to confirm the diagnosis. However, several test are recommended in asymptomatic patients. National Institute for Health and Care Excellence (NICE) also advice HbA_{1c} results not to be used on some patients such as pregnant women (or women who are 2 months postpartum) and those living with HIV. Table 2.1 shows different diagnostic cut-off points recommended by the American Diabetes Association (ADA), International Expert Committee (IEC), WHO and National

Institute for Health and Care Excellence (NICE) (Barnett, 2006; NICE, 2017c; IEC, 2009; WHO, 2006; Roberts et al., 2017).

It is highly recommended for all asymptomatic adults aged 40-74 years in England to participate in diseases screening by the National Health Service (NHS) through the National Health Check program. Patients receive invitations from their GP every 5 years for screening of several selected diseases (NHS, 2019b). However, the screening of T2DM is not currently done nationwide as it is not publicly funded and currently not recommended by the National Screening Committee (NSC) citing no available evidence to supports the program Government (2022); NHC (2022). Some pharmacies provide T2DM screening for a premium.

Table 2.1: Diabetes Diagnosis Cut-Off Markers

DM Type	**Test Used**	**Criteria for Diagnosis** WHO	ADA	IEC	NICE	**T2DM Incidence**
IGT	FPG 2HrPG OGT	<7 7.8-11.1 7-11.1	NR 7.8-11.1 7-11.1		<7 7.8-11.1 7-11.1	0.045
IFG	FPG	6.1-6.9	5.6-6.9	5.6-6.9	6.1-6.9	WHO: 0.047 ADA: 0.036
HbA_{1c} 'at risk' range	HbA_{1c} (%)	6.0-6.4	5.7-6.4	6.0-6.4	6.0-6.4	WHO: 0.036
DM-II	FPG 2HrPG* HbA_{1c} (%)	≥ 7 ≥ 11.1 ≥ 6.5	≥ 7 ≥ 11.1 ≥ 6.5	 ≥ 11.1 ≥ 6.5	≥ 7 ≥ 8.5 ≥ 6.5	NA

NR: Not Required; FPG: Fasting Plasma Glucose Test; OGT: Oral Glucose Tolerance Test; 2HrPG: 2Hr Plasma Glucose Test. Diagnosis units are in mmol/L. Incidence rates are per person-year. * NICE recommends testing after 90 minutes instead of 2 hours

Treatment

Treatment of T2DM includes insulin or non-insulin medical interventions. Initially they comprise *change of life-style*, which includes diet and exercise, followed by *non-insulin medication* when *change of life style* alone fails to control the blood glucose levels. There are 7 main types of T2DM oral drugs, see Table D.2. Apart from these, there

are injectable incretin mimetics which mimic the action of glucose-like-peptide 1 (GLP-1) hormone and have been used since 2005. Dipetidyl peptidease - 4 DPP-4 inhibitors is another group of oral medication that work in the same way as incretin mimetics. These work by inhibiting the action of dipetidyl peptidease - 4 (DPP-4), an enzyme which destroys the hormone incretin. Incretins help in the production of more insulin only when it is needed and at the same time reduce the amount of glucose produced by the liver when it is not needed. In the event that non-insulin therapy fails, insulin with or without non-insulin therapy is used. About a quarter of DM patients receive insulin medication and almost every child diagnosed with DM is treated using insulin (Fox and Kilvert, 2007). Table D.3 briefly describes different types of insulin drugs (Diabetes UK, 2017).

2.4 Management of T2DM

The management of T2DM includes all facets of life of a diabetic individual: social, mental, spiritual and physical. Patients also need to be well informed about T2DM, its self management and possible complications and their causes (NICE Guideline NG28) (Fox and Kilvert, 2007). Patients are advised to keep a healthy life-style which includes a healthy diet, regular exercise, stopping smoking, reducing or stopping alcohol intake, weight management, adherence to medication and regular medical reviews. In medical reviews, T2DM markers which include blood pressure (BP), total, non-high and high density lipoprotein (HDL) and HbA_{1c} are monitored and screened for complications such as cardiovascular disease (CVD) (Fox and Kilvert, 2007; Matthews et al., 2008) and (NICE Guideline NG28).

Metformin is currently the first line drug therapy. At first, its dosage is increased over several weeks to minimise the risk of gastrointestinal side effects. Patients are prescribed a modified-release metformin if they experience gastrointestinal side effects. If their estimated glomerular filtration rate (eGFR) is less than 45ml/minute/1.73m^2,

metformin dose is reviewed. If eGFR is less than 30ml/minute/1.73m^2, metformin is either stopped or prescribed with caution of a risk of a sudden deterioration in kidney function.

Figure 2.2 shows a typical management of T2DM. After diagnosis and life style changes have failed to control the blood glucose level, usually metformin is given as the first line drug. In the case of it being intolerant, DPP-4, pioglitazone, sulfonylurea can be used as monotherapy. Caution is exercised as pioglitazone is associated with risk of heart attack, bladder cancer and bone fracture especially for the aged. It should not be offered to T2DM patients who have a heart failure or history of heart failure, hepatic impairment, diabetic ketoacidosis, current or history of bladder cancer and uninvestigated macroscopic haematuria. Sodium-glucose contransporter 2 SGLT-2 can also be used only if DPP-4 can be prescribed when sulfonylurea or pioglitazone can not be prescribed. If monotherapy fails, a dual therapy is prescribed using the monotherapy drug and one of the others. In the case of metformin being the monotherapy drug, empagliflozin, canagliflozine, dapagliflozine and alogliptin can be used in dual therapy when sulfonylurea is contraindicated or the person is at a significant risk of hypoglycaemia or its consequences. Triple therapy is then administered when the dual therapy fails by adding another drug, in the case of metformin being tolerant. If the triple therapy fails, insulin therapy with or without non-insulin drugs is then prescribed. In case of metformin being intolerant, insulin therapy is administered after the dual therapy fails.

GLP-1 mimetic can be used along metformin and sulfonylurea, in triple therapy, in cases where the T2DM patient has : (i) BMI greater than 35 kg/m^2 (adjusted according to ethnicity) or (ii) BMI less than 35 kg/m^2 and insulin therapy would have significant occupational implications or weight loss would benefit after significant obesity-related co-morbidities.

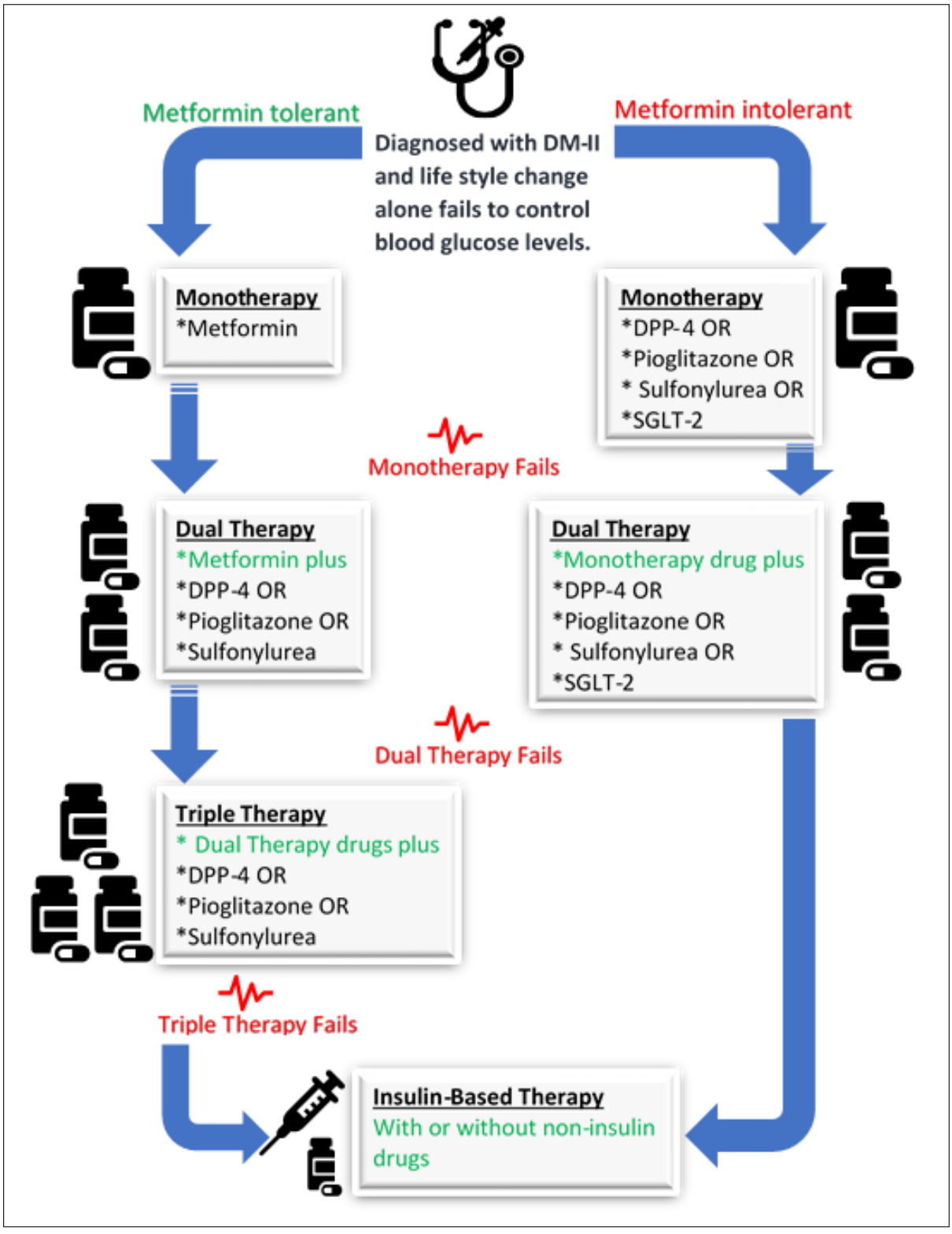


Figure 2.2: T2DM Therapy Management.

Source: (NICE, 2017b; Fox and Kilvert, 2007; Matthews et al., 2008)

2.4.1 Complications of DM and their management

T2DM is associated with serious complications including stroke, heart disease, blindness, kidney disease and amputations leading to disability and early death. Generally, half

of the T2DM individuals showed signs of complications when diagnosed (Diabetes UK, 2016b). Complications can develop 5 to 6 years before diagnoses and the actual onset of T2DM can be 10 years earlier. Proper T2DM management reduces complication risks.

Eye Disease

Globally, 2.6% of blindness is attributed to DM (Bourne et al., 2013). Retinopathy affects the blood vessels connected to the retina and accounts for about 7% of the people registered blind in England and Wales and about 60% of T2DM individuals have some degree of retinopathy within 20 years of diagnosis (Diabetes UK, 2016b). T2DM patients have nearly 50% increased risk of developing glaucoma especially if they have high blood pressure and up to 3 times increased risk of developing cataracts both of which can lead to blindness (Diabetes UK, 2016b). Drugs that can be administered include Aflibercept solution for injection, ranibizumab, dexamethazone intravitreal implant and fluocinolone acetonide intravitreal implant. However NICE does not recommend the use of aflibercept solution for injection and dexamethazone intravitreal implant.

Erectile dysfunction ED

erectile dsyfunction (ED) or impotence is one of the most common sexual problems experienced by men. About 35 to 90% of men with DM have ED according to an international literature review by Malavige and Levy (2009). Statistics with respect to sexual dysfunction in women living with T2DM are not available as this area is under-researched (Diabetes UK, 2016b).

Gastroparesis

It is a disorder that slows or stops the movement of food from the stomach to the small intestines. It can be treated alternating the use of Erythomycin and Metoclopramide or Domperidone. Metoclopramide can cause neurological effects such as short-

term extrapyramidal disorders. Medical and Healthcare Products Regulatory Authority (MHRA) notes that Domperidone is associated with small risk of serious cardiac side effects.

Neuropathy

T2DM can damage the nerves that transmit impulses to and from the brain and spinal cord, to the muscles, skin, blood vessels and other organs. Chronic painful neuropathy is the most common and estimated to affect 26% of DM patients. Cardiovascular autonomic neuropathy CAN affects the nerves that control the heart and the blood vessels and T2DM patients who develop it have a higher mortality risk than those without (Diabetes UK, 2016b). Medication prescribed include amitriptyline, duloxetine, gabapentin or pregabalin as initial treatment except in trigeminal nevralgia cases. There are also drugs that can be offered but in a specialist settings which include, morphine, cannabis sativa extract, lacosamide, oxacarbazine and topiramate (NICE Guideline CG173). Neuropathy affects up to 50% of DM patients.

Chronic kidney disease CKD

Kidney disease accounts for 11% of deaths in T2DM individuals (Diabetes UK, 2016b). chronic kidney disease (CKD) can be screened through testing for proteinuria, haematuria or renal ultrasound. Proteinuria is the presence of abnormal quantities of protein in the urine due to glomeruli diseases and urine tract infections among others. Haematuria is the presence of blood in the urine caused by urinary tract infection, kidney stones and bladder cancer among others. Albumin to creatine ratio ACR or protein to creatinine ratio (PCR)) can be used for testing for proteinuria, but albumin to creatinine ratio (ACR) is recommended to be used, especially for DM patients. According to NICE ACR greater than 3mg/mmol indicates the presence of CKD. ACR or PCR is calculated by measuring the amount of albumin or protein loss in the urine of patients with a GFR of less than 60ml/min/1.73m^2. Reagent strips are mostly recommended when testing for

haematuria. Risk is detected if the reagent strip result is 1+ or more. CKD severity is classified into 5 categories and it can be progressive.

Diabetic Foot Problems

Diabetic patients should be advised of the importance of basic foot care and footwear for the prevention of diabetes foot problems. Adult T2DM patients should be assessed for the diabetes foot problems at least annually (NICE, 2016a). Table D.4 in Appendix D gives the risk factors ascertained in assessing T2DM patients for diabetes foot problems.

DM is the most common cause of lower limb amputations. About 7,400 leg, toe or foot amputations happen in England per year, translating to 140 amputations per week or 20 per day (PHE, 2016a). DM individuals are 30 times more likely to have an amputation than the general population (Khanolkar et al., 2008).

Lower limb peripheral arterial disease LL PAD

lower limb peripheral arterial disease (LL PAD) is another T2DM complication. Peripheral artery disease PAD is tested by the examination of legs and feet for ulcerations, femoral, popliteal and foot pulses; and measuring ankle branchial pressure index (ABI). The latter is done manually using a Doppler probe of suitable frequency in preference to an automated system. The index is calculated for each leg by dividing the highest ankle blood pressure by the highest arm blood pressure (NICE, 2012a). If the index is not within the normal range 0.9-1.4, the T2DM individual is at risk of LL PAD (Stanford Medicine, 2018). peripheral arterial disease (PAD) patients have an increased risk of mortality from CVD, mainly due to the increased risk of heart attack and stroke.

Non-alcoholic fatty liver disease NAFLD

non-alcoholic fatty liver disease (NAFLD) is another possible complication of T2DM. Alcohol history should be taken to rule out possibilities of alcohol-related liver disease.

NICE discourages use of routine liver blood tests to rule out NAFLD. In assessing the risk, liver ultrasound should be done to test for NAFLD in children and young people (aged between 1 and 18) with T2DM (NICE, 2016b). Patients taking statins are adviced by NICE to continue taking medication except when liver enzymes double within 3 months of starting statins. For assessing advanced liver fibrosis in adults, enhanced liver fibrosis test (ELFT) is used. Adults are diagnosed to have advanced liver fibrosis if their ELFT score is equal to or greater than 10.51 and have NAFLD. Cirrhosis should be monitored in adults diagnosed with NAFLD and advanced liver fibrosis (NICE, 2016b). Diagnosis of cirrhosis can be established through blood tests, ultrasound scan, computerised tomography (CT) scan or magnetic resonance imaging (MRI) scan, liver biopsy and endoscopy (NHS, 2017). *Pioglitazone* or *vitamin E* are given to adults with advanced liver fibrosis but only in secondary or tertiary care settings (NICE Guideline NG49).

Cardiovascular Disease (CVD)

CVD includes heart disease, stroke, and all other diseases of the heart and circulation, such as hardening and narrowing of the arteries supplying blood to the legs, which is known as peripheral vascular disease (PVD). DM individuals are twice at risk of developing CVD compared with non DM individuals (Diabetes UK, 2016b). CVD is the major cause of death among DM individuals accounting for 52% in people with T2DM (Diabetes UK, 2016b). T2DM patients also have higher risk of developing stroke than the general population (Diabetes UK, 2016b).

Adults aged 85 years and above are at high risk of CVD because of age alone, especially those who smoke or have raised BP (NICE, 2017a). One of the tools that can be used for the assessment of a 10 year risk of CVD is the QRISK2. It is used for the primary prevention of CVD in all people aged less than or equal to 84 years. The tool can not be used for people with T1DM and those with eGFR less than $60/\text{ml}/\text{min}/1.73m^2$ or albuminuria. People above the age of 40 are encouraged to have their CVD risk reviewed

on an ongoing basis (NICE, 2017a).

Threat to oral health is another complication as T2DM patients have an increased risk of inflammation of the tissues surrounding the teeth (periodontitis) which is a major cause of teeth loss and is associated with an increased risk of CVD in people with poor blood glucose control (IDF, 2015).

2.5 Epidemiology of T2DM

WHO estimated the global prevalence of DM to have increased from 4.7% in 1980 to 8.5% in 2014. The prevalence has been rising in the low- and middle-income countries.

In 2015 alone, about 415 million people worldwide were living with DM and an estimated 1.6 million deaths were attributed to DM. These numbers increased by 29% and reduced by 6.5%,respectively, in 2020 (WHO, 2017c; IDF, 2021). This disease is now among the top 7 leading causes of death in the world. With the prevalence rising in low- and middle-income countries, the majority of deaths from DM occur in these countries due to limited access to basic technologies needed by diabetic patients in primary health care settings. The International Diabetes Federation (IDF) in 2015 estimated 193 million undiagnosed adults, aged 20-79 years, are living with DM. About 318 million were estimated to have IGT. IDF (2015) also estimated a total expenditure between US$673 billion and US$1.2 trillion in health care in 2015 due to DM.

IDF projected a total of 642 million diabetic cases in 2040 if control measures are not taken and people continue with unhealthy life styles. The substantive increase in the prevalence is expected to be driven by the ageing population (Roberts et al., 2017). Bagust et al. (2002) estimated not more than 3% increase in the total UK population between 2000 and 2060 but an increase of 11% in the population aged 35 years and above by 2030. The global prevalence of T2DM cases would increase by approximately 20% in 2036 and DM related complications would rise by 20-30% between 2035 and 2045 (Bagust et al., 2002). The cost of health care related to treatment and complications

is estimated to increase by 12% of the global DM health expenditure. Figure 2.3 shows an estimated increase of 55% of DM cases in 2040 worldwide. In the UK the cost of T2DM will present a serious clinical and financial challenge in the near future (Bagust et al., 2002). In 2015, the global prevalence of DM was 1 in 11 adults and in 2040 it is

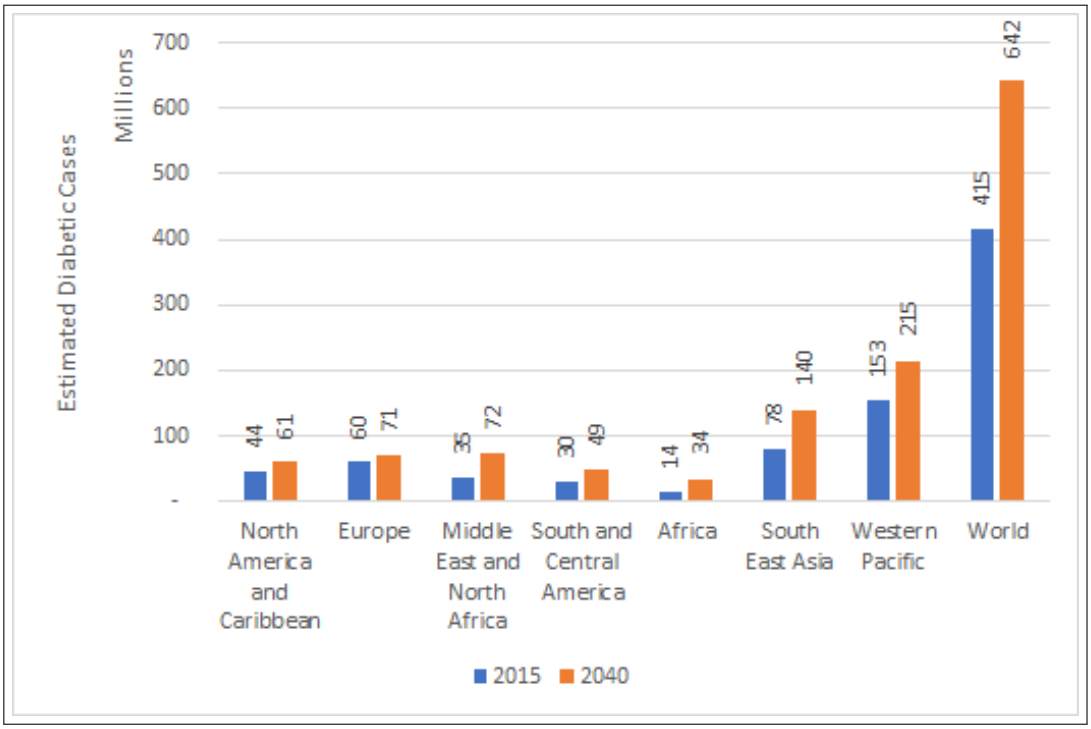


Figure 2.3: Number of diabetes cases by IDF Region

Source: IDF (2015)

estimated to be 1 in 10 adults. Globally, 269.7 million in the urban and 145.1 million in the rural population had DM. The numbers are estimated to increase to 477.9 million and 163.9 million, by 2040 respectively (IDF, 2015). It is estimated that 90% of diabetic patients have T2DM.

Top Ten Countries with the Highest Numbers of DM Patients and Highest Expenditure

China had the highest estimated number of diabetic patients in 2015, 109.6 million followed by India with 69.2 million patients. Countries with the highest costs were US, US$320 billion, followed by China, US$51 billion, as shown in Table 2.2, (IDF, 2015) .

Table 2.2: Top ten countries with the highest numbers of T2DM patients and the highest expenditure

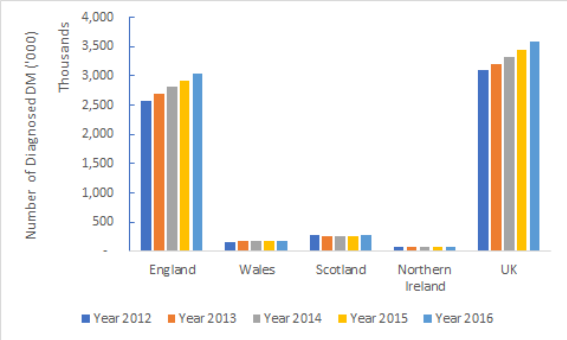
Country or Territory	**DM prevalence x** 10^6		**Expenditure x** 10^9	
	Number	*Rank*	*US$*	*Rank*
China	109.6	1	51	2
India	69.2	2		
USA	29.3	3	320	1
Brazil	14.3	4	22	5
Russian Federation	12.1	5	14	8
Mexico	11.5	6		
Indonesia	10	7		
Egypt	7.8	8		
Japan	7.2	9	29	4
Bangladesh	7.1	10		
Germany			35	3
France			19	6
Canada			17	7
UK			13	9
Italy			12	10

Source: IDF (2015)

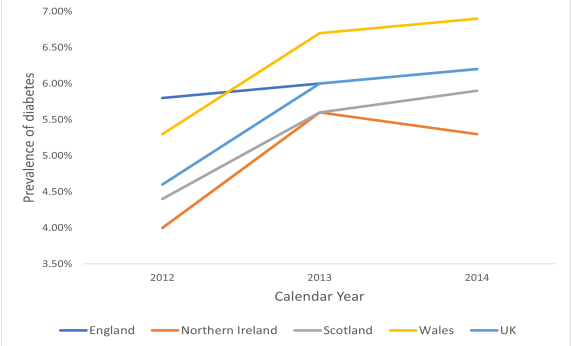
2.5.1 DM Statistics in the UK

Between 2012 and 2014 the prevalence of DM in the UK has increased from 4.6% to 6.2%. The total number of DM patients increased by more than 16%, from 3 094 681 in 2012 to 3 590 501 in 2016. England has the highest number of DM cases throughout the period but its prevalence is not far from the other three states. DM cases in England constitute 84% of the UK cases. The DM prevalence for England and the UK at large in 2013 and 2014 were similar, 6% and 6.2%, respectively. The prevalence of T2DM in Wales from 2012 to 2014 was higher than that of the UK. While the prevalence of DM increased in all the other three nations in 2014, it declined in Northern Ireland (Figure 2.4). Of the four states in the UK, Wales is highly affected and Northern Ireland had the lowest prevalence (Diabetes UK, 2013, 2014, 2015a,b, 2016b). Overall, there has been an increase in the number of diagnosed DM cases throughout the four states. In England alone, about 1 in 4 undiagnosed adults is believed to be diabetic translating to 940

000 people living with diabetes unknowingly. The prevalence is higher in males (9.6%) than in females (7.6%) in England (PHE, 2016b). Public Health England (PHE) also estimated higher prevalence in people of South Asian and black ethnic groups (15.2%) compared to those from white, mixed or other ethnic groups (8.0%); and it estimated about £10 billion medical costs per year related to diabetes of which £8.8 billion per year was due to T2DM in England alone.



(a) Number of DM Cases



(b) DM prevalence

Figure 2.4: DM and it prevalence in UK.
Source: (Diabetes UK, 2013, 2014, 2015a,b, 2016b)

T2DM Complications Statistics in UK

Table 2.3 shows the number of complications and prevalence for T2DM patients in England and Wales (NHS, 2017). It demonstrates that T2DM patients are 2 to 2.5 times more at risk of developing CVD. The complications with the highest rates were angina and HF with 3.1% and 3.2% prevalence, respectively. According to National Health Service (NHS) in 2017, the risk of being admitted at a hospital among T2DM patients by age at admission was exponential from the age of 35-39. It can be seen that non-DM patients had higher CVD admission rates than T2DM and other diabetic patients, in the 0-54 years and 85+ years age groups and T2DM and other diabetic patients had high CVD admission rate in the 55-84 years age group compared to non-DM patients.

Table 2.3: One year prevalence of CVD complications in T2DM.

Complication		**Type 2 and Other***	
		Complication Prevalence	*Prevalence (%)*
CVD	Angina	56,046	*3.1*
	Myocardial Infarction	14,762	*0.8*
	Heart Failure	56,817	*3.2*
	Stroke	19,350	*1.1*
Diabetes Specific Complications	MajorAmputation	1,322	*0.1*
	Minor Amputation	2,918	*0.2*
	Renal Replacement Therapy (ESKD)	11,006	*0.6*

Note: * Other types of diabetes except for DM-I.
Source: NHS (2017): England and Wales

2.6 Conclusion

The incidence of DM is increasing worldwide. Less than two decades ago the total number of deaths caused by DM was less than a million in the world. This increased to 1.6 million in 2015. Since 2010, DM has been among the top ten causes-of-death worldwide. The complications of T2DM include the top four causes of death, including stroke. As T2DM constitutes 70-90% of all DM cases, it is of paramount importance to quantify its impact on longevity (Hiam et al., 2017). T2DM is associated with poor life style and its prevalence is on the rise. It is therefore, important to demonstrate its impact on longevity and morbidity through survival techniques and translate the results into statistical and actuarial use.

Chapter 3

Review of the Life Expectancy in T2DM Individuals

This chapter is based on existing and published T2DM literature retrieved through searches on Google Scholar, PubMed, ScienceDirect and Scopus whose outcome was all-cause mortality. Keywords for literature search were, "type 2 diabetes", "diabetes mellitus II", "T2DM", "survival","all-cause mortality", "mortality" and "life expectancy". The topics and abstracts were used to find relevant existing literature. The questions that the literature search intended to answer are "the survival of T2DM patients after first diagnosis" and " the life expectancy of T2DM patients after diagnosis" compared to people without diabetes at the time of entry into the study as to assess the impact of T2DM on all-cause mortality. To remain abreast with recent studies, alerts on newly published papers were created.

3.1 All-Cause Mortality and Life Expectancy in T2DM Individuals

In as much as there have been significant medical advancements, the all-cause mortality hazards associated with type 2 diabetes mellitus (T2DM) are still considerably

higher than those of the general population (Barr et al., 2008). Though there have been numerous papers on diabetes, research on the association of diabetes with total life expectancy and mortality still remains limited, see Table 3.1. These associations are not easily established as different risk factors can lead to unexpected consequences in life expectancy (Franco et al., 2007).

Franco et al. (2007) study constituting 9033 (5% T2DM) patients aged 50 years and above, in the United States (US) found that several confounders such as total cholesterol, presence of left ventricular hypertrophy, arthritis, ankle edema, and pulmonary disease had no significant effect on the hazard ratios (HRs) for cardiovascular disease (CVD) morbidity and all-cause mortality. They also estimated the total sex stratified all-cause mortality adjusted hazard ratio (aHR) of T2DM patients to be 1.26[0.78-2.02] and 1.31[0.8-2.10] for male and female T2DM patients with no CVD, respectively. However, apart form the small study population size, their T2DM participants were much older than non-diabetics (mean age 68 vs 59).

The study by Wright et al. (2016) on 1 095 984 (17.2% had T2DM, diagnosed from 1998 to 2015) patients in the UK with an average age of 61 using the CPRD, formerly General Practice Research Database (GPRD), found T2DM to be associated with a higher all-cause mortality risk compared to non-diabetics, with a HR 2.19 [2.16-2.21], after adjusting for age, gender, ethnicity, deprivation and calendar year. Their study also found that T2DM patients of BAME origin had at least 26% reduced all-cause mortality risk when compared to T2DM patients of white origin. Another study by Taylor et al. (2013) using the GPRD constituting 87 098 (25% T2DM) participants with an average age of 55.1 for both T2DM participants and non-diabetics and index date from 2004 to 2010, also found T2DM to be associated with more than double the risk of all-cause mortality when compared to non-diabetics after adjusting for smoking status only, HR 2.07 [1.95-2.2].

The life expectancy of men and women aged 50 years and above with diabetes mellitus (DM) in the US, living in Framingham, Mass, between 1948 and 1951, was 7.5

years (95% CI: 5.5-9.5) and 8.2 years (95% CI: 6.7-10.4) less than of their non-diabetic counterparts (Franco et al., 2007). Franco et al. (2007) used the Poisson regression and Gompertz distribution to calculate the HRs. They then used the HRs to calculate the life tables stratified by sex and presence of DM that were used for the computations of the differences in life expectancy. In a later study, the United Kingdom (UK) men and women with T2DM, age 40 years in 1998, lost 5.4 and 6.3 years of life, respectively, compared to non T2DM persons after adjusting for age, sex, ethnicity, deprivation and calender year(Wright et al., 2016).

Table 3.1: Previous observational studies on longevity in T2DM patients

Observational Studies [All-Cause Mortality]										
Author	Country	Database	Entry Dates	Variables adjusted for	Study Size		Mean Age (years)	Matching	Measure	Findings
					T2DM	non-DM				
Mulnier et al. (2006)	UK	GPRD	1 January 1992	age	44 230	219 797	67	1:5 (birth year and gender)	HR	1.93 [1.89-1.97]
Lind et al. (2013)	UK, Canada	THIN, Ontario	1996-2009	age and gender	①,①	②	②	No	MRR	①
Taylor et al. (2013)	UK	GPRD	1 January 2004	smoking status	21 798	65 300	55.1	1:3 (age,gender and GP)	HR	2.12 [2-2.25]
Wright et al. (2016)	UK	CPRD	1998 to 2015	age, gender, IMD and entry year	187 968	908 016	61.8	1:5 (birth year (± 2 years),gender,GP and entry date)	HR	2.19[2.16-2.21]
Kim et al. (2017)	Korea	NHI	2003-2005	③	749,161	5 078 988	69.5	No	HR	1.46 [1.45–1.47]
Almdal et al. (2004)	Denmark	CCHS	1976-1978	④	301	7 651	not specified	age	HR	⑤

① Qualifying participants were selected by calendar year from 1996 to 2009. ② Reported mean age among participants in the UK was 62.7 years and 63 years in the 2006 and 2009 cohorts, respectively. ③ Adjusted for age, sex, classes of national health insurance system, place of residence, HTN, dyslipidemia, ischemic heart disease, CeVD, HF, peripheral occlusive arterial disease, COPD, CKD and depression. ④ Adjusted for tobacco consumption, physical activity, alcohol consumption, body mass index, and triglyceride and total cholesterol levels. ⑤ HRs reported by age group and gender. Ranged from 1.7 to 2.7. ⑥ mortality rate ratio (MRR)

Arora (2017) showed that life expectancy at birth is positively correlated with spending on health care per person, up to a limit, after which the benefit in life expectancy diminishes. Thus, medical therapies have limitations once a certain level of spending is exceeded. Extra spending on hospitals and specific surgery may not have any added benefits for life expectancy but may provide better quality of life. Life expectancy reflects not just health spending but the life style, such as alcohol consumption, smoking, education, socio-economic factors and so on.

Franco et al. (2007) reported that variables such as total cholesterol and presence of left ventricular hypertrophy, arthritis, ankle edema, or pulmonary disease at entry were found not to alter the HRs of mortality of T2DM patients to controls. However, Franco et al. (2007) noted the limitations of their data, that were based on the period 1950s-1980s. They may have underestimated the true association of DM with $\mathring{e}_{50}$ because the incidence of T2DM has been on the increase in recent decades. Mortality due to ischaemic heart disease (IHD) and stroke is 2 to 4 times higher in DM than in non - diabetic population (Bilous and Donnely, 2010).

Taylor et al. (2013), using data from the General Practice Research Database (GPRD) (2004-2010), found that T2DM patients have a two-fold increased risk of all-cause mortality (hazard ratio (HR): 2.12 95% CI: 2.0-2.25), adjusted for smoking. Women were found to have a higher HR than men, and patients younger than 55 years of age were reported to have a higher HR than those aged 56 years and above when compared to those without DM.

Lind et al. (2013) studied the mortality trends of diabetic patients aged 20 years and above in Canada and the UK using the Ontario and The Health Improvement Network (THIN) databases, respectively. The total number of subjects under study was 8 757 772 in 1996 and increased to 12 696 305 in 2009. The study showed the same mortality trends per 1000 person-years for both DM and non-DM patients. They reported a decline in the mortality rates. They also reported high mortality rate ratios for the young T2DM patients compared to the old patients for both countries.

Mulnier et al. (2006) reported a similar mortality risk pattern among the young T2DM patients (HR: 3.35 [95% CI: 2.86-3.93] for men aged 35-54 years and 3.07 [95%CI: 2.37-3.97] for women of the same age) when compared to patients without T2DM as opposed to the old patients aged 85-89 years (HR: 1.44 [1.3-1.6] for men and 1.65 [1.52-1.78] for women). Thus the younger the age the higher is the mortality risk among T2DM patients when compared to their non-diabetic counterparts.

Currie et al. (2010b) based on the GPRD data (1986-2008) studied the survival effect of glycated haemoglobin (HbA_{1c}) in T2DM patients. They found that people with lower HbA_{1c} levels or higher HbA_{1c} had higher risks of all-cause mortality compared to people with HbA_{1c} levels between 7.41%-7.48%.

According to Kontopantelis et al. (2015) in a study of 246 544 T2DM patients based from GPRD (2006 to 2012), about 61% were taking metformin and 32% were using sulfonylurea. The least used was acarbose (0.5%). Though their study did not show the mortality and morbidity risks associated with the drugs, their distribution is of vital importance when analysing clinical outcomes.

In a 20-year follow-up prospective study by Almdal et al. (2004), based on 13 105 T2DM patients from the Copenhagen City Heart Study (CCHS), from 1976 to 1997, the relative risk (RR) of death for both men and women was 1.5 to 2 compared to non T2DM patients. In women, the age groups that were reported to have higher RR of 1.7 to 2.7, were the under 55s and the 55-64 year olds compared to non-diabetic women. Men in the under 55s age group were reported to have a 2.5 times higher total mortality rate compared to non-diabetic men. Almdal et al. (2004) reported that smoking, hypertension, high triglyceride level, high alcohol intake and low ($\leq$20kg/m^2) and high ($\geq$30kg/m^2) body mass index (BMI) independently increased the risk of death in women while in men the above variables excluding high tryglycerides and high BMI ($\geq$30kg/m^2) had the same effect. High physical activity and high total cholesterol levels were reported to decrease the mortality risk (Almdal et al., 2004).

The Spanish study by Salinero-Fort et al. (2018) based on a period from 2007 to 2012 reported that male T2DM patients with a BMI less than 23kg/m^2 had the highest all-cause mortality (adjusted hazard ratio (aHR): 2.78 [95% CI: 1.72-4.49]. The study showed a U-shaped mortality curve among men with T2DM with respect to BMI. Overall a BMI less than 23kg/m^2 was reported to have a higher significant effect on mortality when compared to a BMI in the range 23-26.8kg/m^2. They concluded that higher BMI was not associated with a higher mortality risk among T2DM women patients. Kontopantelis et al. (2015) and Soriano et al. (2015) have similar findings. They reported that there was less significant association between higher BMI and higher mortality among female T2DM patients than in patients without T2DM. They adjusted for age, gender, diabetes duration, CVD, complications of DM, smoking, blood pressure, high density lipoprotein (HDL)-cholesterol, low density lipoprotein (LDL)-cholesterol, triglycerides, HbA_{1c}, albuminuria, treatment of hypertension, treatment of diabetes, statins and antiplatelet drugs.

A similar study based on the GPRD by Mulnier et al. (2006) of T2DM patients aged between 35 and 89 years in 1992, with a follow up period of less than 7 years, also reported a U shaped HR plot in respect to BMI showing that decrease from the normal BMI range was associated with an increased risk of mortality. However, they noted that different studies in the past had conflicting conclusions on the effect of obesity or BMI on mortality.

The prescription of medication to patients is done with the full intent of achieving an optimal treatment outcome. Optimal treatment outcome has been defined as as an improvement in the underlying condition in the absence of drug-related problems,(Johnson and Booman, 1996). Studies by Johnson and Booman (1996) and Watanabe et al. (2018) found out that the cost of drug-related morbidity and mortality in the ambulatory setting in the US was considerable. However, they acknowledged that a large proportion of drug-related morbidity was preventable. Hence, there has been several diabetes pharmacosurveillence studies in the UK and other countries. Currie et al. (2010b) noted that

T2DM patients using insulin-based therapy had a higher risk of mortality compared to those on oral combination therapy such as metformin plus sulphonylurea with an HR of 1.49 (95% CI: 1.39-1.59). Soriano et al. (2015) reported an increased risk of mortality with an increase in the number of prescribed drugs among T2DM patients. The study did not specify if they included study subjects using antidiabetic drugs only. We therefore assume that their study included other non-diabetic drugs. This can also be due to more advanced T2DM in patients on more drugs in both studies. Toulis et al. (2017) retrospectively studied all-cause mortality and CVD incidence outcomes in a cohort of 22 124 T2DM patients from the THIN database, January 2013 to September 2015. The study has validated the results from an earlier study of empagliflozin clinical outcomes, using the drug dapagliflozin. Both drugs are sodium-glucose contransporter 2 (SGLT-2) inhibitors. They found out that patients using dapagliflozin had a lower all-cause mortality risk compared to patients not using SGLT-2 inhibitors (adjusted incidence relative ratio (aIRR): 0.5[95% CI: 0.33-0.75]). The data were matched for age, sex, BMI, T2DM duration and smoking and adjusted for age, sex, BMI, smoking, HbA_{1c}, duration of diabetes, systolic blood pressure (SBP), lipid-lowering medication, insulin use, estimated glomerular filtration rate (eGFR), social deprivation index, presence of hypertension and Charlson comorbidity index.

Zimmerman et al. (2017) studied 105 856 patients from the Cleveland Clinic in Ohio, US, between 2005 and 2014, and found that T2DM patients with or without a history of CVD using glucose-like-peptide 1 (GLP-1) receptor drugs, were overall at a lower risk of mortality and CVD. They adjusted for age, gender, race, income, BMI, LDL, cholesterol, eGFR, diabetes complications severity index (DCSI), acute myocardial infarction (AMI) history, cerebrovascular accident (CVA) history, coronary heart disease (CHD) history, hypertension, smoking status, use of statins and diabetic medications. Their study also reported that newly diagnosed T2DM patients using GLP-1 had an all-cause mortality HR of 0.80 [95% CI: 0.47-1.34]) compared to those using other drugs.

Several studies have shown that management of T2DM by intensifying treatment in-

creases the risk of mortality among T2DM patients (ADVANCE Collaborative Group, 2008; UKPDS Group and others, 1998; ACCORD Study Group, 2008). According to NICE (2015), intensive treatment is a treatment of a T2DM patient with more than one drug in the case where monotherapy fails to reduce the HbA_{1c} level to below 6.5%. T2DM patients under an intensive treatment have their blood glucose levels more frequently monitored than those on standard treatment. The treatment dosage may also be increased. For example, in the Action in Diabetes and Vascular Disease: Preterax and Diamocron Modified Release Controlled Evaluation (ADVANCE) study by ADVANCE Collaborative Group (2008), patients on intensive treatment were examined every 3 months while those on standard treatment were examined every 6 months.

A meta-analysis by Boussageon et al. (2011) included 13 studies of intensive treatment performed between January 1950 and July 2010 constituting 34 533 T2DM patients of whom 60% were men. Intensive treatment was found to increase the risk of mortality by 19% compared to standard treatment. In their meta-analysis, 39% of patients already had a CVD at baseline which explained the high overall all-cause mortality risk.

Boussageon et al. (2011) reported that using intensive treatment increased the risk of CVD related deaths among T2DM patients compared to those using standard treatment (RR: 1.11 [95% CI: 0.86-1.43]). They reported a 43% increase in CVD-caused deaths in intensive treatment cohort compared to the standard treatment cohort. They also reported that 50% of T2DM patients with myocardial infarction (MI) die before receiving medication.

In a study by Hayes et al. (2013) from the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model 2 project, based on 5 102 UKPDS patients aged 25-65 years, recruited between 1977 and 1991 and followed up for 20 years, several demographic, risk and event history factors were used to develop a model that predicts the lifetime health outcomes of T2DM patients. For all-cause mortality, they used logistic models for patients with complication(s) and proportional hazards Gompertz survival models in patients with no complications. Blindness and ulcers were found not to be as-

sociated with mortality. They also reported that smoking was an *independent significant* predictor of mortality but HbA_{1c} and systolic blood pressure (SBP) were not. Duration of DM was also found to be significantly associated with mortality. They reported that about 60% of T2DM patients would have died after 25 years.

Hayes et al. (2013) performed a sensitivity analysis of the risk factors by adding or subtracting 1 standard deviation (SD) to their values. Increasing/decreasing by 1 standard deviation (SD) the values of risk factors had a significant effect on the life expectancy of T2DM patients. Increasing SBP by 1 SD decreased the life expectancy by about 0.5 years while decreasing SBP by 1 SD increased the life expectancy by about 0.4 years. LDL, HbA_{1c}, white blood cells count and heart rate were also reported to have a similar significant effect on all-cause mortality hazard as SBP. A slight increase of the BMI was reported to cause a serious reduction in life expectancy compared to a slight decrease, whereas slight change in haemoglobin and atrial fibrillation (AF) were reported not to have a significant association with life expectancy. Using the UKPDS Outcomes Model 2 (UKPDS-OM2), they predicted life expectancies of 25.1 years, 17.7 years and 11.7 years for T2DM patients aged 50-54 years, 60-64 years and 70-74 years, respectively. These LEs were higher than those predicted in UKPDS-OM1 by 5.1 years, 3.8 years and 1.8 years, respectively (Hayes et al., 2013). When they compared their results with the UKPDS Outcomes Model 1 (UKPDS-OM1), they reported that UKPDS-OM1 had higher outcome incidences compared to Model 2 (UKPDS-OM2). They included eGFR, micro- or - macroalbuminuria, heart rate and white blood cell counts which were not included in UKPDS-OM1. In addition, they included the haemoglobin in the Model 2. This was based on the results reported by Go et al. (2006) that haemoglobin is an independent predictor of mortality in congestive heart disease (CHF) patients.

Jeong et al. (2017) suggested that depression was significantly associated with a higher mortality risk among 1 043 089 T2DM patients aged 30 years and above from a Korean database, with a study period between 2003 and 2013. The mortality HRs were high in all the age groups. The risk was higher among men or young age groups (Jeong et al.,

2017). The reported total mortality HR accounted with depression was 1.43 [95% CI: 1.41-1.46] compared with those without depression. Depressed T2DM patients, aged between 30 and 39 years, had the highest mortality HR of 2.81 [95% CI:2.334-3.384] compared with non depressed T2DM patients. The HR decreased with an increase in age. The differences in the mortality of T2DM patients with and without depression are reported to be increasing throughout the 21 years of follow-up.

3.2 Cause-Specific Mortality among T2DM Patients

3.2.1 CVD-Related Mortality among T2DM Patients

CVD accounts for about 70% of deaths among T2DM patients (Franco et al., 2007). The excess mortality between T2DM and non-T2DM patients is largely attributed to CVD (Leal et al., 2008). CVD increases the risk of dying among T2DM patients (HR: 2.2 [95% CI: 1.74-2.84] for women and 1.7 [95% CI: 1.38-2.07] for men) compared with non-T2DM patients with CVD. T2DM patients aged 50 years without CVD at baseline were reported to have lived 7.1 and 6.8 years more than T2DM patients with CVD, in men and women, respectively (Franco et al., 2007). Taylor et al. (2013) found that T2DM patients with CVD had a threefold increased risk of CVD mortality (HR: 3.28 95%CI: 2.91-3.70), adjusted for smoking, compared to those without T2DM.

The risk of CVD deaths increased with DM duration (Taylor et al., 2013). From a study of Spanish subjects (2007-2012), 51.9% deaths among T2DM individuals were attributed to CVD (Salinero-Fort et al., 2018). A GPRD based study of 246 544 T2DM patients, from 2006 to 2012, by Kontopantelis et al. (2015) showed that those with macrovascular complications were at a higher risk of dying from CHD (HR: 2.14 [95% CI: 1.88-2.44]) and cerebrovascular disease (CeVD) (HR: 1.6 [1.41-1.82]). In the same study, men were found to be at a higher risk of dying from CHD with an HR of 1.5

[95% CI: 1.33-1.7] compared to women.

Smoking T2DM patients from the UKPDS with a SBP of 180, total/ HDL cholesterol ranging from 4 to 8, aged 75 years and with an HbA_{1c} ranging between 6%-10% lived 1.2 years less, on average, than non-smokers (Leal et al., 2008). From their findings, the higher the HbA_{1c}, SBP and cholesterol (Total:HDL) ratio the lower was the life expectancy at any age. Improving modifiable risk factors such as HbA_{1c} increases T2DM patients' longevity. Kontopantelis et al. (2015) also reported U shaped HRs when plotted against HbA_{1c}, SBP, distolic blood pressure (DBP) and cholesterol. From their findings the category levels of HbA_{1c}, total cholesterol, SBP and DBP with the lowest mortality risks were 7.25-7.75%, 3.5-4.5mmol/L, 135-145mmHg and 82.5-87.5mmHg, respectively.

3.2.2 Kidney Disease

As discussed in section 1.4.1, chronic kidney disease (CKD) accounts for 11% of deaths in T2DM individuals (Diabetes UK, 2016b). Wright et al. (2016) showed that renal failure had a mortality HR of 3.33 [95% CI: 3-3.69] in T2DM patients compared to those without T2DM. The eGFR can be used to measure the damage to the kidney and can be used as a predictor of mortality. In a study of T2DM patients, aged 20-89 years in 2000-2005, from the THIN database (UK), patients with an eGFR $\geq$ 60mL/min had the highest survival rates throughout the 11 years of follow up. The lower the eGFR, the lower was the survival rate (Soriano et al., 2015). However, due to the bad recording of ethnicity in the THIN database, the study's eGFR calculated using the Cockcroft-Gault formula, may have underestimated eGFR, particularly for the blacks (Soriano et al., 2015).

3.2.3 Dementia

Dementia is a condition in which there is a decline in memory, thinking, behaviour and ability to carry out day-to-day activities. Dementia can be classified as, Alzheimer, vas-

cular dementia (VD), dementia with Lewy bodies (DLB), and diseases that contribute to frontotemporal dementia (WHO, 2017). Kim et al. (2017) study of 749 161 patients with T2DM only, 388 636 with chronic liver disease (CLD) only, 122 590 with both T2DM and CLD and 5 080 631 controls with neither T2DM nor CLD, all aged 60 years and above at 2003 to 2005 and followed up to 2013, reported that T2DM alone increased all-cause mortality among dementia patients (aHR: 1.46[95% CI: 1.45-1.47]). The presence of both diseases further increased the mortality risk compared with the control group (aHR: 1.67[95% CI: 1.65-1.69]). They used data from the National Health Insurance Service of Korea and adjusted for age, sex, classes of national health insurance system, place of residence, hypertension, dyslipidemia, IHD, CVD, heart failure (HF), peripheral arterial disease (PAD), chronic obstructive pulmonary disease (COPD), CKD and depression.

3.2.4 Effect of Lifestyle on Mortality among T2DM Patients

Leal et al. (2008) using the UKPDS Outcomes model, reported that the life expectancy of a non-smoking T2DM male patients aged 55 years, $\mathring{e}_{55}$, was between 3.6 and 9.4 years less than UK male general population life expectancy at 55 years, while those who smoked lived between 5.5 and 11.5 years less. This means that with controlled modifiable risk factors, the difference in LE was smaller. T2DM patients who smoked and had an SBP of 120, total/ HDL cholesterol ratio ranging from 4 to 8, aged 55 years and with HbA_{1c} ranging from 6%-10% were found to live 2 years less, on average, than non-smoking T2DM patients (Leal et al., 2008).

3.2.5 Effect of Ethnicity on Mortality among T2DM Patients

The study by Wright et al. (2016) used abridged period life tables based on Chiang II method to estimate the life expectancy of T2DM and non-T2DM patients. Wright et al. (2016) found that those of Asian origin had a lower adjusted mortality due to cancer,

aHR of 0.43[95% CI: 0.36-0.51] and respiratory disease mortality, aHR of 0.60[0.48-0.76], compared with diabetic whites. Their study showed that white patients lost more years of life due to T2DM than T2DM patients of Asian origin and black persons. Wright et al. (2016) also found that T2DM patients of Asian origin in the UK had a lower mortality HR followed by Black and then White patients. Interestingly, those of Asian origin aged 65 years and above, with T2DM had up to 1.1 years more of life expectancy than their counterparts without T2DM. At the age of 40, it was reported that white men and women with T2DM loose 5 and 6 years of life expectancy compared to those without T2DM, respectively (Wright et al., 2016). Wright et al. (2016) reported a lower adjusted CVD mortality among T2DM patients of Asian origin (aHR 0.82[0.75-0.89]) compared with white T2DM patients.

3.2.6 Other Cause-Specific Mortality among T2DM Patients

Wright et al. (2016) reported that T2DM persons had an aHR of cancer-specific mortality of 1.63 [95% CI: 1.60-1.67], respiratory diseases (1.84 [95% CI: 1.79-1.89]), diseases of the nervous system (1.48 [95% CI: 1.39-1.58]) and diseases of digestive system (2.16 [95% CI: 2.06-2.27]), compared to non-T2DM patients.

3.3 Conclusion

All the existing relevant studies estimated an increased all-cause mortality risk in patients with T2DM compared to people without diabetes using observational data. All the studies in Table 3.1 matched T2DM to people without diabetes except for Kim et al. (2017). However, with the exception of two studies, previous studies did not include the general practice in matching study patients. Thus, they did not account for risks shared by patients from the same general practice. Lind et al. (2013)'s unmatched study resulted in having older T2DM patients compared to non-diabetics which may have had an effect on the increased HR. With the exception of Kim et al. (2017), all the studies

in Table 3.1 did not adjust for comorbidities. The only adjusted for variables were age, gender, smoking status or entry year. This can cause the overestimation of the impact of T2DM on all-cause mortality. In addition, studies by Mulnier et al. (2006) and Taylor et al. (2013) did not adjust for the duration of diabetes an important variable as their selection was not at the incidence of T2DM. Severe medical conditions were included in some studies without adjusting for them which could potentially have caused the HR estimates to be overestimated. The differences in estimated HR across studies are as a result of differences in the study population, study period and medical advancements, and poor adjustment. These gaps could have produced biased estimates of the HRs. This study will adjust for a number of socio-demographic, lifestyle and medical conditions (excluding severe conditions) and perform time-variant survival analysis that was not used in previous studies to estimate the impact of T2DM on all-cause mortality and life expectancy.

Chapter 4

Statistical Methods

4.1 Survival Analysis

Survival analysis is a statistical technique used for analysing time-to-event data on study subjects $i = 1, 2, 3, \cdots, n$, who are followed up over a period from time of entry up to τ_{max}, or until an exit occurs at t_i. Though the origins of survival analysis are attributed to the early developments of life tables, it has been used mostly in engineering (Miller et al., 1998), medical and health fields. Exits are in the form of an event of interest occurring or censoring.

Censoring means that a subject's survival time is not known or incomplete. Due to censoring, classic models (such as standard ordinary linear models) are not appropriate as they can not handle it. Subjects can be censored in three ways: left, interval and (or) right censoring. Subsequent chapters will discuss both censored and uncensored time-to-event data.

In survival analysis, the dependent variable consists of two elements: (1) follow-up time, t_i, and (2) exit status, $\zeta_i = \{0, 1\}$, where 0 and 1 denote censoring or an event occurring, respectively. Subjects that outlive the maximum follow-up time become right censored at τ_{max}. In survival analysis it is of interest to describe the association of

a factor(s) of interest (e.g. diabetes) to the time to event, in the presence of several covariates, such as age, sex and smoking. There are a number of models that can be used to analyse the relationship of a set of covariates with the survival time. These techniques include parametric, semi-parametric and non-parametric methods. However, before we discuss these survival methods we shall discuss survival and hazard functions as they are key in survival analysis for describing survival time distributions.

4.1.1 Survival Function

Survival function calculates the probability that an event of interest has not occurred by a given time, t. Let $T \in [0, \infty)$ be a continuously distributed random variable (r.v.) that denotes the time of event of interest with cumulative distribution function (c.d.f) $F_T(t)$ defined as,

$$F_T(t) = \mathbb{P}\{T \leq t\} = \int_0^t f(s)\, ds \quad \forall\, t \geq 0,$$

where $f(t)$ is the respective probability density function (p.d.f). Its survival function, $S(t) = (S_T(t)$, is given by

$$S_T(t) = 1 - F_T(t) = \mathbb{P}\{T > t\} = \int_t^\infty f(s)\ ds \quad \forall\, t \geq 0.$$

For a positive-valued r.v. T, $F_T(0) = 0$ and $F_T(\infty) = 1$, implying that $S_T(0) = 1$ and $S_T(\infty)$=0, respectively (Bowers et al., 1997). Thus $S_T(t)$ is a monotonous non-increasing function.

4.1.2 Hazard Function

The hazard function or instantaneous rate of occurrence of an event estimates the potential that an event under study will occur, at time t, given that the subject has survived up to t. In actuarial science and demographics it is better known as the *force*

of mortality. It is an alternative characterisation of the distribution of T, defined as,

$$\lambda(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}\{t \leq T < t + \delta t | T \geq t\}}{\delta t}.$$

The numerator can be interpreted as the ratio of the probability that T is in the interval $[t, t+\delta t]$ to the probability of $T \geq t$ for a small δt. The first probability can be represented mathematically by $f(t)\delta t$ and the second as $S(t)$. This reduces $\lambda(t)$ to,

$$\lambda(t) = \lim_{\delta t \to 0} \frac{f(t)\delta t}{S(t)} \cdot \frac{1}{\delta t} = \frac{f(t)}{S(t)}. \tag{1}$$

Since $S(t) = 1 - F_T(t)$, it is easy to see that $\lambda(t)$ is the first derivative of $\log(1 - F_T(t))$. Thus, $\lambda(t) = -\frac{\partial}{\partial t} \log(S(t))$. A related quantity is the cumulative hazard function (c.h.f) $\Lambda(t)$. For a continuous r.v. T,

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -\log(S(t)), \quad \forall\, t > 0.$$

Thus,

$$S(t) = e^{-\int_0^t \lambda(u)\, du} = e^{-\Lambda(t)}. \tag{2}$$

The hazard function, $\lambda(t)$ is non-negative and can be of any shape. In particular, it can be a constant, increasing, decreasing, bathtub or bump shaped function. (Klein and Moeschberger, 2003). If T is a discrete r.v. defined at t_i, $i = 1, 2, 3...$, then $\lambda(t_i)$ is given by,

$$\lambda(t_i) = \mathbb{P}\{T = t_i | T \geq t_{i-1}\} = \frac{q(t_i)}{S(t_{i-1})}, \quad i = 1, 2, 3...$$

where $q(t_i) = S(t_{i-1}) - S(t_i)$. The value $q(t_i)$ can be interpreted as the probability of dying between time t_{i-1} and t_i after surviving for at least t_{i-1}. Hence $\lambda(t_i)$ can be written as,

$$\lambda(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})}. \tag{3}$$

Taking note that $S(t)$ can be defined as a product of conditional survival probabilities,

$$S(t) = \prod_{t_i \leq t} \frac{S(t_i)}{S(t_{i-1})},$$

the survival function of a discrete r.v. can be defined with respect to $\lambda(t)$ as follows (Klein and Moeschberger, 2003),

$$S(t) = \prod_{t_i \leq t} \Big(1 - \lambda(t_i)\Big). \tag{4}$$

4.2 Non-Parametric Models

The hazard function $\lambda(t)$ as defined in section 4.1.2, requires a p.d.f to be known. However, in empirical data analysis, the p.d.f may not be known and needs to be estimated. Several non-parametric estimators are available among which the Kaplan-Meier estimator (KME) is the most popularly used estimator. Another often used estimator is the Nelson-Aalen estimator (NAE), that estimator can be used to estimate the cumulative hazard function $\Lambda(t)$. These estimators can be used to compare survival experience of two or more groups of study subjects and to obtain univariate descriptive statistics for survival data, including the median survival time.

Popular descriptive statistics used in survival analysis include the mean hazard rate, $\bar{\lambda}$. However, $\bar{\lambda}$ does not provide rates at a particular point in time of the study for the comparison of the groups under study. The KME is a solution to the limitations of these descriptive statistics. It estimates the survival probabilities at time point t using the product limit formula.

Let N_i denote the number of subjects at risk and d_i, the number of subjects experiencing an event at time t_i. The KME is defined as

$$\hat{S}(t) = \prod_{t_i \leq t} \Big(1 - \frac{d_i}{N_i}\Big). \tag{5}$$

In the case of uncensored data, $N_i = N_{i-1} - d_{i-1}$ while in the case of censored data, $N_i = N_{i-1} - d_{i-1} - C_{i-1}$ where C_i is the number of censored subjects at a time t_i. The variance of the KME may be derived from several different approaches. One often used approach is the *delta method*. It uses the first-order Taylor series expansion. Hosmer et al. (2008) clearly discussed the computation of the variance of $\hat{S}(t)$ by first log-transforming the KME to give the variance of the log of the survival function as,

$$\widehat{\text{Var}}\Big[\ln\Big(\hat{S}(t)\Big)\Big] = \sum_{t_i \leq t} \frac{d_i}{N_i\Big(N_i - d_i\Big)}. \tag{6}$$

The most used and "traditional" formula derived from another application of the *delta method* calculates the variance of the exponentiated variable. Given a variable X, its mean denoted by μ_X and variance Var(X),

$$\text{Var}\Big(e^X\Big) \approxeq \Big(e^{\mu_X}\Big)^2 \cdot \text{Var}(X).$$

Re-writing $\hat{S}(t)$ as $\exp\big[\ln\big[\hat{S}(t)\big]\big]$ and letting μ_X be approximated by $\ln\big[\hat{S}(t)\big]$ and Var(X) be given by equation (6), the familiar Greenwood's formula (Hosmer et al., 2008; Kleinbaum and Klein, 2012) for the pointwise variance of $\hat{S}(t)$ is obtained:

$$\widehat{\text{Var}}\big[\hat{S}(t)\big] = \big(\hat{S}(t)\big)^2 \cdot \sum_{t_i \leq t} \frac{d_i}{N_i\Big(N_i - d_i\Big)}. \tag{7}$$

The KME is asymptotically normally distributed and we can therefore compute its pointwise confidence interval (CI) using (7) and the $z_{1-\alpha/2}$ critical values. However, it is known that this approach may lead to CIs having values outside the range [0,1]. In addition, Hosmer et al. (2008) reported that the normal approximation may not hold for small to moderate sample sizes. Kalbfleisch and Prentice (2002) solved the problem by suggesting that the CI should be based on taking the log of -log $(\hat{S}(t))$. This

transformation has a range (-∞, ∞) and maps to [0,1]. Letting $\hat{\nu}(t) = \log\left[-\log\left(\hat{S}(t)\right)\right]$,

$$\widehat{\text{Var}}\left[\hat{\nu}(t)\right] = \frac{1}{\left[\log\left(\hat{S}(t)\right)\right]^2} \cdot \sum_{t_i \leq t} \frac{d_i}{N_i\Big(N_i - d_i\Big)}. \tag{8}$$

Using (equation (8)), the CI for $\hat{\nu}(t)$ is given by,

$$\hat{\nu}(t) \pm z_{1-\alpha/2} \cdot \widehat{\text{Var}}(\hat{\nu}(t)).$$

Denoting the lower and upper bounds of this pointwise CI by $\hat{C}_l$ and $\hat{C}_u$, the CI of $\hat{S}(t)$ is found by the back transformation as (Hosmer et al., 2008),

$$\Big(\exp\left(-\exp\left(\hat{C}_u\right)\right), \exp\left(-\exp\left(\hat{C}_l\right)\right)\Big). \tag{9}$$

In non-parametric survival analysis, one has to consider testing the differences of survival functions across p groups. To test $H_0 : S_1 = S_2 = S_3 = \cdots = S_p$, against the alternative $H_1 : S_g \neq S_h$ for some $g \neq h$, g, h =1,2,$\cdots$,p, the log-rank statistic (LRS) is used with one group arbitrarily chosen as the reference group. Let O_g and E_g be the numbers of observed and expected events in group g, under H_0, $t_{g(w)}$, w =1, 2, $\cdots$,k, be the distinct failure times and $n_{g(w)}$, $d_{g(w)}$ and $e_{g(w)}$ be the risk set, number of observed failures and expected number of failures in the gth group at the wth ordered failure time, respectively. Assuming the same survival in p groups and denoting $n_{(w)} = \sum_{g=1}^{p} n_{g(w)}$, $d_{(w)} = \sum_{g=1}^{p} d_{g(w)}$ and $e_{g(w)} = \left(\frac{n_{g(w)}}{n_{g(w)}+n_{g(w)}}\right)(d_{g(w)} + d_{h(w)})$ (Kleinbaum and Klein, 2012). The difference between observed and expected numbers of failures is,

$$OE_g = O_g - E_g = \sum_{w=1}^{k} \left(d_{g(w)} - e_{g(w)}\right).$$

Denote also by $O = (O_1, O_2, \cdots, O_p)$ and $E = (E_1, E_2, \cdots, E_p)$ vectors of observed and expected values of numbers of failures in the p groups and denote,

$$\mathbf{u} = (OE_1, OE_2, \cdots, OE_{p-1})'$$

with $(p-1)\,\mathrm{x}\,(p-1)$ covariance matrix $\mathbf{V}$ constituted by the variance of OE_g, v_{gg}, and covariances $v_{gh} = \mathrm{Cov}(OE_g, OE_h)$ calculated as,

$$v_{gg} = \sum_{w=1}^{k} \frac{n_{g(w)}(n_{(w)} - n_{g(w)})d_{g(w)}(n_(w) - d_(w))}{n_(^2w)(n_(w) - 1)},$$
$$v_{gh} = \sum_{w=1}^{k} \frac{-n_{g(w)}n_{h(w)}d_f(n_(w) - d_(w))}{n_(^2w)(n_(w) - 1)},\ g \neq h$$

then the log-rank statistic (LRS) for p groups is defined by

$$LRS = \mathbf{u'V^{-1}u} \tag{10}$$

and has a χ^2_{p-1} distribution under H_0. The LRS is based on assumptions similar to the Kaplan Meier survival curves: (1) censoring is independent of prognosis, thus, the survival patterns of the groups under study are the same, (2) the survival probabilities are the same for subjects recruited early and late in the study, implying that the hazard functions for the groups are proportional, and (3) the events happened at the times specified.

Other tests used to compare survival curves include the Wilcoxon, Tarone-Ware, Peto and Flemington-Harrington test. These are variations of the LRS derived by applying different weights at the wth event time. The general weighted LRS test is given by (Kalbfleisch and Prentice, 2002),

$$LRS_{weighted} = \frac{\Big(\sum_{w=1}^{k} \eta(t_{(w)})(d_{g(w)} - e_{g(w)})\Big)^2}{\sigma^2\Big(\sum_{\forall w} \eta(t_{(w)})(d_{g(w)} - e_{g(w)})\Big)}.$$

For the alternative test statistics, the weights $\eta(t_{(w)})$ are defined as

Method	$\eta(t_{(w)})$
Wilcoxon	$n_(w)$
Tarone-Ware	$\sqrt{n_(w)}$
Peto	$\tilde{s}(t_{(w)}) = \prod_{t_{(i)} \leq t_{(w)}} (\frac{n_g + 1 - d_g}{n_g + 1})$
Flemington-Harrington	$\hat{S}(t_{(w)})^{\rho} \cdot [1 - \hat{S}(t_{(w)})]^{\gamma}$

When $\eta(t_{(w)})$=1 or $\rho = 0$ and $\gamma = 0$ in the case of the Flemington-Harrington (FH) test, $LRS_{weighted}$ reduces to the LRS test. The Wilcoxon and Tarone-Ware tests puts more weight to the earlier event times. The Peto test uses the survival estimate $\tilde{s}(t_{(w)})$ which is calculated over all the groups under the study, while the FH test uses the KME $\hat{S}(t_{(w)})$. The Peto and FH tests are more flexible than the Wilcoxon and Tarone-Ware tests. In the case of the FH, if $\rho = 1$ and $\gamma = 0$, more weight is assigned to the earlier survival times when $\hat{S}(t_{(w)})$ is closer to 1. If $\rho = 0$ and $\gamma = 1$, more weight is assigned to the later survival times.

The choice of the test statistic depends on what period of study the effect of an exposure is believed to be more pronounced or whether one believes that hazards are non-proportional. Kalbfleisch and Prentice (2002) suggest that one should make an a priori decision on which statistical test to use rather than fish for a desired p-value.

4.3 Semi-Parametric Models

Even though the KME estimates the survival function, it does not adjust for confounding variables. Its limitations can be addressed by semi-parametric models. By design, semi-parametric regression models have a fully parametric regression structure but leave the functional dependence on time unspecified (Hosmer et al., 2008). Of interest in our research is the Cox proportional hazards (PH) model.

4.3.1 Cox proportional hazards (PH) model

The Cox PH model makes fewer assumptions than fully parametric models and it allows for comparisons of survival times of two or more study subject groups while adjusting for several covariates. In addition, unlike the parametric models, it does not make assumptions about the baseline hazard function. The Cox proportional hazard (PH) model assumes that the hazard functions are proportional over time (that is, it assumes a constant relative hazard). If we let $\mathbf{X} = (X_1, X_2, X_3, \cdots, X_p)$ be the covariate matrix in the Cox PH model and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \cdots, \beta_p)$ be the effects of each covariate, then the hazard function at time t is modelled as

$$\lambda(t, \mathbf{X}) = \lambda_0(t) \cdot e^{\mathrm{X}'\boldsymbol{\beta}}. \tag{11}$$

The baseline hazard function $\lambda_0(t)$ is independent of covariates presented on the right hand side of the equation. The advantage of the Cox PH model is that one need not to worry about the baseline hazard $\lambda_0(t)$ when comparing two or more groups as the hazard ratio does not depend on $\lambda_0(t)$. Given two study groups with covariate matrices $\mathbf{X}$ and $\mathbf{X}^*$, the hazard ratio is given by,

$$\hat{\lambda}_{HR} = \exp\left[\sum_{i=1}^{p} \beta_i\big(X_i^* - X_i\big)\right]. \tag{12}$$

The general survival function for the semi-parametric model, given a vector of parameters β'=$(\beta_1, \beta_2, \cdots, \beta_p)$ and matrix of covariates $\mathbf{X}$, is defined by (Hosmer et al., 2008),

$$S(t, \mathbf{X}, \boldsymbol{\beta}) = e^{-r(\mathrm{X}, \boldsymbol{\beta}) \cdot \Lambda_0(t)} = \big[e^{-\Lambda_0(t)}\big]^{r(\mathrm{X}, \boldsymbol{\beta})} = \big[S_0(t)\big]^{r(\mathrm{X}, \boldsymbol{\beta})}, \tag{13}$$

where $r(\mathrm{X}, \boldsymbol{\beta})$ is a regression function for covariates $\mathbf{X}$ and $\Lambda_0(t)$ is the cumulative baseline hazard function. The Cox model has the regression function $r(\mathrm{X}, \boldsymbol{\beta}) = \exp(\mathrm{X}'\boldsymbol{\beta})$, so that equation (13) becomes,

$$S(t, \mathbf{X}, \boldsymbol{\beta}) = \big[S_0(t)\big]^{\exp\left(\mathrm{X}'\boldsymbol{\beta}\right)}. \tag{14}$$

To estimate the coefficients $\boldsymbol{\beta}$, Cox(1972) proposed the use of *partial likelihood function.* This function, assuming that there are no time-tied events, is defined by

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp\left(\mathrm{X}_i'\boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathrm{X}_j'\boldsymbol{\beta}\right)} \right]^{\zeta_i},$$

where $R(t_j)$ is the risk set at time t_j. The above equation can be modified to include subjects with $\zeta_i = 1$ only, giving

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\mathrm{X}_i'\boldsymbol{\beta}\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\mathrm{X}_j'\boldsymbol{\beta}\right)}$$

with m denoting the number of failures and for ordered failure times $t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(\omega)}$, with $t_{(\omega)}$ being the last failure time. Differentiating $l_p = \log(L_p)$ with respect to $\beta_k, k = 1, 2, \cdots, p$ yields (Hosmer et al., 2008),

$$\begin{aligned} U(\boldsymbol{\beta}) &= \frac{\partial l_p(\boldsymbol{\beta})}{\partial \beta_k} \\ &= \sum_{i=1}^{m} \left\{ x_{ik} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} \cdot exp\mathrm{X}_j' \cdot \boldsymbol{\beta}}{\sum_{j \in R(t_{(i)})} exp\mathrm{X}_j' \cdot \boldsymbol{\beta}} \right\} \\ &= \sum_{i=1}^{m} \left\{ x_{(ik)} - \bar{x}_{w_i k} \right\}, \end{aligned} \tag{15}$$

where $x_{(ik)}$ denotes the value of covariate x_k for the subject i with observed ordered survival time $t_{(i)}$ and

$$\begin{aligned} \bar{x}_{w_i k} &= \sum_{j \in R(t_{(i)})} \eta_{ij}(\boldsymbol{\beta}) \cdot x_{jk}, \text{ where} \\ \eta_{ij}(\boldsymbol{\beta}) &= \frac{e^{\mathrm{X}_j' \cdot \boldsymbol{\beta}}}{\sum_{j \in R(t_{(i)})} e^{\mathrm{X}_j' \cdot \boldsymbol{\beta}}}. \end{aligned}$$

Through numerical methods such as the Newton-Raphson algorithm, β can be obtained through n iterations as:

$$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} + \widehat{\boldsymbol{Var}}(\hat{\boldsymbol{\beta}}^{(n)}) \boldsymbol{U}(\boldsymbol{\beta}^{(n)}).$$

The iterative process starts by initialising $\hat{\boldsymbol{\beta}}^{(0)} = 0$. The process is repeated until there is stability in the partial log-likelihood; $l_p(\hat{\boldsymbol{\beta}}^{(n+1)}) = l_p(\hat{\boldsymbol{\beta}}^{(n)})$ .The variance of estimates $\hat{\beta}$ is computed from the *observed information matrix*, $\mathbf{I}(\boldsymbol{\beta})$, evaluated at $\hat{\beta}$. $\mathbf{I}(\boldsymbol{\beta})$ is a p x p matrix defined as,

$$\mathbf{I}(\boldsymbol{\beta}) = -\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}, \tag{16}$$

whose diagonal elements are computed from

$$\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} \eta_{ij}(\boldsymbol{\beta}) \big(x_{jk} - \bar{x}_{w_i k}\big)^2$$

and off-diagonal entries are given by

$$\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_l} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} \eta_{ij}(\boldsymbol{\beta}) \big(x_{jk} - \bar{x}_{w_i k}\big)\big(x_{jl} - \bar{x}_{w_i l}\big).$$

Hence the covariance matrix of the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is given by the inverse of $\mathbf{I}(\hat{\boldsymbol{\beta}})$ given in equation (16) as

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}.$$

To assess the significance of the model, several methods can be used which include the Wald test, the score test and the partial log-likelihood-ratio (LR) test. In LR test, a comparison is made of nested models, l_p and l_{p-k} to assess the significance of including k covariates in the model, given $p - k$ covariates. The LR statistic is calculated as

$$\mathrm{LR} = 2\big[l_p(\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}) - l_{p-k}(\hat{\boldsymbol{\beta}}_{\boldsymbol{p-k}})\big]. \tag{17}$$

The LR has a χ^2_k distribution with k degrees of freedom.

4.3.2 Assessment of the Cox Model

In modelling the data, the main goal is to develop a model that best fits the empirical data. Methods used to assess the adequacy of a fitted PH model are similar to those for other regression models. These include *subject-specific diagnostics evaluations on leverage and influence on the fit of the PH model* and *computation of the goodness-of-fit measures*. Additionally, the PH assumption needs to be tested.

Residuals are central in assessing the adequacy of regression models. In survival analysis, the definition of a residual is different from the other regression models due to censoring, time-to-event as the outcome variable and the use of partial likelihood. This makes it more difficult than in other regression models. However, several residual definitions have been developed to evaluate aspects of a fitted PH model.

The *Schoenfeld residuals*, $\hat{r}_{ik}$, developed by Schoenfeld in 1982, of the kth covariate for the ith subject are derived from equation (15) again are given by

$$\hat{r}_{ik} = \begin{cases} x_{(ik)} - \bar{x}_{w_i k}, & \eta_i = 1 \\ 0, & \eta_i = 0 \end{cases}, \tag{18}$$

$$\text{where } \bar{x}_{w_i k} = \frac{\sum_{j \in R(t_{(i)})} x_{jk} \cdot \exp\left(\mathrm{X}_j' \cdot \hat{\boldsymbol{\beta}}\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\mathrm{X}_j' \cdot \hat{\boldsymbol{\beta}}\right)}$$

The Schoenfeld residuals, $\hat{r}_{ik}$, are defined on the non-censored subjects just as the components of the likelihood equations of $\hat{\boldsymbol{\beta}}$. Therefore, for the residuals,

$$\sum \hat{r}_{ik} = 0 \ ,$$

since $\hat{\boldsymbol{\beta}}$ is calculated by equating equation (15) to zero. Grambsch and Therneau (1994) introduced the *scaled Schoenfeld residuals* calculated from scaling the Schoenfeld residuals by their variances. The scaled Schoenfeld residuals provide a better diagnostic power. Let $\hat{\boldsymbol{r}}_{\boldsymbol{i}}' = (\hat{r}_{i1}, \hat{r}_{i2}, \cdots, \hat{r}_{ip})$ be the vector of p Schoenfeld residuals defined in equation (18)

for the ith subject. Given a p x p covariance matrix of the residuals for the ith subject, the vector of the scaled Schoenfeld residuals is given by

$$\hat{\boldsymbol{r}}_i^* = \left[\widehat{\text{Var}}(\hat{\boldsymbol{r}}_i)\right]^{-1} \cdot \hat{\boldsymbol{r}}_i$$

Grambsch and Therneau (1994) also suggested that $[\hat{\text{Var}}(\hat{\boldsymbol{r}}_i)]^{-1}$ can be approximated by $m \cdot \hat{\text{Var}}(\hat{\boldsymbol{\beta}})$. With m being the number of failing subjects, they proposed to have

$$\hat{\boldsymbol{r}}_i^* = m \cdot \hat{\text{Var}}(\hat{\boldsymbol{\beta}}) \cdot \hat{\boldsymbol{r}}_i. \tag{19}$$

There are also other types of residuals that can be used, which include the *martingale, deviance* and *score* residuals. For a clear understanding of these one can refer to Hosmer et al. (2008).

Assessing the PH Assumption

The PH assumption is important in the interpretation and use of the fitted PH model. To assess the model, Hosmer et al. (2008) recommended (1) computation of covariate-specific tests and (2) scaled Schoenfeld residuals plots. The results of (1) and (2) should support each other. The evaluation examines the extent to which $\log\left[\Lambda(t, \mathbf{X}, \boldsymbol{\beta})\right]$ functions of g groups are equidistant from each other over time, t. The proportional hazard (PH) assumption defines the model as a function of time and states that only the baseline hazard is important in assessing the variation over time (Hosmer et al., 2008). Grambsch and Therneau (1994) proposed an alternative to the Cox model that allows for the effect of covariates to change over time. Coefficients of the Cox model are modified by,

$$\beta_j(t) = \beta_j + \gamma_j g_j(t)\,, \tag{20}$$

where $g_j(t)$ is a specified function of time, t. Using equation (20), Grambsch and Therneau (1994) showed that scaled Schoenfeld residuals in equation (19), for the jth

covariate at time t, have a mean approximated by,

$$\mathbb{E}\big[r_j^*(t)\big] = \gamma_j \cdot g_j(t).$$

Plotting $r_j^*(t)$ over time may be used to visually assess if $\gamma_j = 0$ and if not, what form of time dependence $g_j(t)$ may have. Function $g_j(t)$ may be defined as $t, \ln(t), \hat{S}_{KM}(t)$ or $rank(t)$, where $\hat{S}_{KM}(t)$ is the KME of the survival function. The PH assumption is satisfied if residuals against $g_j(t)$ are randomly scattered around zero. When the PH assumption has been violated, extended Cox models can be used.

Extended Cox models

Extended Cox models include the stratified models and time-variant effects models with or without time-variant covariates. In a stratified Cox model, violating covariates are used for stratification. In the case of more than one violating variable, these variables can be combined into one new covariate that is then used for stratification. This will result in g hazard functions where g is the number of levels in the stratified covariate.

Time-variant effects Cox models are based on effects that change with time. The assumed function of time is informed by scaled Schoenfeld residuals plots. For the time-variant covariates Cox model, each covariate is updated at selected points in time during the follow-up period. All these and other available models that include accelerated failure time models are discussed in detail by Hosmer et al. (2008),Klein and Moeschberger (2003) and other survival analysis books.

4.3.3 Estimation of the Baseline Survival and Hazard Functions

Most software packages provide the estimation of the baseline survival function, $S_0(t)$. According to Hosmer et al. (2008), the method used is to mimic the arguments that

lead to the KME of the survival function. Thus, the quantity $\hat{\alpha}_i = 1 - d_i/n_i$ is the key estimator of the conditional survival function at observed ordered times $t_{(i)}$. Using the fact that $\alpha_i = S(t_i|t_{i-1}) = S(t_i)/S(t_{i-1})$ and denoting $\alpha_{0i} = S_0(t_i)/S_0(t_{i-1})$ we have,

$$\frac{S(t_{(i)}, \mathbf{X}, \boldsymbol{\beta})}{S(t_{(i-1)}, \mathbf{X}, \boldsymbol{\beta})} = \left[\frac{S_0(t_{(i)})}{S_0(t_{(i-1)})}\right]^{\exp(\mathrm{X}'\boldsymbol{\beta})} = \alpha_{0i}^{\exp(\mathrm{X}'\boldsymbol{\beta})}.$$

Using the maximum likelihood methods conditioned on $\hat{\boldsymbol{\beta}}$ and letting $\hat{\theta}_l = e^{\mathbf{X}'\hat{\boldsymbol{\beta}}}$, the conditional baseline survival probability is estimated by solving,

$$\sum_{l \in D_i} \frac{\hat{\theta}_l}{1 - \alpha_{0l}^{\hat{\theta}_l}} = \sum_{l \in R_i} \hat{\theta}_l, \tag{21}$$

where D_i denotes subjects at risk with survival times, equal to t_i and R_i denoting the subjects at risk at the ordered survival time t_i (Hosmer et al., 2008). In the case of no tied survival times, the solution to equation (21) is given by,

$$\hat{\alpha}_{0i} = \left[1 - \frac{\hat{\theta}_i}{\sum_{l \in R_i} \hat{\theta}_l}\right]^{\hat{\theta}_l^{-1}}.$$

In the case of tied survival times, Equation (21) is solved by iterative methods. The estimator of the baseline survival function is then given by,

$$\hat{S}_0(t) = \prod_{t_{(i)} \leq t} \hat{\alpha}_{0i}.$$

Baseline hazard functions can also be estimated from the conditional survival probabilities as,

$$\hat{h}_0(t_{(i)}) = 1 - \hat{\alpha}_{0i},$$

but these individual pointwise estimators are typically very unstable. The solution to this is the cumulative baseline hazard function because it is less noisy. It is computed as,

$$\hat{H}_0(t) = -\ln[\hat{S}_0(t)],$$

and

$$\hat{H}(t, \mathbf{X}, \hat{\beta}) = -\ln\left[\hat{S}(t, \mathbf{X}, \hat{\beta})\right] = -e^{\mathrm{X}'\hat{\boldsymbol{\beta}}} \ln\left[\hat{S}_0(t)\right].$$

4.4 Parametric Survival Models

These models make use of classic parametric distributions with parameters dependent on selected covariates. Usually the shape parameter is kept constant while the scale parameter depends on the covariates. This general approach assumes time-invariant effects. For time-varying effects, the shape depends on covariates whose effects are time-variant. Generally, any distribution defined for $t \in \{0, \infty\}$ can be a survival distribution. For distributions defined on $z \in \{-\infty, \infty\}$, survival time can be defined as $t = \exp(z)$ (Rodrıguez, 2010). Table 4.1 gives the survival and hazard functions for the exponential, Weibull, Gompertz and log-logistics distributions often used in survival applications.

Table 4.1: Survival and Hazard Functions

Distribution	S(t)	$\lambda(t)$
Exponential	$e^{-\alpha t}$	α
Weibull	$e^{-(\alpha t)^{\upsilon}}$	$\alpha^{\upsilon}\upsilon t^{\upsilon-1}$
Gompertz	$e^{-\upsilon(e^{\alpha t}-1)}$	$\upsilon\alpha e^{\alpha t}$
Log-Logistic	$[1+(t/\alpha)^{\upsilon}]^{-1}$	$1 - \frac{(\upsilon/\alpha)(t/\alpha)^{\upsilon-1}}{(1+(t/\alpha)^{\upsilon})}$

4.5 Survival Models with Frailty

Frailty models are extensions of PH models which incorporate unobserved random effects. These models add a multiplicative frailty random effect to the PH model. The frailty random effect is usually drawn from power variance distribution functions. These distributions are a subfamily of the exponential family and they have a special mean-variance relationship. Given a r.v. Z from a power variance distribution $G(z)$ having parameters $\mu = \mathbb{E}[Z]$ and $\phi = \sigma^2$ the dispersion, its mean-variance relation is of the

form $\text{Var}(Z) = \phi\mu^{\tau}$. The parameter $\tau \in \{-\infty, \infty\}$ is the power parameter. A univariate frailty in the Cox model setting can be introduced as,

$$\begin{aligned}\Lambda(t|\mathbf{X}, \boldsymbol{\beta}, Z) &= Z\Lambda(t)e^{\mathbf{X}'\boldsymbol{\beta}} \\ &= Z\Lambda(t|\mathbf{X}).\end{aligned}$$

Then the population survival function can be calculated from the conditional expectation

$$\begin{aligned}\hat{S}(t|\mathbf{X}) &= \mathbb{E}\big[\hat{S}(t|\mathbf{X}, \boldsymbol{\beta}, Z)\big] \\ &= \mathbb{E}\big[e^{-Z\Lambda(t|\mathbf{X})}\big] \\ &= L_{trans}(\Lambda(t|\mathbf{X})),\end{aligned} \tag{23}$$

where $L_{trans}(\Lambda(t|\mathbf{X}))$ is the Laplace transform of Z at $\Lambda(t|\mathbf{X})$. The gamma distribution is a popular choice for frailty distribution $G(z)$, with shape parameter θ and scale parameter θ for identifiability, because it is convenient for computational and analytical analysis, as it is easy to derive the survival, p.d.f and the hazard function in closed form. Its mean $\mathbb{E}(Z) = 1$ and $\text{Var}(Z) = 1/\theta$. However, it assumes late time events.

Under the frailty model, an individual is at high risk of an event if $Z > 1$ and lower risk if $Z < 1$. In a shared frailty model, every subject i in cluster j has a shared risk Z_j. When frailty latent variable Z_j follows the $Gamma(\theta, \theta)$ distribution, the conditional survival marginal function is

$$\hat{S}(t|\mathbf{X}) = \big[1 + \frac{1}{\theta}\Lambda(t|\mathbf{X})\big]^{-\theta}. \tag{24}$$

The estimates of θ,β, $\Lambda(t|\mathbf{X})$ are then obtained using the expectation maximisation (EM) algorithm. EM algorithm is the main method used for estimation in frailty models under frequentist approach (Wintrebert, 2007). It also provides means of maximizing complex likelihoods.

4.6 Parametric Double-Cox Model with Frailty

In this model, the latent variable, Z, is assumed to be from a gamma distribution and the hazard function, $\lambda(t)$, is from a Gompertz or Weibull parametric distribution. In the study, $\lambda(t)$ was assumed to follow the Gompertz distribution, selected as the "best" fit, with υ and α as shape and scale parameters. Denoting

$$\alpha^* = \alpha e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{scale}}} \text{ and } \upsilon^* = \upsilon e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{shape}}},$$

the cumulative hazard $\Lambda(t|\mathbf{X})$ is

$$\Lambda(t|\mathbf{X}) = \frac{\alpha^*}{\upsilon^*}(e^{\upsilon^* t} - 1).$$

Thus, the model can be written as follows,

$$\Lambda(t|\mathbf{X}, \mathbf{Z}) = \mathbf{Z}\Lambda(t|\mathbf{X}) \tag{25}$$

Parameters of the Cox model with frailty are estimated through the marginal likelihood function which can be written as (Wintrebert, 2007),

$$L_m = \prod_{j,i} \left[\lambda(t_{ji}|\mathbf{X}_{ji})\right]^{\delta_{ji}} \cdot \frac{\theta^{D_j+2\theta} \cdot \Gamma(D_j + \theta) \cdot [L_{trans}(\Lambda(t_{ji}|\mathbf{X}_{ji}))]^{-D_j}}{\Gamma(\theta)},$$

for each subject i in cluster j with $D_j = \sum_i \delta_{ji}$ where $\delta_{ji} \in \{0, 1\}$ is an event indicator. This model with re-parametrised parameters given above was first introduced by Begun et al. (2019) who then developed an **R** package for the model that can be accessed on the GitHub (Begun, A. and Begun, F., 2022). By re-parametrising the scale parameter, Begun et al. (2019) allowed for the inclusion of variables with time-varying effects in their model. Their program uses the bootstrap method to calculate the CIs of the estimates. It currently models hazards with Weibull or Gompertz functions.

4.7 Life expectancy calculation

By their design, survival models calculate survival probabilities conditioned on the individual being alive at the time of entry. Thus, $\hat{S}(t, \mathbf{X}, \boldsymbol{\beta}) \approxeq 1$ at entry. By definition, life expectancy, $\mathring{e}_x$, is the expected remaining life of a person at a given age. It is the integral of the conditional survival probabilities up to the maximum age, ω. Mathematically,

$$\mathring{e}_{(t,\mathbf{X},\boldsymbol{\beta})} = \frac{\int_t^{\omega} \hat{S}(t, \mathbf{X}, \boldsymbol{\beta})\, dt}{\hat{S}(t, \mathbf{X}, \boldsymbol{\beta})}.$$

The calculation of the life expectancy was implemented in the **R** Shiny App that I developed which can be accessed online (Ncube, N, 2022). The application makes use of predefined **R** functions such as *integrate*. Life expectancy and survival probabilities ratios can then be calculated by dividing the respective values for people with T2DM and without diabetes.

4.8 Model Assessment

The assessment of the overall performance, discrimination and external validations of the final models was performed using Harrell's concordance statistic, Negative likelihood, and Akaike information criteria (AIC). In reality, goodness-of-fit statistics may be highly sensitive to influential data points and hence it is a good practice to assess models using more than one statistic. There are several goodness-of-fit statistics that can be used, but the study has made use of these three because they are widely used and are the only goodness-of-fit statistics implemented in the parametric-Double-Cox model with frailty package by Begun, A. and Begun, F. (2022).

4.8.1 Harrel's concordance statistic

The Harrel's concordance statistics, C_H, in survival Cox models was popularised by Harrell Jr et al. (1996) and it has been reported to be the most frequently used measure of goodness-of-fit in survival models (Therneau and Atkinson, 2020). The concordance is defined as the fraction of concordant pairs. Thus, it is the probability that predicted values are ordered in the same direction as the observed values for all combinations of pairs of subjects. By interpretation, for a Cox model if the risk score $\mathbf{X}\boldsymbol{\beta}$ is lower, the predicted survival $\hat{\phi}$ is longer and hence the definitions of concordant and discordant pairs are flipped. Let C, D, T_x, T_y and T_{xy} be a count of pairs that are concordant, discordant pairs, tied pairs on the predictor values (and not on observed values), tied pairs on observed but not predictor values and tied on both, respectively. Using Somers' d statistics

$$d = \frac{C - D}{C + D + T_x},$$

the concordance is defined as $(d + 2)/2$ with $d \in \{0, 1\}$. A concordance of 1 stands for a perfect discrimination between two randomly selected subjects (Therneau and Atkinson, 2020). A value below 0.5, over 0.7 or over 0.8 indicates a very poor, good or strong model, respectively. C_H is asymptotically normally distributed and its 95% CI can be calculated as $\hat{C}_H \pm 1.96 * \sigma_H/\sqrt{n}$ where n is the sample size (Uno et al., 2011). The Somers' d ignores ties in observed values but gives a score of $\frac{1}{2}$ to pairs tied with predictor values and not tied in observed values. In survival analysis, pairs that cannot be ranked with certainty are also ignored, for example censored people at a given time. For stratified models, observations in different strata(s) are also ignored. One of the advantages of this statistic is that, it is well defined not just for survival models, but for logistic and ordinary linear regression also (Therneau and Atkinson, 2020).

4.8.2 Negative Log-likelihood

The negative log-likelihood (negLogLik) is a cost function that is used as a loss for machine learning models. It can be used to assess the performance of the model. This is calculated by multiplying the partial log-likelihood explained in Section 4.3.1 by -1. A penalisation ψ (a positive tuning parameter) can be added in the loss function in minimising the negative log-likelihood to estimate effects as $\min(-l_p(\beta) + \frac{\psi}{\beta})$. The lower the negLogLik, the better model performance when compared to another comparable model. Thus, the less the number of censored subjects, the better the predictive power of the model.

4.8.3 AIC

AIC is the other mathematical method used to evaluate how well a model fit the data. AIC is defined as (Akaike, 1974),

$$AIC = 2(-l_p + p),$$

where p, the penalisation factor, is the number of coefficients estimated. AIC becomes non-zero as the sample size increases. The lower the AIC the better the model when compared to another model.

4.9 Handling Missing Data

Data usually include missing values of importance to the modelling of specific risk factors. This section discusses several methods that have been used to handle missing data. Addressing the challenges of missing data began around 1987, although there were influential attempts before then (Graham, 2009).The first two monographs on the subject were written by Roderick and Rubin (1987) and Rubin (1987) (Graham, 2009).

The underlying motive of data imputation is to improve uncertainty and reduce biases. However, before any data imputation, it is important to understand why data is missing.

4.9.1 Missing Data Mechanisms

Knowing why data are missing helps in choosing the proper data imputation method. Let $\mathbf{Y}$ be a matrix with p covariates for n subjects. Define a missing values indicator n x p matrix, $\mathbf{M}$, with entries $m_{ik} \in \{0, 1\}$ for each k^{th} variable, $k = 1, 2, 3 \cdots, p$ for the i^{th} subject, where 0 indicates a missing value and observed and missing data by $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{miss}$, respectively, such that $\mathbf{Y} = \mathbf{Y}_{obs} \cup \mathbf{Y}_{miss}$. There are three main reason why datasets may have missing values.

Missing completely at random (MCAR)

Data is said to be MCAR if and only if the probability of missing values is the same for all cases. This implies that the causes of the missing data are unrelated to the data. Thus,

$$\mathbb{P}(\mathbf{M} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}) = \mathbb{P}(\mathbf{M} = \mathbf{0}).$$

Missing at random (MAR)

Data is said to be MAR if and only if the probability of missing values depends on other observed variables,

$$\mathbb{P}(\mathbf{M} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}) = \mathbb{P}(\mathbf{M} = \mathbf{0}|\mathbf{Y}_{obs}).$$

Missing not at random (MNAR)

Data is said to be MNAR if the probability of missing values depends on the behaviour of the subjects even after other observed variables are taken into account. Thus,

$$\mathbb{P}(\mathbf{M} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}) = \mathbb{P}(\mathbf{M} = \mathbf{0}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}).$$

4.9.2 Techniques for Handling Missing Data

Excluding Missing Data

It is the most commonly used technique. It includes complete-case analysis and available-case analysis. In complete-case analyses, all cases with missing data are excluded. It can be used without loosing any information if data is MCAR and missing data is less than 5% (Graham, 2009; Rubin, 1987; Roderick and Rubin, 1987). However, it may lead to loss of power and to biases if the subjects with missing data differ systematically from subjects with non-missing data. More cases with missing data are inevitably excluded in models with many variables. Graham (2009) promotes the use of better methods regardless of proportion of missing values.

In available-case analysis, different aspects of a study are analysed using different subsets of the dataset. A drawback of this method is possible inconsistencies in the different analyses. This method also arises when variables with missing data are excluded from the analysis (*complete-variable analysis*). This may lead to omission of important variables in the study.

Simple approaches that return all data

Rather than excluding subjects with missing data or excluding variables with missing data, simple imputation can fill missing values. Methods include imputation using mean or median, last-value carried forward, information from related observations or nearest-

neighbourhood information, indicator for missing unordered categorical variable data and imputations based on logical rules. This may result in distortion of the distribution of imputed variables.

Random Imputation of one variable

1. **Simple random imputation** This method ignores useful information from other variables in $\mathbf{Y}$ and imputes missing data in a variable $Y_{miss} \subset \mathbf{Y}$ using randomly selected values from its observed set of values.

2. **Deterministic imputation by regression** This method regresses Y_{miss} on other fully observed variables on $\mathbf{Y}_{obs}$. The missing data is then imputed using the predicted values from the resulting model.

3. **Random regression imputation** This method builds up from the deterministic imputation by regression approach. This method handles uncertainty by adding the prediction error into the model.

Imputation of several missing variables

In reality, more than one variable may have missing values. Hence the values of all variables need to be viewed as a multivariate outcome whose elements are subject to missing data. Routine multivariate and iterative regression imputations are methods used in this technique.

1. **Routine multivariate imputation** imputes missing data in several variables by fitting a multivariate model. In practice, readily available models are used such as multinomial distribution or t-distribution for continuous outcomes.

2. **Iterative regression imputation** builds on the univariate methods discussed in the previous section by performing iterative regressions on the variables with missing data. The imputation is performed by first initialising the missing data in $\mathbf{Y}_{miss}$

and then regressing Y_i by other variables in $\mathbf{Y}_{miss}$ and $\mathbf{Y}_{obs}$. Missing predictor values in $\mathbf{Y}_{miss}$ are replaced with current imputed data in subsequent regressions until approximation converges. This method allows for inclusion of interactions and is easier to understand than most joint modelling methods. However, it requires consistency of separate regressions. Its limitation is that the outcome will not, in general, correspond to any joint probability model in the imputed variables. Simple diagnostics can be made by comparing histograms and scatter plots of the variable's observed and imputed data.

4.9.3 Multiple Imputation (MI)

This method is based on the Monte-Carlo (MC) method and uses observed data to impute missing data. Rather than having one imputed dataset, multiple imputations can be done on the missing data to reflect uncertainty of the model. Statistical analysis are performed on each imputed dataset. The estimates are then pooled together using the Rubin's rules to give the final estimates and their confidence intervals. The two popularly used methods are the joint modeling (JOMO) and fully conditional specification (FCS) multiple imputations.

JOMO assumes that data can be described by a multivariate distribution. The most widely used distribution is the multivariate normal distribution. Since missing data can occur in any of $\mathbf{Y_{miss}}$ variables, the distribution from which imputation is to be drawn differs from row to row. The FCS also known as chained-equations method requires an imputation model for each incomplete variable. It then imputes each incomplete variable iteratively. The MI process can be represented graphically as,

Using the Rubin's method, the estimate $\hat{\beta} = \hat{\beta}(\mathbf{Y_{obs}}, \mathbf{Y_{mis}})$ is obtained from $\mathbf{m}$ imputed data analyses as the mean

$$\hat{\beta}_{MI} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_{\mathbf{i}}.$$

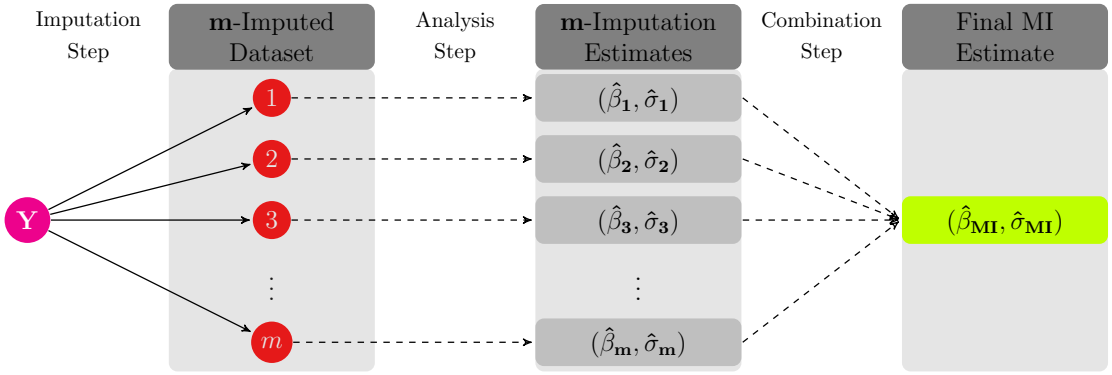


Figure 4.1: Multiple Imputation Algorithm

The variance of $\hat{\beta}_{MI}$ is calculated using estimates of the average of within-imputation variances ($\hat{\mathbf{V}}_i^w$) and the between-imputation variance ($\hat{\mathbf{V}}^b$) given by

$$\bar{\mathbf{V}}^w = \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{V}}_i^w$$
$$\hat{\mathbf{V}}^b = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\beta}_\mathbf{i} - \bar{\beta}\right)^2.$$

The pooled estimated variance of $\hat{\beta}_{MI}$ is given by

$$\hat{\sigma^2}_\mathbf{MI} = (1 + \frac{1}{m})\hat{\mathbf{V}}^b + \bar{\mathbf{V}}^w.$$

The distribution of $\hat{\beta}_{MI}$ is approximated by the **t**-distribution as follows,

$$\frac{(\hat{\beta}_{MI} - \beta)}{\sqrt{\hat{\sigma}_{MI}(\hat{\beta}_{MI})}} \sim \mathbf{t}_v,$$

with the degrees of freedom v given by

$$v = (m-1)\Big[\frac{\hat{\sigma^2}_\mathbf{MI}}{\hat{\sigma^2}_\mathbf{MI} - \bar{\mathbf{V}}^w})\Big]^2.$$

The total pooled variance, $\hat{\sigma^2}_\mathbf{MI} = \mathbf{T}_m$ of the estimated effects on the imputed data is dependent on the number of imputations, m , and the percentage of missing values γ when compared to the ideal total variance, $\mathbf{T}_\infty$ as shown in equation (4.1) (Van Buuren,

2018),

$$\mathbf{T}_m = (1 + \frac{\gamma}{m}) \cdot \mathbf{T}_\infty. \tag{4.1}$$

In each imputation, iterations are performed until it appears a convergence has been achieved. Although no more than 5 iterations are usually necessary, additional iterations can be performed if it appears that the average imputed values have not converged. Accuracy of imputations depend on the information in the dataset. Thus a dataset of completely independent variables with no correlation will yield inaccurate imputations. Earlier researchers recommended 3 to 5 imputations while others suggested 5 to 10 imputations to be sufficient (Austin et al., 2021). However, an analyst is ideally interested in selecting m such that the pooled estimated coefficients and standard errors would not vary much across repeated applications of multiple imputations (Austin et al., 2021). Several imputation packages are available in R software which include **mice, jomo, pan, amelia, mi, Himsc** and **missForest**. In this study the **jomo** packed was used.

4.10 Conclusion

In this chapter, survival techniques relating to this study have been discussed together with imputation and model assessment methods to be used. In particular, the parametric double-Cox model with frailty was discussed and will be used in this study with $\lambda(t)$ being a reparametised Gompertz hazard function with υ^* and α^* as parameters. The following chapter describes the research methodology.

Chapter 5

Study Methodology

5.1 Data Source

This observational matched cohort study is based on the United Kingdom (UK) EHR data from The Health Improvement Network (THIN) database owned by Cegedim. The database was set up in 2002 and stores pseudonymised patient records. These records are regularly collected from all the general practices that use the Vision clinical system that was developed by the In Practice Systems (INPS). The collected data is then submitted to Cegedim and access to the data is provided by Quantiles IMS. The THIN data can be acquired in two ways. The user can either apply for a subset of the database or full database which can be granted after approval and having paid for data. For research users holding the full database, each research study has to be approved by the THIN Scientific Research Council. This research project (THIN Project ID TL038) was approved by the Scientific Research Council (Approval Number 16THIN095). The research protocol is appended in Appendix F. In addition, the study was ATAS cleared (ATAS certificate reference number 282753).

According to QuintilesIMS (2017) as at the end of December 2016, over 730 practices had contributed to THIN database and 385 Vision practices were active. As of January 2017, THIN had data from over 711 practices with a total of 15.6 million patients of

whom over 3 million were registered with active practices and could be prospectively studied (QuintilesIMS, 2017). The remainder are said to have either left the practices or were deceased. Most of these practices have contributed data for more than 20 years and THIN presents a longitudinal view of the population of UK (QuintilesIMS, 2017). Active patients in THIN represent about 6% of the total UK population. The following sub-sections will describe the THIN database schema and discuss its relevance.

5.1.1 Data Structure

THIN is made up of seven main tables and eight look-up tables. Figure 5.1 shows the main seven tables of the THIN database and their description. The eight look-up tables in the THIN database are *staff*, *postcode variable indicators (PVI)*, *pack size*, *drug codes*, *dosage*, *medical codes*, *anonymised comments* and *Additional Health Data (AHD)*. The description of these look up tables is provided in Table D.6.

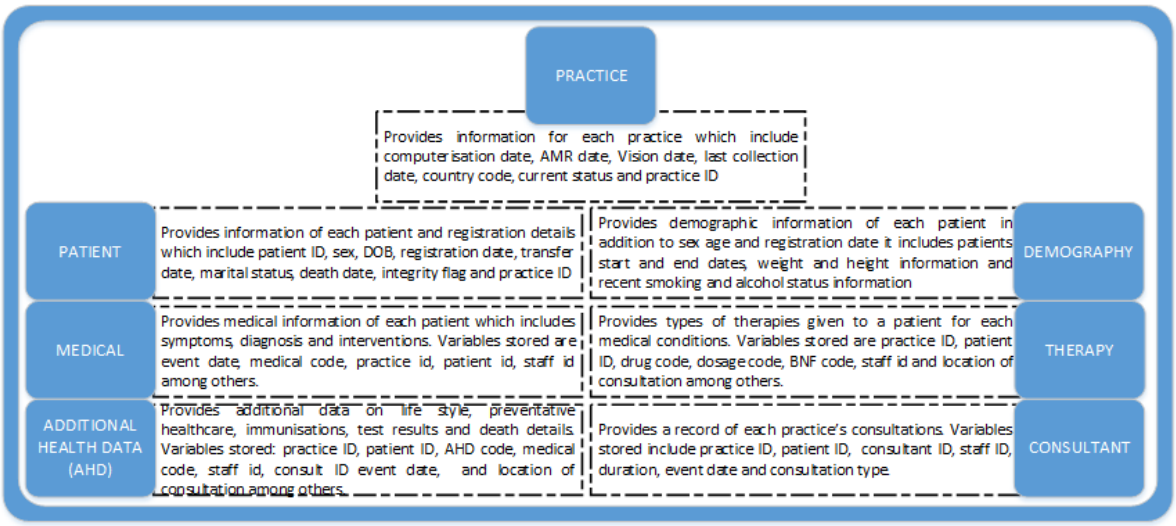


Figure 5.1: Description of the Main Seven Tables in the THIN Database

5.1.2 Relevance of The Health Improvement Network (THIN)

The use of EHRs has been a pivotal advancement in epidemiological studies. Several studies have reported that THIN generalises to the UK population with respect to

demographics and crude prevalence/ rates of major events which include major medical conditions and death (Blak et al., 2011; Lewis et al., 2007). Lewis et al. (2007) in a validation study of THIN reported that the THIN's data collected outside of the General Practice Research Database (GPRD) was as valid as the GPRD data.

In a comparison study between THIN and the UK population, the difference of the total crude prevalence of major conditions was -0.1 percentage points, Blak et al. (2011). The crude prevalence for diabetes mellitus (DM), cancer, coronary heart disease (CHD), chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), dementia, stroke/transient ischemic attack (TIA) and heart failure (HF) were reported to be 3.5%, 0.9%, 2.5%, 1.6%, 0.6%, 1.9% and 0.9% in the THIN database compared, to 3.7%, 0.9%, 3.7%, 2.3%, 1.5%, 0.4%, 1.7% and 0.8% in the UK population, respectively.

According to Blak et al. (2011), the percentage distribution of THIN patients with respect to sex and age group in 2009 was similar to the UK population. For the younger ages it was reported that the maximum difference was about 1 percentage point among males aged 20-24 years. With respect to exit information (deaths and transfers), Hall (2009) reported that THIN death data was reliable. Hall (2009) also pointed out that the transfer-out dates were usually later than death dates. However, they reported that the cause of death was not always recorded electronically. Table D.5 in the Appendix summaries the strengths and weaknesses of using THIN data.

5.1.3 Justification of using THIN

As discussed in the previous sub-section, the THIN database is clearly suitable for the study of longevity and morbidity risks in the presence of type 2 diabetes mellitus (T2DM). The database represents 6% of the UK's population, its strengths outweigh its weaknesses and its population age range is appropriate for studies related to T2DM. In addition, the crude prevalence of the main health conditions is similar to the UK general population. Hence the results of the study will be relevant, reliable and can be

generalised to the UK population.

5.2 Selection Criteria

The subset of THIN database used for this research included records that have been flagged by THIN as acceptable for the purposes of any research. It also includes patients who had acceptable visiting gaps and with a medical record for at least 10 years before their diagnosis at a practice. The dataset included T2DM patients diagnosed between 1 January 1984 and 31 December 2016. Incidence sampling was used in selecting T2DM patients. All medical conditions in this study were identified using THIN "medcodes" extracted from the ClinicalCodes.org website (ClinicalCodes.org, 2021) or from database search informed by the WHO ICD-11 disease classifications (WHO, 2021). The included medical codes for T2DM are presented in Section D.1. Subsequent sections and chapters may use "entry date" to refer to either "diagnosis date" in case of T2DM or "selection date" in case of people without diabetes.

Patients with other severe health conditions such as stroke, heart failure, myocardial infarction, peripheral vascular disease (PVD), lower limb amputations, all forms of cancer, all forms of dementia, all forms of cognitive impairment and CKD stages 3 to 5 before entry date were excluded as they could modify the effects of T2DM on mortality and morbidity. Medical professionals involved in the the research were instrumental in identifying and selecting medical codes for inclusion and exclusion. Patients whose death date, transfer date from GP or acceptable mortality reporting (AMR)[1] date were earlier than the diagnosis date of T2DM were also excluded. Only patients with at least 12 months of registration with a GP were included. T2DM patients were then matched by age, GP and gender to at most three controls as at the date of their diagnosis who met the inclusion and exclusion criteria. Matching was done to broadly balance the

[1]AMR date is a variable date in the THIN database used for the selection of cases which helps reduce potential biases in disease occurrence and guaranteeing that no "immortal" periods are present in the study data.

distribution of people with T2DM to those without diabetes. All selected patients with no matched controls were excluded. The initial total size the study population before further exploratory analysis was 362 082 individuals as shown in Figure C.7. After further elimination, the final total size of the study population is 221 182. The subsequent sections and Chapter 6 explain the reasons of the reduction in the study population.

T2DM patients were matched to a maximum of three people without diabetes and without any of the conditions in the exclusion criteria at the first date of T2DM incidence.

5.2.1 Selection and Information Biases

Cases were matched to controls of the same age, sex and GP to reduce for selection or information bias. It is possible that some data may be missing not at random (MNAR), particularly lifestyle variables, due to patients leaving the GP or not frequently visiting the GP which makes it a challenge to know whether the patient is healthy or receiving treatment. It is near impossible to correct for this. Incidence-prevalence bias, a form of selection bias, was avoided by using identical inclusion and exclusion criteria on T2DM and people without diabetes at study entry.

Exclusion of severe medical conditions diagnosed before study entry was done to reduce their effects on the association of T2DM and all-cause mortality. This was performed on both T2DM and people without diabetes. Disease identification bias, a form of information bias, was avoided by collating relevant medical codes from published previous studies from the www.ClinicalCodes.org website.

In addition, two medical practitioners identified and validated relevant medical conditions from the collated list.Another form of information bias, surveillance bias, was avoided by making use of diagnosis codes and not drug prescription, for example, as some antidiabetes drugs are used for other conditions. As diabetes is an insidious disease, it is possible to have undiagnosed cases who may be identified as people without

diabetes. However, THIN still remains relevant and can be generalised to the UK as explained in Section 5.1.2 and Section 5.1.3.

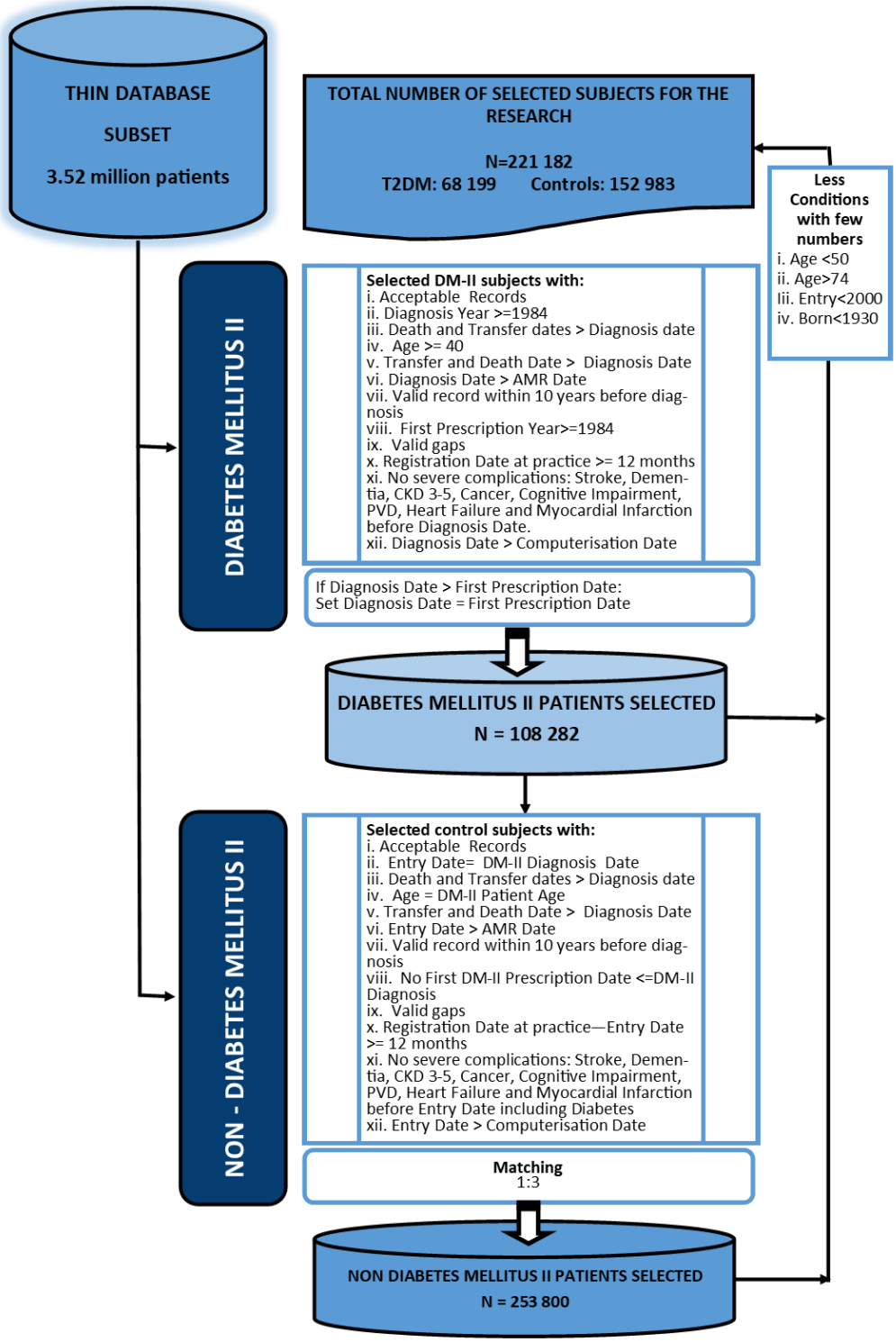


Figure 5.2: Data extraction process and final study population size

5.3 Study Variables of Interest

The variables used in the study included demographic, life style and medical/health factors. These are shown in the Table 5.1 below.

Table 5.1: Study Variables

Group	Variables
Demographic	Age, year of birth, gender, ethnicity and general practice identification
Life Style	body mass index (BMI), Smoking status, alcohol consumption and deprivation index
Comorbidities	Amputation, cancer, cognitive impairment (including dementia), CKD stages 3 to 5, heart failure, hypertension, hypercholesterolaemia, myocardial infarction, PVD and stroke (TIA)
Bio-markers	systolic blood pressure (SBP), distolic blood pressure (DBP), HDL, blood lipid ratio
Drugs	Antidiabetic drugs, antihypertensive drugs and statins
Survival	Time to event, start and stop age, and status

5.3.1 Demographic Covariates

The patients were grouped into three age bands, 40-49, 50-59 and 60+ years. Both females and males were included in the study. The birth cohorts were classified in three decades as 1930-1939, 1940-1949 and 1950-1960. Entry year variable was split into 3 categories, namely, 2000-2004, 2005-2009 and 2010-2016. The practice number is a unique identifier for each general practice in the UK using the Vision software. The GP number was included as a frailty variable.

Ethnicity is an important factor in assessing the risk of getting T2DM and the survival of patients with T2DM. Sadly, it was poorly recorded in THIN database especially for older patients. As a result, only 3.1% of the study subjects had their ethnicity recorded. Hence, ethnicity was not considered in the study.

5.3.2 Life Style Covariates

The BMI was divided into three categories, *normal weight*, *overweight* and *obese*. BMI classification followed the cut-offs given in Table 5.2 (NICE, 2014). Values of BMI<13 or BMI>100 were considered to be missing as they were not within the acceptable range, QuintilesIMS (2017). Underweight patients were merged into the normal weight group due to small numbers. The smoking status was categorised as *never*, *former* and smoker.

Table 5.2: BMI Classification

BMI Group	BMI Range	
	From	To
Underweight	13	< 18.5
Normal Weight	18.5	< 25
Overweight	25	< 30
Obese	30	≤ 100

Previous studies have reported that, on average, health status in deprived areas is poorer and the use of health facilities higher compared to less deprived areas (Reijneveld et al., 2000). Hence it is important to include a socio-economic indicator in the survival study. The THIN database has several deprivation indices for patients, characterising the socio-economic status (SES) of their areas. These include the Townsend deprivation index (TDI), index of multiple deprivation (IMD) and Mosaic (QuintilesIMS, 2017). The TDI was chosen for the study, as it had considerably less missing values in the study population compared to the IMD and is comparable across the UK. The data available on Mosaic index had almost the same proportion of missing values in our study as the TDI. However, TDI was chosen based on its use by the Office of National Statistics (ONS) in their census data. TDI is described in more detail in appendix B.

5.3.3 Co-morbidities and Severity of T2DM Covariates

The diseases in Table 5.1 were recorded as at the entry date and for the study to be generalisable to the UK population, patients with life non-limiting conditions were

included. The T2DM drug therapy was used as a proxy to the severity of T2DM. The higher the intensity of the drug therapy, the greater the severity was assumed. The window period from diagnosis to the first drug initiation was set to 6 months to increase the number of cases with a particular therapy.

5.3.4 Bio-makers

The hypertension bio-makers SBP and DBP in addition to the diagnosis code were used to identify hypertensive patients. Patients diagnosed with hypertension and either a DBP$\geq$ 90 or SBP$\geq$ 140 (NICE, 2019) were classified as hypertensive. Hypertensive patients were further classified as living with a treated (having antihypertensive drugs) or untreated hypertension (no antihypertensive drugs) as at or before study entry.

Patients with a hypercholesterolaemia diagnosis or with a total cholesterol (=blood lipid ratio*HDL)$\geq$ 5 (NHS, 2019a) were classified as having hypercholesterolaemia before or as at entry. As with hypertension, these patients were further classified as living with treated or untreated hypercholesterolaemia. Though HbA_{1c} is an important variable, it was excluded due to more than 60% missing values at baseline.

5.3.5 Time to event

Two time-scales were used in the study, time to event and start and stop age within the study. Time to event was calculated as the difference between the event date and entry date in years. The start age is the age at study entry and the stop age is the time the selected person exited the study based on the event status. The event status was classified as 0 (censored) for no event and 1 (death) for occurrence of an event. Patients who survived beyond 2016 or left the GP before 2017 were categorised as censored.

5.3.6 Prescriptions

All antidiabetic drugs were included and categorised into No prescription within 6 months, First line, Second line, Third line, Insulin only and Insulin and other drugs. These followed the T2DM therapy management as recommended by NICE and summarised in Figure 2.2. Intensified therapies were included as the data showed that some patients were on intensified treatment within 6 months from diagnosis. More than 64% of T2DM patients were prescribed metformin as with or without other antidiabetes drugs. Antihypertensive drugs and statin were used to group people with hypertension or hypercholesterolaemia into "no diagnosis", "with treatment" or "untreated" categories.

5.4 Data extraction and statistical analysis

The study used Stractured Query Language (SQL) in SQL Server and Visual Basic for Applications (VBA) in MS Excel to extract and perform validation on the data before applying exploratory and survival modelling tools in R software. First, the classic Cox model and its extensions were applied which included time splitting. The Gompertz-double-Cox model with frailty was finally used to estimate the all-cause mortality hazards. Kaplan-Meier plots were also performed. The full-case (complete-case) analysis was performed first to find significant covariates at 5% significance level and their interactions at 1% significance level using the backward elimination method. Due to the large sample size, significance level of 1%, was used to obtain interactions that contributed the most to the model.

Multi-level multiple imputation was performed on missing data using the JOMO method and the related package in **R** software. A parametric Cox model with frailty for each kth imputed full dataset was performed. The Rubin's rule was then used to combine the estimates, and their variance and their 95% CIs were then calculated. For-

est and survival plots for selected subpopulations were then created and analysed after which internal and external validation was performed to check for overall performance and discrimination and external agreement using the estimated AIC, negative likelihood and concordance statistics as measures of the model's goodness-of-fit. Comparisons with previously reported studies were used for external validation. Estimates of time invariant effects were presented in forest plots while time variant effects were presented in plots of effects versus time.

5.5 Conclusion

This chapter described the study methodology which included the source of data (THIN), covariates of interest and their description, selection criteria and statistical tools. The final population selected for the study consisted of 68 199 T2DM patients and 152 983 non-diabetics. The following chapter provides a descriptive analysis of the study population including unadjusted hazard ratios of selected covariates and Kaplan-Meier plots.

Chapter 6

Exploratory and Unadjusted Hazards Analysis

6.1 Introduction

This chapter describes the prevalence of selected covariates in the study population and related survival curves using descriptive analysis such as cross-tabulation, Kaplan-Meier (KM) plots and unadjusted hazard ratios. Excluded severe conditions include ischaemic stroke, MI and CVD, whereas patients with milder conditions such as TIA were retained. The exclusion list is specified in Section 5.2. It makes use of weighted average prevalence, which was computed using equation (E.1) in Appendix E, to compare study population with the United Kingdom (UK) population. The demographic composition of the study population is first described, followed by the distribution of lifestyle factors and medical conditions.

6.2 Demographic Composition of the Study Population

The initial study population consisted of 362 082 patients born before 1961. Of these 108 282 were cases and 253 800 controls. The minimum, average and maximum ages of the study participants at entry were 40, 62 and 97 years, respectively. The proportions of study subjects aged 65 years and above were 39.4% and 39.7% for cases and controls, respectively. These proportions were more than 5 percentage points higher than the weighted average proportion for the UK (34.4%) for those aged 65 years and above, for the period 1984 to 2016. This is because of higher incidence of type 2 diabetes mellitus (T2DM) in the older ages compared to younger ages Diabetes UK (2016a). Figure 6.1 and Table 6.1 show that T2DM is mostly diagnosed between the ages 55 to 64 and 60 to 69 years for men and women, respectively. Figure 6.1 also shows that male cases had a somewhat better unadjusted survival than male controls up to 20 years of follow-up. The differences in mortality risk were more pronounced in males than females. However, these are unadjusted mortality risks. In addition, the lower risk may be attributed to the higher percentages of people with untreated HTN and HCL among non-diabetics at study entry when compared to T2DM patients. These are risk factors of CVD such as stroke, which may increase the risk of all-cause mortality. Other increased incidences of diseases among non-diabetics are cancer, cognitive impairment (including dementia) and CKD stages 3 to 5 for the over-60s. This is depicted in Figure 6.4. Without adjusting for these risk factors, their effect on all-cause mortality are masked leading to the underestimated effect of T2DM on all-cause mortality. Chapter 7 provides an analysis of the fully adjusted model to assess the impact of T2DM on mortality risk.

Table 6.1: Number and Proportions of Subjects by age group, gender and case-control status at study entry

Age Group	Cases (Number(%))		Controls (Number(%))	
	Male	Female	Male	Female
40-44	1,594 (2.57)	1,061 (2.29)	4,000 (2.88)	3,006 (2.62)
45-49	5,090 (8.2)	2,889 (6.25)	11,934 (8.59)	7,760 (6.76)
50-54	9,572 (15.43)	5,603 (12.12)	20,894 (15.04)	14,157 (12.33)
55-59	11,680 (18.83)	7,332 (15.86)	25,414 (18.29)	17,943 (15.62)
60-64	11,831 (19.07)	7,888 (17.06)	26,307 (18.93)	19,301 (16.81)
65-69	9,989 (16.1)	7,630 (16.5)	22,546 (16.23)	18,515 (16.12)
70-74	6,979 (11.25)	6,444 (13.94)	16,152 (11.62)	16,087 (14.01)
75-79	3,657 (5.89)	4,307 (9.31)	8,319 (5.99)	10,821 (9.42)
80+	1,651 (2.66)	3,085 (6.67)	3,386 (2.44)	7,258 (6.32)
Total	62,043 (100)	46,239 (100)	138,952 (100)	114,848 (100)

6.3 Selected life style factors and medical conditions

6.3.1 Life style factors

Smoking

The overall prevalence of smoking in the study population at entry was 18.63% when subjects with missing smoking status were included and 21.97% when excluded. These prevalences were 2.66 and 0.274 percentage points higher when compared to the weighted average of smoking prevalence among those aged 35 years and above in the UK in 1984 to 2016.

The prevalence of current smokers among controls (19.43%) was somewhat higher than that of cases (18.12%). Also the prevalence of former smokers was more in cases (27.53%) than controls (19.61%). It can also be seen (Table D.8) that the prevalence of smoking in both cases and controls decreased by age at entry. This can be attributed to the fact that older people more frequently seek medical help than those younger. In so doing, smokers get advised to stop smoking and improve on their life styles.

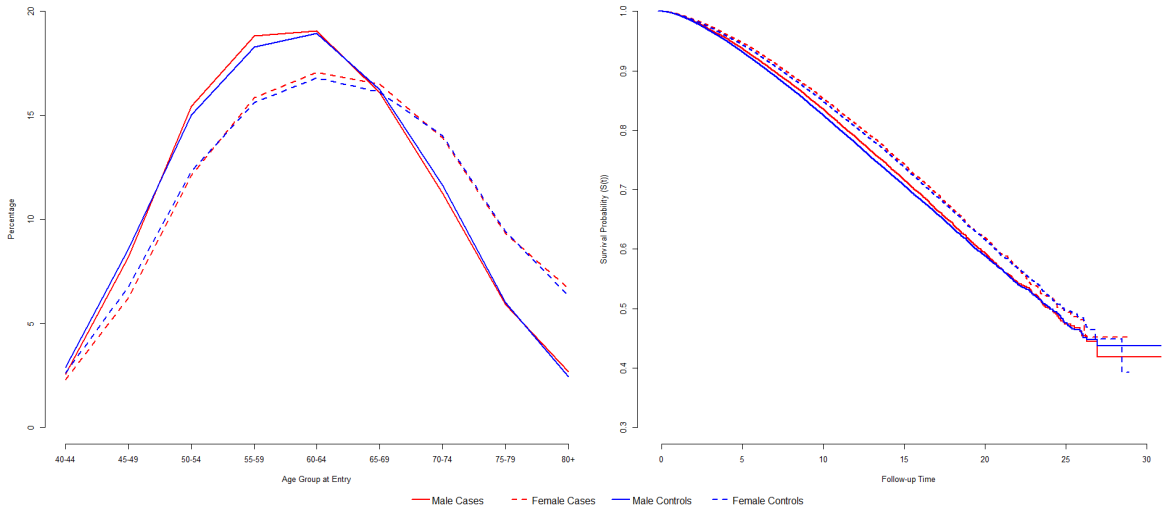


Figure 6.1: Composition of the study population at entry and their survival by case-control status, gender and age group

Alcohol Consumption

The prevalence of current alcohol consumers at entry was 58.7% and 55.3% for men and women aged 45 years and above, respectively, as shown in Table D.9. These were 12.2 percentage points less and 0.98 percentage points higher than the weighted average prevalence of alcohol consumption in men (70.9%) and women (54.33%) in the UK, respectively (ONS, 2018; Robinson et al., 2011; Steven, 2012). Table D.9 also shows that alcohol consumption prevalence by age group was somewhat lower among cases than controls. However, cases had more former alcohol consumers than controls as consuming alcohol would increase the risk of hypoglycaemia or hyperglycaemia; it may also interfere with positive effects of antidiabetic drugs WHO (2014); Diabetes UK (2019).

Obesity

71% percent of the study population had body mass index (BMI) recorded at entry. 31% percent of the study population in total were overweight and 24% were obese as shown in Table 6.2. Table 6.2 and Tables D.10 to D.11 show that obesity was considerably higher

in cases than controls. These findings agree with previous studies which demonstrated that T2DM patients are more likely to be obese, compared to non-diabetic patients (Daousi et al., 2006).

Table 6.2: BMI Classifications

BMI Group	BMI Range From	BMI Range To	Study Population (Number (%)) Cases	Study Population (Number (%)) Controls
Underweight	-	< 18.5	251 (0.23%)	2, 551 (1.01%)
Normal Weight	18.5	< 25	10, 162 (9.39%)	74, 824 (29.50%)
Overweight	25	< 30	31 164 (28.80%)	79, 408 (31.31%)
Obese	30	< 35	50, 380 (46.55%)	36, 070 (14.22%)
Missing			16, 267 (15.03%)	60, 795 (23.97%)

Deprivation

The study population was mostly constituted by individuals from areas classified as Townsend deprivation index (TDI) 1 (least deprived), 2 and 3 (medium deprived). Table D.12 shows that, 24% (21.05% of cases and 26.40% of controls) of the study population were from less deprived areas and 10.94% were from the most deprived areas. The overall survival worsened by the level of deprivation for both cases and controls as shown in Figure 6.2. However, a comparison of people with T2DM and without diabetes at entry by deprivation showed that cases from less deprived areas were at a higher risk of mortality, with an unadjusted hazard ratio (HR) of 1.17 [1.12,1.21], but the opposite was true for the most deprived areas (unadjusted HR 0.72[0.62,0.76]). Thus using unadjusted HRs, the more deprived the area was, the better was the hazard of mortality among cases when compared to controls as shown in Figure 6.2. The overall median survival time was 23.65 years for cases and 23.48 years for controls.

6.3.2 Medical Conditions

Several diseases which included amputation, cancer and cardiovascular disease (CVD) were selected for the study as mentioned in chapter 5. Figure 6.3 shows that only

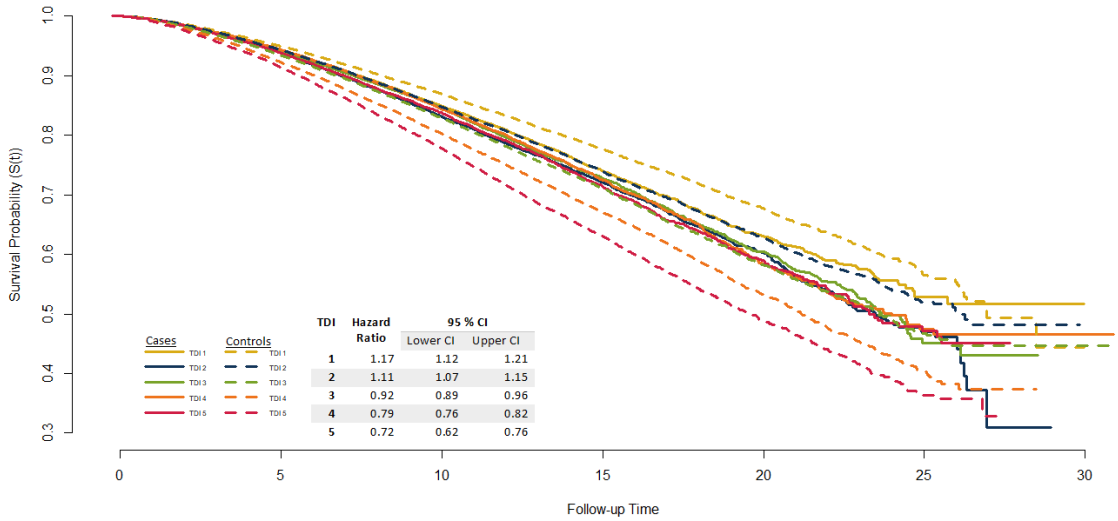


Figure 6.2: Survival of Study Subjects by Townsend Deprivation Index at Entry

10.5%, 16%, 12% and 1.1% of cases had heart failure (HF), myocardial infarction (MI), peripheral vascular disease (PVD) and transient ischemic attack (TIA), respectively, at diagnosis of T2DM. When compared to controls, cases had similar proportions of amputation, cognitive impairment, PVD or TIA. However, there were significant differences in HF and MI, with cases having higher proportions of the diseases. Both cases and controls had more than 50% people with hypertension.

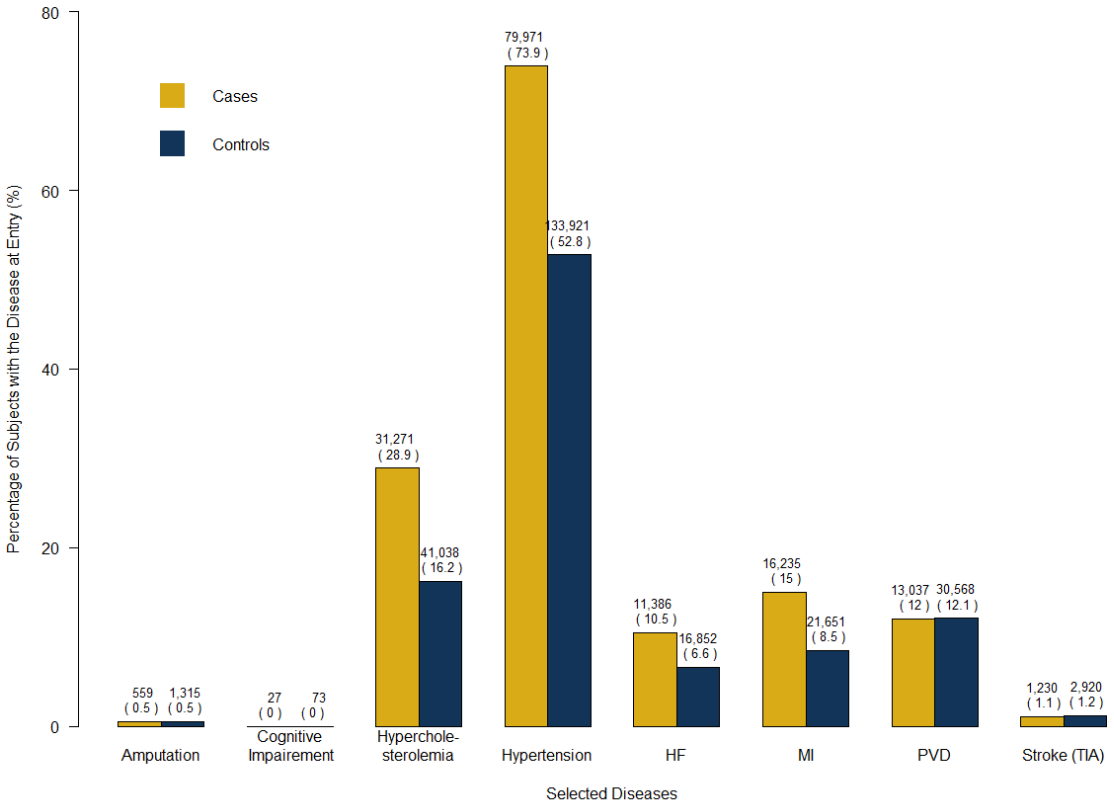


Figure 6.3: Number and percentage (in brackets) of study participants with the selected diseases at study entry

Amputation

Throughout the follow-up period, the annual amputation prevalence among subjects was on the increase from 0.26% in 1984 to 0.5% in 2016, *see* Tables D.32 to D.34. Normally, cases are at a higher risk of an amputation compared to controls. Cases diagnosed at 40-54 years of age, i.e. with earlier diagnosis, had the highest prevalence of amputations due to a longer exposure to the T2DM risks, compared to controls of the same age group and other age groups in cases as shown in Figure C.14. This figure also shows that cases were at a higher risk of amputations. Figure C.15 shows that both female and male cases had higher risks of amputation than their control counterparts. For both cases and controls, females had lower amputation risk than males.The unadjusted amputation HR in cases diagnosed at 40-49 years was 4.04 [3.11,5.24] compared to controls while in

those diagnosed at 70 years and above it was 1.95 [1.64,2.32], see Figure 6.4.

Cancer

The total incidence of all cancers was 35.9% throughout the follow-up period. It increased by age group at diagnosis, in cases from an incidence of 8.46% among the 44-49 years age group to 23.67% in the 75-79 years age group. Controls had the same experience up to the age group 65-69 years as shown in Table D.18. Patients in the 40-44 years age group had the lowest prevalence of cancer in both cases and controls. Overall, cases had considerably lower prevalence of cancer than controls. It is possible that this can be attributed to changes in life style and/or use of medications such as metformin by cases to control blood glucose levels. Figure C.14 shows that cases had a lower risk of any cancer compared to controls. Male cases had a higher risk of any form of cancer compared to female cases as depicted in Figure C.15. Figure 6.4, also shows that though the risk was less in cases compared to controls in all ages, it increased with age at entry.

Chronic kidney disease CKD Stages 3 to 5

About 21.9% of the cases and 24.5% of controls were diagnosed with chronic kidney disease (CKD) stages 3 to 5 in the follow-up period. The prevalence of the CKD stages 3 to 5 increased with age in both cases and controls as presented in Table D.21. Figure C.15 shows that both male and female cases had a lower risk of CKD stages 3 to 5 up to 15 years of follow-up than their control counterparts, after which both male and female controls had slightly lower risk. It can also be noted from Figure C.14, that risk of CKD stages 3 to 5 increased with age in both cases and controls, with cases having a lower risk in age groups 40-49 and 50-59. The unadjusted HR for CKD 3 to 5 was significantly higher for cases in the age group 40-49 years at entry (1.23[1.14,1.32]) and lower for the age groups 60-69 years (0.93[0.91,0.96]) and 70 years and above (0.91[0.89,0.93]) when compared to controls.

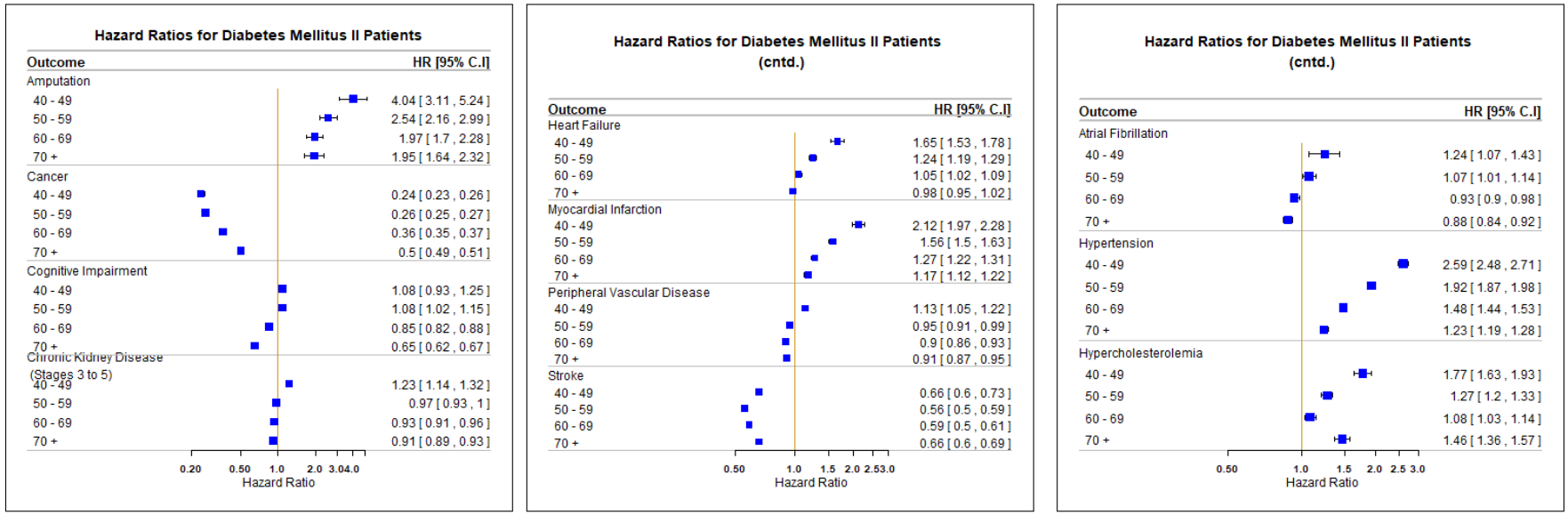


Figure 6.4: Hazard ratios of selected further morbidities in people with T2DM compared to people without T2DM

CVD

The prevalence of stroke at entry in cases and controls was 1.14% and 1.15%, respectively as shown in Table D.17. The same table shows that MI had the highest prevalence of 15% among cases at entry, followed by PVD (12%) and HF (10.5%), whereas the prevalence of PVD was higher in controls (12.1%). MI and HF were more prevalent in cases than controls at entry. Figure C.15 and Figure C.14 show that the risk of atrial fibrillation (AF) was similar and increased by age-group at entry in both cases and controls. The same figures illustrate that cases had a lower risk of PVD and TIA compared to controls but it was not the same for MI, HF, hypercholesterolaemia and hypertension. The risk of hypertension and hypercholesterolaemia were higher in cases soon after diagnosis for both males and females. Only 29% and 16% of cases and controls, respectively, had hypercholesterolaemia at entry. Treated hypercholesterolaemia constituted 27% and 10% in cases and controls, respectively.

Over 25% of the UK population were reported to be hypertensive in 2016 (Kennard and O'Shaughnessy, 2016). Table D.13 shows that the prevalence of high blood pressure (BP) was higher in cases than controls. Overall, patients with high BP constituted 48.7% of the study population. When subjects are split by case - control status and age group at entry, prevalence of high BP was higher in cases than in controls (Table D.14). It can also be noted that prevalence increased by age in both cases and controls and the difference of BP prevalence between cases and controls reduced with age. More than 73% and 52.8% of cases and controls, respectively, were hypertensive at entry. Of these, 83.1% and 59.8% of cases and controls were prescribed antihypertensive drugs. 53.4% of cases had pre-high BP and not hypertensive compared to 43.6% of controls. More than 83% and 94% of cases and controls with missing BP had no diagnosis of HTN, respectively. Only 18% of cases with high BP at entry had untreated hypertension compared to 44% in controls. The unadjusted hazard ratios (HRs) for MI associated with T2DM were significantly higher while HRs for stroke were significantly lower in all age groups compared to controls. T2DM females had higher unadjusted hazards of

CVD than males when compared to their control counterparts as shown in Table 6.3.

Cognitive Impairment (including Dementia)

There were only 27 cases and 73 controls with cognitive impairment at entry. This is because we excluded all forms of dementia and severe cognitive impairment in our study design. Hence the prevalence of cognitive impairment at entry was very close to zero. Controls were at a higher risk of developing cognitive impairment compared to cases as depicted in Figure C.15. For both cases and controls, females were at a higher risks of cognitive impairment compared to males. Figure C.14 also shows that there was a negligible difference in risk between cases and controls for people aged less than 60 years in the first ten years of follow-up. However, for people aged 60 years and above, controls had a higher risk compared to cases. The risk gaps for the participants of 60 years and above increased with time in the follow-up period. The risk was significantly higher in the age group 50-59 and lower in the age groups 60-69 and 70 years and above, with unadjusted HRs of 1.08[1.02,1.15], 0.82[0.82,0.88] and 0.65[0.62,0.97], respectively.

6.4 Final Study Population

The initial study population included some very sparse data with very low prevalence (less than 5%). For example, only 0.01% and 3.7% of the population born in 1930-1939 and 1940-1949,respectively, were aged 40-49 years at entry, though the overall prevalence for this age group at entry was high (35.4%) in the 1950-1960 birth cohort. For this reason, participants aged less than 50 years were excluded from the study as possible outliers. Individuals aged above 74 years at entry were also excluded due to low numbers. Participants with pre-existing medical conditions such as TIA and amputations were also excluded due to very low numbers. The distribution of the final study population is shown in Table 6.4 and the average of age at entry was 61 years.

Table 6.3: Unadjusted hazard ratios of total mortality and selected morbidities associated with T2DM by age group

Age Group	Overall Hazard Ratios (Confidence Intervals)								
	All-cause mortality	Morbidities							
		Amputation	Cancer	Cognitive Impairment	CKD 3 to 5	Heart Failure	Myocardial Infarction	Peripheral Vascular Disease	Stroke
40 - 44	0.99 (0.83 , 1.18)	6.53 (3.93 , 10.84)	0.25 (0.22 , 0.28)	1.28 (0.94 , 1.74)	1.59 (1.36 , 1.85)	1.81 (1.57 , 2.09)	2.39 (2.08 , 2.76)	1.25 (1.09 , 1.44)	0.75 (0.62 , 0.92)
45 - 49	0.91 (0.82 , 1.01)	3.31 (2.43 , 4.5)	0.24 (0.22 , 0.26)	1.03 (0.87 , 1.21)	1.12 (1.03 , 1.23)	1.58 (1.45 , 1.73)	2.02 (1.85 , 2.2)	1.09 (0.99 , 1.18)	0.63 (0.56 , 0.7)
50 - 54	0.97 (0.91 , 1.04)	3.13 (2.46 , 3.97)	0.24 (0.23 , 0.25)	1.05 (0.95 , 1.16)	1.02 (0.96 , 1.08)	1.32 (1.24 , 1.41)	1.75 (1.65 , 1.87)	0.99 (0.93 , 1.06)	0.58 (0.53 , 0.63)
55 - 59	0.95 (0.91 , 1)	2.09 (1.67 , 2.62)	0.28 (0.27 , 0.29)	1.1 (1.02 , 1.18)	0.94 (0.9 , 0.98)	1.19 (1.12 , 1.25)	1.44 (1.36 , 1.52)	0.92 (0.87 , 0.98)	0.55 (0.52 , 0.59)
60 - 64	0.95 (0.91 , 0.99)	1.97 (1.61 , 2.41)	0.32 (0.31 , 0.34)	0.95 (0.89 , 1.01)	0.96 (0.92 , 0.99)	1.08 (1.03 , 1.14)	1.3 (1.24 , 1.37)	0.87 (0.83 , 0.92)	0.57 (0.53 , 0.6)
65 - 69	0.93 (0.9 , 0.97)	1.96 (1.57 , 2.43)	0.39 (0.38 , 0.41)	0.78 (0.74 , 0.82)	0.91 (0.88 , 0.95)	1.03 (0.98 , 1.08)	1.23 (1.17 , 1.3)	0.92 (0.88 , 0.98)	0.61 (0.58 , 0.65)
70 - 74	0.99 (0.95 , 1.03)	1.98 (1.57 , 2.5)	0.47 (0.45 , 0.49)	0.68 (0.64 , 0.72)	0.93 (0.89 , 0.96)	1.03 (0.98 , 1.08)	1.15 (1.09 , 1.22)	0.89 (0.83 , 0.94)	0.65 (0.61 , 0.69)
75 - 79	0.96 (0.92 , 1.01)	1.93 (1.41 , 2.64)	0.53 (0.51 , 0.56)	0.65 (0.61 , 0.69)	0.92 (0.88 , 0.96)	0.95 (0.89 , 1.01)	1.19 (1.1 , 1.29)	0.92 (0.85 , 1)	0.65 (0.6 , 0.7)
80 +	1 (0.95 , 1.05)	1.9 (1.2 , 3.01)	0.54 (0.51 , 0.58)	0.56 (0.51 , 0.6)	0.82 (0.77 , 0.87)	0.89 (0.81 , 0.97)	1.13 (1.01 , 1.27)	0.97 (0.86 , 1.09)	0.68 (0.62 , 0.74)
Male Hazard Ratios									
40 - 44	1.04 (0.85 , 1.28)	5.28 (2.97 , 9.41)	0.21 (0.17 , 0.25)	1.28 (0.94 , 1.74)	1.52 (1.23 , 1.88)	1.61 (1.35 , 1.91)	1.94 (1.64 , 2.3)	1.2 (0.99 , 1.46)	0.61 (0.48 , 0.79)
45 - 49	0.9 (0.8 , 1.02)	3.14 (2.24 , 4.39)	0.22 (0.2 , 0.24)	1.03 (0.87 , 1.21)	0.96 (0.85 , 1.08)	1.47 (1.32 , 1.63)	1.77 (1.6 , 1.95)	1.09 (0.97 , 1.22)	0.53 (0.46 , 0.6)
50 - 54	0.97 (0.9 , 1.05)	3.13 (2.4 , 4.09)	0.22 (0.21 , 0.24)	1.05 (0.95 , 1.16)	0.92 (0.85 , 1)	1.24 (1.15 , 1.35)	1.56 (1.44 , 1.68)	0.96 (0.88 , 1.05)	0.5 (0.46 , 0.56)
55 - 59	0.96 (0.9 , 1.02)	2.02 (1.56 , 2.61)	0.26 (0.25 , 0.28)	1.1 (1.02 , 1.18)	0.94 (0.88 , 1)	1.12 (1.05 , 1.19)	1.34 (1.26 , 1.43)	0.94 (0.88 , 1.01)	0.49 (0.45 , 0.54)
60 - 64	0.94 (0.89 , 0.99)	2.04 (1.61 , 2.59)	0.32 (0.3 , 0.33)	0.95 (0.89 , 1.01)	0.93 (0.88 , 0.98)	1.05 (0.99 , 1.12)	1.22 (1.15 , 1.3)	0.89 (0.83 , 0.95)	0.51 (0.47 , 0.54)
65 - 69	0.92 (0.88 , 0.97)	2.2 (1.69 , 2.86)	0.38 (0.37 , 0.4)	0.78 (0.74 , 0.82)	0.92 (0.88 , 0.97)	1 (0.94 , 1.06)	1.18 (1.1 , 1.26)	0.97 (0.9 , 1.04)	0.6 (0.56 , 0.65)
70 - 74	0.97 (0.93 , 1.02)	2.14 (1.59 , 2.87)	0.46 (0.43 , 0.48)	0.68 (0.64 , 0.72)	0.94 (0.89 , 0.98)	1.01 (0.94 , 1.09)	1.09 (1 , 1.18)	0.91 (0.84 , 0.99)	0.63 (0.59 , 0.69)
75 - 79	0.97 (0.91 , 1.03)	2.51 (1.65 , 3.8)	0.52 (0.49 , 0.56)	0.65 (0.61 , 0.69)	0.94 (0.88 , 1.01)	0.98 (0.89 , 1.08)	1.11 (0.99 , 1.25)	0.93 (0.82 , 1.05)	0.64 (0.58 , 0.71)
80 +	1 (0.92 , 1.09)	1.74 (0.86 , 3.53)	0.52 (0.47 , 0.58)	0.56 (0.51 , 0.6)	0.85 (0.77 , 0.95)	0.88 (0.76 , 1.01)	1.09 (0.9 , 1.31)	0.97 (0.79 , 1.18)	0.68 (0.58 , 0.8)
Female Hazard Ratios									
40 - 44	0.82 (0.59 , 1.15)	11.27 (3.77 , 33.71)	0.31 (0.25 , 0.38)	1.28 (0.94 , 1.74)	1.69 (1.35 , 2.12)	2.24 (1.75 , 2.87)	3.96 (3 , 5.23)	1.33 (1.08 , 1.63)	1.07 (0.78 , 1.49)
45 - 49	0.91 (0.76 , 1.1)	3.71 (1.71 , 8.09)	0.28 (0.25 , 0.31)	1.03 (0.87 , 1.21)	1.41 (1.24 , 1.61)	1.89 (1.59 , 2.24)	2.92 (2.45 , 3.47)	1.09 (0.95 , 1.26)	0.93 (0.76 , 1.13)
50 - 54	0.94 (0.83 , 1.06)	2.7 (1.56 , 4.68)	0.27 (0.25 , 0.3)	1.05 (0.95 , 1.16)	1.18 (1.09 , 1.29)	1.46 (1.3 , 1.64)	2.34 (2.08 , 2.64)	1.05 (0.94 , 1.17)	0.75 (0.65 , 0.85)
55 - 59	0.92 (0.84 , 1.01)	2.18 (1.37 , 3.47)	0.3 (0.28 , 0.32)	1.1 (1.02 , 1.18)	0.95 (0.89 , 1.02)	1.32 (1.2 , 1.45)	1.67 (1.51 , 1.84)	0.89 (0.81 , 0.98)	0.67 (0.6 , 0.74)
60 - 64	0.96 (0.89 , 1.03)	1.71 (1.16 , 2.52)	0.33 (0.31 , 0.35)	0.95 (0.89 , 1.01)	1 (0.95 , 1.05)	1.12 (1.03 , 1.21)	1.44 (1.32 , 1.57)	0.85 (0.78 , 0.93)	0.67 (0.61 , 0.74)
65 - 69	0.94 (0.88 , 1)	1.46 (0.99 , 2.17)	0.4 (0.38 , 0.42)	0.78 (0.74 , 0.82)	0.92 (0.87 , 0.96)	1.05 (0.97 , 1.14)	1.31 (1.2 , 1.42)	0.87 (0.8 , 0.94)	0.62 (0.57 , 0.68)
70 - 74	1 (0.94 , 1.06)	1.72 (1.17 , 2.52)	0.48 (0.45 , 0.51)	0.68 (0.64 , 0.72)	0.93 (0.88 , 0.97)	1.04 (0.96 , 1.13)	1.23 (1.12 , 1.34)	0.86 (0.79 , 0.94)	0.66 (0.61 , 0.72)
75 - 79	0.94 (0.88 , 1)	1.3 (0.79 , 2.15)	0.53 (0.49 , 0.57)	0.65 (0.61 , 0.69)	0.91 (0.86 , 0.96)	0.91 (0.82 , 0.99)	1.26 (1.13 , 1.41)	0.91 (0.81 , 1.01)	0.65 (0.59 , 0.72)
80 +	0.99 (0.92 , 1.05)	1.98 (1.08 , 3.63)	0.54 (0.49 , 0.6)	0.56 (0.51 , 0.6)	0.81 (0.75 , 0.87)	0.89 (0.79 , 0.99)	1.13 (0.98 , 1.31)	0.96 (0.83 , 1.12)	0.68 (0.61 , 0.76)

Table 6.4: Study Population.

	Entry Cohort					
	2000-2004		2005-2009		2010-2016	
Description	**T2DM**	**Diabetic-free**	**T2DM**	**Diabetic-free**	**T2DM**	**Diabetic-free**
Selected (n)	25 333	64 334	27 403	61 155	15 463	27 494
Birth Cohort (n, %)						
1930-1939	10 980 (28.10)	28 120 (71.90)	6 413 (30.20)	14 848 (69.80)	893 (34.70)	1 678 (65.30)
1940-1949	11 521 (28.30)	29 231 (71.70)	12 225 (30.70)	27 605 (69.30)	6 324 (35.50)	11 474 (64.50)
1950-1960	2 832 (28.90)	6 983 (71.10)	8 765 (30.90)	18 702 (69.10)	8 246 (36.50)	14 342 (63.50)
Males (%)	58.70	56.80	59.70	56.50	60.30	57.00
Deaths (n)	4 425	11 891	3 150	7 308	996	1 848
Age (%)						
50-59	54.82	54.83	54.74	54.79	55.25	55.12
60-74	65.88	65.89	66.15	66.19	65.96	65.96
mean (in years)	61.22	61.26	61.36	61.51	61.65	61.64
standard deviation	6.44	6.44	6.75	6.75	6.31	6.36
Mean Follow-up Time (years)	10.95	11.14	7.65	7.85	3.98	4.17
Smoking Status (%)						
Nonsmoker	45.05	45.36	45.43	49.82	45.02	51.01
Former	27.60	18.12	34.66	26.04	35.10	27.70
Smoker	20.56	21.81	19.26	21.10	19.81	20.61
Missing	6.79	14.71	0.65	12.04	0.06	0.70
Townsend Index (%)						
Less Deprived	21.63	27.17	21.94	26.89	20.74	26.69
2	20.36	22.93	20.38	22.93	20.09	22.34
3	20.19	18.78	20.1	19.07	19.88	19.10
4	18.35	15.36	18.00	15.01	19.04	15.02
Most Deprived	12.88	9.70	12.92	9.59	13.06	9.56
Missing	6.59	5.06	6.66	6.51	7.19	7.29
BMI (%)						
Normal Weight	8.79	30.71	8.41	31.14	7.87	30.79
Overweight	30.49	31.31	29.36	35.46	27.85	37.23
Obese	45.12	13.02	54.44	18.36	58.84	21.99
Missing	15.61	24.95	7.79	15.04	5.44	9.99
mean (kgs/m^2)	31.25	26.38	32.04	26.92	32.65	27.31
standard deviation	5.85	4.49	6.21	4.82	6.57	5.07
Medical Conditions (%)						
HTN						
No Diagnosis	23.33	46.75	23.54	45.34	27.99	48.12
Treated	67.76	32.98	63.34	32.31	52.6	27.56
Untreated	8.92	20.26	13.13	22.35	19.42	24.31
HCL						
No Diagnosis	77.51	90.21	58.50	74.39	50.04	61.84
Treated	21.79	7.55	39.55	15.73	44.32	18.05
Untreated	0.71	2.24	1.95	9.88	5.65	20.11
AF	10.96	9.71	8.06	7.70	4.99	5.42
HF	11.67	6.85	11.24	6.99	9.58	6.25
MI	16.5	8.15	14.39	7.95	12.42	6.78
PVD	11.66	11.65	11.9	12.34	12.53	12.75

* n stands for number of selected individuals. % is for percentage.

The variables with missing values in the final study population were smoking status (6.5% in total), TDI (18.8%) and BMI (7%). The missing values were more pronounced in people without diabetes.

6.5 Summary

This chapter has shown that the composition of the study population is representative of the UK population aged 40 years and above and the results of the study can be reasonably generalised to the UK population. The final total number of the study participants was 221 182 after excluding variables and conditions with very low prevalence. The average age at entry was 61 years. Through the use of the KM survival plots it appears that cases have a better mortality risk than controls before adjusting for covariates. It was observed that selected morbidity risks, and complications increased with the age at T2DM diagnosis except for cancer, when compared to controls. Mortality and morbidity risks were seen to change with respect to age at entry, gender and case-control status except for AF when split by gender. Finally, this chapter has presented some exploratory and unadjusted results that may be informative and will be validated by modelling the all-cause mortality hazards presented in the following chapter when baseline variables are adjusted for.

Chapter 7

Survival Models for type 2 diabetes mellitus

7.1 Introduction

The previous chapter presented the descriptive analysis of the study data, which confirmed the choices of the variables and interactions to be included in the study analysis. This chapter describes and compares results from two Gompertz-Cox survival models with frailty (Model A and Model B)[1] that were used to estimate the effect of type 2 diabetes mellitus (T2DM) on all-cause mortality using complete case dataset which constituted 69.64% of records and 66.19% of deaths and the full imputed dataset. Model A assessed the all-cause mortality risk associated with diagnosis of T2DM at two distinct age categories which are 50 to 59 or 60 to 74 years of age at study entry compared to all non-diabetics. Model A can be presented mathematically by the equation below having the age at diagnosis as a factor at three levels ("Control","T2DM at 50-59" and "T2DM at 60-74"). Model B has the same mathematical model as Model A except that age at diagnosis variable is continuous and centred on the average age at entry (68.89 years) in

[1]Adjusted for age at study entry, year of birth, gender, BMI, smoking status, TDI, AF, HF, MI, PVD, HCL and HTN and the following interactions: T2DM indicator with MI and smoking status, BMI with smoking and birth cohort with smoking.

the 1930-1939 birth cohort. Both models use follow-up time scale The centering of the continuous age at diagnosis variable in Model B was done as to obtain a baseline hazard function similar to that in Model A. Both Model A and B, are of the mathematical form given by Equation (25) explained in Section 4.6.

$$\Lambda(t|\mathbf{X}) = \mathbf{Z}\frac{\alpha e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{scale}}}}{\upsilon e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{shape}}}}(e^{\upsilon e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{shape}}}\cdot t} - 1).$$

In developing Model B, quadratic and cubic polynomials on age at diagnosis were applied and it was found that their effects were statistically insignificant ($p - values \geq 0.05$). Hence, age in Model B was included in linear form. The results of both the models with quadratic and cubic functions are shown in Table D.29 and Table D.30. Interactions with age, including gender and age, were also found to be statistically insignificant.

As a reminder, the baseline cohort in both models corresponded to female individuals in the 1930-1939 birth cohort (aged 60-74 years in Model A and 68.89 years in Model B), who were non-smokers, with normal BMI, from medium deprivation area (TDI =3), with no diagnosis of AF, HF, PVD, MI, HTN and HCL at entry. A brief description of the analysis leading to the development of the two models is given in the following section.

7.2 Conventional Cox model analysis

A complete case dataset included observations with non-missing values in BMI, TDI and BMI. The Cox proportional hazards PH model was then fitted using both the complete case and imputed data, and significant covariates were first selected by backward elimination using significance levels of 5% for main effects and 1% for interactions. Using the *cox.zph* function in **R**, the results showed that among the significant covariates, birth cohort, HTN and HCL had time-variant effects as shown in Table D.31 and scaled Schoenfeld residual plots in Figure C.9. A test for interactions between the follow-up time and violating covariates was then performed but no results were produced as it

was extremely time expensive.

The method of splitting the time at follow-up times 4 and 7 years was then attempted to address the violation of the PH assumption. Unfortunately, the PH assumption was violated again. The estimated frailty effect for the general practice was also found to be insignificant.

Another method that could have been attempted was the stratification of the covariates violating the PH assumption, but this would have excluded their effects on all-cause mortality. Hence, the stratification method was not explored in the study. The Gompertz-Cox models with frailty were then fitted to the full case data, which also included the conventional Cox PH assessment. The effects of birth cohort, AF and HTN were found to be time-variant.

7.2.1 Imputation of missing values

To determine which missing pattern to assume, logistic regressions with missing status as an outcome were performed for each variable. The p-values were assessed for significance at 5% to determine if the data was MAR. It was determined, by diving the count of models with p-values less that 0.05 by total number of regressions performed, that the data was missing at MAR with a probability of 83.33%[2]. Hence a multiple imputation of missing values was performed.

A **jomo** package in R software was used to impute the missing data, using the developed complete case model. First, the *jomo.MCMCchain* function was used to check for convergence of imputations and to determine the number of imputations (m) needed. Eight imputations were determined to be sufficient. However, 15 imputations were performed using the *jomo* function. The prevalences of the imputed characteristics showed prevalence within $\pm$ 1 percentage points compared to the complete case data, as shown in Figure C.10. With an overall average percentage of missing data $\gamma = 9.47\%$ in this

[2]The R code used can be viewed in Figure C.8

study, the total variance of variables with missing data, $\mathbf{T}_m$, using equation (4.1) was only 0.63% larger than ideal variance, $\mathbf{T}_\infty$. The survival models on the full case and imputed data are presented in the following sections.

7.3 Fitted Survival Models

The best fitting survival distribution was selected from several parametric distributions using the *flexsurvreg* function in R. Based on the lowest AIC (Figure C.11), the Gompertz distribution was used for baseline hazards. Parameters of the baseline Gompertz distribution of the variables with time-varying effects and gamma frailty estimates were estimated using complete case and imputed datasets (Table 7.1). In both models, estimates based on the imputed data were statistically similar to those based on the complete case dataset. However, the estimates' standard errors were improved in both models by using imputed data. The goodness-of-fit concordance statistics were similar in imputed and complete case datasets. Since the complete case and imputed datasets produced similar models, subsequent sections and chapters will describe estimates based on imputed data.

The baseline population had a Gompertz hazard function with scale and shape parameters estimated to be $a = 0.00673[0.00624, 0.00725]$ and $b = 0.0874[0.0813, 0.0939]$ for Model A and $a = 0.0068[0.0063 - 0.0073]$ and $b = 0.001[0.0097 - 0.0011]$ for Model B using imputed data. For both models, the 1940-1949 and 1950-1960 birth cohorts reduced the baseline shape parameter. However, the estimated 1950-1960 shape modification effect was statistically insignificant (p-value = 0.156) in Model B. AF and treated HTN increased the baseline shape parameter by about 30% while untreated HTN significantly reduced the shape estimate, in both models by at least 11%. On the log-scale, an increase in shape implied an increase in the intercept estimate and vice-versa. Thus, *pari-pasu*, an increase in shape parameter reduced the survival median and mode. However, the shape modifying variables also had significant scale effects or HRs. Thus, the

overall effect of these variables can be easier understood graphically, as described in subsequent paragraphs. The frailty variance, σ^2, was estimated to be 0.14 [0.12 , 0.16] and 0.15 [0.13,0.17] in Models A and B, respectively.

In Model A, being diagnosed with T2DM at an age between 50 to 59 years was associated with an increase of 20.6% in all-cause mortality hazard , while a diagnosis at older ages, 60 to 74 years, increased all-cause mortality hazard by 51.7% compared to compared to people without diabetes as shown in Figure 7.1. Model B estimated T2DM to be associated with a HR of all-cause mortality of 1.424[1.36,1.49] compared to non-diabetics. A single year increase in age at entry was associated with 9.5% increase in HR of all-cause mortality.

Furthermore, Model A estimated the following HRs for covariates, males 1.38[1.34,1.41] compared to females, obesity 1.16[1.1,1.23] compared to normal weight, smoking 2.56[2.4,2.73] and former smoking 1.67[1.56,1.79] compared to non-smokers, less deprived areas 0.83[0.79,0.86] and most deprived areas 1.18[1.13,1.23] compared to medium deprived areas. Living with pre-existing MI was associated with a HR of 1.38[1.32,1.44], whereas for HF and PVD HRs were 1.18[1.13,1.24] and 1.09[1.06,1.13] compared to individuals without the disease, respectively. Model B estimated similar time-invariant effects whose deviation from Model A ranged from 0.003 to 0.141. An earlier study reported that the estimates of coefficients from grouped data analysis were within one standard error of those from the continuous data analysis (McKeague and Zhang, 1996). This is expected when the baseline hazard has variations that are moderate and the respective covariate effects are mild [3] (McKeague and Zhang, 1996).

An interesting finding was that the all-cause mortality hazard associated with smoking increased for the younger birth cohorts. The all-cause mortality hazards associated with smoking in the 1940-1949 and 1950-1960 birth cohorts were 1.137[1.066,1.213] and 1.195[1.09,1.311], respectively, compared to smokers in the 1930-1939 birth cohort.

To compare and assess the association of the birth cohort with HR of all-cause mor-

[3] A variable whose effect changes insignificantly when the model is changed.

tality estimated by Models A and B, an age of 65 years at entry was used to plot the all-cause mortality hazards of Model B in females without diabetes in Figure 7.2(a). For AF and HTN, Figure 7.2(b-c), the plotted all-cause mortality hazards are for females without diabetes in the 1930-1939 birth cohort. The effect of AF or HTN was similar across all birth cohorts and in all people with and without T2DM.

Figure 7.2 shows a lower mortality risk for the age grouped population (Model A) compared to an individual mortality risk (Model B). This is the main principle used in group insurance products as it diversifies risks through large numbers thereby reducing insurance premiums paid by an individual, hedging against business risk and making their product prices competitive. For example, Figure 7.2 (a) shows that a 65 year old at entry in the 1950-1960 birth cohort had the highest all-cause mortality hazards throughout follow-up compared to the same birth cohort when the risk was shared. It can be noted that the 1930-1939 birth cohort showed an opposite result though the differences reduced with time. The differences between shared mortality hazards and individual hazards increased in younger birth cohort. The same sub-figure shows that differences in mortality hazards increased between the 1940-1949 and 1930-1939 birth cohorts during follow-up while decreasing between the 1950-1959 and 1940-1949 birth cohorts in the shared mortality risk Model A. For example, the reduced all-cause mortality in the 1950-1960 birth cohort compared to the 1940-1949 cohort was 27.9%, 19.8% and 12.1% less at 5, 10 and 15 years of follow-up, respectively. As with Model B, the differences were small and decreased with follow-up time between the 1930-1939 and 1940-1949 birth cohorts but increased between the 1950-1959 and 1940-1949 birth cohorts. Model B estimated that individual mortality hazards increased in later birth cohorts.

Individuals with pre-existing AF were estimated to have a lower mortality hazard compared to those without AF at study entry. However, their mortality hazards increased with time and became higher after 8 years and 9 years in Models A and B, respectively, for a person aged 65 years at entry (Figure 7.2 (b)).

Figure 7.2(c) shows untreated HTN to be associated with increased all-cause mortality hazards from study entry compared to people without HTN or with treated HTN. However, due to age as an independent hazard and the diminishing number of people at risk, the all-cause mortality risk increased among people with treated HTN, such that after 17 years (Model A) and 13 years (Model B), individuals with treated HTN had higher mortality hazards compared to those with untreated HTN. After 9 years and 7 years of follow up, individuals with treated HTN had increased mortality hazards compared to people without HTN in Model A and Model B, respectively.

Figure 7.3 compares the effect of these covariates with time-variant effects by birth cohort and T2DM indicator using Model A. The figure shows that effects of HTN and AF were similar in all birth cohorts in people with or without T2DM. In comparison to the baseline population, people with treated HTN at diagnosis of T2DM had lower mortality hazards than people with pre-existing untreated HTN at diagnosis of T2DM. Survival curves of hazards shown in Figure 7.3 are depicted in Figure C.13.

Another important finding was an increased all-cause mortality hazard at older ages in the recent birth cohort , see Figures 7.2 and 7.3. In Figure 7.3(a), a diagnosis of T2DM at 60-74 years of age in the 1930-1939 cohort showed higher mortality hazards than in people without diabetes in the 1940-1949 birth cohort throughout the follow-up period and compared to people diagnosed with T2DM at ages 50-59 years, in the 1940-1949 birth cohort, after 13 years of follow-up. A diagnosis of T2DM at 50-59 years in the 1950-1959 cohort, showed similar mortality hazards to people without diabetes in the 1940-1949 cohort at the end of follow-up and possibly higher after the follow-up period.

7.4 Conclusion

T2DM is associated with an increased all-cause mortality hazard that increases with age at diagnosis as estimated in Model A. Diagnosis of T2DM at ages 50-59 has an adjusted HR of 1.206 while at ages 60-74 is 1.517 when compared to all non-diabetic patients.

The youngest 1950-1960 birth cohort had an increased all-cause mortality hazards in later follow-up times compared to older birth cohorts. The next chapter discusses the estimated life expectancies and years of life lost due to T2DM using Model B and a modified Model A. Currently, a manuscript on Model A has been published in a peer reviewed Elsevier Diabetes Epidemiology and Management journal and can be accessed on https://doi.org/10.1016/j.deman.2022.100065.

Table 7.1: Scale and shape parameter estimates at baseline and for time-variant covariates in survival models A and B for full case and imputed data.[1]

Gompertz-Cox model with frailty parameter estimates												
	Model A						**Model B**					
	Imputed			Complete Case			Imputed			Complete Case		
Parameter	Estimate	95% CI	p-value	Estimate	95% CI	p-value	Estimate	95% CI	p-value	Estimate	95% CI	p-value
1000a (scale)	6.73	6.24 - 7.25	<1e-16	6.64	6.07 - 7.25	<1e-16	6.83	6.34-7.35	<1e-16	6.51	5.97 - 7.11	<1e-16
100b (shape)	8.74	8.13 - 9.39	<1e-16	8.10	7.31 - 8.98	<1e-16	10.34	9.75-10.96	<1e-16	9.96	9.19 - 10.8	<1e-16
Covariates Shape Parameters												
Year of Birth												
1930-1939	1			1								
1940-1949	0.71	0.66 - 0.77	<1e-16	0.69	0.62 - 0.77	<1e-16	0.83	0.78-0.88	<1e-16	0.82	0.76 - 0.88	<1e-16
1950-1960	0.87	0.77 - 0.97	<0.0186	0.89	0.764 - 1.03	0.122	0.94	0.85-1.03	0.156	0.94	0.84 - 1.06	0.314
AF	1.36	1.27 - 1.47	<1e-16	1.36	1.24 - 1.49	<1e-16	1.33	1.25-1.42	<1e-16	1.33	1.23 - 1.44	<1e-16
HTN												
None	1			1								
Treated	1.30	1.21 - 1.40	<1e-16	1.37	1.24 - 1.52	<1e-16	1.29	1.21-1.37	<1e-16	1.32	1.22 - 1.43	<1e-16
Untreated	0.84	0.76 - 0.94	0.0015	0.89	0.77 - 1.03	0.105	0.89	0.82-0.97	<0.0058	0.91	0.82 - 1.02	0.105
Covariates Scale Parameters												
Year of Birth												
1930-1939	1			1								
1940-1949	0.71	0.67 - 0.76	<1e-16	0.733	0.68 - 0.79	<1e-16	1.4	1.31-1.49	<1e-16	1.43	1.32 - 1.54	<1e-16
1950-1960	0.47	0.43 - 0.52	<0.0036	0.45	0.40 - 0.50	<1e-16	1.69	1.53-1.86	<1e-16	1.61	1.43 - 1.81	<1e-16
AF	0.78	0.73 - 0.84	<1e-16	0.83	0.76 - 0.90	<1e-16	0.74	0.68-0.79	<1e-16	0.78	0.71 - 0.85	<1e-16
HTN												
None	1			1								
Treated	0.80	0.76 - 0.84	<1e-16	0.81	0.76 - 0.86	<1e-16	0.75	0.71-0.79	<1e-16	0.76	0.72 - 0.81	<1e-16
Untreated	1.52	1.47 - 1.6	<1e-16	1.48	1.38 - 1.58	<1e-16	1.42	1.34-1.49	<1e-16	1.40	1.31 - 1.50	<1e-16
Frailty Estimate												
Variance (σ^2)	0.14	0.12 - 0.16	<1e-16	0.12	0.1 - 0.14	<1e-16	0.15	0.13 - 0.17	<1e-16	0.13	0.11 - 0.15	<1e-16
Goodness of Fit												
Concordance (C_ρ)	0.754			0.748			0.746			0.739		
Standard deviation of C_ρ	0.002			0.002			0.002			0.002		
Log likelihood	-145, 150.22			-96, 410.95			-143, 902.60			-95, 559.04		
AIC	290, 386.43			192, 907.9			287, 885.19			191, 198.08		

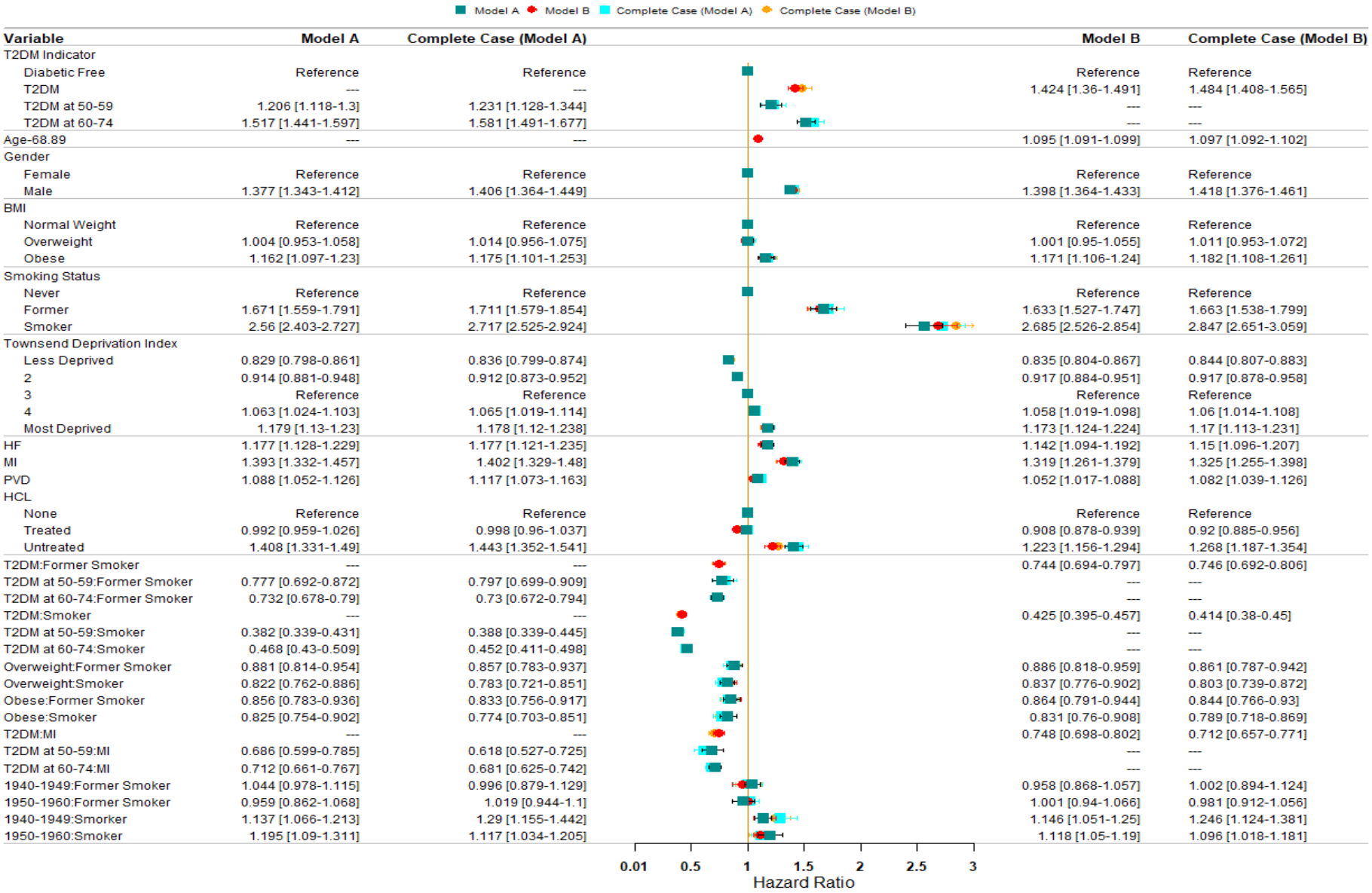


Figure 7.1: Estimated hazard ratios in Models A and B for complete case and imputed data.

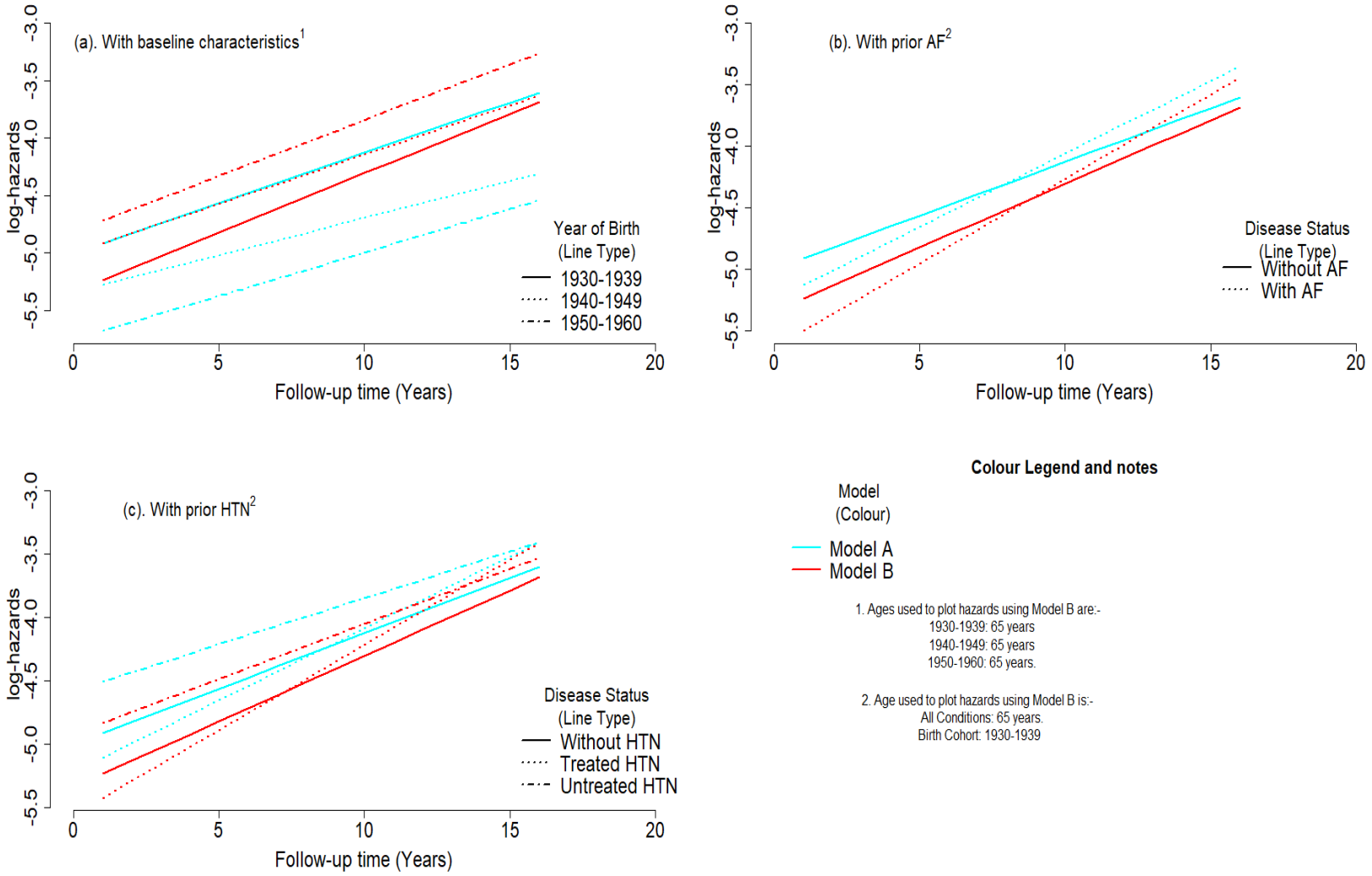


Figure 7.2: Comparison of the all-cause mortality hazards depicting time-varying effects from Models A and B for complete case and imputed data.

Models were adjusted for age, year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and their interactions.

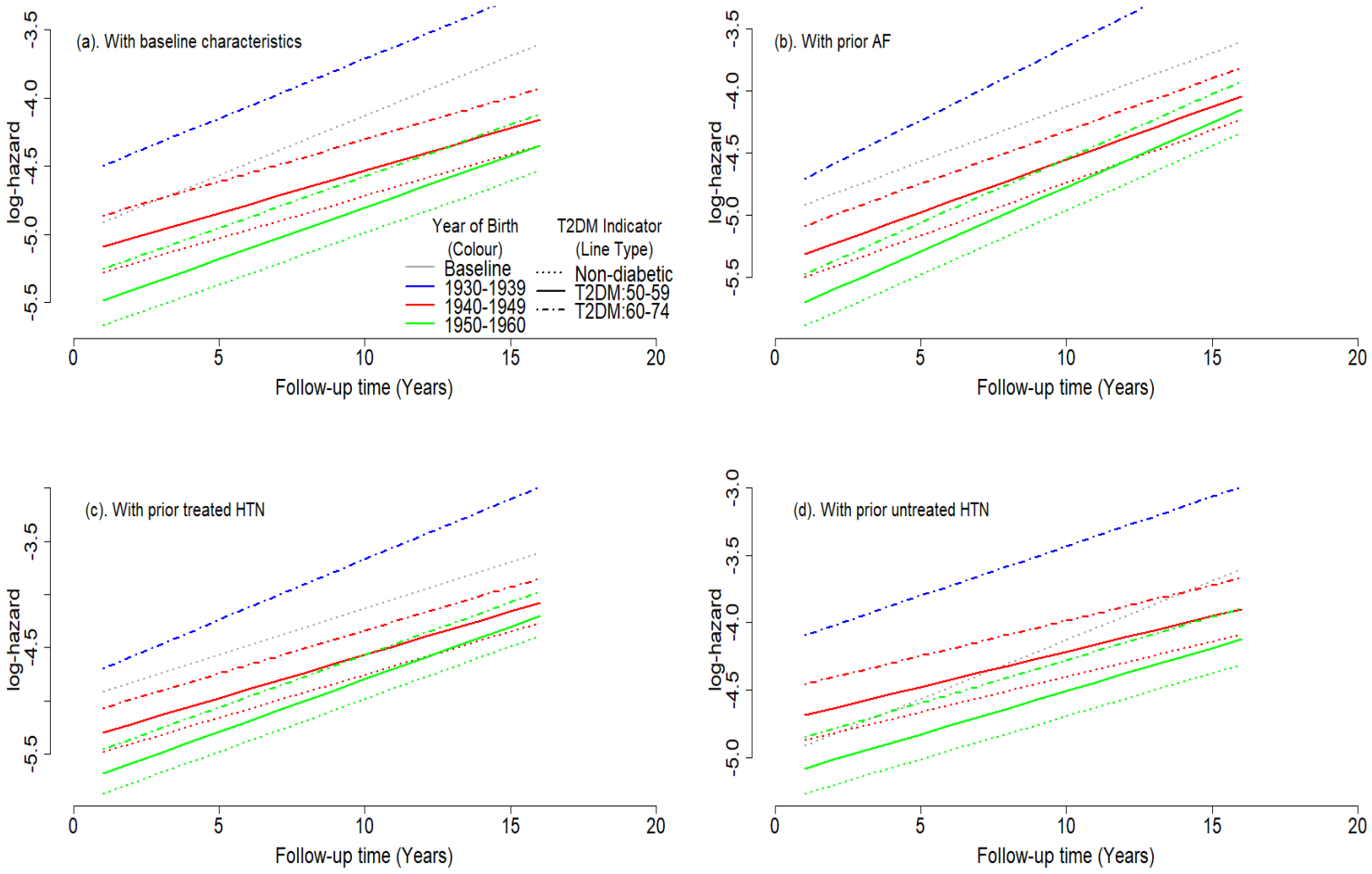


Figure 7.3: Comparison of estimated all-cause mortality risk among people with similar medical conditions by birth cohort and T2DM status using Model A

Model adjusted for age, year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and their interactions.

Chapter 8

Life expectancy results

8.1 Introduction

This chapter is a continuation of the previous chapter, but discussing life expectancy (LE) outcomes calculated using Model B presented in Chapter 7 and a new Model C whose mathematical model is similar to Model B except that it uses a timescale of age at entry and exit. The cumulative hazard function is given in the equation below, with t being valued at t_{start} representing age at study entry and t_{stop}, the study exit age.Life expectancies are calculated using equation provided in Section 4.7 Model C was developed to complement Model A in regards to estimation of LE at a particular age. This chapter presents Model C and the estimated life expectancies for individuals with selected socio-demographic and medical characteristics.

$$\Lambda(t_{start}, t_{stop}|\mathbf{X}) = \mathbf{Z}\frac{\alpha e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{scale}}}}{\upsilon e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{shape}}}}(e^{\upsilon e^{\mathbf{X}'\boldsymbol{\beta}_{\boldsymbol{shape}}}\cdot t} - 1).$$

8.2 Estimates in Model C

Model C used the entry age and exit age instead of the follow-up time used in Model A to estimate the survival model coefficients. Table 8.1 shows that Model C was a

better fit compared to Model A with a C_ρ=0.810 compared to 0.754. Table 8.1 shows major differences in the baseline hazards function and shape modifying covariates and their scale estimates. This is mainly due to the differences in time scales to estimate the baseline hazard functions. In Models A and B, follow-up time is used to determine the risk set regardless of age, while in Model C only people alive at or after a given age constitute the risk set. This explains why Model C is a better fit than Models A and B. The effect of untreated HTN is especially pronounced in Model C with a HR of nearly three compared to people without HTN. Figure 8.1 shows HRs for time-invariant covariates in Model C. They are similar to those in Models A and B except for T2DM Indicator and HCL variables. The differences between Models C and A are in the effects of diagnosis at 50-59 years and 60-74 years of age, respectively. The effect of being diagnosed at 50-59 years in Model C is 0.364 higher than in Model A while being diagnosed at 60-74 years is 0.157 less in Model C than Model A. The effects of having a treated HCL and untreated HCL are 0.092 and 0.198 less in Model C compared to Model A, respectively.

Table 8.1: Scale and shape parameter estimates at baseline and for time-variant covariates in survival models A and C

Parameter estimates for the Gompertz-Cox model with frailty.						
	Model A			**Model C**		
Parameter	Estimate	95% CI	p-value	Estimate	95% CI	p-value
1000a (scale)	6.73	6.24 - 7.25	<1e-16	0.012	0.008-0.017	<1e-16
100b (shape)	8.74	8.13 - 9.39	<1e-16	9.44	9-9.91	<1e-16
			Exponentiated Covariates Shape Parameters			
Birth Cohort						
1930-1939	1			1		
1940-1949	0.71	0.66 - 0.77	<1e-16	1.04	1.01-1.079	<0.0099
1950-1960	0.87	0.77 - 0.97	<0.0186	1.24	1.19-1.29	<1e-16
AF	1.36	1.27 - 1.47	<1e-16	1.12	1.07-1.18	<1e-16
HTN						
None	1			1		
Treated	1.30	1.21 - 1.40	<1e-16	1.18	1.11-1.25	<1e-16
Untreated	0.84	0.76 - 0.94	0.0015	0.88	0.83-0.94	<0.0001
			Exponentiated Covariates Scale Parameters			
Year of Birth						
1930-1939	1			1		
1940-1949	0.71	0.67 - 0.76	<1e-16	0.95	0.74-1.21	<0.6507
1950-1960	0.47	0.43 - 0.52	<0.0036	0.40	0.30-0.54	<1e-16
AF	0.78	0.73 - 0.84	<1e-16	0.38	0.25-0.57	<1e-16
HTN						
None	1			1		
Treated	0.80	0.76 - 0.84	<1e-16	0.26	0.17-0.41	<1e-16
Untreated	1.52	1.47 - 1.6	<1e-16	2.97	2.00-4.43	<1e-16
			Frailty Estimate			
Variance (σ^2)	0.14	0.12 - 0.16	<1e-16	0.003	0.002-0.004	<1e-16
			Goodness of Fit			
Concordance (ρ)		0.754			0.810	
Concordance (std.)		0.002			0.002	
Log likelihood		-145150.22			-144981.88	
AIC		290386.43			290049.76	

Models were adjusted for age, year of birth, gender, BMI, smoking status, TDI, AF, HF, MI, PVD, HCL and HTN and interactions. std. stands for standard deviation.

The ratio of total LE between T2DM cases and controls for people aged 50, 55, 60, 65 and 70 years at entry ranged from 0.87 to 0.92 in Model C and 0.91 to 0.97 in Model A as shown in Table 8.2. The differences in LE of cases and controls were more pronounced in the 1940-1949 and 1950-1960 cohorts. Ratios of LE were calculated by dividing the life expectancy of the cases by the life expectancy of controls at the same

age and with similar characteristics. The effect of time-scale had a significant impact on the estimation of LE. Estimations on the LE using Model B were higher than those given by Model C. As Model C had better goodness-of-fit, it is advised to use it rather than Model B.

Table 8.2: Comparison of life expectancies for females with and without T2DM and ratios of LE of cases to controls at given ages using Models B and C.

Birth Cohort	Model	T2DM Indicator	Age				
			50	55	60	65	70
1930-1939	Model B	Control			29.44	28.01	25.70
		T2DM			26.89	25.99	24.27
		Ratio			0.91	0.93	0.94
		YLL			2.55	2.03	1.43
	Model C	Control			29.42	25.02	20.83
		T2DM			27.06	22.81	18.81
		Ratio			0.92	0.91	0.90
		YLL			2.36	2.21	2.02
1940-1949	Model B	Control	38.30	36.72	34.40	31.38	27.77
		T2DM	35.31	34.28	32.55	30.08	26.92
		Ratio	0.92	0.93	0.95	0.96	0.97
		YLL	2.99	2.44	1.85	1.3	0.85
	Model C	Control	36.51	31.86	27.35	23.02	18.95
		T2DM	32.56	28.04	24.88	20.71	16.84
		Ratio	0.89	0.88	0.91	0.90	0.89
		YLL	3.95	3.82	2.47	2.31	2.11
1950-1960	Model B	Control	34.34	33.62	32.21	29.97	26.94
		T2DM	31.22	30.86	29.97	28.22	25.84
		Ratio	0.91	0.92	0.93	0.94	0.96
		YLL	3.12	2.76	2.24	1.75	1.1
	Model C	Control	31.83	27.17	22.68	18.44	14.52
		T2DM	28.24	23.82	20.42	16.34	12.63
		Ratio	0.89	0.88	0.90	0.89	0.87
		YLL	3.59	3.35	2.26	2.1	1.89

Model adjusted for age, year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and interactions.

Life expectancies from Model B and C are different due to the differences in timescales. It should be noted that life expectancies provided in from both models are cohort based and of stochastic nature due to adjusted time-variant variables' effects. In the UK, the official LE are produced by ONS using a period-life table method. As it is known, period-

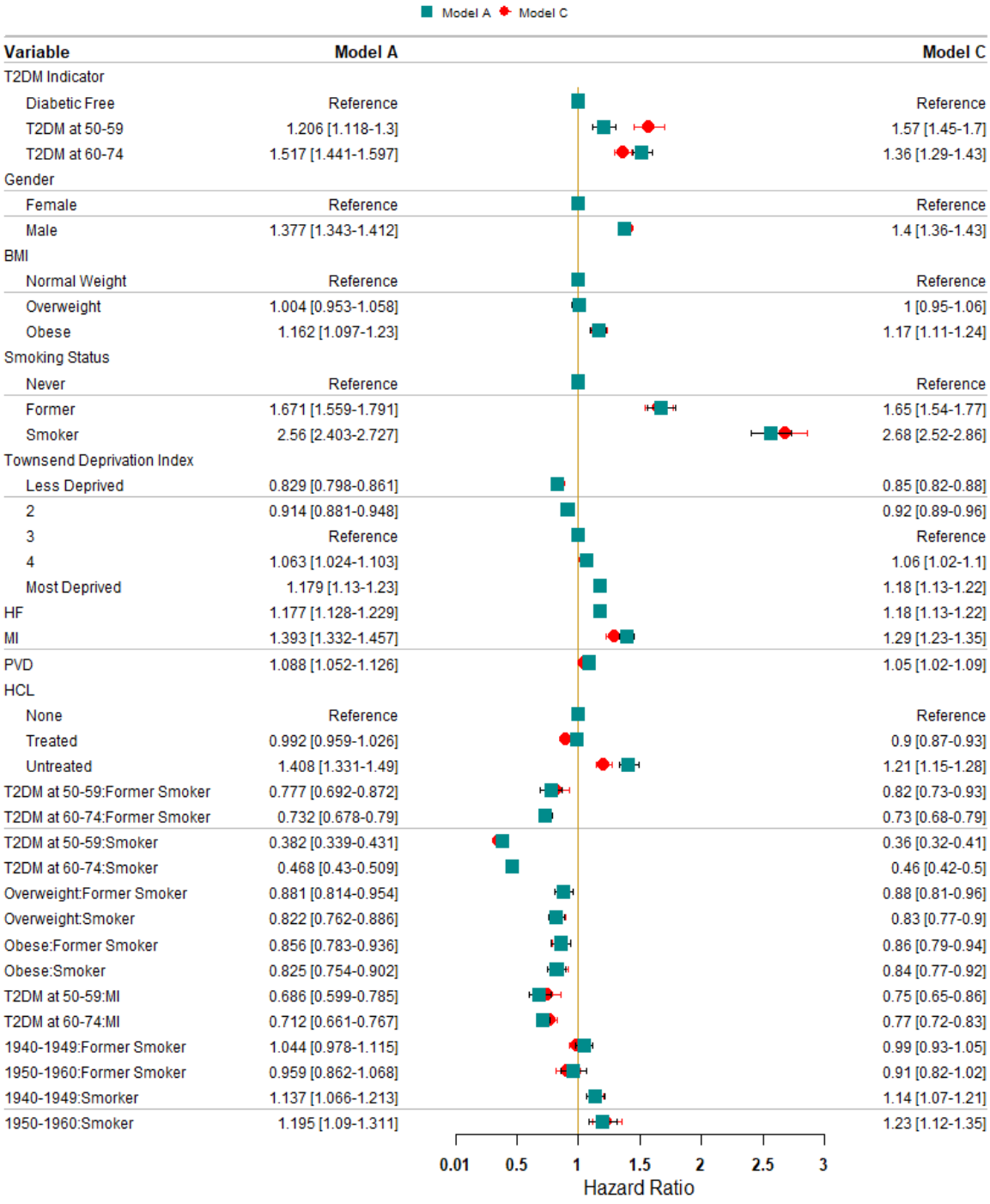


Figure 8.1: Estimated all-cause hazard ratios of time-invariant covariates in Model C compared to Model A

Models adjusted for age, year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and interactions.

life tables do not account for improvements in mortality and risk factors like life style and year of birth (ONS, 2019b). Hence, they are deterministic by nature. It is expected that ONS may slightly differ from LEs estimated by this study and it is possible for period life tables to produce underestimated LE for a group of people born in a given year. For example, ONS reported that in their analysis for people born in 1950, the period life table method would have underestimated by 9 to 10 years (ONS, 2019b). Another possible reason for the differences in LE is ONS calculates average LE while this study is risk based. Confidence intervals for life expectancies estimated in this study will be done in future works on the R Shiny app through the bootstrap method on the data. This will involve producing several parameter estimates from which the average LE together with its 95% CI will be calculated. There are two ways that this can be done (1) parametric bootstrap on estimates (using multivariate normal distribution) and (2) non-parametric bootstrap on data. The later method will involve computing N model estimates from N data samples from the study dataset, compute N life expectancies from each model followed by finding the empirical 95% CIs. Both methods are highly time-consuming.

The effect of the exclusion criteria on the estimated LE is conjectured to be minimal as it was performed on both T2DM patients and people without diabetes at study entry. A comparison of estimates from a dataset with no exclusions with this study's estimates could have been done to calculate the bias. However, due to time constraints this was not performed. The maximum age of 100 years was used based the possible maximum age predicted by the model. This is in line with the maximum age used by ONS (ONS, 2022). The maximum age of follow-up in the study was about 88 giving 12 years of life expectancy forecast. This is less than 50 years used by actuaries in actuarial valuations.

The ratio of the LE of people with baseline characteristics and T2DM to people without diabetes (Figure 8.2a), or smokers (Figure 8.2c) at entry has a convex shape over age while that of former smokers (Figure 8.2b) has a concave shape. The comparison of T2DM and controls who were non-smokers or smokers at entry and born in 1950-1960,

with all other baseline characteristics constant, showed increased differences in LE compared to the 1930-1939 and 1940-1949 birth cohorts. The differences in LEs increased with birth cohort as shown in Section 8.3.1. Thus, while the survival of a T2DM patient diagnosed at age x at age $x+t$ decreased with time, the differences in LEs of cases and controls who had baseline characteristics or were smokers increased and then decreased with time.

However, T2DM patients who were former smokers had higher and exponentially increasing relative risk of survival over time, while the case-control ratios in life expectancies were greater than 1 with time. Life expectancy ratios increased with birth cohort for people who were former smokers at entry. Using model C, the total life expectancy for people with type 2 diabetes mellitus (T2DM), diagnosis at ages 50, 55, 60, 65, and 70 ranged from 12.63 to 32.56 years, while LEs for those without diabetes ranged between 14.52 and 36.51 years. The years of life lost (YLL) associated with T2DM decreased by age at diagnosis, for example cases in the 1950-1960 birth cohort and diagnosed at 50 years of age lost 3.59 years while a case diagnosed at 70 years of age lost 1.89 years. YLL by cases in 1940-1949 birth cohort at different ages of diagnosis were higher than YLL by cases in the 1930-1939 and 1950-1960 birth cohorts.

To compare the results to the UK population, ONS LE calculator based on the national life tables derived from the 2018-2020 data was used. Using 2019 to calculate the age of study subjects and the middle year of the birth cohort LE in the study can be compared to the UK population.

According to ONS, the LE of a female born in 1959, aged 60 years in 2019, is 87 years with a 5.9% chance of reaching 100 years. This is 5.21 years lower than the LE estimated by Model B but 4.32 years higher than in Model C. For a female born in 1949, aged 70 years in 2019, ONS estimated LE to be 88 years with a 4.8% chance of reaching 100 years. This is 9.77 years lower and 0.95 years lower than LEs estimated by Models B and C, respectively. Broadly, the difference in LE estimated by Model C is reduced when compared to ONS estimates than Model B.

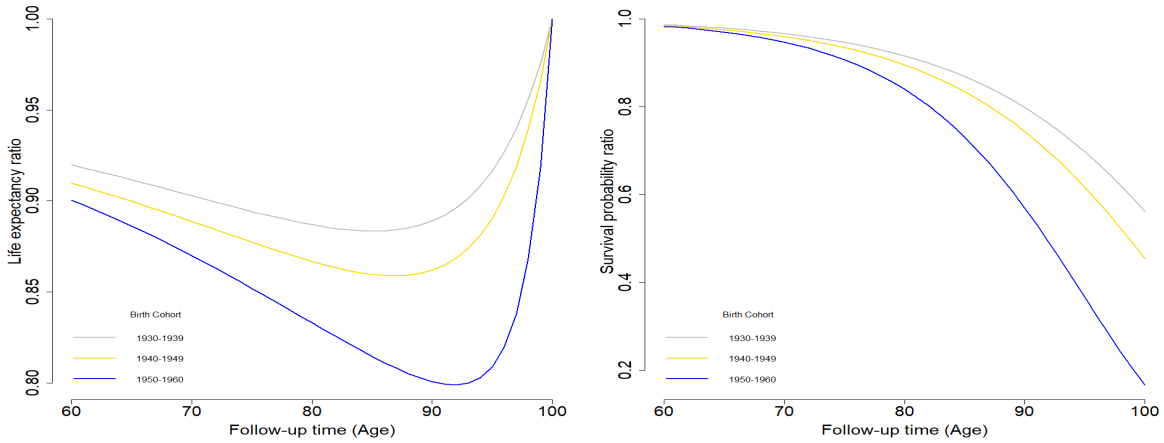
The following section describes the **R Shiny App** that was developed and used to calculated the life expectancy, survival probability at a given age and respective ratios (Ncube, N, 2022).

8.3 Actuarial Translation R Shiny Package

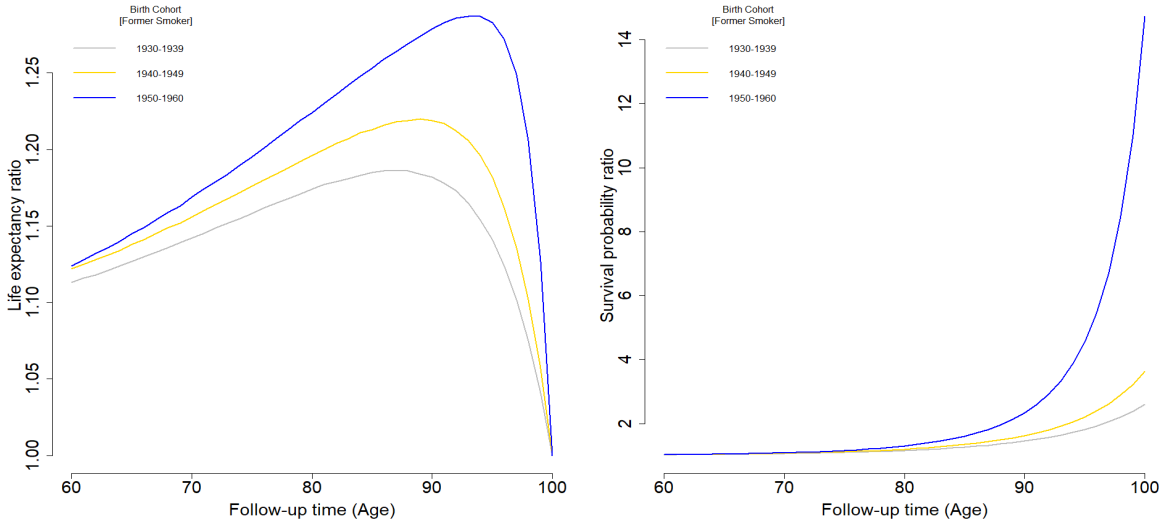
The first version of the **R** package which translated survival models into actuarial models and life expectancy was developed using the **R Shiny App**. It can be accessed on https://njabulo-ncube.shinyapps.io/actuarialtranslation/. The package makes it easy for **R** programmers to build interactive web apps. It makes it easy for stakeholders to engage with **R** packages without having to learn the software. This package allows users to calculate life expectancies after specifying models' variables. Model B and Model C are used to make these calculations and plots. Users can select either of the two, however Model C is recommended as it is a better fit, compared to Model B. The package has 3 sections which are General Settings, Model Variables and Model Estimates/Results. The variables section includes all the models' variables with optional values.

8.3.1 General Settings

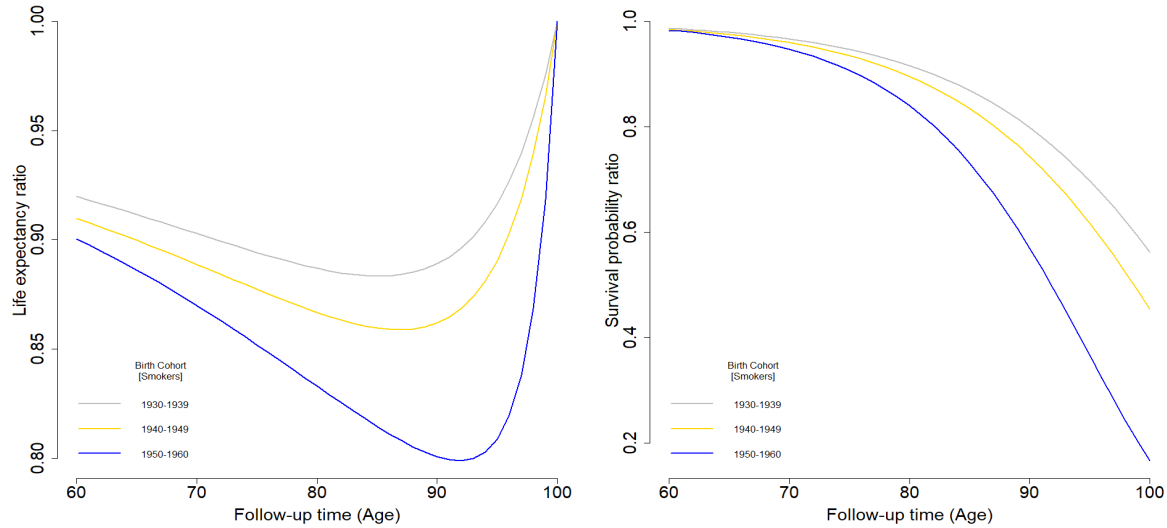
This section allows the user to select general settings that specify how LE results will be calculated and displayed. After selecting the medical condition, the application loads the relevant model to be used to calculate the outcomes. Table 8.3 describes all of general settings. Figure 8.3 shows the snapshot of the screen with the general setting.



(a) People with baseline characteristics at entry



(b) Former Smokers at Entry



(c) Smokers at Entry

Figure 8.2: Visualisation and comparison of life expectancy ratios of cases and controls by birth cohort using Model C

Table 8.3: Description of the package's general settings

Setting	Description	Values
Medical Condition	This drop-down list allows user to select the all-causes mortality model to be used.	• Type 2 diabetes mellitus • Hormone Replacement Therapy
Graph Function	This setting allows the user to plot either a survival or hazard function	• Hazard • Cumulative Hazard • Survival
Log Scale	This setting allows the user to specify whether the hazards should be on a log-scale.	• No • Yes
Age Variable	The user can either use Model B by selecting "Continuous" or Model C by selecting "Grouped"	• Grouped • Continuous
Distribution	This is built-in setting that shows the underlying distribution used by the models.	• Gompertz for T2DM • Weibull for HRT.

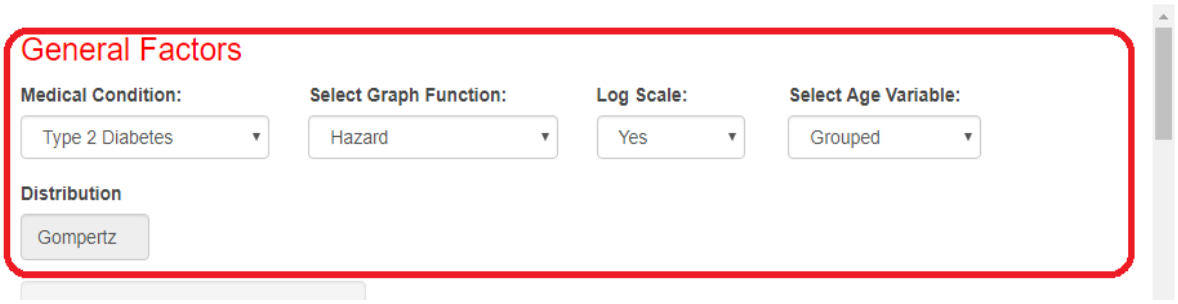


Figure 8.3: Actuarial Model - General Setting Screen

8.3.2 Model Covariates

The model covariates are placed to the right side of the tab panel in the package. To compare the cases to controls the user has to select "Yes" from the Has Medical Condition drop down list. Figure 8.4 shows the covariates section (in red colour) in the package. Users can use the application to calculate hazard ratios and LEs by specifying the person's characteristics.

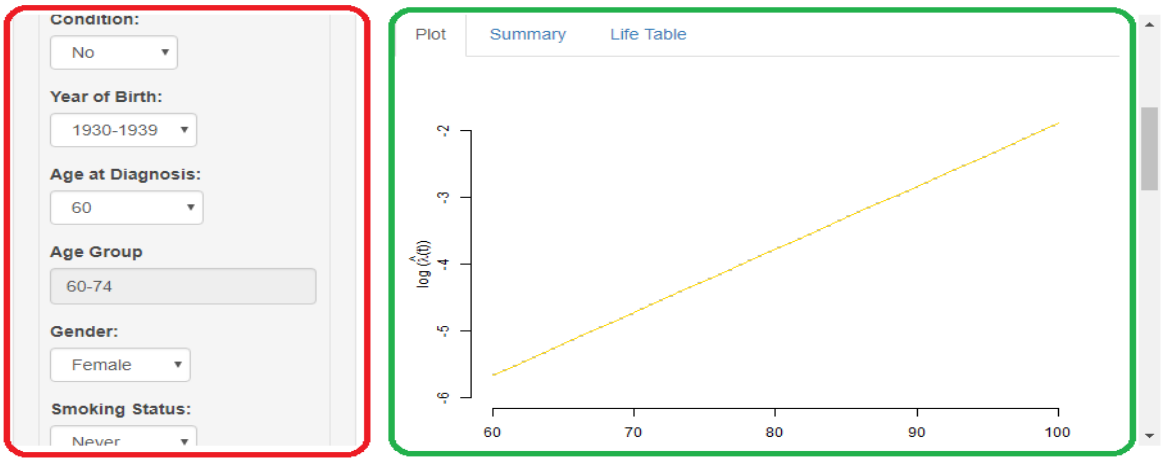


Figure 8.4: Actuarial Model - Covariates

8.3.3 Model Estimates

This section has 3 menu tabs. The first one plots the specified hazard ratios graph on log-scale for a person with selected conditions (Figure 8.4, in green colour). The second shows in summary the life expectancy for people with specified conditions and

model parameters used (Figure 8.5). The third and final tab shows the table with life expectancy by age at follow-up and ratios of survival probabilities and life expectancy (Figure 8.6).

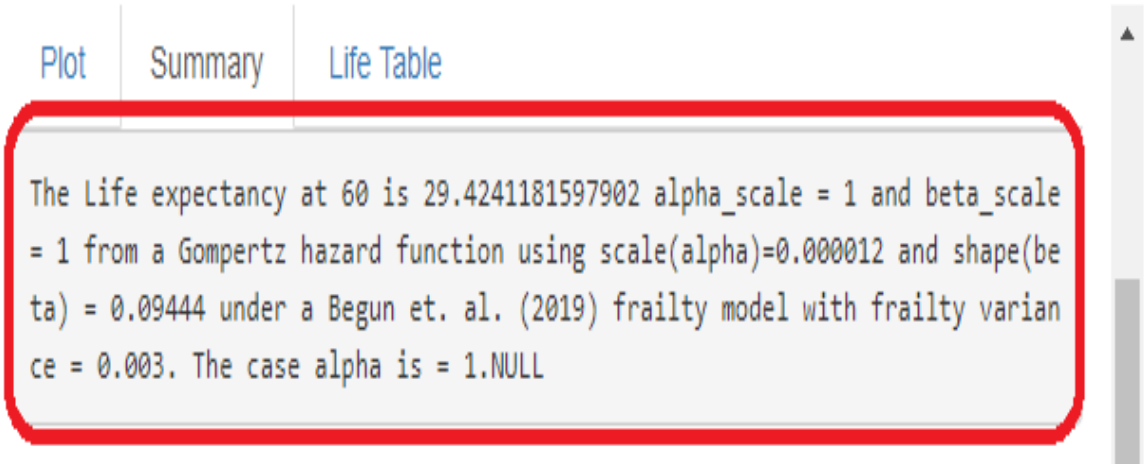


Figure 8.5: Actuarial Model - Summary

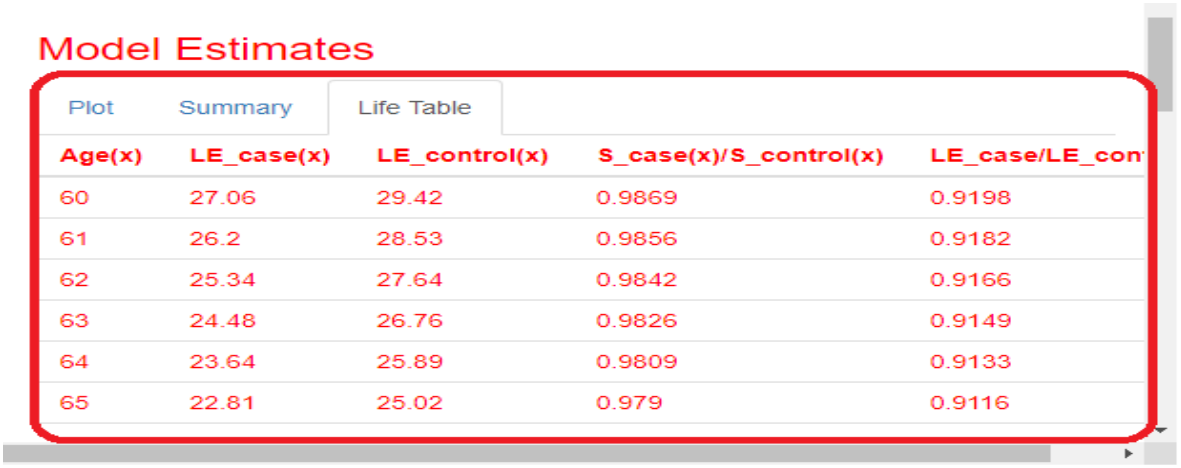


Figure 8.6: Actuarial Model - life expectancy and ratios of LE and survival functions for cases and controls

8.4 Conclusion

The chapter introduced Model C to complement Model A in calculating LE. Model B estimated higher LEs than Model C due to the choice of follow-up time-scale that has an impact on the estimation of the baseline $\hat{S}(t)$. Model C has a better fit than Model B and due to the variability of age at entry of individuals in the study. People with T2DM

and baseline characteristics at diagnosis lost up to a maximum of about 10% of life expectancy when compared to people of the same age without T2DM. The differences in life expectancies reduced at older ages. The following chapter discusses the findings of the study.

Chapter 9

Discussion

This study estimated the impact of T2DM diagnosis on all-cause mortality after adjusting for several socio-demographic, lifestyle and medical factors using survival modelling of EHR in the UK. This chapter discusses the newly developed models and estimated life expectancies. First, the main findings are summarised. Second, the strengths and limitations of this study are reviewed. Third, the aims and implications of this study are addressed. Finally, the overall conclusion is given.

9.1 Main Findings

For this study, medical records from the period of 2000 to 2016, inclusive, from GPs contributing to THIN database were used. People with the first incidence of T2DM were selected and matched to at most three controls. All people included in the study had no history at entry of any conditions in the exclusion list provided in Section 5.2. To estimate the effect of T2DM diagnosis on all-cause mortality, three survival models, Models A, B and C were developed. Models A and B were based on time from diagnosis as the time-scale while Model C used participants' age as the time-scale. Model B was developed to take into account continuous age while Model C complements Model A in estimating life expectancies.

9.1.1 Impact of T2DM on all-cause mortality

The study found that people with T2DM had higher all-cause mortality hazards than people without diabetes and these hazards increased with age at diagnosis when time from diagnosis of T2DM was used as time-scale but declined when age was used as the time-scale. Age as time-scale has been advocated for compared to time from diagnosis as it provides less biased estimates (Hurley, 2015; Columbia Public Health, 2022; Kom et al., 1997). However, the all-cause mortality hazards associated with T2DM estimated in the study were lower than previously reported estimates (Almdal et al., 2004; Mulnier et al., 2006; Taylor et al., 2013). This may be explained by the inclusion of a wider range of people with T2DM and adjustment for more variables compared to previous studies. Furthermore, previous studies did not adjust for age at diagnosis as was implemented in our models.

A previous study by Lind et al. (2013) found a declining trend in relative mortality risk among people with T2DM compared to those without diabetes between 1996 and 2009. The decline in all-cause mortality hazards associated with T2DM has previously been attributed to medical advancement, improved T2DM guidelines and management (ACCORD Study Group, 2008; Currie et al., 2007; Toulis et al., 2017; UKPDS Group and others, 1998; Valentine et al., 2015; Zimmerman et al., 2017). However, this study found the all-cause mortality relative risk associated with T2DM to be constant across all birth cohorts and follow-up time.

There has been a number of previous studies on T2DM in the UK, but they focused mainly on pharmacosurveillance. The impact of T2DM on all-cause mortality compared to controls has not been extensively studied. The previous relevant studies adjusted for at most 3 variables (age, gender, smoking status or entry year), (Lind et al., 2013; Mulnier et al., 2006; Taylor et al., 2013). The exclusion of important variables such as the birth year can have impact on the study population's estimated survival prospects across all years increasing relative risk bias and confounding Hurley (2015).

Additionally, previous studies had relatively short follow-up time and HTN was modelled as a time-invariant factor. This study has shown that the hazard on all-cause mortality associated with HTN had a significant shape effect and hence it should be modelled as time-variant. The study also found that the hazards due to smoking increased in the later birth cohorts.

Another important new finding of this study is the increase in all-cause mortality hazards at later ages in the 1950-60 birth cohort for individuals with or without T2DM.

9.1.2 Impact of T2DM on life expectancy

The life expectancy of people with T2DM was reduced after diagnosis and it was further somewhat reduced in later birth cohorts. Broadly, the ratio of LE between cases and controls decreased with age or time after diagnosis.

This current study also found that life expectancy was reduced in people diagnosed with T2DM at a younger age. In as much as there has been medical advances, it can not be inferred with certainty that these advances had an effect on the relative life expectancy of people with T2DM. The differences in the shape of life expectancy ratio over follow-up time scale across all birth cohorts can be conjectured to be due to the socio-political and socio-economical conditions the different cohorts were exposed to and their life styles. However, there can be other serious risk factors impacting life expectancy in people with and without T2DM.

Based on Model B, life expectancy was higher for the 1940-49 birth cohort compared to the preceding and succeeding birth cohorts. This finding is supported by ONS findings in their life expectancy projections done in 2018 and preceding years. However, their methodology does not allow for future assumed changes in mortality rates (ONS, 2019a). Life style changes were found to contribute positively to life expectancy in T2DM patients. For example, former smokers without diabetes at entry had a reduced life expectancy compared to former smokers with T2DM. Healthy life styles at T2DM

diagnosis resulted in improved survival. This improvement among T2DM patients increased by year of birth. However, the opposite was true for smokers. It can be concluded that life expectancy ratio is dependent on the life style of an individual.

When compared to people without diabetes and no other risk factors, T2DM patients with no other risk factors a reduced LE ranging from 0.1 years to 6 years. This concurred with what was reported by Walker et al. (2018); Wright et al. (2016). However, any additional mortality risks would decrease LE in people with and without T2DM. Another secondary finding from this study was the decline in life expectancy among recent birth cohorts. A further study that includes people born after 1960 would help in ascertaining this finding.

This finding supports a recent study by Rashid et al. (2021) that found that the number of middle-layer super output areas (MSOA) in England with a decline in life expectancy in women increased by 262% in 2014-2019 compared to 2010-2014, out of the total of 6791 MSOAs. Though the study by Rashid et al. (2021) was based on the England population, England had a respectable 84.2% of the UK population. In addition, they reported that the detrimental in mortality trends in the UK began from 2010.

9.2 Strengths and Limitations

9.2.1 Strengths

The population studied in this research was drawn from the EHRs and was representative of the UK population (Blak et al., 2011; Hippisley-Cox and Coupland, 2010). The matched controls were drawn from the same source population, hence valid comparisons were made with no selection bias. The selection and follow-up periods were 17 years, this means that a considerable number of death events could be observed providing for more accurate estimations of life expectancies. This also provided the insight into the

past and the present well-being of the study population. Adjusting for multiple risk factors in our models and developing time-variant effects models, it became possible to identify potential longevity risks.

By matching cases to controls and adjusting for a multiple risk factors, this large study minimised selection biases (such as sick-user bias and healthy-user bias), information bias and bias by indication. Furthermore, dependencies among patients in a GP were accounted for by including the GP as a latent (frailty) variable. Imputed data were validated against the complete data and no differential outcomes were found as a result of imputation.

The Cox PH model is the standard method of survival analysis used in assessing survival prospects of a population. By its very nature, the Cox model has a strong assumption on PH. The model by Begun et al. (2019) is a generalisation of the Cox model that allows modelling of the baseline hazards shape parameter to adjust for time-variant variables while retaining the Cox model structure. This model has a better statistical power in estimating time-variant hazards.

Finally, with respect to retirement planning and computation of reserves of insurance products, the findings of this study are informative for actuarial valuations, financial planning of retirement for individuals, insurance products pricing and social security administration.

9.2.2 Limitations

Though the study included a considerable number of statistically significant variables, these were not exhaustive. Hence, there may be potential residual confounding due to unavailable data such as changes in therapy or lifestyle, severity of smoking and other unrecorded variables. The exclusion of ethnicity and antidiabetic drugs in our models due to high percentage of missing values, limited the study in assessing their impact on all-cause mortality. According to ONS (2021), the proportion of BAME population

has been on the increase in the UK. Assuming uniform increases, between 1991 to 2011 the proportion of BAME increased by 8.1 percentage points per year and then by 7.1 percentage points per year between 2011 and 2019. As the BAME population is on the increase in the UK, it is important for GPs to mandatorily capture this variable. Another variable that was excluded due to a high proportion of missing values was the HbA_{1c}. As a measure that is used to diagnose diabetes, it is imperative that this variable be captured from the onset of diagnosis. High missing values in HbA_{1c} restrict research into effects of severity of T2DM at diagnosis. The life expectancy confidence intervals have not been provided in this study but will be included in future works on the R Shiny application.

9.3 Conclusions

In conclusion, the hazard ratios (HRs) associated with T2DM were somewhat lower than previously reported, at 1.21-1.52. This was due to the differences in the selection period, selection by first incidence in this study and not presence of disease as in other previous studies, poor adjustment in previous studies, and/ or differences in study population. However, the hazards were constant across all birth cohorts demonstrating a lack of progress in reducing relative risks of mortality associated with T2DM over time. Pre-existing medical conditions and, in particular, untreated HTN and smoking increased the mortality hazards and their effects increased in later birth cohorts for both people with and without T2DM. In addition, this study has found that the mortality hazards were higher at older ages in the younger birth cohort. The poor mortality experience in the 1950-1960 birth cohort, merits further research on individuals born after 1960 to explore the increased all-cause mortality hazards in recent birth cohorts. A diagnosis of T2DM reduced life expectancy by 1.1 to 3.95 years compared to people without diabetes at entry. An assessment of the life expectancy of people with first incidence of T2DM that includes confidence intervals and significant other medical complications (such as CKD 3-5, cancer and dementia), that were excluded from this study will be a subject

of further research.

Bibliography

ACCORD Study Group (2008). Effects of intensive glucose lowering in type 2 diabetes. New England Journal of Medicine, 358(24):2545–2559.

ADVANCE Collaborative Group (2008). Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. New England Journal of Medicine, 358(24):2560–2572.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716–723.

Ali, O. (2013). Genetics of type 2 diabetes. World journal of diabetes, 4(4):114.

Almdal, T., Scharling, H., Jensen, J. S., and Vestergaard, H. (2004). The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: a population-based study of 13 000 men and women with 20 years of follow-up. Archives of Internal Medicine, 164(13):1422–1426.

Arora, S. (2017). The relationship between health spending and life expectancy. Longevity Bulletin, Issue(10):15–17.

Austin, P. C., White, I. R., Lee, D. S., and van Buuren, S. (2021). Missing data in clinical research: a tutorial on multiple imputation. Canadian Journal of Cardiology, 37(9):1322–1331.

Bagust, A., Hopkinson, P., Maslove, L., and Currie, C. (2002). The projected health care burden of Type 2 diabetes in the UK from 2000 to 2060. PubMed, 19(4):1–5.

Barnett, A. H. (2006). Best Practice and Research Compendium. Elsivier Limited.

Barr, E. L., Zimmet, P. Z., and Shaw, J. E. (2008). Mortality and life expectancy associated with diabetes. The Epidemiology of Diabetes Mellitus, Second Edition, pages 603–625.

Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., and Salhi, Y. (2012). Understanding, modelling and managing longevity risk: key issues and main challenges. Scandinavian actuarial journal, 2012(3):203–231.

Begun, A., Kulinskaya, E., and MacGregor, A. J. (2019). Risk-adjusted cusum control charts for shared frailty survival models with application to hip replacement outcomes: a study using the njr dataset. BMC medical research methodology, 19(1):217.

Begun, A. and Begun, F. (2022). Parametric gamma-frailty models with non-proportional hazard functions [computer software].

Bilous, R. and Donnely, R. (2010). Handbook of Diabetes *4th Edition*. Wiley-Blackwell.

Blak, B., Thompson, M., Dattani, H., and Bourke, A. (2011). Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Journal of Innovation in Health Informatics, 19(4):251–255.

Blood Pressure UK. What do the numbers mean? https://www.bloodpressureuk.org/your-blood-pressure/understanding-your-blood-pressure/what-do-the-numbers-mean/. Accessed: 11/03/2022.

Bourne, R. R., Stevens, G. A., White, R. A., Smith, J. L., Flaxman, S. R., Price, H., Jonas, J. B., Keeffe, J., Leasher, J., and Naidoo, K. (2013). Causes of vision loss worldwide, 1990–2010: a systematic analysis. The Lancet Global Health, 1(6):e339–e349.

Boussageon, R., Bejan-Angoulvant, T., Saadatian-Elahi, M., Lafont, S., Bergeonneau, C., Kassaï, B., Erpeldinger, S., Wright, J. M., Gueyffier, F., and Cornu, C. (2011).

Effect of intensive glucose lowering treatment on all cause mortality, cardiovascular death, and microvascular events in type 2 diabetes: meta-analysis of randomised controlled trials. BMJ, 343:d4169.

Bowers, N., Gerber, H., Hickman, J., Jones, D., and Nesbitt, C. (1997). Actuarial Mathematics. Society of Actuaries.

CIPR-NAIC (2018). Longevity risk. www.naic.org/cipr_topics/topic_longevity_risk.htm. Accessed: 27/04/2018.

ClinicalCodes.org (2021). Clinicalcodes. https://clinicalcodes.rss.mhs.man.ac.uk. Accessed: 08/11/2021.

Columbia Public Health (2022). Time-to-event data analysis. https://www.publichealth.columbia.edu/research/population-health-methods/time-event-data-analysis. Accessed: 16/03/2022.

Currie, C. J., Gale, E., and Poole, C. D. (2010a). Estimation of primary care treatment costs and treatment efficacy for people with type 1 and type 2 diabetes in the united kingdom from 1997 to 2007. Diabetic Medicine, 27(8):938–948.

Currie, C. J., Peters, J. R., Tynan, A., Evans, M., Heine, R. J., Bracco, O. L., Zagar, T., and Poole, C. D. (2010b). Survival as a function of hba1c in people with type 2 diabetes: a retrospective cohort study. The Lancet, 375(9713):481–489.

Currie, C. J., Poole, C. D., Tetlow, T., Holmes, P., and McEwan, P. (2007). The outcome of care in people with type 1 and type 2 diabetes following switching to treatment with either insulin glargine or insulin detemir in routine general practice in the uk: A retrospective database analysis. Current Medical Research and Opinion, 23(sup1):S33–S39.

Daousi, C., Casson, I., Gill, G., MacFarlane, I., Wilding, J., and Pinkney, J. (2006). Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. Postgraduate medical journal, 82(966):280–284.

Diabetes UK (2013). Facts and figures: Diabetes prevalence 2012. diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevelence-2012. Accessed: 08/11/2017.

Diabetes UK (2014). Facts and figures: Diabetes prevalence 2013. diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevelence-2013. Accessed: 08/11/2017.

Diabetes UK (2015a). Facts and figures: Diabetes prevalence 2014. diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevelence-2014. Accessed: 08/11/2017.

Diabetes UK (2015b). Facts and figures: Diabetes prevalence 2015. diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevelence-2015. Accessed: 08/11/2017.

Diabetes UK (2016a). Diabetes in the UK 2010: Key statistics on diabetes. www.diabetes.org.uk/resources-s3/2017-11/diabetes_in_the_uk_2010.pdf. Accessed on 08/08/2019.

Diabetes UK (2016b). Facts and stats. diabetes.org.uk/professionals/position-statements-reports/statistics. Accessed: 08/11/2017.

Diabetes UK (2017). About insulin. www.diabetes.co.uk/about-insulin.html. Accessed: 08/11/2017.

Diabetes UK (2019). Diabetes and Alcohol. https://www.diabetes.co.uk/diabetes-and-alcohol.html. Accessed: 15/08/2019.

EC (2007). Solvency II Glossary. http://ec.europa.eu/internal_market/insurance/docs/solvency/impactassess/annex-c08d_en.pdf. Accessed: 27/04/2018.

Eurostat (2017). Mortality and life expectancy statistics. ec.europa.eu/eurostat/statistics-explained/index.php/Mortality-and-life-expectancy-statistics. Accessed: 12/11/2017.

Eurostat (2021). Mortality and life expectancy statistics. ec.europa.eu/eurostat/statistics-explained/index.php/Mortality-and-life-expectancy-statistics. Accessed: 12/09/2021.

Falkous, C. (2017). Current Changes in Life Expectancy in UK. Longevity Bulletin, Issue(10):6–9.

Fox, C. D. and Kilvert, A. D. (2007). Type 2 Diabetes. London.

Franco, O. H., Steyerberg, E. W., Hu, F. B., Mackenbach, J., and Nusselder, W. (2007). Associations of diabetes mellitus with total life expectancy and life expectancy with and without cardiovascular disease. Archives of Internal Medicine, 167(11):1145–1151.

Go, A. S., Yang, J., Ackerson, L. M., Lepper, K., Robbins, S., Massie, B. M., and Shlipak, M. G. (2006). Hemoglobin level, chronic kidney disease, and the risks of death and hospitalization in adults with chronic heart failure: the Anemia in Chronic Heart Failure: Outcomes and Resource Utilization (ANCHOR) Study. Circulation, 113(23):2713–2723.

Government, U. (2022). Population screening programmes: detailed information. https://www.gov.uk/topic/population-screening-programmes. Accessed: 29/09/2022.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60:549–576.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3):515–526.

Hall, G. C. (2009). Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiology and drug safety, 18(2):120–131.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine, 15(4):361–387.

Hayes, A., Leal, J., Gray, A., Holman, R., and Clarke, P. (2013). UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia, 56(9):1925–1933.

Hiam, L., Dorling, D., Harrison, D., and McKee, M. (2017). What caused the spike in mortality in england and wales in january 2015? Journal of the Royal Society of Medicine, 110(4):131–137.

Hippisley-Cox, J. and Coupland, C. (2010). Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney® Scores. BMC Family Practice, 11(1):49.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Wiley-Interscience Publication.

Hurley, M. A. (2015). A reference relative time-scale as an alternative to chronological age for cohorts with long follow-up. Emerging themes in epidemiology, 12(1):1–12.

IDF (2015). IDF Atlas 2015. https://www.idf.org/e-library/epidemiology-research/diabetes-atlas/13-diabetes-atlas-seventh-edition.html. Accessed: 22/05/2018.

IDF (2021). IDF Atlas 2021. https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf. Accessed: 08/03/2021.

IEC (2009). Interbational Expert Committee Report on the Role of the HbA_{1c} Assay in the Diagnosis of Diabetes. Diabetes Care, 32(7):1327–1333.

IFoA (2011). Longevity Bulletin: Is life ecpectancy the life span we should expect. https://www.actuaries.org.uk/search/site. Accessed: 22/05/2018.

IFoA (2016). Longevity Bulletin: Antimicrobial resistance. https://www.actuaries.org.uk/search/site/. Accessed: 22/05/2018.

IFoA (2017). Longevity Bulletin: Longevity: Is the tide turning? https://www.actuaries.org.uk/search/site/. Accessed: 28/09/2021.

Jeong, J., Um, Y. H., Ko, S., Park, J., Park, J., Han, K., Ko, K., et al. (2017). Depression and mortality in people with type 2 diabetes mellitus, 2003 to 2013: a nationwide population-based cohort study. Diabetes & Metabolism Journal, 41(4):296–302.

Johnson, J. and Booman, L. (1996). Drug-related morbidity and mortality. Journal of Managed Care Pharmacy, 2(1):39–47.

Kalbfleisch, J. D. and Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. Wiley-Interscience Publication.

Kennard, L. and O'Shaughnessy, K. M. (2016). Treating hypertension in patients with medical comorbidities. BMJ.

Khanolkar, M., Bain, S., and Stephens, J. (2008). The diabetic foot. QJ Moed, 101(9):685–695.

Kim, H. M., Lee, Y.-h., Han, K., Lee, B.-W., Kang, E. S., Kim, J., and Cha, B.-S. (2017). Impact of diabetes mellitus and chronic liver disease on the incidence of dementia and all-cause mortality among patients with dementia. Medicine, 96(47):e8753.

Klein, J. K. and Moeschberger, M. L. (2003). SURVIVAL ANALYSIS: Techniques for Censored and Truncated Data. Springer.

Kleinbaum, D. G. and Klein, M. (2012). SURVIVAL ANALYSIS: A Self-Learning Text. Springer.

Kom, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. American journal of epidemiology, 145(1):72–80.

Kontopantelis, E., Springate, D. A., Reeves, D., Ashcroft, D. M., Rutter, M., Buchan, I., and Doran, T. (2015). Glucose, blood pressure and cholesterol levels and their relationships to clinical outcomes in type 2 diabetes: a retrospective cohort study. Diabetologia, 58(3):505–518.

Leal, J., Gray, A. M., and Clarke, P. M. (2008). Development of life-expectancy tables for people with type 2 diabetes. European Heart Journal, 30(7):834–839.

Lewis, J. D., Schinnar, R., Bilker, W. B., Wang, X., and Strom, B. L. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiology and drug safety, 16(4):393–401.

Lim, S., Eckel, R. H., and Koh, K. K. (2018). Clinical implications of current cardiovascular outcome trials with sodium glucose cotransporter-2 (SGLT2) inhibitors. Atherosclerosis, 272:33–40.

Lind, M., Garcia-yz, L., Booth, G., Cea-Soriano, L., Shah, B., Ekeroth, G., and Lipscombe, L. (2013). Mortality trends in patients with and without diabetes in Ontario, Canada and the UK from 1996 to 2009: a population-based study. Diabetologia, 56(12):2601–2608.

Lu, J. (2017). Summary on the Longevity Bulletin Issue 10. Longevity Bulletin, Issue(10):4–5.

Malavige, L. S. and Levy, J. C. (2009). Erectile dysfunction in diabetes mellitus. The Journal of Sexual Medicine, 6(5):1232–1247.

Matthews, D. et al. (2008). Diabetes. Oxford University Press.

McKeague, I. W. and Zhang, M.-J. (1996). Fitting cox's proportional hazards model using grouped survival data. In Lifetime Data: Models in Reliability and Survival Analysis, pages 227–232. Springer.

Miller, R. G., Gong, G., and Munoz, A. (1998). Survival Analysis. Wiley-Interscience Publication.

Mulnier, H., Seaman, H., Raleigh, V., Soedamah-Muthu, S., Colhoun, H., and Lawrenson, R. (2006). Mortality in people with type 2 diabetes in the UK. Diabetic Medicine, 23(5):516–521.

Ncube, N (2022). Parametric gamma-frailty models with non-proportional hazard functions [computer software] - life expectancy calculation.

NHC (2022). Uk nsc recommendations. https://view-health-screening-recommendations.service.gov.uk/. Accessed: 29/09/2022.

NHS (2017). Cirrhosis. www.nhs.uk/conditions/cirrhosis. Accessed: 30/01/2018.

NHS (2019a). Cholesterol levels: High cholesterol. https://www.nhs.uk/conditions/high-cholesterol/cholesterol-levels/. Accessed: 29/04/2020.

NHS (2019b). Nhs health check. https://www.nhs.uk/conditions/nhs-health-check/. Accessed: 29/09/2022.

NICE (2012a). Peripheral arterial disease: diagnosis and management (2012): NICE Guideline CG147. pathways.nice.org.uk/pathways/lower-limb-peripheral-arterial-disease. Accessed: 22/05/2018.

NICE (2012b). Preventing type 2 diabetes: risk identification and interventions for individuals at high risk. https://www.nice.org.uk/guidance/ph38. Accessed: 22/05/2018.

NICE (2014). Obesity: identification, assessment and management. www.nice.org.uk/guidance/cg189/chapter/. Accessed: 08/10/2019.

NICE (2015). Type 2 diabetes in adults: management. https://www.nice.org.uk/guidance/ng28/resources/type-2-diabetes-in-adults-management-pdf-1837338615493. Accessed: 21/05/2018.

NICE (2016a). Diabetes Foot Problems: Prevention and Management (2016): NICE Guideline NG19. pathways.nice.org.uk/pathways/foot-care-for-people-with-diabetes. Accessed: 22/05/2018.

NICE (2016b). Non-alcoholic fatty liver disease (NAFLD): assessment and management (2016): NICE Guideline NG49. pathways.nice.org.uk/pathways/non-alcoholic-fatty-liver-disease. Accessed: 22/05/2018.

NICE (2017a). Identification and assessment of cardiovascular disease risk. pathways.nice.org.uk/pathways/cardiovascular-disease-prevention. Accessed: 23/02/2018.

NICE (2017b). Managing blood glucose in adults with type 2 diabetes. pathways.nice.org.uk/pathways/type-2-diabetes-in-adults. Accessed: 24/02/2018.

NICE (2017c). Type 2 diabetes: prevention in people at high risk. www.nice.org.uk/guidance/PH38/chapter/Recommendations/risk-assessment. Accessed: 26/01/2018.

NICE (2019). Hypertension in adults: diagnosis and management (guideline ng136). https://www.nice.org.uk/guidance/NG136. Accessed: 29/04/2020.

NICE (2021). What are the risk factors? https://cks.nice.org.uk/topics/diabetes-type-2. Accessed: 29/09/2022.

ONS (2016). National life tables UK:2013–2015. ons.gov.uk/peoplepopulationandcommunity/. Accessed: 07/03/2017.

ONS (2017). Deaths registered in England and Wales (series DR). ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/. Accessed: 22/02/2018.

ONS (2018). Health and social care. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare. Accessed: 27/01/2019.

ONS (2019a). National population projections, mortality assumptions: 2018-based. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies. Accessed: 24/03/2022.

ONS (2019b). Period and cohort life expectancy explained: December 2019. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/methodologies/periodandcohortlifeexpectancyexplained. Accessed: 01/11/2022.

ONS (2020). Life expectancy at birth and selected older ages. https://www.ons.gov.uk/peoplepopulationandcommunity/. Accessed: 07/03/2022.

ONS (2021). Deaths registered in England and Wales (series DR). https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/. Accessed: 06/03/2022.

ONS (2022). Life expectancy calculator. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/articles/lifeexpectancycalculator/2019-06-07. Accessed: 15/12/2022.

PHE (2016a). Diabetes footcare activity profiles. Using the average annual number of amputations per year from 2012-2015.

PHE (2016b). Diabetes Prevalence Model. https://www.gov.uk/government/publications/diabetes-prevalence-estimates-for-local-populations. Accessed: 22/05/2018.

PHE (2016c). Technical document for the diabetes prevalence model for England 2016. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612307/Diabetesprevalencemodeltechnicaldocument.pdf. Accessed: 22/05/2018.

QuintilesIMS (2017). THIN Data Guide for Researchers Version 1701. QuintilesIMS.

Rashid, T., Bennett, J. E., Paciorek, C. J., Doyle, Y., Pearson-Stuttard, J., Flaxman, S., Fecht, D., Toledano, M. B., Li, G., Daby, H. I., et al. (2021). Life expectancy and risk of death in 6791 communities in england from 2002 to 2019: high-resolution spatiotemporal analysis of civil registration data. The Lancet Public Health.

Reijneveld, S. A., Verheij, R. A., and de Bakker, D. H. (2000). The impact of area deprivation on differences in health: does the choice of the geographical classification matter? Journal of Epidemiology & Community Health, 54(4):306–313.

Roberts, S., Barry, E., Craig, D., Airoldi, M., Bevan, G., and Greenhalgh, T. (2017). Preventing type 2 diabetes: systematic review of studies of cost-effectiveness of lifestyle programmes and metformin, with and without screening, for pre-diabetes. BMJ Open, 7(11):e017184.

Robinson, S., Harris, H., and Steven, D. (2011). Smoking and drinking among adults, 2009. http://doc.ukdataservice.ac.uk/doc/6716/mrdoc/pdf/2009_report.pdf. Accessed: 24/02/2019.

Roderick, JA, L. and Rubin, D. (1987). Statistical analysis with missing data. New York: John Wiley & Sons.

Rodrıguez, G. (2010). Parametric survival models. Rapport technique, Princeton: Princeton University.

Rubin, D. (1987). Multiple imputation for nonresponse in surveys John Wiley.

Salinero-Fort, M. Á., San-Andrés-Rebollo, F. J., Jiménez-García, R., and de Miguel-Yanes, J. M. (2018). Body mass index and all-cause mortality among type 2 diabetes mellitus patients. European Journal of Internal Medicine, 47:e28.

Scobie, I. and Katherine, S. (2014). Diabetes Mellitus. Health Press Limited.

Soriano, L. C., Johansson, S., Stefansson, B., and Rodríguez, L. A. G. (2015). Cardiovascular events and all-cause mortality in a cohort of 57,946 patients with type 2 diabetes: associations with renal function and cardiovascular risk factors. Cardiovascular Diabetology, 14(1):38.

Stanford Medicine (2018). Measuring and Understanding the Ankle Brachial Index (ABI). www.stanfordmedicine25.stanford.edu/the25/ankle.html. Accessed: 30/01/2018.

Steven, D. (2012). General life style survey overview, 2010. http://doc.ukdataservice.ac.uk/doc/6716/mrdoc/pdf/2010_report.pdf. Accessed: 24/02/2019.

Taylor, K. S., Heneghan, C. J., Farmer, A. J., Fuller, A. M., Adler, A. I., Aronson, J. K., and Stevens, R. J. (2013). All-cause and cardiovascular mortality in middle-aged people with type 2 diabetes compared with people without diabetes in a large UK primary care database. Diabetes Care, 36(8):2366–2371.

Therneau, T. and Atkinson, E. (2020). 1 the concordance statistic.

Toulis, K. A., Willis, B. H., Marshall, T., Kumarendran, B., Gokhale, K., Ghosh, S., Thomas, G. N., Cheng, K. K., Narendran, P., Hanif, W., et al. (2017). All-cause mortality in patients with diabetes under treatment with dapagliflozin: a population-based, open-cohort study in The Health Improvement Network Database. The Journal of Clinical Endocrinology & Metabolism, 102(5):1719–1725.

UCL (2013). THIN Data: Strengths and Limitations. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database/pros-cons. Accessed: 18/05/2018.

UKPDS Group and others (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). The Lancet, 352(9131):837–853.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine, 30(10):1105–1117.

Valentine, W., Curtis, B., Pollock, R., Van Brunt, K., Paczkowski, R., Brändle, M., Boye, K., and Kendall, D. (2015). Is the current standard of care leading to cost-effective outcomes for patients with type 2 diabetes requiring insulin? a long-term health economic analysis for the uk. Diabetes research and clinical practice, 109(1):95–103.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

Walker, J., Colhoun, H., Livingstone, S., McCrimmon, R., Petrie, J., Sattar, N., and Wild, S. (2018). Type 2 diabetes, socioeconomic status and life expectancy in scotland (2012–2014): a population-based observational study. Diabetologia, 61(1):108–116.

Watanabe, J. H., McInnis, T., and Hirsch, J. D. (2018). Cost of prescription drug–related morbidity and mortality. Annals of Pharmacotherapy, 52(9):829–837.

WHO (2006). Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. http://www.who.int/diabetes/publications/diagnosis_diabetes2006/en/. Accessed:22/05/2018.

WHO (2011). Use of glyserated haemoglobin (hba_{1c}) in the diagnosis of diabetes mellitus. abbrevated report of a WHO consultation. http://www.who.int/diabetes/publications/report-hba1c_2011.pdf. Accessed: 22/05/2018.

WHO (2014). Global status seport on alcohol and health 2014. https://books.google.co.uk/. Accessed: 15/08/2019.

WHO (2017). Dementia. http://www.who.int/mediacentre/factsheets/fs362/en/. Accessed: 11/04/2018.

WHO (2017a). Fact Sheets 310. who.int/mediacentre/factsheets/fs310/en/. Accessed: 03/11/2017.

WHO (2017b). Fact Sheets 312. who.int/mediacentre/factsheets/fs312/en. Accessed: 02/11/2017.

WHO (2017c). The top 10 causes of death. www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 11/09/2021.

WHO (2021). International classification of diseases 11th revision. https://icd.who.int/en. Accessed: 08/11/2021.

Wintrebert, C. (2007). Statistical modelling of repeated and multivariate survival data. PhD Thesis. Leiden University Medical Center, Leiden.

Wright, A. K., Kontopantelis, E., Emsley, R., Buchan, I., Sattar, N., Rutter, M. K., and Ashcroft, D. M. (2016). Life expectancy and cause-specific mortality in type 2 diabetes: a population-based cohort study quantifying relationships in ethnic subgroups. Diabetes Care, page dc161616.

Yudkin, J. S. and Montori, V. M. (2014). The epidemic of pre-diabetes: the medicine and the politics. The BMJ, 349:e4485.

Zimmerman, R. S., Hobbs, T. M., Wells, B. J., Kong, S. X., Kattan, M. W., Bouchard, J., Chagin, K. M., Yu, C., Sakurada, B., Milinovich, A., et al. (2017). Association of glucagon-like peptide-1 receptor agonist use and rates of acute myocardial infarction, stroke and overall mortality in patients with type 2 diabetes mellitus in a large integrated health system. Diabetes, Obesity and Metabolism, 19(11):1555–1561.

Appendices

Appendix A

THIN Medical Codes

Type 2 diabetes mellitus (T2DM) readcodes that were used in this study were downloaded from the Clinical Codes website (www.clinicalcodes.org). The Readcodes collected from this website included other diabetes mellitus (DM) types that were excluded from the study as advised by the medical professionals in our research team. For the purpose of this study the readcodes used are listed in Table D.7.

Table A.1: T2DM readcodes used.

Readcode	Description	Source	Included
C11y000	Steroid induced diabetes	Clinicalcodes	YES
C314.11	Renal diabetes	Clinicalcodes	YES
9N4p.00	Did not attend diabetic retinopathy clinic	Clinicalcodes	YES
66At111	Type 2 diabetic dietary review	Clinicalcodes	YES
C109K00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus	Clinicalcodes	YES
C109J12	Insulin treated Type II diabetes mellitus	Clinicalcodes	YES
C109J11	Insulin treated non-insulin dependent diabetes mellitus	Clinicalcodes	YES
C109J00	Insulin treated Type 2 diabetes mellitus	Clinicalcodes	YES
C109H12	Type 2 diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C109H11	Type II diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C109H00	Non-insulin dependent d m with neuropathic arthropathy	Clinicalcodes	YES
C109G11	Type II diabetes mellitus with arthropathy	Clinicalcodes	YES
C109G00	Non-insulin dependent diabetes mellitus with arthropathy	Clinicalcodes	YES
C109F12	Type 2 diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C109F11	Type II diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C109F00	Non-insulin-dependent d m with peripheral angiopath	Clinicalcodes	YES
C109E12	Type 2 diabetes mellitus with diabetic cataract	Clinicalcodes	YES
66AH200	Conversion to insulin by diabetes specialist nurse	Clinicalcodes	YES
C109E11	Type II diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C109E00	Non-insulin depend diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C109D12	Type 2 diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C109G12	Type 2 diabetes mellitus with arthropathy	Clinicalcodes	YES
U602317	[X] Adverse reaction to glipzide	Clinicalcodes	YES
66At100	Type II diabetic dietary review	Clinicalcodes	YES
ZC2C800	Dietary advice for diabetes mellitus	Clinicalcodes	YES
ZC2CA00	Dietary advice for type II diabetes	Clinicalcodes	YES
ZRB5.00	Diabetes treatment satisfaction questionnaire	Clinicalcodes	YES
ZRB5.11	DTSQ - Diabetes treatment satisfaction questionnaire	Clinicalcodes	YES
ZRB6.00	Diabetes wellbeing questionnaire	Clinicalcodes	YES
C10FC00	Type 2 diabetes mellitus with nephropathy	Clinicalcodes	YES
C10FH11	Type II diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C10FH00	Type 2 diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C10FG11	Type II diabetes mellitus with arthropathy	Clinicalcodes	YES
C10FG00	Type 2 diabetes mellitus with arthropathy	Clinicalcodes	YES
C10FF11	Type II diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C10FF00	Type 2 diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C10FE11	Type II diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C10FE00	Type 2 diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C10FD11	Type II diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C10FC11	Type II diabetes mellitus with nephropathy	Clinicalcodes	YES
C10FK00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus	Clinicalcodes	YES
C10FB11	Type II diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C10FB00	Type 2 diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C10FA11	Type II diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C10FA00	Type 2 diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C10F911	Type II diabetes mellitus without complication	Clinicalcodes	YES
C10F900	Type 2 diabetes mellitus without complication	Clinicalcodes	YES
C10F711	Type II diabetes mellitus - poor control	Clinicalcodes	YES
C10F700	Type 2 diabetes mellitus - poor control	Clinicalcodes	YES
C10F611	Type II diabetes mellitus with retinopathy	Clinicalcodes	YES
C10FD00	Type 2 diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C10FR00	Type 2 diabetes mellitus with gastroparesis	Clinicalcodes	YES
13AC.00	Diabetic weight reducing diet	Clinicalcodes	YES
250 AM	MATURITY ONSET DIABETES (MELLITUS)	Clinicalcodes	YES
250 AL	MATURITY ONSET DIABETES(MELLITUS) NON-IN	Clinicalcodes	YES
250 AK	MATURITY ONSET DIABETES MELLITUS INSULIN	Clinicalcodes	YES
250 AA	NIDDM (NON-INSULIN DEPENDENT DIABETES)	Clinicalcodes	YES
L180600	Pre-existing diabetes mellitus; non-insulin-dependent	Clinicalcodes	YES
66AV.00	Diabetic on insulin and oral treatment	Clinicalcodes	YES
66A4.00	Diabetic on oral treatment	Clinicalcodes	YES
C10z100	Diabetes mellitus; adult onset; + unspecified complication	Clinicalcodes	YES
66AH000	Conversion to insulin	Clinicalcodes	YES
C10FJ00	Insulin treated Type 2 diabetes mellitus	Clinicalcodes	YES
66Ao.00	Diabetes type 2 review	Clinicalcodes	YES
C10FJ11	Insulin treated Type II diabetes mellitus	Clinicalcodes	YES
C10FQ00	Type 2 diabetes mellitus with exudative maculopathy	Clinicalcodes	YES
C10FP00	Type 2 diabetes mellitus with ketoacidotic coma	Clinicalcodes	YES
C10FN11	Type II diabetes mellitus with ketoacidosis	Clinicalcodes	YES
C10FN00	Type 2 diabetes mellitus with ketoacidosis	Clinicalcodes	YES
C10FM11	Type II diabetes mellitus with persistent microalbuminuria	Clinicalcodes	YES
C10FM00	Type 2 diabetes mellitus with persistent microalbuminuria	Clinicalcodes	YES
C10FL11	Type II diabetes mellitus with persistent proteinuria	Clinicalcodes	YES
C10FL00	Type 2 diabetes mellitus with persistent proteinuria	Clinicalcodes	YES
C10FK11	Hyperosmolar non-ketotic state in type II diabetes mellitus	Clinicalcodes	YES
C10F500	Type 2 diabetes mellitus with gangrene	Clinicalcodes	YES
C10y100	Diabetes mellitus; adult; + other specified manifestation	Clinicalcodes	YES
C10F600	Type 2 diabetes mellitus with retinopathy	Clinicalcodes	YES
679R.00	Patient offered diabetes structured education programme	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
889A.00	Diab mellit insulin-glucose infus acute myocardial infarct	Clinicalcodes	YES
8CA4100	Pt advised re diabetic diet	Clinicalcodes	YES
8CP2.00	Transition of diabetes care options discussed	Clinicalcodes	YES
8CR2.00	Diabetes clinical management plan	Clinicalcodes	YES
8CS0.00	Diabetes care plan agreed	Clinicalcodes	YES
8H4e.00	Referral to diabetes special interest general practitioner	Clinicalcodes	YES
8Hj3.00	Referral to DAFNE diabetes structured education programme	Clinicalcodes	YES
6761	Diabetic pre-pregnancy counselling	Clinicalcodes	YES
8I3k.00	Insulin therapy declined	Clinicalcodes	YES
8I57.00	Patient held diabetic record declined	Clinicalcodes	YES
9360	Patient held diabetic record issued	Clinicalcodes	YES
C10F.00	Type 2 diabetes mellitus	Clinicalcodes	YES
C10F411	Type II diabetes mellitus with ulcer	Clinicalcodes	YES
C10F400	Type 2 diabetes mellitus with ulcer	Clinicalcodes	YES
C10F311	Type II diabetes mellitus with multiple complications	Clinicalcodes	YES
C10F300	Type 2 diabetes mellitus with multiple complications	Clinicalcodes	YES
C10F211	Type II diabetes mellitus with neurological complications	Clinicalcodes	YES
C10F200	Type 2 diabetes mellitus with neurological complications	Clinicalcodes	YES
C10F111	Type II diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C10F100	Type 2 diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C10F011	Type II diabetes mellitus with renal complications	Clinicalcodes	YES
679L.00	Health education - diabetes	Clinicalcodes	YES
C10F.11	Type II diabetes mellitus	Clinicalcodes	YES
C10F511	Type II diabetes mellitus with gangrene	Clinicalcodes	YES
C10E912	Insulin dependent diabetes maturity onset	Clinicalcodes	YES
C10D.11	Maturity onset diabetes in youth type 2	Clinicalcodes	YES
C10D.00	Diabetes mellitus autosomal dominant type 2	Clinicalcodes	YES
1434	H/O: diabetes mellitus	Clinicalcodes	YES
14P3.00	H/O: insulin therapy	Clinicalcodes	YES
3882	Diabetes well being questionnaire	Clinicalcodes	YES
C10F000	Type 2 diabetes mellitus with renal complications	Clinicalcodes	YES
C109.11	NIDDM - Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
E11.7	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.8	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.9	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C104100	Diabetes mellitus; adult onset; with renal manifestation	Clinicalcodes	YES
66AH100	conversion to insulin in secondary care	Clinicalcodes	YES
C105100	Diabetes mellitus; adult onset; + ophthalmic manifestation	Clinicalcodes	YES
C103y00	Other specified diabetes mellitus with coma	Clinicalcodes	YES
C103100	Diabetes mellitus; adult onset; with ketoacidotic coma	Clinicalcodes	YES
C109211	Type II diabetes mellitus with neurological complications	Clinicalcodes	YES
C109.00	Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
E11.4	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C109.12	Type 2 diabetes mellitus	Clinicalcodes	YES
C109.13	Type II diabetes mellitus	Clinicalcodes	YES
C109000	Non-insulin-dependent diabetes mellitus with renal comps	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C109011	Type II diabetes mellitus with renal complications	Clinicalcodes	YES
C109012	Type 2 diabetes mellitus with renal complications	Clinicalcodes	YES
C109100	Non-insulin-dependent diabetes mellitus with ophthalm comps	Clinicalcodes	YES
C10P111	Type 2 diabetes mellitus in remission	Clinicalcodes	YES
C109112	Type 2 diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C102100	Diabetes mellitus; adult onset; with hyperosmolar coma	Clinicalcodes	YES
C107100	Diabetes mellitus; adult; + peripheral circulatory disorder	Clinicalcodes	YES
8I2P.00	sulphonylureas contraindicated	Clinicalcodes	YES
C10P100	Type II diabetes mellitus in remission	Clinicalcodes	YES
C10FP11	Type II diabetes mellitus with ketoacidotic coma	Clinicalcodes	YES
ZV6DB00	[v]admitted for conversion to insulin	Clinicalcodes	YES
C106100	Diabetes mellitus; adult onset; + neurological manifestation	Clinicalcodes	YES
E11.6	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.5	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.0	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.1	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.2	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.3	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C109200	Non-insulin-dependent diabetes mellitus with neuro comps	Clinicalcodes	YES
C109912	Type 2 diabetes mellitus without complication	Clinicalcodes	YES
C109212	Type 2 diabetes mellitus with neurological complications	Clinicalcodes	YES
C109111	Type II diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C109712	Type 2 diabetes mellitus - poor control	Clinicalcodes	YES
C100112	Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
C100111	Maturity onset diabetes	Clinicalcodes	YES
C100100	Diabetes mellitus; adult onset; no mention of complication	Clinicalcodes	YES
C109D11	Type II diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C109612	Type 2 diabetes mellitus with retinopathy	Clinicalcodes	YES
C109911	Type II diabetes mellitus without complication	Clinicalcodes	YES
C109700	Non-insulin dependent diabetes mellitus - poor control	Clinicalcodes	YES
C109A00	Non-insulin dependent diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C109A11	Type II diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C109B00	Non-insulin dependent diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109B11	Type II diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109B12	Type 2 diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109C00	Non-insulin dependent diabetes mellitus with nephropathy	Clinicalcodes	YES
C109C11	Type II diabetes mellitus with nephropathy	Clinicalcodes	YES
C109C12	Type 2 diabetes mellitus with nephropathy	Clinicalcodes	YES
C109D00	Non-insulin dependent diabetes mellitus with hypoglyca coma	Clinicalcodes	YES
C109900	Non-insulin-dependent diabetes mellitus without complication	Clinicalcodes	YES
C109600	Non-insulin-dependent diabetes mellitus with retinopathy	Clinicalcodes	YES
C109300	Non-insulin-dependent diabetes mellitus with multiple comps	Clinicalcodes	YES
C109312	Type 2 diabetes mellitus with multiple complications	Clinicalcodes	YES
C109400	Non-insulin dependent diabetes mellitus with ulcer	Clinicalcodes	YES
C109411	Type II diabetes mellitus with ulcer	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C109412	Type 2 diabetes mellitus with ulcer	Clinicalcodes	YES
C109500	Non-insulin dependent diabetes mellitus with gangrene	Clinicalcodes	YES
C109511	Type II diabetes mellitus with gangrene	Clinicalcodes	YES
C109711	Type II diabetes mellitus - poor control	Clinicalcodes	YES
C107400	NIDDM with peripheral circulatory disorder	Clinicalcodes	YES
66o2.00	Diabetic on non-insulin injectable medication	Clinicalcodes	YES
C101100	Diabetes mellitus; adult onset; with ketoacidosis	Clinicalcodes	YES
66o5.00	Diabetic on oral treatment and glucagon-like pepti	Clinicalcodes	YES
C109611	Type II diabetes mellitus with retinopathy	Clinicalcodes	YES
C101000	Diabetes mellitus; juvenile type; with ketoacidosis	Clinicalcodes	YES
C109512	Type 2 diabetes mellitus with gangrene	Clinicalcodes	YES
66A3.00	Diabetic on diet only	Clinicalcodes	YES

Appendix B

Townsend Deprivation Index Calculation

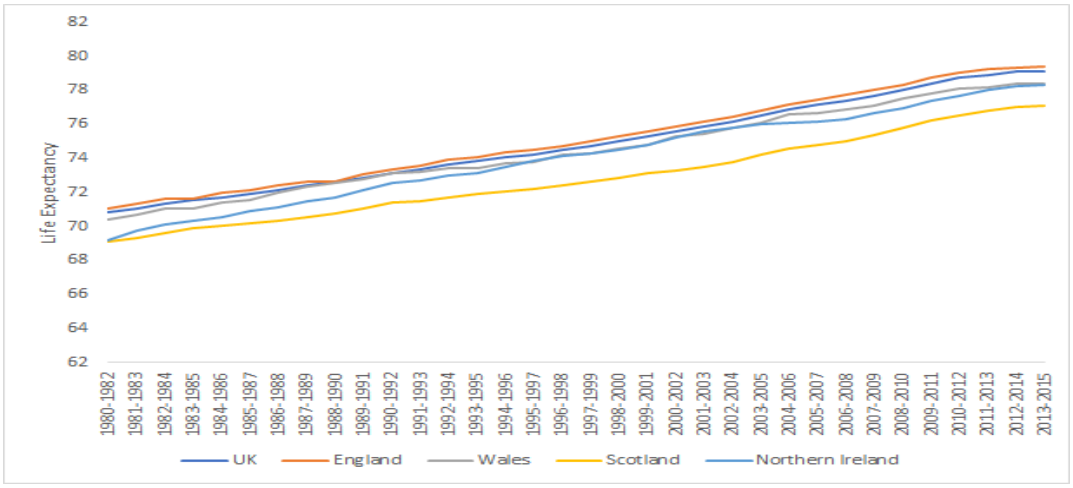
The Townsend deprivation index (TDI) is one of the several indices used to measure deprivation in the UK. Other indices are the Index of Multiple Deprivation and the Mosaic Index. The TDI is computed as a weighted index from four indicators, which are Unemployment, Non-car ownership, Non-house ownership and Overcrowding. The TDI used by the THIN database are based on the 2001 Census data. The indicators are computed as follows:

$$
\begin{aligned}
unemployment &= \frac{UnemployedActiveLabourForce}{TotalActiveLabourForce} \cdot 100 \\
non - carownership &= \frac{Numberofhouseholdswithnocar}{TotalNumberofHouseholds} \cdot 100 \\
non - houseownership &= \frac{Numberofhouseholdstenantoccupied}{TotalNumberofHouseholds} \cdot 100 \\
overcrowding &= \frac{Numberofhouseholdsovercrowded}{TotalNumberofHouseholds} \cdot 100
\end{aligned}
$$

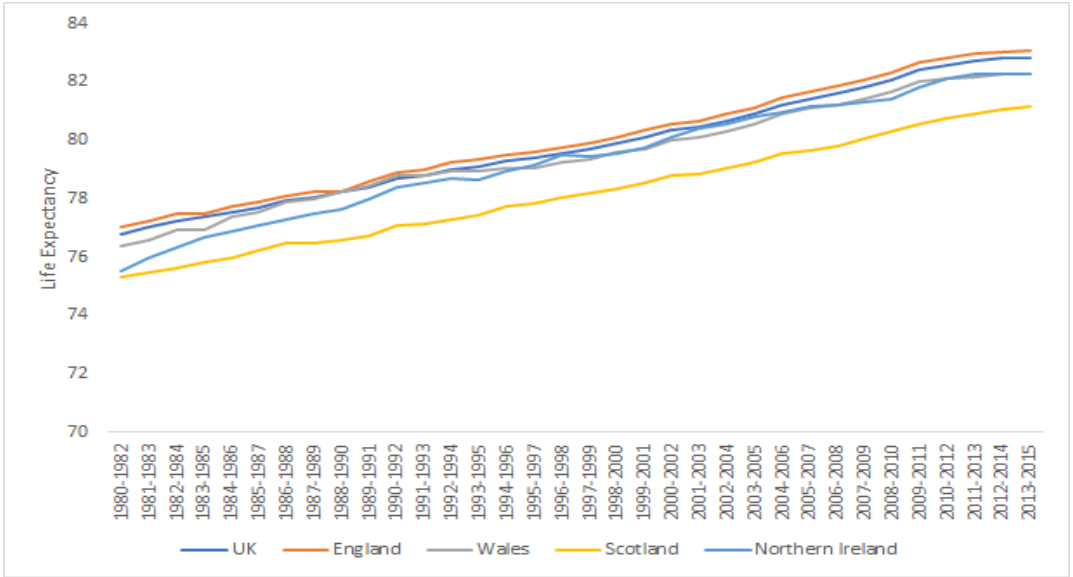
Their standard Z-scores are calculated, weighted and summed up to give the TDI score. THIN uses equally weighted Z-scores which are based on Office of National Statistics (ONS)'s 2001 census method. Unemployment and overcrowding indicators are first log transformed before their Z-scores are calculated to reduce skewness. The scores are then grouped into TDI quintiles from 1 (least deprived) to 5 (most deprived). Due to the fact that the standard Z-scores are centred on the mean zero, an area with TDI score greater than zero fall in the deprived category whereas a score less than zero is in the affluent category.

Appendix C

Additional Figures

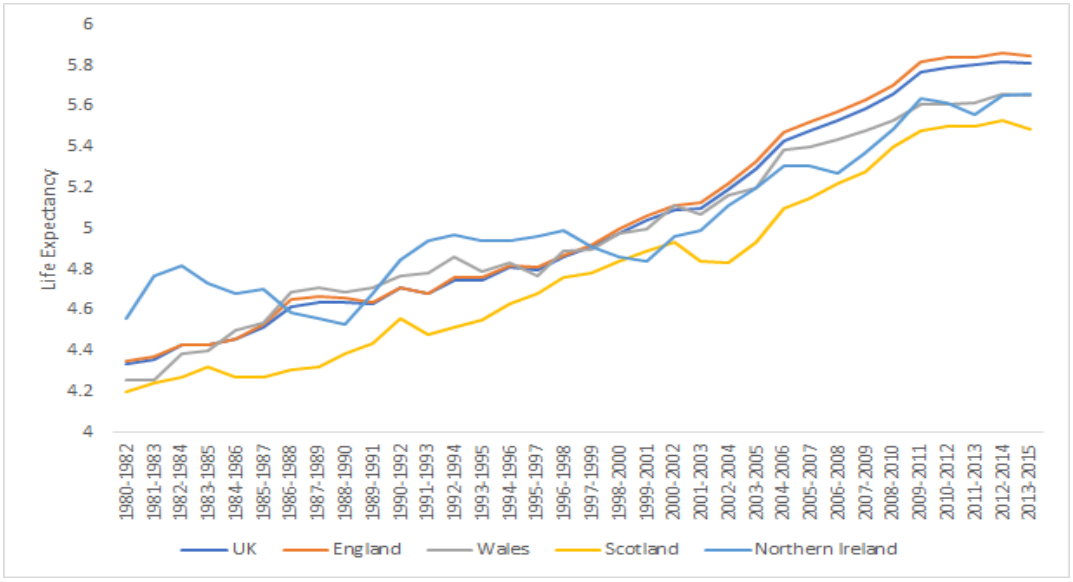


(a) Males

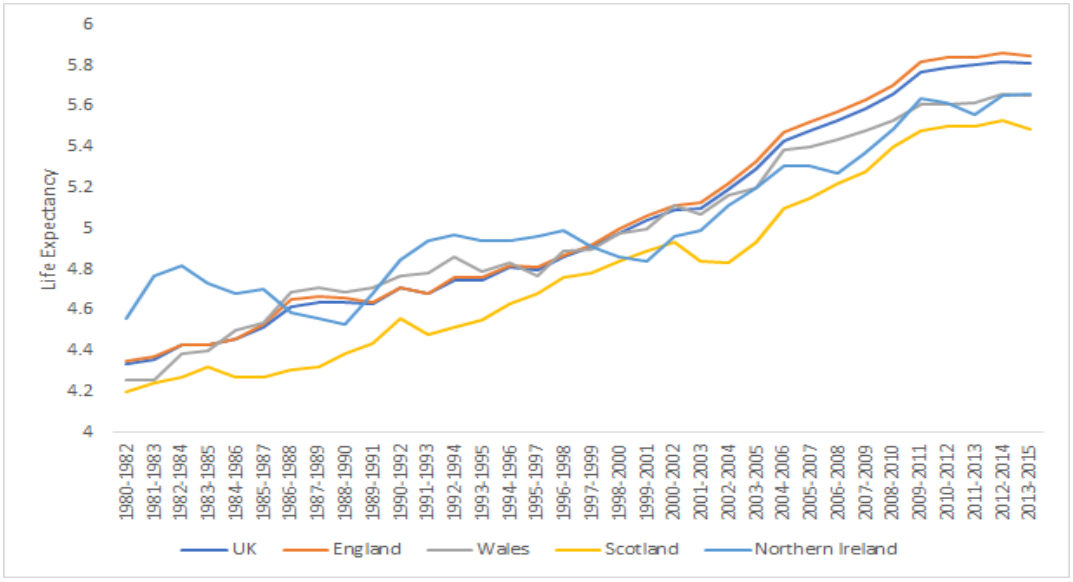


(b) Females

Figure C.1: Life Expectancies in the UK: 1980-1982 to 2013-2015: $\mathring{e}_0$



(a) Males



(b) Femalee

Figure C.2: Life Expectancies in the UK: 1980-1982 to 2013-2015: $\mathring{e}_{85}$

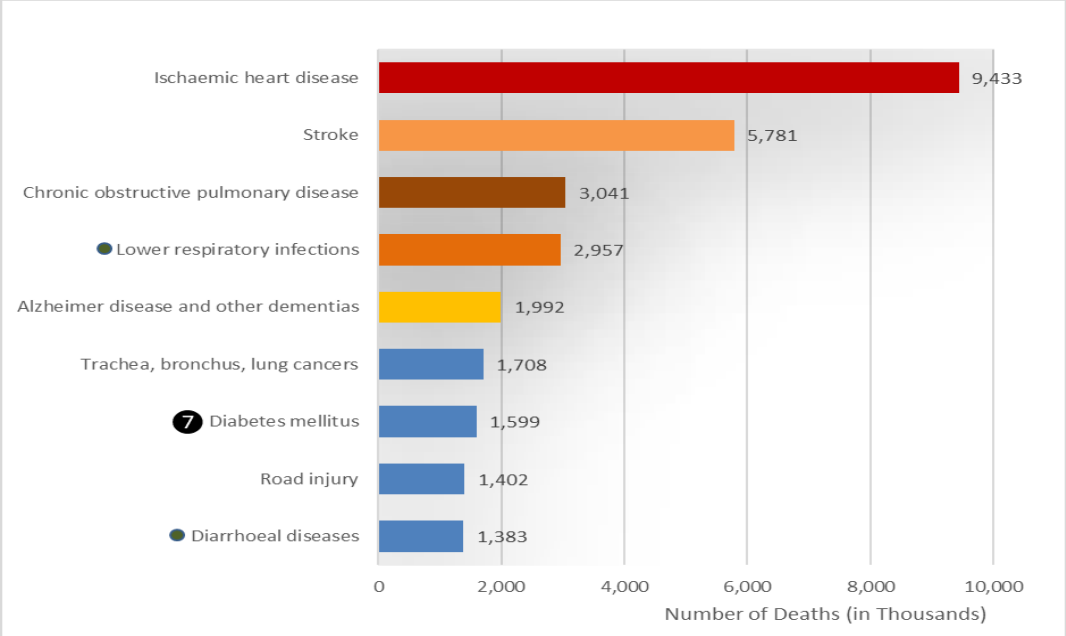


Figure C.3: Top Ten Causes of Death in the World 2019.

Source: (WHO, 2017a)

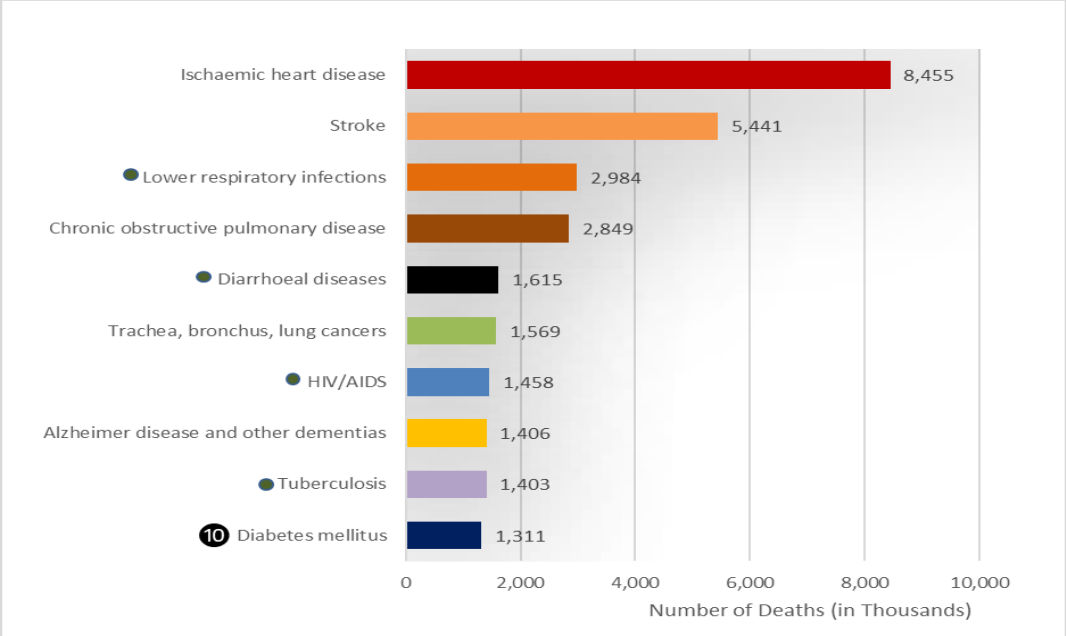


Figure C.4: Top Ten Causes of Death in the World 2010.

Source: (WHO, 2017a)

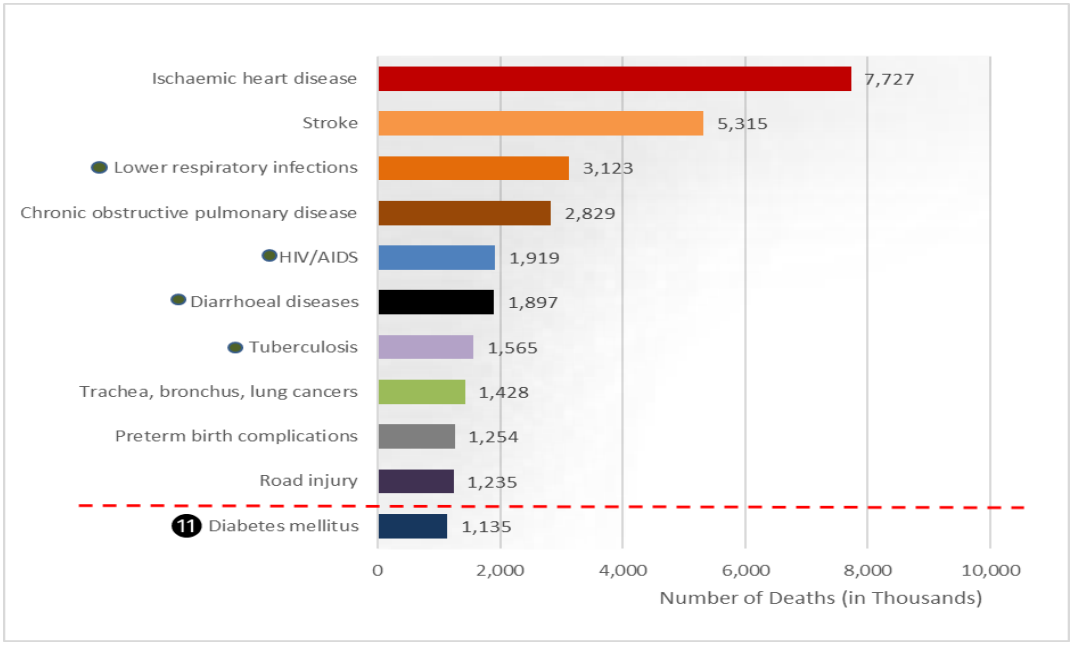


Figure C.5: Top Ten Causes of Death in the World 2005.

Source: (WHO, 2017a)

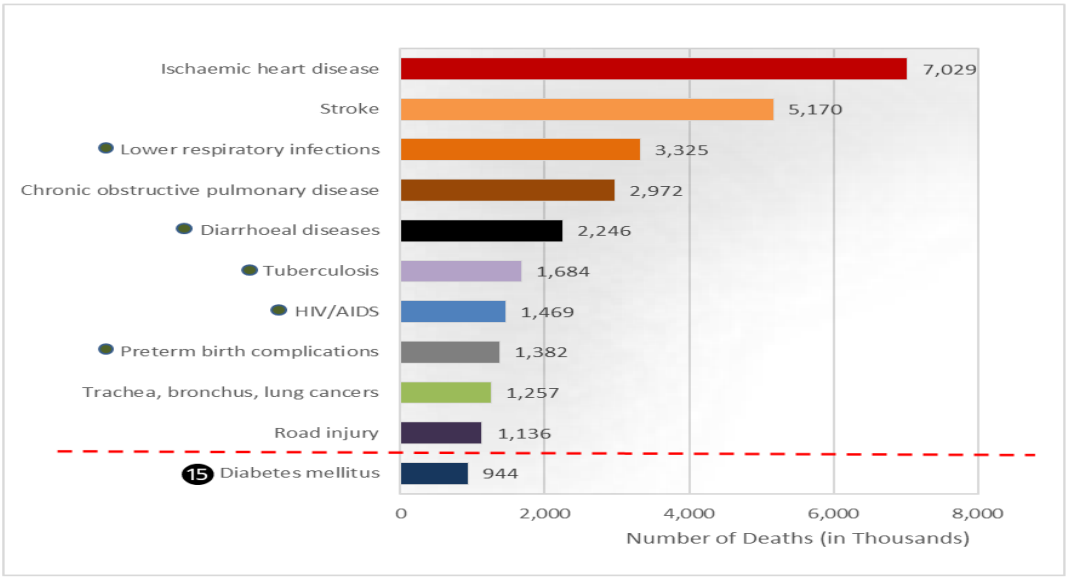


Figure C.6: Top Ten Causes of Death in the World 2000.

Source: (WHO, 2017a)

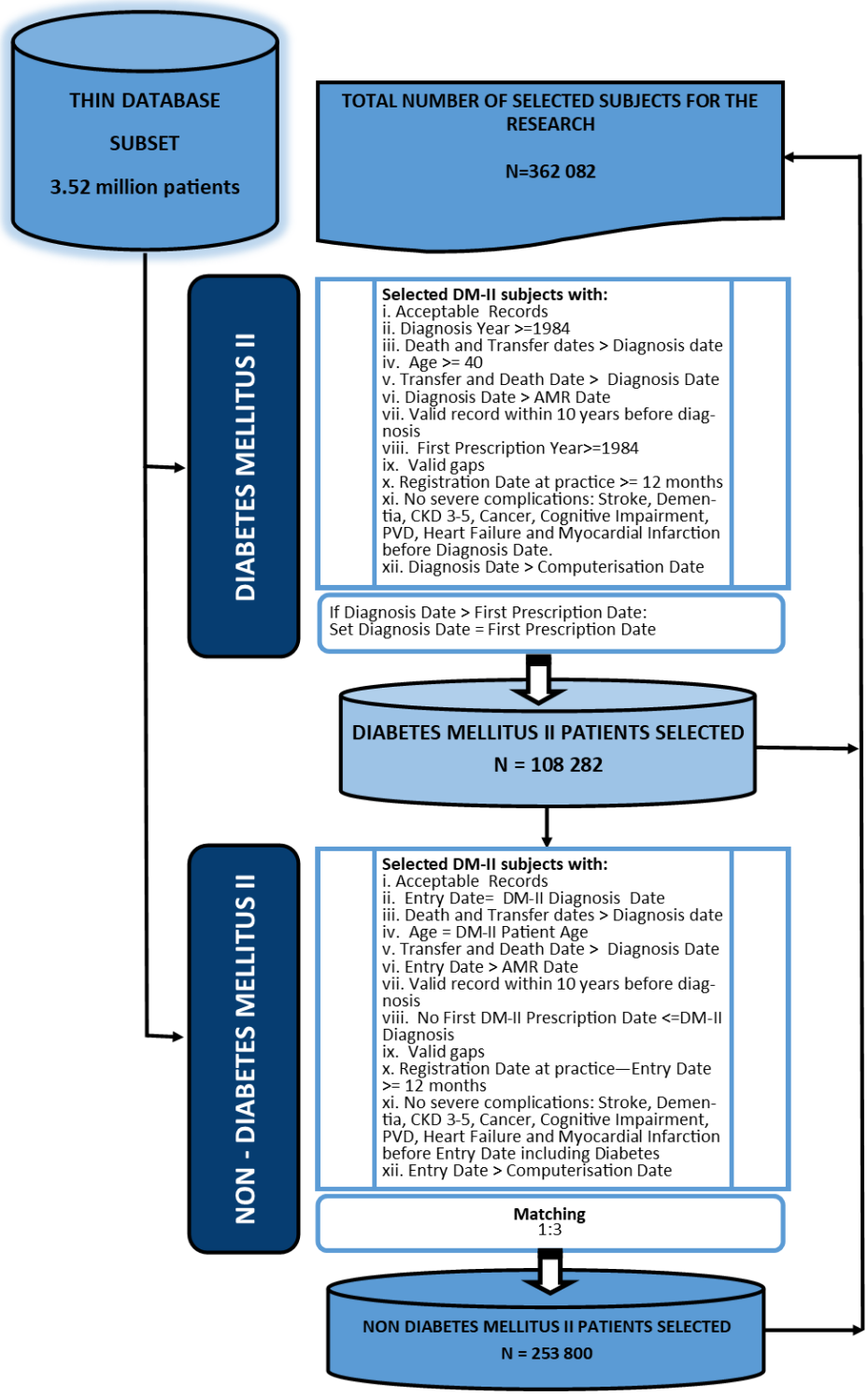


Figure C.7: Data Extraction Process

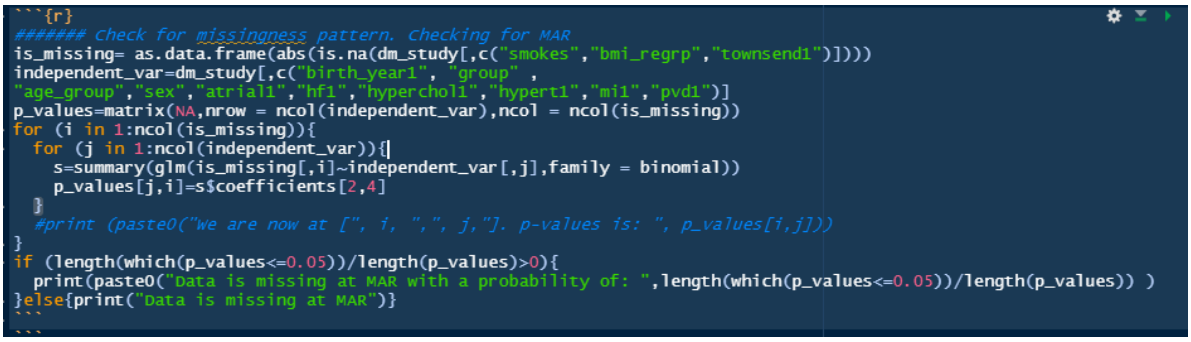
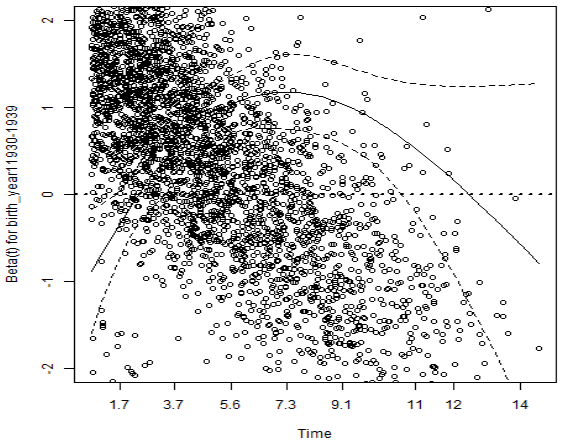
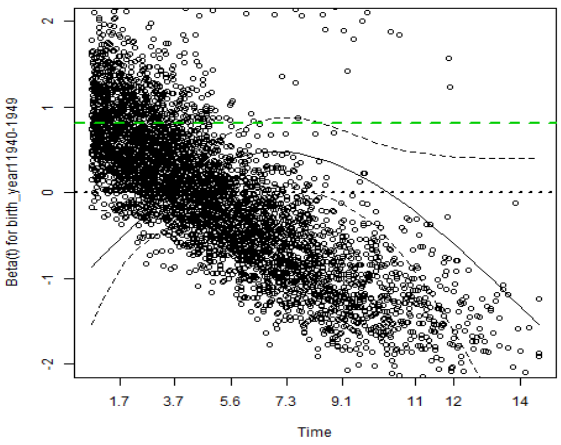


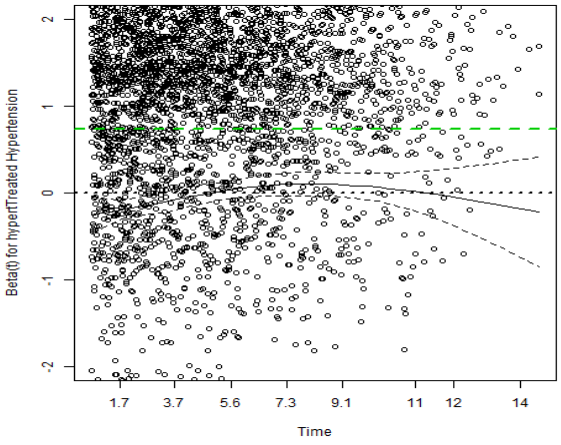
Figure C.8: Code for checking MAR pattern



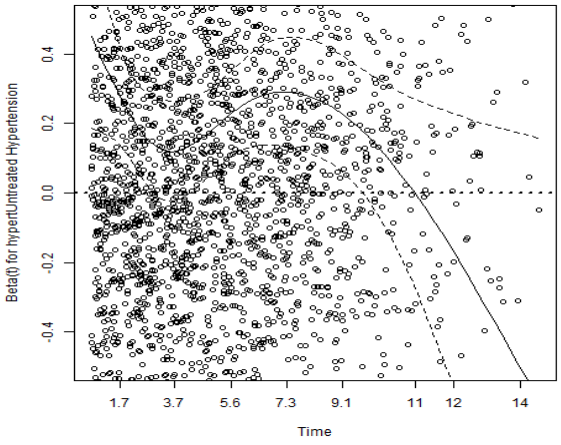
(a) Birth Cohort 1930-1939



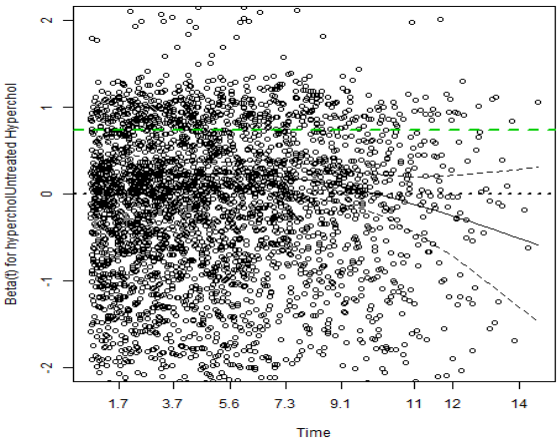
(b) Birth Cohort 1940-1949



(c) Treated Hypertension



(d) Untreated Hypertension



(e) Untreated Hypercholesteromia

Figure C.9: Plots demonstrating violation of the PH assumptions for HTN, HCL and birth cohort.

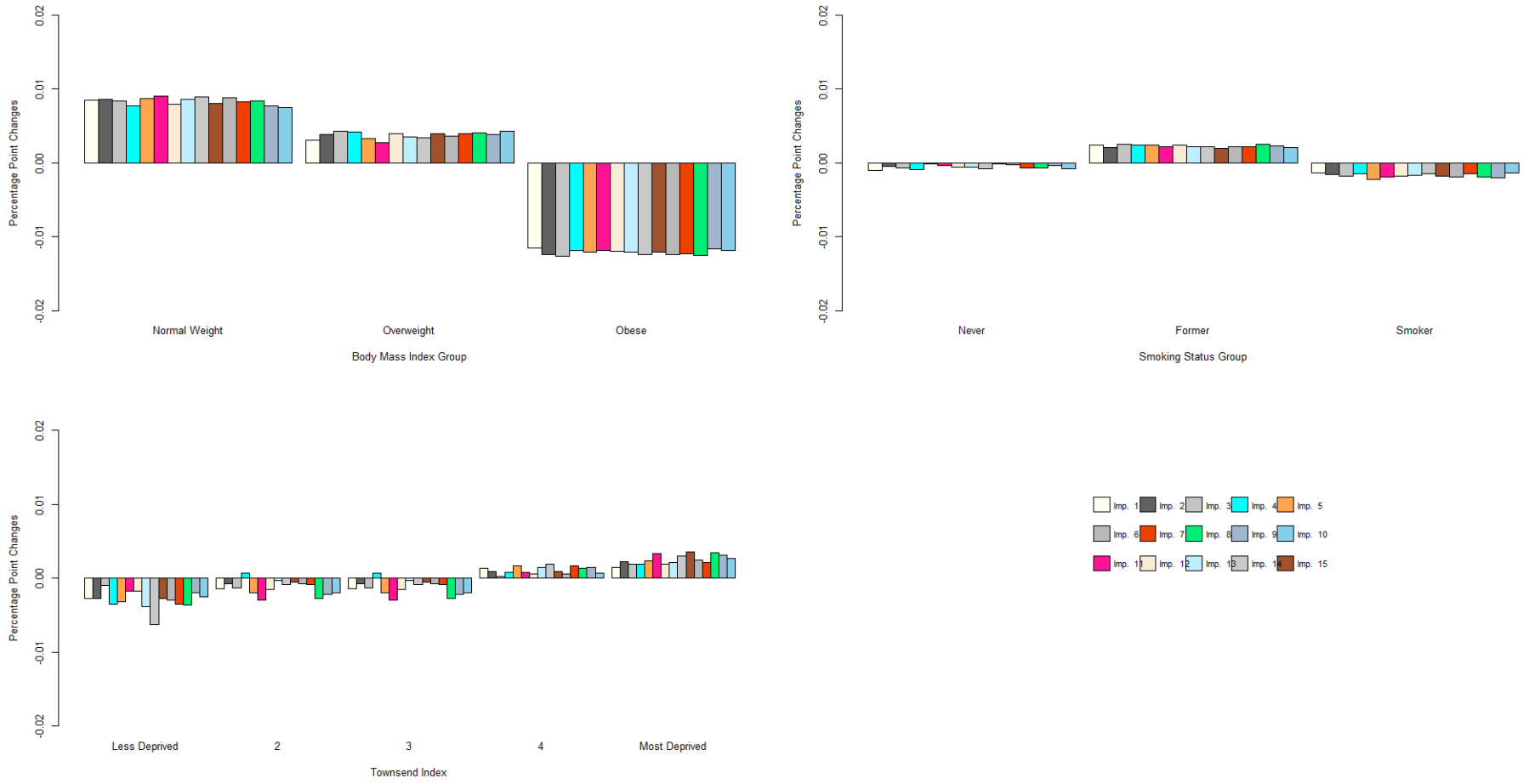


Figure C.10: Comparisons of the distribution of body mass index (BMI), TDI and smoking status in complete cases and imputed data

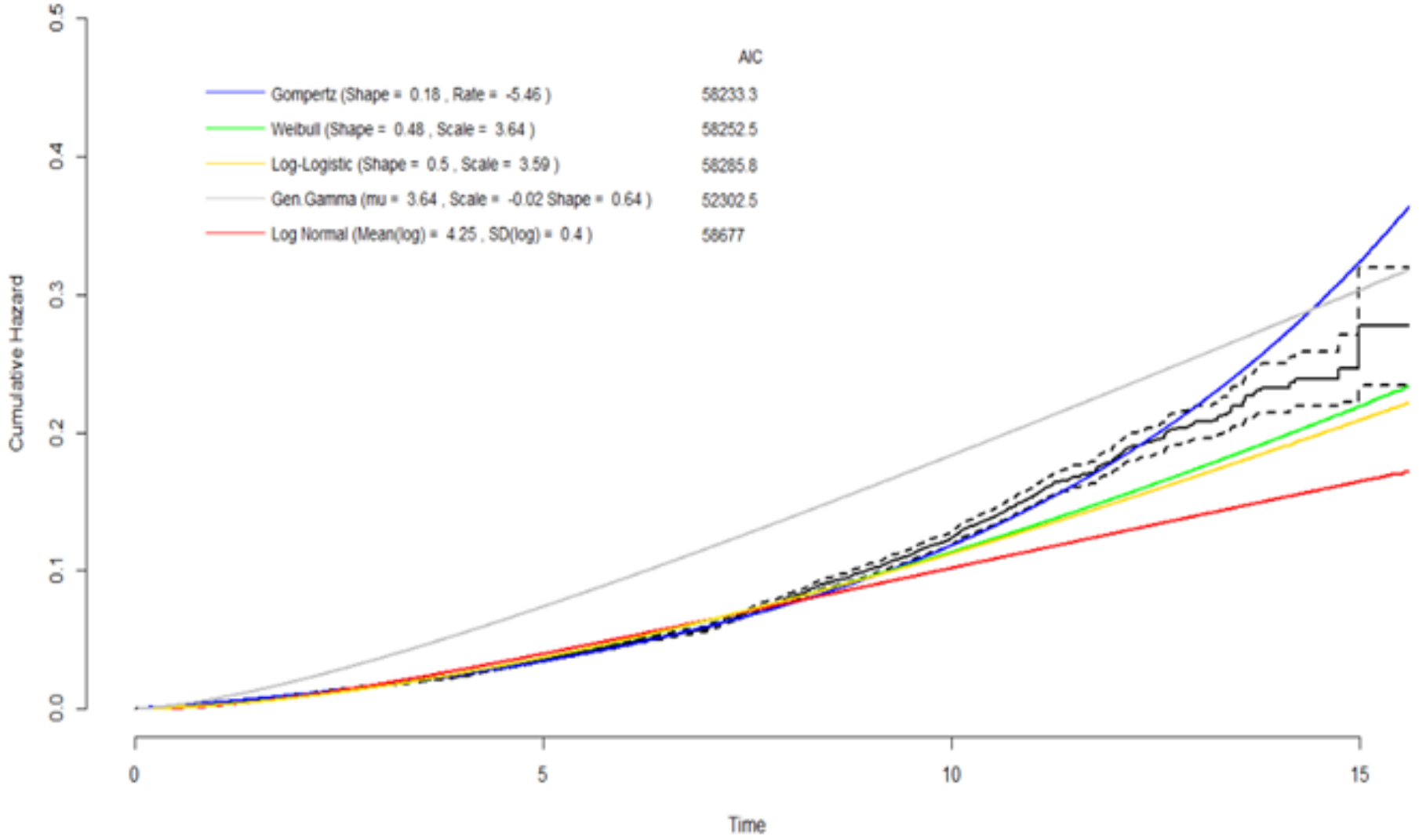


Figure C.11: Selection of the best fitting hazard function. Estimated observed hazards and their 95% CIs are in black.

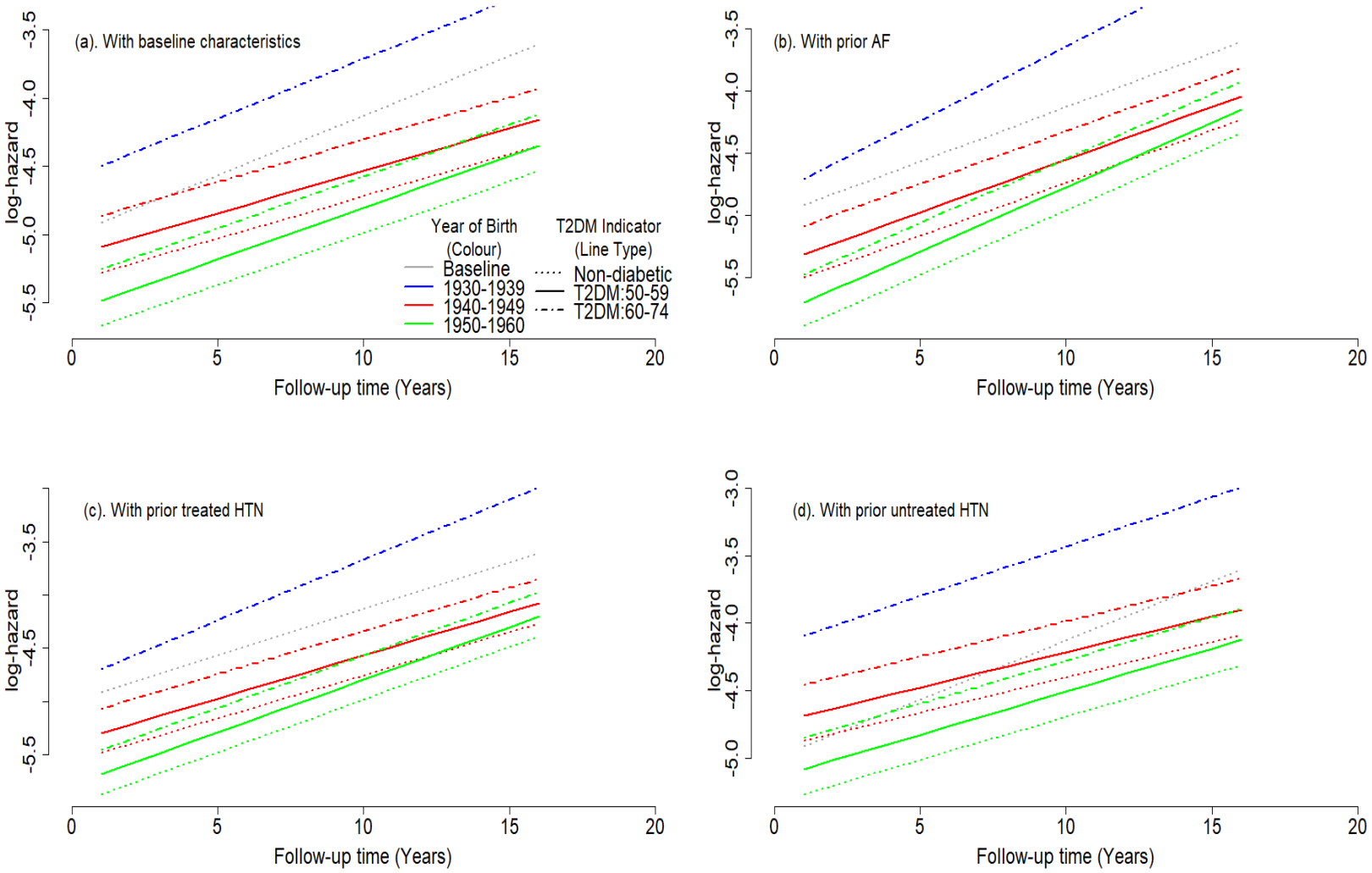


Figure C.12: Comparison of estimated all-cause mortality risk among people with similar medical conditions by birth cohort and T2DM status using Model A

Model adjusted for age, year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and their interactions.

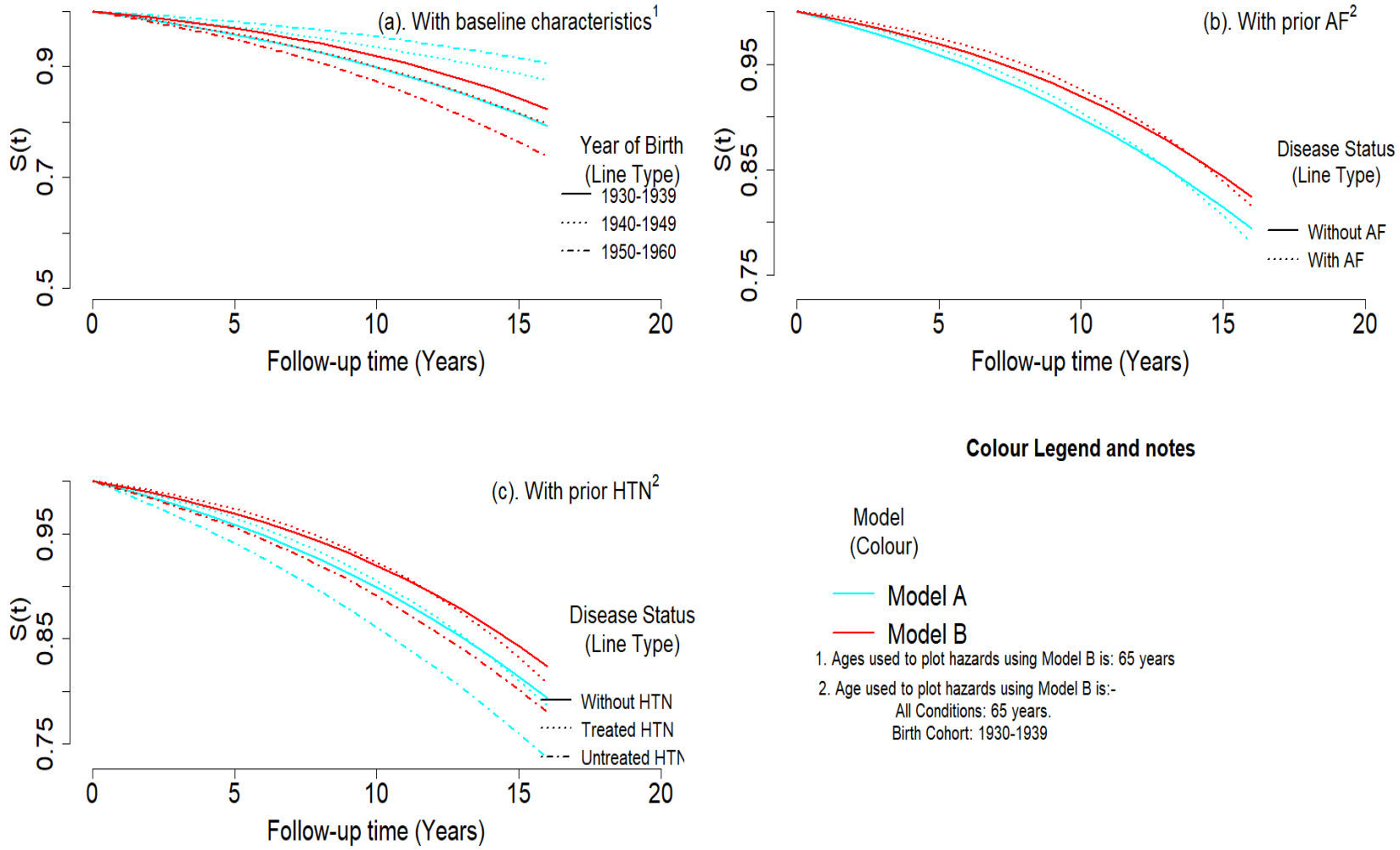


Figure C.13: Comparison of estimated survival probabilities among people with similar medical conditions by birth cohort and T2DM status using Model A and Model B

Models adjusted for age at diagnosis (Model A) or continuous age at entry (Model B), year of birth, gender, BMI, smoking status,TDI, AF, HF, MI, PVD, HCL, HTN and their interactions.

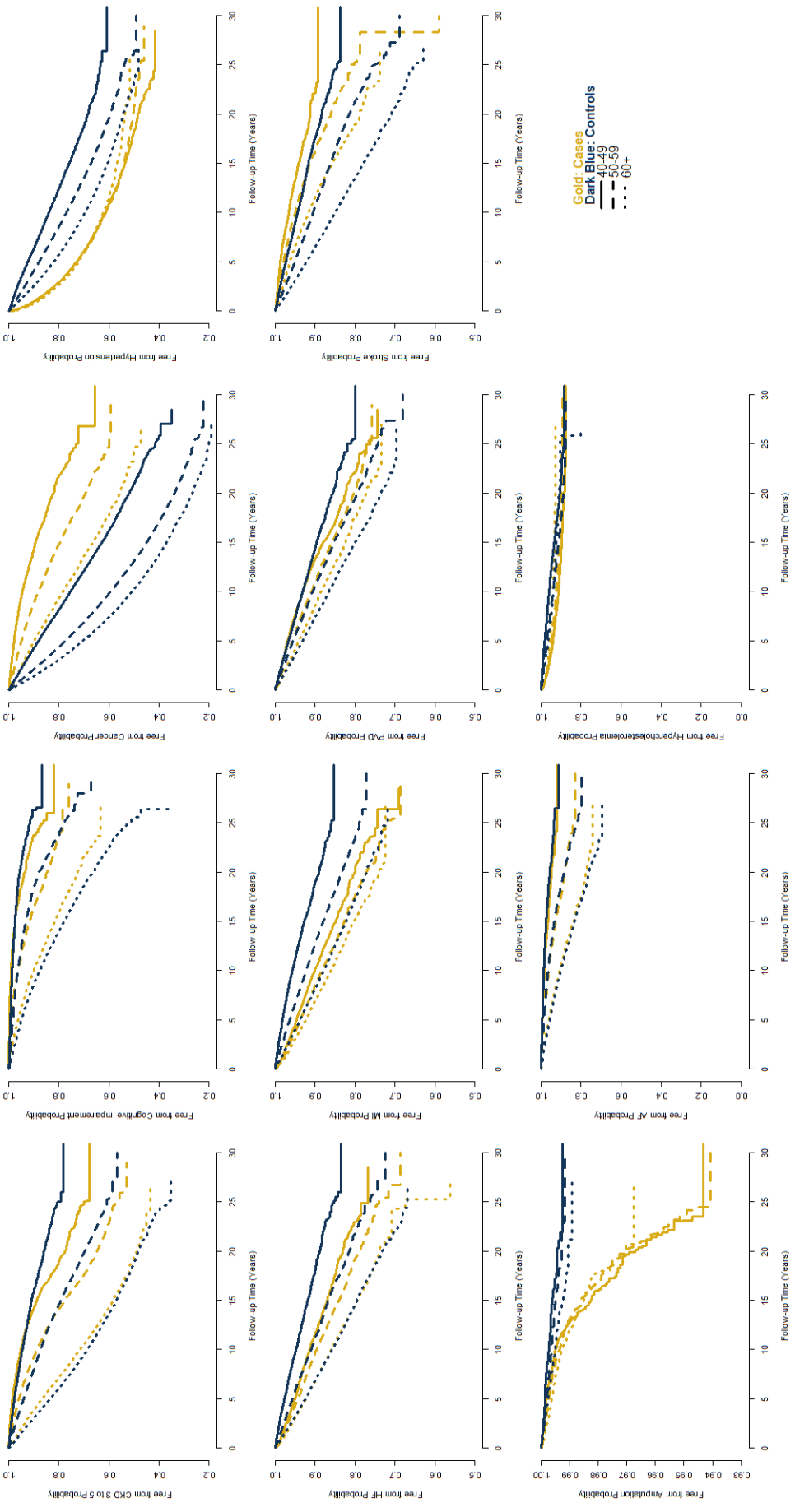


Figure C.14: KM plots of time to selected morbidities for people with or without T2DM

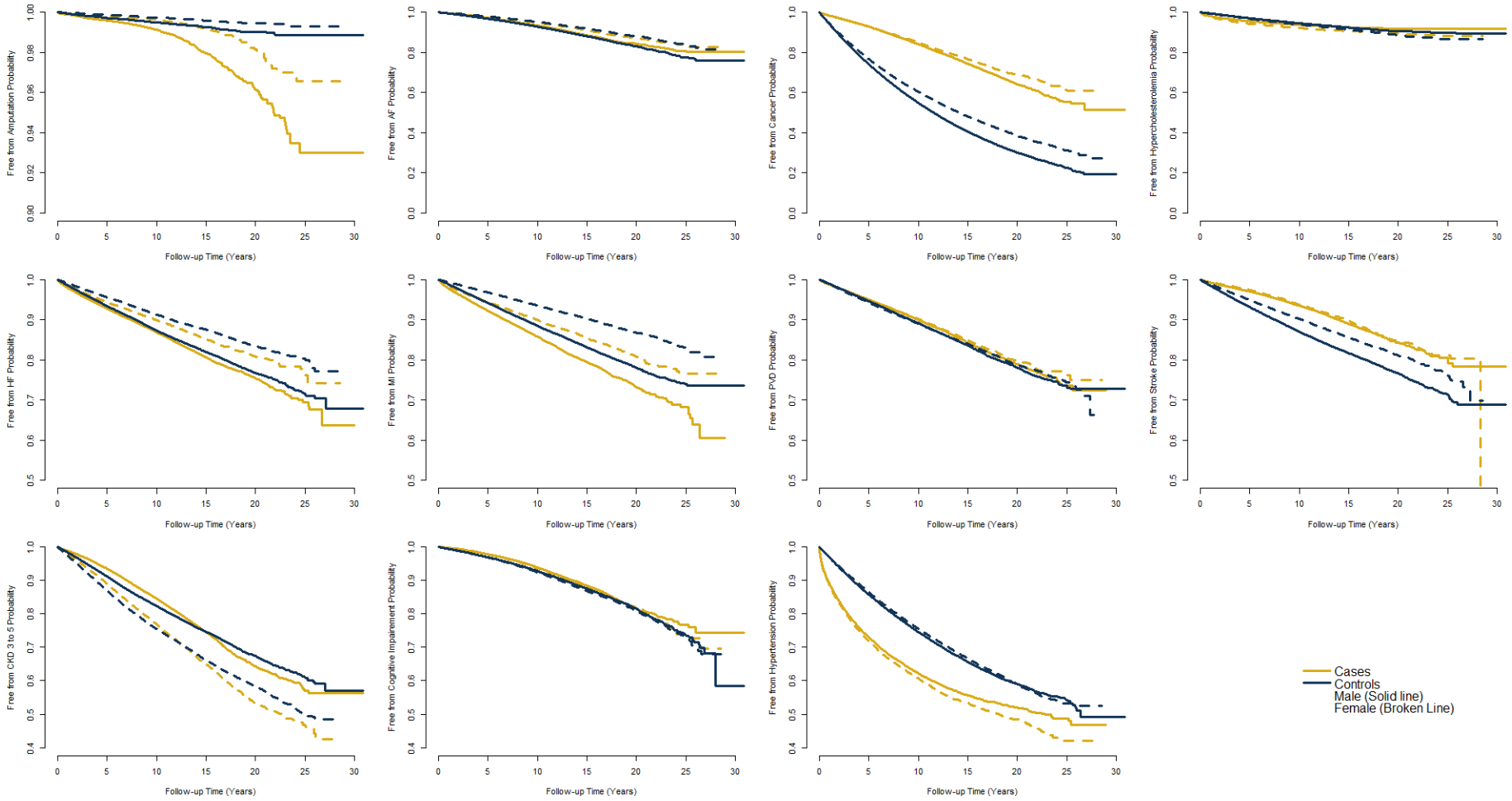


Figure C.15: KM plot of time to selected morbidities for people with or without T2DM by gender

Appendix D

Additional Tables

Table D.1: Causes, Prevention and Symptoms of T1DM and T2DM

Charactristics	***Causes***	***Prevention***	***Symptoms***
		T1DM	
Previously known as insulin-dependant, juvenile or childhood-onset. Characterised by deficiency in insulin production and requires daily administration of insulin.	Not Known	NOT preventable with current knowledge	Polyuria - excessive excretion of urine Polydipsia - excessive thirst Constant hunger Weight loss Vision changes, Fatigue
		T2DM	
Previously known as non-insulin-dependant or adult-onset. Comprises the majority of people with diabetes around the world. Arises due to the body's ineffective use of insulin. Until recently, it was a disease of the adults, but is now occurring increasingly frequently in children.	Largely a result of excessive body weight and physical inactivity.	Proper healthy diet Not smoking Avoiding alcohol Exercising	Similar to those of DM-I but are often less marked.

Table D.2: Main Types of non-insulin antidiabetic drugs used in T2DM Therapy

Group	**Description**	**Examples**
Sulphonylureas	increases the amount of natural insulin produced by the pancreas	Gliclazide, Glibenclamide, Glipizide, Glimepiride, Tolbutamide
Biguanides	reduce the release of glucose from the liver and increasing the uptake of glucose into the muscles	Metformin (Glucophage®)
Thiazolidinediones	target IR and are used by people who have been unable to control their blood glucose levels with metformin or sulphonylurea. Also avalaible in combinations with metformin as Avandamet® and Actophus Met®	*Rosiglitazone (Avandin®), Pioglitazone (Actos®)
Alpha Glucosidase Inhibitor	slows digestion of carbohydrate in the intestines and suppresses the rise of blood glucose after meals	Acarbose®, Glucobay®
Prandial (mealtime) Glucose Regulators, Meglitinide Analogues	stimulate the release of insulin from the pancreas and are taken with meals. Can be used with or without metformin	Prandin/ Repaglinide, Nateglinide®
Sodium-glucose contransporter 2 (SGLT-2)	reduces glucose reabsorption and increases urinary glucose excretion by reversibly inhibiting SGLT-2 in the renal proximal convoluted tubule.	Dapagliflozin, Canagliflozin, Empagliflozin
Dipetidyl peptidease - 4 (DPP-4) inhibitors	increases insulin secretion and reduces glucagon secretion by inhibiting DPP-4	Alogliptin, Linagliptin, Saxagliptin, Sitagliptin, Vildagliptin
Glucagon-Like peptide - 1 (GLP-1) receptor	attaches to, and activates, the GLP-1 receptor to increase insulin secretion, suppresses secretion of glucagon, and slows down gastric emptying. Its an injectable drug.	Albiglutide, Dulaglutide, Exenatide, Liraglutide, Lixisenatide, Semaglutide

* Restricted by the USA Food and Drug Administration due to increased myocardial infarction (MI) risk (Lim et al., 2018).

Table D.3: The Four Types of Insulin

Type	**On-set**	**Peak Time**	**Duration**	Examples
Rapid-acting (Lispro)	Reaches blood within 15 mins	30 to 90 mins later	5hrs	Humalog, Novorapid
Short-acting (Regular)	Reaches blood within 30 mins	2 - 4 hrs later	4 - 8hrs	Actrapid, Humulin, Hypurin, Velosulin
Intermediate-acting (NPH and Lente)	Reaches blood within 2 to 6hrs	4 - 14hrs later	14 - 20hrs	Humulin, Hypurin, Insuman
Long-acting (Ultralente)	Reaches blood within 6 to 14hrs	Has no peak or very small peak 10 - 16hrs	20-24hrs	Lantus, Levemir, Hypurin, Tresiba

Table D.4: Risk assessment of diabetes-related foot problems

Diabetes Foot Problems Risk			
Low	**Moderate**	**High**	**Active**
No Risk Factors except callus alone	Deformity OR	Previous Ulceration OR	Ulceration OR
	Neuropathy OR	Previous Amputation OR	Spreading Infection OR
	Non-Critical Limb Ischaemia	On Renal Replacement Therapy OR	Critical Limb Ischaemia OR
		Neuropathy and Non-Critical Limb Ischaemia together OR	Gangrene OR
		Neuropathy in Combination with Callus and / or Deformity OR	Suspicion of Acute Charcot Arthropathy or unexplained hot, red, swollen foot with or without pain
		Non-Critical Limb Ischaemia in Combination with Callus and / or Deformity.	

Table D.5: Strengths and Weaknesses of The Health Improvement Network (THIN) Database

Strengths	**Weaknesses**
Big database representing 6% of the UK population	Only reflect patient's data that is deemed relevant to the patient's care
Allows study of all patients with a selected medical condition	Its use requires an experienced data manager
Control subjects can be selected from the same source population	When exposure and outcomes are both rare there is a likelihood of power problems
Data is collected in a non-intervention way and therefore reflects 'real life'	Not appropriate for studies with emphasis on ethnicity, occupation, employment and / or socio-economic status as it is not available at a patient level
Data is continually updated	Non-compliance to medication prescriptions may be an issue for drug-related exposures
Reduction of costs and time spent in data collection	Not appropriate for studies where data is primarily related to secondary care
Amenable to most epidemiological study designs such as cohorts, case-control and case-series studies	over-the-counter (OTC) drugs are not usually recorded
Can be used to study relatively rare exposures or outcomes	Only abnormal lab test results (values) may have been entered
Several practices are linked to the Hospital Episode Statistics (HES) data	

Source: UCL (2013)

Table D.6: Description of linked-tables in THIN

Look-up table	Description
Staff	Provides information on the gender and roles of staff members at a GP
Postcode Variable Indicators	Provides anonymous patient postcode-linked, socio-economic status, ethnicity and environmental indices.
Pack size	Provides the quantity in a pack of drugs.
Drug Codes/ BNF	Provides the codes for the drugs using a BNF format.
Dosage	Gives the dosage of the drug prescription including the frequency and strength.
Medical Codes	Has the codes used to identify medical conditions.
AHD	Contains information on lifestyle, preventive healthcare, immunisation, test results and death details.

D.1 THIN Medical Codes

T2DM readcodes used in this study were collected from the Clinical Codes. Readcodes collected from this website included other diabetes mellitus (DM) types which were excluded from the study as advised by the medical professionals in the study team.

Table D.7: T2DM readcodes used.

Readcode	Description	Source	Included
C11y000	Steroid induced diabetes	Clinicalcodes	YES
C314.11	Renal diabetes	Clinicalcodes	YES
9N4p.00	Did not attend diabetic retinopathy clinic	Clinicalcodes	YES
66At111	Type 2 diabetic dietary review	Clinicalcodes	YES
C109K00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus	Clinicalcodes	YES
C109J12	Insulin treated Type II diabetes mellitus	Clinicalcodes	YES
C109J11	Insulin treated non-insulin dependent diabetes mellitus	Clinicalcodes	YES
C109J00	Insulin treated Type 2 diabetes mellitus	Clinicalcodes	YES
C109H12	Type 2 diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C109H11	Type II diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C109H00	Non-insulin dependent d m with neuropathic arthropathy	Clinicalcodes	YES
C109G11	Type II diabetes mellitus with arthropathy	Clinicalcodes	YES
C109G00	Non-insulin dependent diabetes mellitus with arthropathy	Clinicalcodes	YES
C109F12	Type 2 diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C109F11	Type II diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C109F00	Non-insulin-dependent d m with peripheral angiopath	Clinicalcodes	YES
C109E12	Type 2 diabetes mellitus with diabetic cataract	Clinicalcodes	YES
66AH200	Conversion to insulin by diabetes specialist nurse	Clinicalcodes	YES
C109E11	Type II diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C109E00	Non-insulin depend diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C109D12	Type 2 diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C109G12	Type 2 diabetes mellitus with arthropathy	Clinicalcodes	YES
U602317	[X] Adverse reaction to glipzide	Clinicalcodes	YES
66At100	Type II diabetic dietary review	Clinicalcodes	YES
ZC2C800	Dietary advice for diabetes mellitus	Clinicalcodes	YES
ZC2CA00	Dietary advice for type II diabetes	Clinicalcodes	YES
ZRB5.00	Diabetes treatment satisfaction questionnaire	Clinicalcodes	YES
ZRB5.11	DTSQ - Diabetes treatment satisfaction questionnaire	Clinicalcodes	YES
ZRB6.00	Diabetes wellbeing questionnaire	Clinicalcodes	YES
C10FC00	Type 2 diabetes mellitus with nephropathy	Clinicalcodes	YES
C10FH11	Type II diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES
C10FH00	Type 2 diabetes mellitus with neuropathic arthropathy	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C10FG11	Type II diabetes mellitus with arthropathy	Clinicalcodes	YES
C10FG00	Type 2 diabetes mellitus with arthropathy	Clinicalcodes	YES
C10FF11	Type II diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C10FF00	Type 2 diabetes mellitus with peripheral angiopathy	Clinicalcodes	YES
C10FE11	Type II diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C10FE00	Type 2 diabetes mellitus with diabetic cataract	Clinicalcodes	YES
C10FD11	Type II diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C10FC11	Type II diabetes mellitus with nephropathy	Clinicalcodes	YES
C10FK00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus	Clinicalcodes	YES
C10FB11	Type II diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C10FB00	Type 2 diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C10FA11	Type II diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C10FA00	Type 2 diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C10F911	Type II diabetes mellitus without complication	Clinicalcodes	YES
C10F900	Type 2 diabetes mellitus without complication	Clinicalcodes	YES
C10F711	Type II diabetes mellitus - poor control	Clinicalcodes	YES
C10F700	Type 2 diabetes mellitus - poor control	Clinicalcodes	YES
C10F611	Type II diabetes mellitus with retinopathy	Clinicalcodes	YES
C10FD00	Type 2 diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C10FR00	Type 2 diabetes mellitus with gastroparesis	Clinicalcodes	YES
13AC.00	Diabetic weight reducing diet	Clinicalcodes	YES
250 AM	MATURITY ONSET DIABETES (MELLITUS)	Clinicalcodes	YES
250 AL	MATURITY ONSET DIABETES(MELLITUS) NON-IN	Clinicalcodes	YES
250 AK	MATURITY ONSET DIABETES MELLITUS INSULIN	Clinicalcodes	YES
250 AA	NIDDM (NON-INSULIN DEPENDENT DIABETES)	Clinicalcodes	YES
L180600	Pre-existing diabetes mellitus; non-insulin-dependent	Clinicalcodes	YES
66AV.00	Diabetic on insulin and oral treatment	Clinicalcodes	YES
66A4.00	Diabetic on oral treatment	Clinicalcodes	YES
C10z100	Diabetes mellitus; adult onset; + unspecified complication	Clinicalcodes	YES
66AH000	Conversion to insulin	Clinicalcodes	YES
C10FJ00	Insulin treated Type 2 diabetes mellitus	Clinicalcodes	YES
66Ao.00	Diabetes type 2 review	Clinicalcodes	YES
C10FJ11	Insulin treated Type II diabetes mellitus	Clinicalcodes	YES
C10FQ00	Type 2 diabetes mellitus with exudative maculopathy	Clinicalcodes	YES
C10FP00	Type 2 diabetes mellitus with ketoacidotic coma	Clinicalcodes	YES
C10FN11	Type II diabetes mellitus with ketoacidosis	Clinicalcodes	YES
C10FN00	Type 2 diabetes mellitus with ketoacidosis	Clinicalcodes	YES
C10FM11	Type II diabetes mellitus with persistent microalbuminuria	Clinicalcodes	YES
C10FM00	Type 2 diabetes mellitus with persistent microalbuminuria	Clinicalcodes	YES
C10FL11	Type II diabetes mellitus with persistent proteinuria	Clinicalcodes	YES
C10FL00	Type 2 diabetes mellitus with persistent proteinuria	Clinicalcodes	YES
C10FK11	Hyperosmolar non-ketotic state in type II diabetes mellitus	Clinicalcodes	YES
C10F500	Type 2 diabetes mellitus with gangrene	Clinicalcodes	YES
C10y100	Diabetes mellitus; adult; + other specified manifestation	Clinicalcodes	YES
C10F600	Type 2 diabetes mellitus with retinopathy	Clinicalcodes	YES
679R.00	Patient offered diabetes structured education programme	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
889A.00	Diab mellit insulin-glucose infus acute myocardial infarct	Clinicalcodes	YES
8CA4100	Pt advised re diabetic diet	Clinicalcodes	YES
8CP2.00	Transition of diabetes care options discussed	Clinicalcodes	YES
8CR2.00	Diabetes clinical management plan	Clinicalcodes	YES
8CS0.00	Diabetes care plan agreed	Clinicalcodes	YES
8H4e.00	Referral to diabetes special interest general practitioner	Clinicalcodes	YES
8Hj3.00	Referral to DAFNE diabetes structured education programme	Clinicalcodes	YES
6761	Diabetic pre-pregnancy counselling	Clinicalcodes	YES
8I3k.00	Insulin therapy declined	Clinicalcodes	YES
8I57.00	Patient held diabetic record declined	Clinicalcodes	YES
9360	Patient held diabetic record issued	Clinicalcodes	YES
C10F.00	Type 2 diabetes mellitus	Clinicalcodes	YES
C10F411	Type II diabetes mellitus with ulcer	Clinicalcodes	YES
C10F400	Type 2 diabetes mellitus with ulcer	Clinicalcodes	YES
C10F311	Type II diabetes mellitus with multiple complications	Clinicalcodes	YES
C10F300	Type 2 diabetes mellitus with multiple complications	Clinicalcodes	YES
C10F211	Type II diabetes mellitus with neurological complications	Clinicalcodes	YES
C10F200	Type 2 diabetes mellitus with neurological complications	Clinicalcodes	YES
C10F111	Type II diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C10F100	Type 2 diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C10F011	Type II diabetes mellitus with renal complications	Clinicalcodes	YES
679L.00	Health education - diabetes	Clinicalcodes	YES
C10F.11	Type II diabetes mellitus	Clinicalcodes	YES
C10F511	Type II diabetes mellitus with gangrene	Clinicalcodes	YES
C10E912	Insulin dependent diabetes maturity onset	Clinicalcodes	YES
C10D.11	Maturity onset diabetes in youth type 2	Clinicalcodes	YES
C10D.00	Diabetes mellitus autosomal dominant type 2	Clinicalcodes	YES
1434	H/O: diabetes mellitus	Clinicalcodes	YES
14P3.00	H/O: insulin therapy	Clinicalcodes	YES
3882	Diabetes well being questionnaire	Clinicalcodes	YES
C10F000	Type 2 diabetes mellitus with renal complications	Clinicalcodes	YES
C109.11	NIDDM - Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
E11.7	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.8	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.9	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C104100	Diabetes mellitus; adult onset; with renal manifestation	Clinicalcodes	YES
66AH100	conversion to insulin in secondary care	Clinicalcodes	YES
C105100	Diabetes mellitus; adult onset; + ophthalmic manifestation	Clinicalcodes	YES
C103y00	Other specified diabetes mellitus with coma	Clinicalcodes	YES
C103100	Diabetes mellitus; adult onset; with ketoacidotic coma	Clinicalcodes	YES
C109211	Type II diabetes mellitus with neurological complications	Clinicalcodes	YES
C109.00	Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
E11.4	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C109.12	Type 2 diabetes mellitus	Clinicalcodes	YES
C109.13	Type II diabetes mellitus	Clinicalcodes	YES
C109000	Non-insulin-dependent diabetes mellitus with renal comps	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C109011	Type II diabetes mellitus with renal complications	Clinicalcodes	YES
C109012	Type 2 diabetes mellitus with renal complications	Clinicalcodes	YES
C109100	Non-insulin-dependent diabetes mellitus with ophthalm comps	Clinicalcodes	YES
C10P111	Type 2 diabetes mellitus in remission	Clinicalcodes	YES
C109112	Type 2 diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C102100	Diabetes mellitus; adult onset; with hyperosmolar coma	Clinicalcodes	YES
C107100	Diabetes mellitus; adult; + peripheral circulatory disorder	Clinicalcodes	YES
8I2P.00	sulphonylureas contraindicated	Clinicalcodes	YES
C10P100	Type II diabetes mellitus in remission	Clinicalcodes	YES
C10FP11	Type II diabetes mellitus with ketoacidotic coma	Clinicalcodes	YES
ZV6DB00	[v]admitted for conversion to insulin	Clinicalcodes	YES
C106100	Diabetes mellitus; adult onset; + neurological manifestation	Clinicalcodes	YES
E11.6	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.5	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.0	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.1	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.2	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
E11.3	non-insulin-dependent diabetes mellitus	Clinicalcodes	YES
C109200	Non-insulin-dependent diabetes mellitus with neuro comps	Clinicalcodes	YES
C109912	Type 2 diabetes mellitus without complication	Clinicalcodes	YES
C109212	Type 2 diabetes mellitus with neurological complications	Clinicalcodes	YES
C109111	Type II diabetes mellitus with ophthalmic complications	Clinicalcodes	YES
C109712	Type 2 diabetes mellitus - poor control	Clinicalcodes	YES
C100112	Non-insulin dependent diabetes mellitus	Clinicalcodes	YES
C100111	Maturity onset diabetes	Clinicalcodes	YES
C100100	Diabetes mellitus; adult onset; no mention of complication	Clinicalcodes	YES
C109D11	Type II diabetes mellitus with hypoglycaemic coma	Clinicalcodes	YES
C109612	Type 2 diabetes mellitus with retinopathy	Clinicalcodes	YES
C109911	Type II diabetes mellitus without complication	Clinicalcodes	YES
C109700	Non-insulin dependent diabetes mellitus - poor control	Clinicalcodes	YES
C109A00	Non-insulin dependent diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C109A11	Type II diabetes mellitus with mononeuropathy	Clinicalcodes	YES
C109B00	Non-insulin dependent diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109B11	Type II diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109B12	Type 2 diabetes mellitus with polyneuropathy	Clinicalcodes	YES
C109C00	Non-insulin dependent diabetes mellitus with nephropathy	Clinicalcodes	YES
C109C11	Type II diabetes mellitus with nephropathy	Clinicalcodes	YES
C109C12	Type 2 diabetes mellitus with nephropathy	Clinicalcodes	YES
C109D00	Non-insulin dependent diabetes mellitus with hypoglyca coma	Clinicalcodes	YES
C109900	Non-insulin-dependent diabetes mellitus without complication	Clinicalcodes	YES
C109600	Non-insulin-dependent diabetes mellitus with retinopathy	Clinicalcodes	YES
C109300	Non-insulin-dependent diabetes mellitus with multiple comps	Clinicalcodes	YES
C109312	Type 2 diabetes mellitus with multiple complications	Clinicalcodes	YES
C109400	Non-insulin dependent diabetes mellitus with ulcer	Clinicalcodes	YES
C109411	Type II diabetes mellitus with ulcer	Clinicalcodes	YES

T2DM readcodes used.

Readcode	Description	Source	Included
C109412	Type 2 diabetes mellitus with ulcer	Clinicalcodes	YES
C109500	Non-insulin dependent diabetes mellitus with gangrene	Clinicalcodes	YES
C109511	Type II diabetes mellitus with gangrene	Clinicalcodes	YES
C109711	Type II diabetes mellitus - poor control	Clinicalcodes	YES
C107400	NIDDM with peripheral circulatory disorder	Clinicalcodes	YES
66o2.00	Diabetic on non-insulin injectable medication	Clinicalcodes	YES
C101100	Diabetes mellitus; adult onset; with ketoacidosis	Clinicalcodes	YES
66o5.00	Diabetic on oral treatment and glucagon-like pepti	Clinicalcodes	YES
C109611	Type II diabetes mellitus with retinopathy	Clinicalcodes	YES
C101000	Diabetes mellitus; juvenile type; with ketoacidosis	Clinicalcodes	YES
C109512	Type 2 diabetes mellitus with gangrene	Clinicalcodes	YES
66A3.00	Diabetic on diet only	Clinicalcodes	YES

Table D.8: Prevalence of smoking by case-control status, gender, age group and smoking status at study entry

Age Group	Case (Number(%))				Control (Number(%))			
	Missing	Former	Never	Current	Missing	Former	Never	Current
				Males				
40-44	404 (25.3)	615 (38.51)	209 (13.09)	369 (23.11)	1,256 (31.31)	1,500 (37.4)	330 (8.23)	925 (23.06)
45-49	767 (15.04)	2,055 (40.29)	1,017 (19.94)	1,262 (24.74)	2,717 (22.7)	4,910 (41.03)	1,445 (12.07)	2,895 (24.19)
50-54	958 (9.99)	3,793 (39.54)	2,429 (25.32)	2,414 (25.16)	3,539 (16.88)	8,773 (41.83)	3,329 (15.87)	5,330 (25.42)
55-59	1,135 (9.71)	4,402 (37.65)	3,621 (30.97)	2,535 (21.68)	3,987 (15.65)	10,225 (40.14)	4,991 (19.59)	6,272 (24.62)
60-64	997 (8.41)	4,307 (36.32)	4,187 (35.31)	2,367 (19.96)	3,690 (13.96)	10,289 (38.93)	6,394 (24.19)	6,055 (22.91)
65-69	873 (8.72)	3,610 (36.05)	3,904 (38.98)	1,628 (16.26)	3,019 (13.32)	8,937 (39.44)	6,304 (27.82)	4,401 (19.42)
70-74	622 (8.88)	2,690 (38.39)	2,846 (40.62)	849 (12.12)	2,111 (12.98)	6,742 (41.45)	4,888 (30.05)	2,523 (15.51)
75-79	358 (9.73)	1,389 (37.75)	1,542 (41.91)	390 (10.6)	1,184 (14.07)	3,531 (41.97)	2,642 (31.4)	1,057 (12.56)
80+	179 (10.73)	708 (42.45)	629 (37.71)	152 (9.11)	484 (14.07)	1,491 (43.33)	1,161 (33.74)	305 (8.86)
Total	6,293 (10.12)	23,569 (37.88)	20,384 (32.77)	11,966 (19.23)	21,987 (15.75)	56,398 (40.39)	31,484 (22.55)	29,763 (21.32)
				Females				
40-44	226 (21.28)	490 (46.14)	102 (9.6)	244 (22.98)	646 (21.47)	1,428 (47.46)	239 (7.94)	696 (23.13)
45-49	364 (12.59)	1,408 (48.7)	405 (14.01)	714 (24.7)	1,182 (15.21)	3,954 (50.87)	856 (11.01)	1,780 (22.9)
50-54	501 (8.94)	2,881 (51.4)	985 (17.57)	1,238 (22.09)	1,730 (12.2)	7,471 (52.71)	1,885 (13.3)	3,089 (21.79)
55-59	610 (8.31)	3,706 (50.49)	1,532 (20.87)	1,492 (20.33)	2,062 (11.47)	9,342 (51.95)	2,800 (15.57)	3,777 (21.01)
60-64	724 (9.16)	3,904 (49.39)	1,781 (22.53)	1,495 (18.91)	2,259 (11.67)	10,072 (52.02)	3,340 (17.25)	3,689 (19.05)
65-69	745 (9.75)	4,033 (52.77)	1,717 (22.47)	1,147 (15.01)	2,277 (12.26)	9,970 (53.69)	3,353 (18.06)	2,968 (15.98)
70-74	690 (10.67)	3,539 (54.75)	1,436 (22.22)	799 (12.36)	2,338 (14.47)	8,801 (54.46)	2,923 (18.09)	2,099 (12.99)
75-79	481 (11.13)	2,495 (57.73)	955 (22.1)	391 (9.05)	1,641 (15.08)	6,142 (56.43)	1,941 (17.83)	1,160 (10.66)
80+	456 (14.73)	1,869 (60.37)	587 (18.96)	184 (5.94)	1,265 (17.3)	4,376 (59.83)	1,164 (15.91)	509 (6.96)
Total	4,797 (10.35)	24,325 (52.51)	9,500 (20.51)	7,704 (16.63)	15,400 (13.37)	61,556 (53.42)	18,501 (16.06)	19,767 (17.16)

Table D.9: Number of participants aged 45 years and above and alive from 1998 by age group, gender, alcohol use at study entry

Age Group	Male (Number(%))				Female (Number(%))			
	Missing	Former	Never	Current	Missing	Former	Never	Current
45-49	168 (25.05)	106 (15.75)	4 (1.33)	338 (57.87)	466 (34.76)	100 (8.52)	20 (1.18)	769 (55.54)
50-54	397 (23.52)	233 (13.3)	20 (2.03)	717 (61.16)	1,008 (32.79)	248 (9.13)	55 (1.44)	1,983 (56.64)
55-59	650 (20.57)	417 (13.4)	36 (2.18)	1,350 (63.85)	1,471 (29.67)	392 (9.14)	113 (1.8)	3,395 (59.39)
60-64	1,073 (22.72)	769 (14.99)	51 (1.92)	2,020 (60.37)	1,892 (28.8)	590 (11.42)	131 (1.56)	4,416 (58.22)
65-69	1,367 (21.43)	1,092 (16.35)	62 (1.48)	2,427 (60.74)	2,094 (28.51)	745 (13.24)	126 (1.43)	5,118 (56.82)
70-74	1,772 (22.48)	1,311 (17.28)	53 (1.47)	2,740 (58.77)	1,860 (29)	784 (15.05)	122 (1.25)	4,725 (54.71)
75-79	1,473 (21.56)	1,253 (19.94)	66 (1.9)	2,320 (56.59)	1,287 (28.84)	619 (17.49)	93 (1.4)	3,187 (52.26)
80+	1,498 (26.84)	1,187 (22.18)	44 (1.78)	1,957 (49.2)	684 (31.07)	311 (19.73)	44 (0.95)	1,596 (48.25)
Total	8,398 (22.6)	6,368 (16.94)	336 (1.76)	13,869 (58.7)	10,762 (29.56)	3,789 (13.73)	704 (1.4)	25,189 (55.31)
Age Group	Cases (Number(%))				Controls (Number(%))			
	Missing	Former	Never	Current	Missing	Former	Never	Current
40-44	803 (30.2)	416 (15.64)	28 (1.05)	1,412 (53.1)	2,411 (34.34)	579 (8.25)	40 (0.57)	3,990 (56.84)
45-49	1,820 (22.77)	1,209 (15.13)	134 (1.68)	4,829 (60.42)	5,597 (28.36)	1,748 (8.86)	162 (0.82)	12,232 (61.97)
50-54	2,828 (18.61)	2,367 (15.57)	365 (2.4)	9,639 (63.42)	8,535 (24.28)	3,361 (9.56)	371 (1.06)	22,879 (65.1)
55-59	3,313 (17.41)	2,945 (15.47)	405 (2.13)	12,370 (64.99)	9,730 (22.39)	4,548 (10.47)	517 (1.19)	28,661 (65.95)
60-64	3,406 (17.24)	3,199 (16.19)	411 (2.08)	12,746 (64.5)	9,870 (21.56)	5,378 (11.75)	571 (1.25)	29,969 (65.45)
65-69	3,001 (17)	3,164 (17.92)	340 (1.93)	11,152 (63.16)	8,744 (21.21)	5,723 (13.88)	489 (1.19)	26,273 (63.72)
70-74	2,419 (17.96)	2,638 (19.58)	226 (1.68)	8,188 (60.78)	7,123 (21.97)	5,147 (15.87)	401 (1.24)	19,754 (60.92)
75-79	1,509 (18.86)	1,810 (22.62)	156 (1.95)	4,526 (56.57)	4,493 (23.28)	3,540 (18.34)	243 (1.26)	11,022 (57.11)
80+	1,091 (22.9)	1,101 (23.11)	93 (1.95)	2,479 (52.04)	2,801 (26.04)	2,214 (20.59)	131 (1.22)	5,609 (52.15)
Total	20,190 (18.6)	18,849 (17.37)	2,158 (1.99)	67,341 (62.04)	59,304 (23.27)	32,238 (12.65)	2,925 (1.15)	160,389 (62.93)

Table D.10: Number of participants by BMI, gender and case-control status at study entry

BMI Group	Cases (Numbers (%))		Controls (Numbers (%))	
	Male	Female	Male	Female
Underweight	92 (0.17)	160 (0.4)	830 (0.81)	1,737 (1.92)
Normal Weight	5,609 (10.65)	4,583 (11.59)	36,856 (35.75)	38,295 (42.24)
Overweight	19,906 (37.79)	11,345 (28.69)	48,224 (46.78)	31,559 (34.81)
Obese I	16,937 (32.15)	11,315 (28.62)	13,905 (13.49)	13,078 (14.42)
Obese II	6,894 (13.09)	6,913 (17.48)	2,551 (2.47)	4,228 (4.66)
Obese III	3,236 (6.14)	5,222 (13.21)	725 (0.7)	1,772 (1.95)
Total	62,212 (100)	46,326 (100)	139,632 (100)	115,224 (100)

Table D.11: Number of participants aged 45 years and above from 1993 by BMI and at Entry, Age Group and Sex

Age Group	Males (Number(%))				Females (Number(%))			
	Underweight	Normal Weight	Overweight	Obese	Underweight	Normal Weight	Overweight	Obese
45-49	11 (3.17)	256 (2.84)	313 (2.49)	275 (3.71)	14 (2.41)	181 (2.52)	94 (1.35)	161 (2.66)
50-54	26 (7.49)	690 (7.64)	839 (6.68)	721 (9.72)	23 (3.96)	347 (4.83)	279 (4.01)	358 (5.91)
55-59	60 (17.29)	1,037 (11.49)	1,485 (11.83)	1,171 (15.79)	58 (9.98)	607 (8.45)	570 (8.19)	616 (10.17)
60-64	67 (19.31)	1,507 (16.69)	2,122 (16.91)	1,393 (18.78)	67 (11.53)	960 (13.36)	922 (13.25)	980 (16.18)
65-69	67 (19.31)	1,744 (19.32)	2,577 (20.53)	1,540 (20.76)	97 (16.7)	1,152 (16.03)	1,253 (18)	1,155 (19.08)
70-74	58 (16.71)	1,734 (19.21)	2,605 (20.75)	1,220 (16.45)	109 (18.76)	1,366 (19.01)	1,428 (20.52)	1,267 (20.92)
75-79	40 (11.53)	1,297 (14.37)	1,754 (13.97)	780 (10.51)	104 (17.9)	1,281 (17.83)	1,342 (19.28)	914 (15.09)
80+	18 (5.19)	762 (8.44)	857 (6.83)	318 (4.29)	109 (18.76)	1,291 (17.97)	1,072 (15.4)	604 (9.98)
Total	347 (1.18)	9,027 (30.8)	12,552 (42.8)	7,418 (25.3)	581 (2.8)	7,185 (34.6)	6,960 (33.5)	6,055 (29.1)

Table D.12: Number of participants by age group, case-control status and deprivation index at study entry

Age Group	Deprivation - Townsend Quintile					
	Missing	Less Deprived 1	2	3	4	Most Deprived 5
			Cases (Number(%))			
40-44	165 (6.21)	478 (17.98)	461 (17.34)	560 (21.06)	560 (21.06)	435 (16.36)
45-49	569 (7.12)	1,515 (18.96)	1,462 (18.29)	1,669 (20.88)	1,567 (19.61)	1,210 (15.14)
50-54	1,036 (6.82)	3,158 (20.78)	2,827 (18.6)	3,042 (20.01)	2,909 (19.14)	2,227 (14.65)
55-59	1,301 (6.84)	4,156 (21.84)	3,752 (19.71)	3,835 (20.15)	3,494 (18.36)	2,495 (13.11)
60-64	1,296 (6.56)	4,341 (21.97)	3,975 (20.11)	4,037 (20.43)	3,600 (18.22)	2,513 (12.72)
65-69	1,148 (6.5)	3,692 (20.91)	3,779 (21.4)	3,589 (20.33)	3,245 (18.38)	2,204 (12.48)
70-74	900 (6.68)	2,915 (21.64)	2,873 (21.33)	2,604 (19.33)	2,529 (18.77)	1,650 (12.25)
75-79	535 (6.69)	1,678 (20.97)	1,729 (21.61)	1,594 (19.92)	1,492 (18.65)	973 (12.16)
80+	313 (6.57)	913 (19.16)	1,041 (21.85)	990 (20.78)	915 (19.21)	592 (12.43)
Total	7,263 (6.69)	22,846 (21.05)	21,899 (20.18)	21,920 (20.20)	20,311 (18.71)	14,299 (13.17)
			Controls (Number(%))			
40-44	396 (5.64)	1,866 (26.58)	1,514 (21.57)	1,365 (19.44)	1,147 (16.34)	732 (10.43)
45-49	1,259 (6.38)	5,514 (27.93)	4,221 (21.38)	3,717 (18.83)	3,028 (15.34)	2,000 (10.13)
50-54	2,192 (6.24)	9,587 (27.28)	7,756 (22.07)	6,619 (18.83)	5,334 (15.18)	3,658 (10.41)
55-59	2,780 (6.4)	11,962 (27.53)	9,804 (22.56)	8,148 (18.75)	6,493 (14.94)	4,269 (9.82)
60-64	2,849 (6.22)	12,474 (27.24)	10,570 (23.08)	8,533 (18.64)	6,865 (14.99)	4,497 (9.82)
65-69	2,593 (6.29)	10,604 (25.72)	9,525 (23.1)	8,032 (19.48)	6,477 (15.71)	3,998 (9.7)
70-74	2,015 (6.21)	8,243 (25.42)	7,293 (22.49)	6,356 (19.6)	5,355 (16.52)	3,163 (9.75)
75-79	1,246 (6.46)	4,630 (23.99)	4,358 (22.58)	3,817 (19.78)	3,270 (16.94)	1,977 (10.24)
80+	678 (6.3)	2,412 (22.43)	2,434 (22.63)	2,163 (20.11)	1,912 (17.78)	1,156 (10.75)
Total	16,008 (6.28)	67,292 (26.40)	57,475 (22.55)	48,750 (19.13)	39,881 (15.65)	25,450 (9.99)
			All Subjects (Number(%))			
40-44	561 (5.8)	2,344 (24.22)	1,975 (20.41)	1,925 (19.89)	1,707 (17.64)	1,167 (12.06)
45-49	1,828 (6.59)	7,029 (25.35)	5,683 (20.49)	5,386 (19.42)	4,595 (16.57)	3,210 (11.58)
50-54	3,228 (6.41)	12,745 (25.32)	10,583 (21.02)	9,661 (19.19)	8,243 (16.37)	5,885 (11.69)
55-59	4,081 (6.53)	16,118 (25.79)	13,556 (21.69)	11,983 (19.18)	9,987 (15.98)	6,764 (10.82)
60-64	4,145 (6.32)	16,815 (25.65)	14,545 (22.19)	12,570 (19.18)	10,465 (15.96)	7,010 (10.69)
65-69	3,741 (6.35)	14,296 (24.28)	13,304 (22.59)	11,621 (19.73)	9,722 (16.51)	6,202 (10.53)
70-74	2,915 (6.35)	11,158 (24.31)	10,166 (22.15)	8,960 (19.52)	7,884 (17.18)	4,813 (10.49)
75-79	1,781 (6.52)	6,308 (23.11)	6,087 (22.3)	5,411 (19.82)	4,762 (17.44)	2,950 (10.81)
80+	991 (6.39)	3,325 (21.43)	3,475 (22.39)	3,153 (20.32)	2,827 (18.22)	1,748 (11.26)
Total	23,271 (6.4)	90,138 (28.4)	79,374 (21.84)	70,670 (19.45)	60,192 (16.56)	39,749 (10.94)

Table D.13: Number of participants by gender, BP and case-control status as at study entry.

BP	Cases (Numbers (%))		Controls (Numbers (%))	
	Male	Female	Male	Female
Missing	3,868 (6.22)	2,680 (5.79)	17,513 (12.54)	10,071 (8.74)
Low	304 (0.49)	281 (0.61)	1,020 (0.73)	1,011 (0.88)
Normal	5,894 (9.47)	4,042 (8.73)	22,957 (16.44)	21,487 (18.65)
Pre-High	15,621 (25.11)	10,921 (23.57)	37,817 (27.08)	30,995 (26.9)
High	36,525 (58.71)	28,402 (61.31)	60,325 (43.2)	51,660 (44.83)
Total	62,212 (100)	46,326 (100)	139,632 (100)	115,224 (100)

Table D.14: Blood pressure status as at study entry by age group and case-control status

Age Group	Cases (Number(%))				Controls (Number(%))			
	Low	Normal	Pre-High	High	Low	Normal	Pre-High	High
40-44	23 (1.05)	430 (19.65)	504 (23.03)	1,231 (56.26)	92 (1.74)	2,443 (46.18)	1,429 (27.01)	1,326 (25.07)
45-49	51 (0.71)	1,017 (14.08)	1,886 (26.11)	4,268 (59.1)	226 (1.38)	5,798 (35.41)	5,266 (32.16)	5,084 (31.05)
50-54	95 (0.67)	1,651 (11.58)	4,010 (28.13)	8,498 (59.62)	320 (1.04)	8,668 (28.11)	10,469 (33.95)	11,376 (36.9)
55-59	98 (0.55)	1,822 (10.15)	4,995 (27.81)	11,043 (61.49)	356 (0.92)	8,800 (22.7)	12,937 (33.37)	16,675 (43.01)
60-64	97 (0.52)	1,745 (9.31)	4,997 (26.66)	11,901 (63.51)	338 (0.82)	7,372 (17.82)	13,232 (31.99)	20,419 (49.37)
65-69	89 (0.53)	1,343 (8.02)	4,330 (25.87)	10,978 (65.58)	325 (0.87)	5,328 (14.21)	11,126 (29.68)	20,710 (55.24)
70-74	72 (0.56)	993 (7.78)	3,057 (23.94)	8,649 (67.72)	205 (0.69)	3,440 (11.64)	7,937 (26.85)	17,978 (60.82)
75-79	42 (0.55)	556 (7.32)	1,766 (23.25)	5,232 (68.88)	113 (0.64)	1,746 (9.88)	4,239 (23.98)	11,576 (65.5)
80+	18 (0.4)	379 (8.38)	997 (22.05)	3,127 (69.17)	56 (0.56)	849 (8.56)	2,177 (21.94)	6,841 (68.94)

Table D.15: BP classification as provided by Blood Pressure UK

BP Group	DBP mmHg Range		SBP mmHg Range	
	From	To	From	To
Low	-	60	-	90
Normal	60	80	90	120
Pre-High	80	89	120	139
High	90	-	140	-

Table D.16: Amputation prevalence at baseline and during follow-up by age-group and case-control at study entry

Age Group	Baseline (Number (%))		Follow-up (Number (%))		Total (Number (%))	
	Case	Control	Case	Control	Case	Control
40-44	8 (0.3)	21 (0.3)	52 (1.96)	21 (0.3)	60 (2.26)	42 (0.6)
45-49	35 (0.44)	74 (0.37)	93 (1.16)	72 (0.36)	128 (1.6)	147 (0.74)
50-54	57 (0.38)	134 (0.38)	154 (1.01)	120 (0.34)	211 (1.39)	254 (0.72)
55-59	93 (0.49)	219 (0.5)	142 (0.75)	165 (0.38)	235 (1.23)	384 (0.88)
60-64	102 (0.52)	244 (0.53)	171 (0.87)	214 (0.47)	273 (1.38)	458 (1)
65-69	102 (0.58)	213 (0.52)	148 (0.84)	185 (0.45)	250 (1.42)	398 (0.97)
70-74	96 (0.71)	205 (0.63)	130 (0.97)	162 (0.5)	227 (1.69)	367 (1.13)
75-79	44 (0.55)	134 (0.69)	70 (0.87)	88 (0.46)	114 (1.42)	222 (1.15)
80+	27 (0.57)	82 (0.76)	33 (0.69)	40 (0.37)	60 (1.26)	122 (1.13)

Table D.17: Prevalence of selected morbidities as at study entry

Disease	Cases (Number (%))	Controls (Number (%))
Amputation	564 (0.52)	1,326 (0.52)
Cancer		
Cognitive impairment	27 (0.02)	73 (0.03)
CKD	1,112 (1.02)	1,855 (0.73)
Dementia		
HF	11,441 (10.54)	17,033 (6.68)
MI	16,304 (15.02)	21,878 (8.58)
PVD	13,105 (12.07)	30,788 (12.08)
Stroke	1,237 (1.14)	2,935 (1.15)

Table D.18: Cancer prevalence at baseline and during follow-up by age group and case-control status at study entry

Age	Baseline (Number (%))		During Follow-up (Number (%))		Total (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44			225 (8.46)	2,102 (29.94)	225 (8.46)	2,102 (29.94)
45-49			737 (9.22)	6,683 (33.86)	737 (9.22)	6,683 (33.86)
50-54			1,668 (10.97)	14,311 (40.72)	1,668 (10.97)	14,311 (40.72)
55-59			2,691 (14.14)	19,979 (45.98)	2,691 (14.14)	19,979 (45.98)
60-64			3,488 (17.65)	22,403 (48.93)	3,488 (17.65)	22,403 (48.93)
65-69			3,683 (20.86)	20,052 (48.64)	3,683 (20.86)	20,052 (48.64)
70-74			3,111 (23.09)	14,904 (45.96)	3,111 (23.09)	14,904 (45.96)
75-79			1,894 (23.67)	8,082 (41.88)	1,894 (23.67)	8,083 (41.89)
80+			912 (19.14)	3,646 (33.9)	912 (19.14)	3,646 (33.9)

Table D.19: Cognitive impairment prevalence by age group and case-control

Age	Baseline (Number (%))		During follow-up (Number (%))		Total (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44			48 (1.81)	20 (0.28)	48 (1.81)	20 (0.28)
45-49			117 (1.46)	101 (0.51)	118 (1.48)	105 (0.53)
50-54			367 (2.41)	329 (0.94)	370 (2.43)	334 (0.95)
55-59			652 (3.43)	582 (1.34)	654 (3.44)	591 (1.36)
60-64			842 (4.26)	910 (1.99)	849 (4.3)	924 (2.02)
65-69			798 (4.52)	1,130 (2.74)	799 (4.53)	1,142 (2.77)
70-74			569 (4.22)	1,117 (3.44)	575 (4.27)	1,129 (3.48)
75-79			311 (3.89)	598 (3.1)	316 (3.95)	608 (3.15)
80+			141 (2.96)	307 (2.85)	143 (3)	314 (2.92)

Table D.20: Dementia prevalence at study entry and during follow-up by age group and case-control

Age	Before (Number (%))		After (Number (%))		ALL (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44	-	-	15 (0.56)	106 (1.51)	15 (0.56)	106 (1.51)
45-49	-	-	83 (1.04)	383 (1.94)	83 (1.04)	383 (1.94)
50-54	-	-	196 (1.29)	991 (2.82)	196 (1.29)	991 (2.82)
55-59	-	-	450 (2.36)	1,885 (4.34)	450 (2.36)	1,885 (4.34)
60-64	-	-	735 (3.72)	3,160 (6.9)	735 (3.72)	3,160 (6.9)
65-69	-	-	1,016 (5.75)	4,533 (10.99)	1,016 (5.75)	4,533 (10.99)
70-74	-	-	1,172 (8.7)	5,197 (16.03)	1,172 (8.7)	5,197 (16.03)
75-79	-	-	948 (11.85)	4,016 (20.81)	948 (11.85)	4,016 (20.81)
80+	-	-	619 (12.99)	2,725 (25.34)	619 (12.99)	2,725 (25.34)

Table D.21: Prevalence of CKD 3 to 5 at study entry and during follow-up by case-control status and age group

Age Group	CKD Stage					
	Cases			Controls		
	CKD 3	CKD 4	CKD 5	CKD 3	CKD 4	CKD 5
	During follow-up (Number(%))					
40-44	224 (8.45)	14 (0.79)	21 (0.79)	389 (5.56)	16 (0.23)	31 (0.44)
45-49	658 (8.25)	22 (0.73)	58 (0.73)	1,543 (7.84)	32 (0.16)	83 (0.42)
50-54	1,477 (9.74)	46 (0.64)	97 (0.64)	3,656 (10.44)	66 (0.19)	149 (0.43)
55-59	2,541 (13.37)	70 (0.76)	144 (0.76)	6,689 (15.44)	119 (0.27)	271 (0.63)
60-64	3,792 (19.24)	97 (0.94)	186 (0.94)	9,797 (21.5)	182 (0.4)	333 (0.73)
65-69	4,268 (24.23)	121 (1.35)	237 (1.35)	11,403 (27.79)	252 (0.61)	440 (1.07)
70-74	4,011 (29.9)	144 (1.67)	224 (1.67)	10,729 (33.3)	302 (0.94)	400 (1.24)
75-79	2,592 (32.55)	104 (1.73)	138 (1.73)	6,707 (35.05)	284 (1.48)	299 (1.56)
80+	1,437 (30.35)	83 (2.01)	95 (2.01)	3,827 (35.97)	249 (2.34)	143 (1.34)

Table D.22: Heart failure prevalence as at study entry and during follow-up by age group and case-control status

Age	Baseline (Number (%))		During follow-up (Number (%))		Total (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44	108 (4.06)	144 (2.05)	319 (12)	473 (6.74)	427 (16.06)	618 (8.8)
45-49	480 (6.01)	524 (2.65)	789 (9.87)	1,325 (6.71)	1,269 (15.88)	1,849 (9.37)
50-54	1,018 (6.7)	1,174 (3.34)	1,368 (9)	2,685 (7.64)	2,386 (15.7)	3,859 (10.98)
55-59	1,643 (8.63)	1,973 (4.54)	1,935 (10.17)	4,274 (9.84)	3,578 (18.8)	6,247 (14.38)
60-64	2,151 (10.88)	2,929 (6.4)	2,239 (11.33)	5,443 (11.89)	4,390 (22.21)	8,372 (18.28)
65-69	2,207 (12.5)	3,386 (8.21)	2,263 (12.82)	5,755 (13.96)	4,470 (25.32)	9,141 (22.17)
70-74	1,928 (14.31)	3,431 (10.58)	1,945 (14.44)	5,077 (15.66)	3,873 (28.75)	8,508 (26.24)
75-79	1,206 (15.07)	2,235 (11.58)	1,161 (14.51)	3,176 (16.46)	2,367 (29.58)	5,411 (28.04)
80+	700 (14.69)	1,237 (11.5)	663 (13.92)	1,830 (17.02)	1,363 (28.61)	3,067 (28.52)

Table D.23: MI prevalence at study entry and during follow-up age-group and case-control status

Age	Baseline (Number (%))		During follow-up (Number (%))		Total (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44	98 (3.69)	47 (0.67)	356 (13.39)	404 (5.75)	454 (17.07)	451 (6.42)
45-49	489 (6.12)	313 (1.59)	923 (11.55)	1,229 (6.23)	1,412 (17.67)	1,542 (7.81)
50-54	1,279 (8.42)	1,018 (2.9)	1,580 (10.4)	2,415 (6.87)	2,860 (18.82)	3,433 (9.77)
55-59	2,183 (11.47)	2,190 (5.04)	1,959 (10.29)	3,719 (8.56)	4,142 (21.76)	5,910 (13.6)
60-64	3,041 (15.39)	3,732 (8.15)	2,187 (11.07)	4,615 (10.08)	5,228 (26.45)	8,347 (18.23)
65-69	3,375 (19.11)	4,706 (11.41)	2,023 (11.46)	4,574 (11.09)	5,398 (30.57)	9,280 (22.51)
70-74	2,935 (21.79)	4,769 (14.71)	1,505 (11.17)	3,690 (11.38)	4,440 (32.96)	8,459 (26.09)
75-79	1,841 (23.01)	3,258 (16.88)	866 (10.82)	1,965 (10.18)	2,707 (33.83)	5,223 (27.06)
80+	1,063 (22.31)	1,845 (17.15)	439 (9.21)	976 (9.07)	1,502 (31.53)	2,821 (26.23)

Table D.24: Peripheral vascular disease prevalence at study entry and during follow-up by age group and case-control status

Age	Before (Number (%))		After (Number (%))		ALL (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44	134 (5.04)	366 (5.21)	288 (10.83)	609 (8.68)	422 (15.87)	975 (13.89)
45-49	463 (5.79)	1,236 (6.26)	717 (8.97)	1,643 (8.32)	1,180 (14.76)	2,879 (14.59)
50-54	1,187 (7.81)	2,806 (7.98)	1,224 (8.05)	2,986 (8.5)	2,411 (15.86)	5,792 (16.48)
55-59	1,881 (9.88)	4,294 (9.88)	1,609 (8.45)	4,193 (9.65)	3,490 (18.34)	8,487 (19.53)
60-64	2,418 (12.24)	5,524 (12.06)	1,798 (9.1)	4,991 (10.9)	4,217 (21.34)	10,515 (22.96)
65-69	2,496 (14.14)	5,849 (14.19)	1,825 (10.34)	4,801 (11.64)	4,321 (24.47)	10,650 (25.83)
70-74	2,206 (16.38)	5,319 (16.4)	1,381 (10.25)	3,852 (11.88)	3,587 (26.63)	9,171 (28.28)
75-79	1,418 (17.72)	3,393 (17.58)	805 (10.06)	2,144 (11.11)	2,223 (27.78)	5,537 (28.69)
80+	902 (18.93)	2,001 (18.61)	381 (8)	906 (8.42)	1,283 (26.93)	2,907 (27.03)

Table D.25: Stroke prevalence at study entry and during follow-up by age group and case-control status

Age	Baseline (Number (%))		During follow-up (Number (%))		Total (Number (%))	
Group	Case	Control	Case	Control	Case	Control
40-44	1 (0.04)	3 (0.04)	128 (4.81)	439 (6.25)	129 (4.85)	442 (6.3)
45-49	19 (0.24)	34 (0.17)	389 (4.87)	1,520 (7.7)	408 (5.11)	1,554 (7.87)
50-54	51 (0.34)	138 (0.39)	760 (5)	3,115 (8.86)	811 (5.34)	3,253 (9.26)
55-59	131 (0.69)	306 (0.7)	1,081 (5.68)	4,595 (10.57)	1,212 (6.37)	4,901 (11.28)
60-64	190 (0.96)	426 (0.93)	1,392 (7.04)	5,824 (12.72)	1,582 (8.01)	6,250 (13.65)
65-69	260 (1.47)	591 (1.43)	1,565 (8.86)	6,091 (14.77)	1,825 (10.34)	6,682 (16.21)
70-74	228 (1.69)	617 (1.9)	1,452 (10.78)	5,455 (16.82)	1,680 (12.47)	6,072 (18.73)
75-79	195 (2.44)	470 (2.44)	964 (12.05)	3,563 (18.46)	1,159 (14.49)	4,033 (20.9)
80+	162 (3.4)	350 (3.25)	630 (13.22)	2,090 (19.43)	792 (16.62)	2,440 (22.69)

Table D.26: Unadjusted hazard ratios of all-cause mortality and selected morbidities by age group and deprivation index at study entry

Age Group	Overal Hazard Ratios (Confidence Intervals)				
	Townsend Deprivation Index				
	1-Less Deprived	2	3	4	5-Most Deprived
40 - 44	1	0.76 (0.46 , 1.24)	0.52 (0.31 , 0.88)	1 (0.64 , 1.56)	1.13 (0.72 , 1.79)
45 - 49	1	1.13 (0.84 , 1.51)	1.2 (0.91 , 1.6)	1.4 (1.06 , 1.84)	1.35 (1 , 1.82)
50 - 54	1	1.09 (0.92 , 1.3)	1.02 (0.85 , 1.21)	1.11 (0.93 , 1.32)	1.34 (1.12 , 1.6)
55 - 59	1	1.18 (1.03 , 1.35)	1.17 (1.02 , 1.34)	1.15 (1 , 1.32)	1.43 (1.24 , 1.65)
60 - 64	1	1.07 (0.95 , 1.2)	1.13 (1.01 , 1.26)	1.02 (0.91 , 1.15)	1.26 (1.12 , 1.42)
65 - 69	1	0.97 (0.87 , 1.07)	0.98 (0.88 , 1.08)	0.98 (0.89 , 1.09)	1.03 (0.92 , 1.16)
70 - 74	1	1.13 (1.02 , 1.25)	1.03 (0.93 , 1.14)	1.04 (0.94 , 1.15)	1.03 (0.92 , 1.16)
75 - 79	1	1.07 (0.96 , 1.2)	1.04 (0.93 , 1.17)	1.02 (0.91 , 1.15)	1.11 (0.97 , 1.26)
80 +	1	1.04 (0.91 , 1.19)	1.07 (0.94 , 1.22)	0.94 (0.82 , 1.08)	0.98 (0.84 , 1.15)
Male Hazard Ratios					
40 - 44	1	0.79 (0.45 , 1.38)	0.54 (0.29 , 0.99)	1.16 (0.7 , 1.92)	1.18 (0.69 , 2.01)
45 - 49	1	1.03 (0.72 , 1.47)	1.36 (0.97 , 1.9)	1.49 (1.07 , 2.07)	1.51 (1.06 , 2.16)
50 - 54	1	1.11 (0.91 , 1.36)	1.1 (0.9 , 1.35)	1.13 (0.92 , 1.38)	1.4 (1.14 , 1.73)
55 - 59	1	1.18 (1.01 , 1.39)	1.23 (1.05 , 1.44)	1.27 (1.08 , 1.5)	1.49 (1.25 , 1.77)
60 - 64	1	1.12 (0.97 , 1.28)	1.18 (1.03 , 1.35)	1.09 (0.95 , 1.26)	1.35 (1.16 , 1.57)
65 - 69	1	1 (0.88 , 1.13)	1.05 (0.92 , 1.19)	1.02 (0.89 , 1.16)	1.11 (0.95 , 1.29)
70 - 74	1	1.22 (1.08 , 1.39)	1.14 (1 , 1.3)	1.09 (0.95 , 1.25)	1.14 (0.97 , 1.33)
75 - 79	1	1.04 (0.9 , 1.21)	1.02 (0.87 , 1.19)	1.06 (0.9 , 1.24)	0.95 (0.78 , 1.14)
80 +	1	1.06 (0.86 , 1.3)	1.09 (0.88 , 1.34)	0.91 (0.73 , 1.14)	1 (0.78 , 1.3)
Female Hazard Ratios					
40 - 44	1	0.67 (0.23 , 1.93)	0.54 (0.2 , 1.5)	0.81 (0.32 , 2.06)	1.23 (0.5 , 3.01)
45 - 49	1	1.36 (0.81 , 2.28)	0.9 (0.53 , 1.55)	1.25 (0.76 , 2.07)	1.12 (0.65 , 1.9)
50 - 54	1	1.05 (0.74 , 1.51)	0.9 (0.63 , 1.28)	1.17 (0.84 , 1.63)	1.39 (0.99 , 1.94)
55 - 59	1	1.21 (0.94 , 1.56)	1.08 (0.83 , 1.4)	1 (0.77 , 1.3)	1.47 (1.13 , 1.9)
60 - 64	1	1.02 (0.84 , 1.25)	1.11 (0.91 , 1.35)	0.99 (0.81 , 1.21)	1.24 (1.01 , 1.52)
65 - 69	1	0.93 (0.78 , 1.1)	0.9 (0.75 , 1.07)	0.98 (0.83 , 1.17)	1.01 (0.84 , 1.21)
70 - 74	1	1.05 (0.89 , 1.23)	0.92 (0.78 , 1.09)	1.04 (0.89 , 1.21)	0.99 (0.83 , 1.18)
75 - 79	1	1.13 (0.95 , 1.34)	1.11 (0.93 , 1.33)	1.07 (0.9 , 1.27)	1.37 (1.14 , 1.65)
80 +	1	1.05 (0.88 , 1.25)	1.09 (0.91 , 1.29)	0.98 (0.82 , 1.17)	1.01 (0.83 , 1.23)

Table D.27: Unadjusted hazard ratios of all-cause mortality by age group and deprivation index for controls

Age Group	Overall Hazard Ratios (Confidence Intervals)				
	Townsend Deprivation Index				
	1-Less Deprived	2	3	4	5-Most Deprived
40 - 44	1	1.31 (0.97 , 1.76)	1.44 (1.06 , 1.94)	2.28 (1.72 , 3.03)	2.49 (1.83 , 3.4)
45 - 49	1	1.13 (0.96 , 1.34)	1.41 (1.2 , 1.67)	1.77 (1.5 , 2.09)	2.69 (2.28 , 3.18)
50 - 54	1	1.22 (1.09 , 1.37)	1.52 (1.36 , 1.69)	1.97 (1.76 , 2.19)	2.46 (2.19 , 2.76)
55 - 59	1	1.21 (1.12 , 1.32)	1.41 (1.29 , 1.53)	1.92 (1.77 , 2.09)	2.15 (1.96 , 2.35)
60 - 64	1	1.17 (1.09 , 1.25)	1.37 (1.27 , 1.47)	1.59 (1.48 , 1.71)	1.88 (1.75 , 2.04)
65 - 69	1	1.06 (1 , 1.13)	1.21 (1.14 , 1.29)	1.4 (1.32 , 1.5)	1.76 (1.64 , 1.88)
70 - 74	1	1.12 (1.05 , 1.19)	1.25 (1.17 , 1.33)	1.34 (1.26 , 1.43)	1.55 (1.44 , 1.67)
75 - 79	1	1.04 (0.97 , 1.11)	1.16 (1.08 , 1.24)	1.2 (1.12 , 1.29)	1.36 (1.26 , 1.48)
80 +	1	1.05 (0.96 , 1.14)	1.05 (0.96 , 1.15)	1.14 (1.04 , 1.24)	1.13 (1.02 , 1.25)
Male Hazard Ratios					
40 - 44	1	1.25 (0.87 , 1.79)	1.46 (1.02 , 2.09)	2.29 (1.62 , 3.22)	2.58 (1.78 , 3.76)
45 - 49	1	1.08 (0.88 , 1.33)	1.39 (1.13 , 1.7)	2.02 (1.66 , 2.46)	2.73 (2.23 , 3.33)
50 - 54	1	1.21 (1.06 , 1.38)	1.61 (1.41 , 1.83)	2.04 (1.79 , 2.32)	2.61 (2.28 , 3)
55 - 59	1	1.28 (1.16 , 1.42)	1.54 (1.39 , 1.7)	2.07 (1.87 , 2.29)	2.35 (2.1 , 2.62)
60 - 64	1	1.21 (1.11 , 1.31)	1.43 (1.31 , 1.56)	1.64 (1.5 , 1.8)	1.99 (1.8 , 2.19)
65 - 69	1	1.06 (0.98 , 1.15)	1.27 (1.17 , 1.37)	1.49 (1.38 , 1.62)	1.88 (1.72 , 2.06)
70 - 74	1	1.12 (1.03 , 1.21)	1.24 (1.15 , 1.35)	1.41 (1.3 , 1.54)	1.65 (1.5 , 1.81)
75 - 79	1	1.13 (1.02 , 1.24)	1.27 (1.15 , 1.4)	1.32 (1.19 , 1.46)	1.54 (1.37 , 1.74)
80 +	1	1.09 (0.95 , 1.25)	1.1 (0.95 , 1.27)	1.24 (1.07 , 1.45)	1.25 (1.04 , 1.49)
Female Hazard Ratios					
40 - 44	1	1.52 (0.89 , 2.61)	1.49 (0.85 , 2.59)	2.49 (1.49 , 4.19)	2.65 (1.52 , 4.62)
45 - 49	1	1.25 (0.93 , 1.68)	1.5 (1.12 , 2.01)	1.37 (1 , 1.87)	2.64 (1.95 , 3.57)
50 - 54	1	1.29 (1.05 , 1.58)	1.4 (1.14 , 1.73)	1.95 (1.6 , 2.38)	2.35 (1.9 , 2.9)
55 - 59	1	1.11 (0.96 , 1.28)	1.21 (1.04 , 1.41)	1.72 (1.49 , 1.99)	1.89 (1.61 , 2.21)
60 - 64	1	1.1 (0.98 , 1.23)	1.3 (1.16 , 1.47)	1.57 (1.4 , 1.76)	1.8 (1.59 , 2.05)
65 - 69	1	1.08 (0.97 , 1.2)	1.19 (1.07 , 1.32)	1.35 (1.22 , 1.5)	1.73 (1.55 , 1.94)
70 - 74	1	1.13 (1.03 , 1.25)	1.3 (1.18 , 1.43)	1.34 (1.21 , 1.47)	1.55 (1.39 , 1.73)
75 - 79	1	0.98 (0.89 , 1.09)	1.09 (0.99 , 1.21)	1.17 (1.05 , 1.29)	1.34 (1.19 , 1.5)
80 +	1	1.03 (0.93 , 1.15)	1.04 (0.93 , 1.16)	1.12 (1.01 , 1.25)	1.12 (0.99 , 1.27)

Table D.28: Prevalence of new cancers over the follow-up period by case-control status and age group at entry

Age	Cases [Number(%)]		Control [Number(%)]	
Group	No	Yes	No	Yes
40-44	2,348 (86.45)	368 (13.55)	1,633 (27.23)	4,363 (72.77)
45-49	6,778 (82.17)	1,471 (17.83)	4,534 (25.52)	13,232 (74.48)
50-54	11,855 (73.92)	4,183 (26.08)	7,642 (21.37)	28,116 (78.63)
55-59	13,528 (63.71)	7,705 (36.29)	9,502 (18.53)	41,771 (81.47)
60-64	13,219 (60.89)	8,490 (39.11)	12,543 (23.21)	41,493 (76.79)
65-69	12,693 (67.86)	6,011 (32.14)	15,874 (34.82)	29,710 (65.18)
70-74	10,802 (74.08)	3,779 (25.92)	17,256 (48.56)	18,280 (51.44)
75-79	6,839 (76.58)	2,091 (23.42)	12,276 (56.96)	9,276 (43.04)
80-84	3,171 (79.24)	831 (20.76)	5,970 (63.48)	3,434 (36.52)
85-89	955 (85.34)	164 (14.66)	1,655 (68.84)	749 (31.16)
90-94	166 (90.22)	18 (9.78)	234 (73.58)	84 (26.42)
95-99	12 (70.59)	5 (29.41)	21 (75)	7 (25)

Table D.29: Estimated effects on Model B with quadratic polynomial on age.

Variables and Descriptive Statistics		Estimates	CI	p-value
Sample size		154045		
Number of non-censored		19604		
Gompertz Parameters				
Scale	1000a	6.649	6.081 - 7.27	0
Shape	100b	9.956	9.18 - 10.799	0
Shape Effects				
Birth Cohort				
	1950-1960	0.942	0.835 - 1.063	0.3338
	1940-1949	0.821	0.762 - 0.885	0
AF	Yes	1.328	1.228 - 1.435	0
HTN				
	Treated	1.316	1.215 - 1.426	0
	Untreated	0.912	0.815 - 1.019	0.1048
Scale Effects				
Birth Cohort				
	1950-1960	1.626	1.444 - 1.83	0
	1940-1949	1.432	1.325 - 1.548	0
T2DM Indicator	Yes	1.485	1.408 - 1.566	0
Age centered at 68.89017				
	(Age-68.89017)^2	1	0.999 - 1	0.1516
	Age-68.89017	1.095	1.086 - 1.105	0
Gender	Male	1.371	1.32 - 1.424	0
Smoking Status				
	Former	1.671	1.545 - 1.807	0
	Smoker	2.853	2.656 - 3.065	0.00E+00
Deprivation (TDI)				
	Less Deprived	0.845	0.807 - 0.883	0
	2	0.917	0.878 - 0.958	1.00E-04
	4	1.06	1.013 - 1.108	0.0111
	Most Deprived	1.17	1.113 - 1.231	0
AF	Yes	0.779	0.713 - 0.852	0
HF	Yes	1.15	1.096 - 1.207	0
HCL				
	Treated	0.919	0.884 - 0.955	0
	Untreated	1.266	1.185 - 1.352	0
HTN				
	Treated	0.764	0.717 - 0.813	0
	Untreated	1.403	1.31 - 1.502	0.00E+00
MI	Yes	1.326	1.257 - 1.4	0
PVD	Yes	1.082	1.039 - 1.126	1.00E-04
BMI				
	Overweight	1.01	0.952 - 1.071	0.7372
	Obese	1.18	1.106 - 1.259	0
Gender:Age centered at 68.89017				
	Male:(Age-68.89017)^2	1.001	1 - 1.001	0.0888
	Male:(Age-68.89017)	1.001	0.991 - 1.011	0.8009
T2DM Indicator:MI	T2DM:Yes	0.711	0.656 - 0.771	0
T2DM Indicator:Smoking Status				
	T2DM:Former	0.747	0.692 - 0.806	0
	T2DM:Smoker	0.413	0.38 - 0.449	0
Smoking Status:BMI				
	Former:Overweight	0.862	0.788 - 0.943	1.20E-03
	Smoker:Overweight	0.803	0.739 - 0.872	0
	Former:Obese	0.845	0.767 - 0.931	7.00E-04
	Smoker:Obese	0.791	0.719 - 0.87	0
Birth Cohort:Smoking Status				
	1950-1960:Former	0.988	0.88 - 1.108	8.33E-01
	1940-1949:Former	0.975	0.906 - 1.05	0.5058
	1950-1960:Smoker	1.234	1.113 - 1.369	1.00E-04
	1940-1949:Smoker	1.093	1.015 - 1.178	0.0186
Frailty(Sigma2)		0.127	0.108 - 0.15	0
Goodness of Fit				
Concordance (se)		0.739 (0.002)		
Loglik		-95554.09		
AIC		191194.19		

Table D.30: Estimated effects on Model B with cubic polynomial on age.

Variables and Descriptive Statistics		Estimates	CI	p-value
Sample size		154045		
Number of non-censored		19604		
Gomperts Parameters				
Scale	1000a	9.887	9.771 - 10.004	0
Shape	100b	8.318	8.121 - 8.52	0
Shape Effects				
Birth Cohort				
	1950-1960	1	0.986 - 1.014	0.9989
	1940-1949	1	0.981 - 1.019	0.999
AF	Yes	1	0.98 - 1.02	0.9992
HTN				
	Treated	1	0.984 - 1.016	1
	Untreated	1	0.985 - 1.016	0.9996
Scale Effects				
Birth Cohort				
	1950-1960	1	0.988 - 1.012	0.9992
	1940-1949	1	0.98 - 1.021	0.9999
T2DM Indicator	Yes	1	0.972 - 1.029	0.9999
Age centered at 68.89017				
	(Age-68.89017)^3	1	1 - 1	0.9996
	(Age-68.89017)^2	1	0.998 - 1.002	0.999
	Age-68.89017	1	0.99 - 1.01	0.9994
Gender	Male	1	0.956 - 1.046	0.9999
Smoking Status				
	Former	1	0.98 - 1.02	0.9998
	Smoker	1	0.978 - 1.022	0.9975
Deprivation (TDI)				
	Less Deprived	1	0.971 - 1.03	0.9993
	2	1	0.971 - 1.029	0.9996
	4	1	0.97 - 1.031	0.9986
	Most Deprived	1	0.968 - 1.034	0.9985
AF	Yes	1	0.962 - 1.04	0.9989
HF	Yes	1	0.981 - 1.019	0.9988
HCL				
	Treated	1	0.964 - 1.038	0.9978
	Untreated	1	0.995 - 1.005	0.9984
HTN				
	Treated	1	0.974 - 1.027	0.999
	Untreated	1	0.968 - 1.033	0.9998
MI	Yes	1	0.971 - 1.03	0.9993
PVD	Yes	1	0.959 - 1.043	0.9998
BMI				
	Overweight	1	0.969 - 1.032	0.9996
	Obese	1	0.966 - 1.036	0.9986
Gender:Age centered at 68.89017				
	Male:(Age-68.89017)^3	1	1 - 1	0.9994
	Male:(Age-68.89017)^2	1	0.998 - 1.002	0.9997
	Male:(Age-68.89017)	1	0.986 - 1.014	0.9994
T2DM Indicator:MI	T2DM:Yes	1	0.983 - 1.018	0.9992
T2DM Indicator:Smoking Status				
	T2DM:Former	1	0.981 - 1.019	0.9997
	T2DM:Smoker	1	0.985 - 1.015	0.9983
Smoking Status:BMI				
	Former:Overweight	1	0.975 - 1.026	0.9999
	Smoker:Overweight	1	0.978 - 1.022	0.999
	Former:Obese	1	0.975 - 1.025	0.9997
	Smoker:Obese	1	0.98 - 1.021	0.9991
Birth Cohort:Smoking Status				
	1950-1960:Former	1	0.993 - 1.007	0.9992
	1940-1949:Former	1	0.986 - 1.015	0.9991
	1950-1960:Smoker	1	0.992 - 1.008	0.9978
	1940-1949:Smoker	1	0.987 - 1.014	0.9985
Frailty(Sigma2)		1	1 - 1	0
Goodness of Fit				
Concordance (se)		1 (0.002)		
Loglik		-99738.24		
AIC		199566.49		

Table D.31: Assessment of the proportionality assumption in the Cox PH model.

Variable	rho	chisq	p
Case-Control [Cases]	0.001	0.008	0.931
Age Group [50-59]	0.003	0.042	0.837
Age Group [60+]	-0.015	1.122	0.289
Birth Year [1930-1939]	0.049	13.646	0.000
Birth Year [1940-1949]	0.028	4.443	0.035
Gender [Male]	-0.001	0.011	0.918
Smokes [Former]	-0.008	0.367	0.545
Smokes [Smoker]	-0.024	2.866	0.090
Townsend [Less Deprived]	-0.017	1.584	0.208
Townsend [2]	-0.013	0.873	0.350
Townsend [4]	-0.010	0.555	0.456
Townsend [Most Deprived]	-0.001	0.003	0.955
HF [Yes]	0.000	0.001	0.981
Hypercholesterolemia [Treated]	0.033	5.858	0.016
Hypercholesterolemia [Untreated]	-0.032	5.510	0.019
Hypertension [Treated]	0.036	6.896	0.009
Hypertension [Untreated]	-0.017	1.455	0.228
MI [Yes]	-0.020	2.116	0.146
PVD [Yes]	-0.004	0.099	0.753
BMI [Overweight]	-0.011	0.606	0.436
BMI [Obese]	-0.003	0.034	0.854
Case-Control [Cases]:Smokes [Former]	0.025	3.357	0.067
Case-Control [Cases]:Smokes [Smoker]	0.021	2.334	0.127
Case-Control [Cases]:Hypercholesterolemia [Treated]	-0.031	5.062	0.024
Case-Control [Cases]:Hypercholesterolemia [Untreated]	-0.015	1.147	0.284
Case-Control [Cases]:MI [Yes]	0.019	1.870	0.171
Case-Control [Cases]:PVD [Yes]	0.005	0.132	0.716
Case-Control [Cases]:BMI [Overweight]	-0.010	0.556	0.456
Case-Control [Cases]:BMI [Obese]	-0.027	3.697	0.055
Age Group [50-59]:Gender [Male]	0.005	0.119	0.730
Age Group [60+]:Gender [Male]	0.014	1.006	0.316
Birth Year [1930-1939]:Gender [Male]	-0.028	4.511	0.034
Birth Year [1940-1949]:Gender [Male]	-0.018	1.779	0.182
Smokes [Former]:BMI [Overweight]	-0.005	0.138	0.710
Smokes [Smoker]:BMI [Overweight]	0.006	0.183	0.669
Smokes [Former]:BMI [Obese]	-0.005	0.149	0.699
Smokes [Smoker]:BMI [Obese]	-0.003	0.061	0.806
Townsend [Less Deprived]:BMI [Overweight]	0.002	0.025	0.876
Townsend [2]:BMI [Overweight]	0.008	0.349	0.554
Townsend [4]:BMI [Overweight]	0.007	0.228	0.633
Townsend [Most Deprived]:BMI [Overweight]	0.003	0.036	0.850
Townsend [Less Deprived]:BMI [Obese]	0.018	1.632	0.201
Townsend [2]:BMI [Obese]	0.012	0.708	0.400
Townsend [4]:BMI [Obese]	0.010	0.534	0.465
Townsend [Most Deprived]:BMI[Obese]	0.008	0.342	0.559
GLOBAL	NA	103.215	0.000

(a) rho is the Pearson product-moment correlation between the scaled Schoenfeld residuals and the KM estimators for each covariate. A p-value<0.05 indicates a violation of the proposrtionality assumption.

Table D.32: Prevalence of selected morbidities at study entry by follow-up year

YEAR	Amputation	Cognitive impairment	CKD	HF	MI	PVD	Stroke
1984	927 (0.26)		300 (0.08)	84 (0.02)	3,196 (0.88)	10,370 (2.85)	107 (0.03)
1985	958 (0.26)		318 (0.09)	94 (0.03)	3,761 (1.03)	11,027 (3.03)	120 (0.03)
1986	992 (0.27)		330 (0.09)	107 (0.03)	4,375 (1.2)	11,744 (3.23)	149 (0.04)
1987	1,024 (0.28)		342 (0.09)	128 (0.04)	5,105 (1.4)	12,504 (3.44)	185 (0.05)
1988	1,062 (0.29)		358 (0.1)	166 (0.05)	5,918 (1.63)	13,319 (3.67)	228 (0.06)
1989	1,103 (0.3)		380 (0.1)	223 (0.06)	6,915 (1.9)	14,357 (3.95)	283 (0.08)
1990	1,152 (0.32)		404 (0.11)	382 (0.11)	8,231 (2.27)	15,982 (4.4)	349 (0.1)
1991	1,205 (0.33)	1 (0)	436 (0.12)	652 (0.18)	9,635 (2.65)	17,800 (4.9)	405 (0.11)
1992	1,255 (0.35)	3 (0)	476 (0.13)	1,130 (0.31)	11,124 (3.06)	19,576 (5.39)	499 (0.14)
1993	1,310 (0.36)	5 (0)	506 (0.14)	2,696 (0.74)	12,973 (3.57)	21,258 (5.85)	603 (0.17)
1994	1,364 (0.38)	7 (0)	558 (0.15)	5,996 (1.65)	14,854 (4.09)	23,041 (6.34)	729 (0.2)
1995	1,417 (0.39)	10 (0)	598 (0.16)	8,685 (2.39)	16,685 (4.59)	24,739 (6.81)	859 (0.24)
1996	1,467 (0.4)	16 (0)	651 (0.18)	10,406 (2.86)	18,464 (5.08)	26,417 (7.27)	1,041 (0.29)
1997	1,509 (0.42)	18 (0)	702 (0.19)	11,889 (3.27)	20,449 (5.63)	28,111 (7.74)	1,298 (0.36)
1998	1,567 (0.43)	19 (0.01)	757 (0.21)	13,398 (3.69)	22,260 (6.13)	29,790 (8.21)	1,561 (0.43)
1999	1,606 (0.44)	24 (0.01)	832 (0.23)	14,792 (4.08)	24,149 (6.66)	31,413 (8.66)	1,852 (0.51)
2000	1,645 (0.45)	25 (0.01)	898 (0.25)	16,455 (4.54)	26,176 (7.23)	33,066 (9.13)	2,177 (0.6)
2001	1,669 (0.46)	27 (0.01)	971 (0.27)	18,611 (5.15)	28,222 (7.81)	34,726 (9.61)	2,550 (0.71)
2002	1,692 (0.47)	27 (0.01)	1,063 (0.3)	20,821 (5.78)	30,057 (8.35)	36,197 (10.06)	2,886 (0.8)
2003	1,730 (0.48)	28 (0.01)	1,150 (0.32)	22,667 (6.33)	31,498 (8.79)	37,472 (10.46)	3,173 (0.89)
2004	1,746 (0.49)	28 (0.01)	1,231 (0.35)	24,013 (6.74)	32,530 (9.14)	38,500 (10.81)	3,369 (0.95)
2005	1,757 (0.5)	30 (0.01)	1,319 (0.37)	24,753 (7.01)	33,214 (9.4)	39,216 (11.1)	3,511 (0.99)
2006	1,763 (0.5)	33 (0.01)	2,000 (0.57)	25,191 (7.2)	33,535 (9.59)	39,604 (11.32)	3,590 (1.03)
2007	1,760 (0.51)	35 (0.01)	2,444 (0.71)	25,524 (7.37)	33,684 (9.73)	39,705 (11.47)	3,648 (1.05)
2008	1,740 (0.51)	36 (0.01)	2,636 (0.77)	25,521 (7.47)	33,413 (9.78)	39,538 (11.57)	3,652 (1.07)
2009	1,718 (0.51)	44 (0.01)	2,692 (0.8)	25,372 (7.54)	33,030 (9.81)	39,258 (11.66)	3,626 (1.08)
2010	1,698 (0.51)	55 (0.02)	2,699 (0.81)	25,082 (7.57)	32,497 (9.81)	38,832 (11.72)	3,599 (1.09)
2011	1,670 (0.51)	61 (0.02)	2,682 (0.82)	24,729 (7.59)	31,856 (9.77)	38,320 (11.75)	3,557 (1.09)
2012	1,637 (0.51)	63 (0.02)	2,635 (0.82)	24,267 (7.58)	31,051 (9.7)	37,585 (11.74)	3,464 (1.08)
2013	1,593 (0.51)	78 (0.02)	2,583 (0.82)	23,709 (7.55)	30,198 (9.62)	36,816 (11.72)	3,379 (1.08)
2014	1,552 (0.5)	84 (0.03)	2,514 (0.82)	23,153 (7.52)	29,407 (9.55)	36,046 (11.7)	3,290 (1.07)
2015	1,519 (0.5)	89 (0.03)	2,457 (0.81)	22,625 (7.49)	28,565 (9.45)	35,289 (11.68)	3,194 (1.06)
2016	1,483 (0.5)	90 (0.03)	2,392 (0.81)	22,061 (7.43)	27,809 (9.36)	34,552 (11.63)	3,109 (1.05)

Table D.33: Incidences of selected morbidities during follow-up

Year	Amputation	Cancer	Cognitive impairment	CKD	Dementia	HF	MI	PVD	Stroke
1987							1 (0)		
1988		3 (0)		1 (0)		1 (0)	4 (0)	1 (0)	1 (0)
1989	2 (0)	21 (0)		4 (0)	1 (0)	6 (0)	9 (0)	9 (0)	3 (0)
1990	2 (0)	93 (0.03)		9 (0)	4 (0)	19 (0.01)	49 (0.01)	42 (0.01)	19 (0.01)
1991	3 (0)	249 (0.07)		17 (0.01)	10 (0)	60 (0.02)	109 (0.03)	120 (0.03)	60 (0.02)
1992	4 (0)	497 (0.14)		39 (0.01)	20 (0.01)	124 (0.04)	216 (0.06)	237 (0.07)	137 (0.04)
1993	15 (0)	926 (0.25)		64 (0.02)	32 (0.01)	351 (0.1)	402 (0.11)	413 (0.12)	281 (0.08)
1994	24 (0.01)	1,513 (0.42)		110 (0.04)	54 (0.02)	839 (0.25)	637 (0.18)	620 (0.17)	463 (0.13)
1995	31 (0.01)	2,339 (0.64)		158 (0)	105 (0.03)	1,356 (0.42)	901 (0.26)	896 (0.25)	691 (0.19)
1996	44 (0.01)	3,446 (0.95)	1 (0)	225 (0.06)	194 (0.06)	1,793 (0.56)	1,237 (0.37)	1,200 (0.34)	1,024 (0.29)
1997	70 (0.02)	4,833 (1.33)	1 (0)	311 (0.09)	312 (0.11)	2,246 (0.72)	1,683 (0.51)	1,584 (0.46)	1,420 (0.4)
1998	98 (0.03)	6,480 (1.78)	2 (0)	424 (0.12)	463 (0)	2,889 (0.94)	2,245 (0.69)	2,049 (0.6)	1,954 (0.55)
1999	128 (0.04)	8,817 (2.43)	6 (0)	600 (0.17)	685 (0.19)	3,663 (1.21)	2,930 (0.92)	2,593 (0.77)	2,605 (0.74)
2000	181 (0.05)	11,997 (3.3)	11 (0)	793 (0.22)	1,010 (0.28)	4,759 (1.6)	3,934 (1.25)	3,406 (1.03)	3,429 (0.99)
2001	229 (0.06)	16,662 (4.59)	23 (0.01)	1,073 (0.3)	1,456 (0.4)	6,576 (0.06)	5,234 (1.7)	4,475 (1.37)	4,452 (1.3)
2002	286 (0.08)	22,579 (6.21)	42 (0.01)	1,520 (0.42)	1,981 (0.55)	9,014 (2.48)	6,736 (2.23)	5,835 (1.82)	5,812 (1.73)
2003	359 (0.1)	30,076 (8.28)	56 (0.02)	2,147 (0.59)	2,725 (0.75)	11,503 (3.17)	8,610 (2.9)	7,284 (2.32)	7,205 (2.17)
2004	430 (0.12)	39,032 (10.75)	84 (0.02)	3,020 (0.83)	3,482 (0.96)	13,731 (3.78)	10,386 (0.09)	8,957 (2.91)	8,728 (2.68)
2005	521 (0.14)	48,246 (13.31)	137 (0.04)	4,539 (1.25)	4,286 (1.18)	15,610 (4.3)	12,105 (3.33)	10,756 (3.56)	10,281 (3.21)
2006	621 (0.17)	58,575 (16.18)	206 (0.06)	25,322 (6.97)	5,418 (1.49)	17,519 (4.82)	13,891 (3.82)	12,617 (4.25)	11,987 (3.82)
2007	703 (0.2)	69,634 (19.28)	274 (0.08)	40,833 (11.24)	6,597 (1.82)	19,230 (5.29)	15,592 (4.29)	14,475 (0.13)	13,760 (4.47)
2008	793 (0.22)	81,687 (22.69)	377 (0.1)	47,961 (13.2)	7,793 (2.14)	20,907 (5.75)	17,139 (4.72)	16,357 (4.5)	15,586 (5.16)
2009	899 (0.26)	93,530 (26.1)	601 (0.17)	53,662 (14.77)	9,187 (2.53)	22,756 (6.26)	18,826 (5.18)	18,132 (4.99)	17,662 (5.95)
2010	966 (0.28)	105,106 (29.52)	939 (0.26)	58,161 (16.02)	10,789 (2.97)	24,249 (6.67)	20,388 (5.61)	19,944 (5.49)	19,794 (0.17)
2011	1,082 (0.32)	116,949 (33.11)	1,168 (0.32)	61,461 (16.95)	12,705 (3.5)	26,027 (7.16)	21,909 (6.03)	21,642 (5.96)	21,957 (6.04)
2012	1,183 (0.35)	128,119 (36.62)	1,505 (0.42)	64,898 (17.92)	14,491 (3.99)	27,650 (7.61)	23,381 (6.43)	23,090 (6.35)	24,140 (6.64)
2013	1,268 (0.38)	138,539 (40.01)	3,168 (0.89)	68,025 (18.83)	16,580 (4.57)	28,989 (7.98)	24,627 (6.78)	24,535 (6.75)	26,282 (7.23)
2014	1,367 (0.42)	147,802 (43.25)	5,315 (1.5)	70,173 (19.49)	18,481 (5.1)	29,983 (8.25)	25,748 (7.09)	25,740 (7.08)	28,227 (7.77)
2015	1,415 (0.44)	155,632 (46.24)	7,244 (2.07)	72,267 (20.17)	20,210 (5.58)	30,587 (8.42)	26,548 (7.31)	26,552 (7.31)	29,905 (8.23)
2016	1,470 (0.47)	161,620 (48.77)	7,787 (2.25)	73,455 (20.63)	21,174 (5.86)	30,939 (8.52)	26,951 (7.42)	27,030 (7.44)	31,023 (8.54)

Table D.34: Prevalence of selected morbidities by Year

Year	Amputation	Cancer	Cognitive impairment	CKD	Dementia	HF	MI	PVD	Stroke
1984	927 (0.26)			300 (0.08)		84 (0.02)	3,196 (0.88)	10,370 (2.85)	107 (0.03)
1985	958 (0.26)			318 (0.09)		94 (0.03)	3,761 (1.03)	11,027 (3.03)	120 (0.03)
1986	992 (0.27)			330 (0.09)		107 (0.03)	4,375 (1.2)	11,744 (3.23)	149 (0.04)
1987	1,024 (0.28)			342 (0.09)		128 (0.04)	5,106 (1.41)	12,504 (3.44)	185 (0.05)
1988	1,062 (0.29)	3 (0)		359 (0.1)		167 (0.05)	5,922 (1.63)	13,320 (3.67)	229 (0.06)
1989	1,105 (0.3)	21 (0.01)		384 (0.11)	1 (0)	229 (0.06)	6,924 (1.91)	14,366 (3.95)	286 (0.08)
1990	1,154 (0.32)	93 (0.03)		413 (0.11)	4 (0)	401 (0.11)	8,280 (2.28)	16,024 (4.41)	368 (0.1)
1991	1,208 (0.33)	249 (0.07)	1 (0)	453 (0.12)	10 (0)	712 (0.2)	9,744 (2.68)	17,920 (4.93)	465 (0.13)
1992	1,259 (0.35)	497 (0.14)	3 (0)	515 (0.14)	20 (0.01)	1,254 (0.35)	11,340 (3.12)	19,813 (5.45)	636 (0.18)
1993	1,325 (0.36)	926 (0.25)	5 (0)	570 (0.16)	32 (0.01)	3,047 (0.84)	13,375 (3.68)	21,671 (5.96)	884 (0.24)
1994	1,388 (0.38)	1,513 (0.42)	7 (0)	668 (0.18)	54 (0.01)	6,835 (1.88)	15,491 (4.26)	23,661 (6.51)	1,192 (0.33)
1995	1,448 (0.4)	2,339 (0.64)	10 (0)	756 (0.21)	105 (0.03)	10,041 (2.76)	17,586 (4.84)	25,635 (7.05)	1,550 (0.43)
1996	1,511 (0.42)	3,446 (0.95)	17 (0)	876 (0.24)	194 (0.05)	12,199 (3.36)	19,701 (5.42)	27,617 (7.6)	2,065 (0.57)
1997	1,579 (0.43)	4,833 (1.33)	19 (0.01)	1,013 (0.28)	312 (0.09)	14,135 (3.89)	22,132 (6.09)	29,695 (8.18)	2,718 (0.75)
1998	1,665 (0.46)	6,480 (1.79)	21 (0.01)	1,181 (0.33)	463 (0.13)	16,287 (4.49)	24,505 (6.75)	31,839 (8.77)	3,515 (0.97)
1999	1,734 (0.48)	8,817 (2.43)	30 (0.01)	1,432 (0.39)	685 (0.19)	18,455 (5.09)	27,079 (7.47)	34,006 (9.38)	4,457 (1.23)
2000	1,826 (0.5)	11,997 (3.31)	36 (0.01)	1,691 (0.47)	1,010 (0.28)	21,214 (5.86)	30,110 (8.32)	36,472 (10.07)	5,606 (1.55)
2001	1,898 (0.53)	16,662 (4.61)	50 (0.01)	2,044 (0.57)	1,456 (0.4)	25,187 (6.97)	33,456 (9.26)	39,201 (10.85)	7,002 (1.94)
2002	1,978 (0.55)	22,579 (6.27)	69 (0.02)	2,583 (0.72)	1,981 (0.55)	29,835 (8.29)	36,793 (10.22)	42,032 (11.68)	8,698 (2.42)
2003	2,089 (0.58)	30,076 (8.39)	84 (0.02)	3,297 (0.92)	2,725 (0.76)	34,170 (9.54)	40,108 (11.19)	44,756 (12.49)	10,378 (2.9)
2004	2,176 (0.61)	39,032 (10.96)	112 (0.03)	4,251 (1.19)	3,482 (0.98)	37,744 (10.6)	42,916 (12.05)	47,457 (13.33)	12,097 (3.4)
2005	2,278 (0.64)	48,246 (13.66)	167 (0.05)	5,858 (1.66)	4,286 (1.21)	40,363 (11.43)	45,319 (12.83)	49,972 (14.15)	13,792 (3.9)
2006	2,384 (0.68)	58,575 (16.74)	239 (0.07)	27,322 (7.81)	5,418 (1.55)	42,710 (12.21)	47,426 (13.56)	52,221 (14.93)	15,577 (4.45)
2007	2,463 (0.71)	69,634 (20.11)	309 (0.09)	43,277 (12.5)	6,597 (1.91)	44,754 (12.93)	49,276 (14.23)	54,180 (15.65)	17,408 (5.03)
2008	2,533 (0.74)	81,687 (23.9)	413 (0.12)	50,597 (14.81)	7,793 (2.28)	46,428 (13.59)	50,552 (14.79)	55,895 (16.36)	19,238 (5.63)
2009	2,617 (0.78)	93,530 (27.79)	645 (0.19)	56,354 (16.74)	9,187 (2.73)	48,128 (14.3)	51,856 (15.41)	57,390 (17.05)	21,288 (6.32)
2010	2,664 (0.8)	105,106 (31.72)	994 (0.3)	60,860 (18.37)	10,789 (3.26)	49,331 (14.89)	52,885 (15.96)	58,776 (17.74)	23,393 (7.06)
2011	2,752 (0.84)	116,949 (35.87)	1,229 (0.38)	64,143 (19.68)	12,705 (3.9)	50,756 (15.57)	53,765 (16.49)	59,962 (18.39)	25,514 (7.83)
2012	2,820 (0.88)	128,119 (40.03)	1,568 (0.49)	67,533 (21.1)	14,491 (4.53)	51,917 (16.22)	54,432 (17.01)	60,675 (18.96)	27,604 (8.62)
2013	2,861 (0.91)	138,539 (44.11)	3,246 (1.03)	70,608 (22.48)	16,580 (5.28)	52,698 (16.78)	54,825 (17.46)	61,351 (19.54)	29,661 (9.44)
2014	2,919 (0.95)	147,802 (47.99)	5,399 (1.75)	72,687 (23.6)	18,481 (6)	53,136 (17.25)	55,155 (17.91)	61,786 (20.06)	31,517 (10.23)
2015	2,934 (0.97)	155,632 (51.49)	7,333 (2.43)	74,724 (24.72)	20,210 (6.69)	53,212 (17.61)	55,113 (18.23)	61,841 (20.46)	33,099 (10.95)
2016	2,953 (0.99)	161,620 (54.42)	7,877 (2.65)	75,847 (25.54)	21,174 (7.13)	53,000 (17.85)	54,760 (18.44)	61,582 (20.73)	34,132 (11.49)

Appendix E

Statistical Formulas

For the ease of comparing the composition of the study population to the United Kingdom (UK) population, the UK age group proportions were weighted. Let N_i^{65+}, T_i^{40+} and w_i be the number of people in age group 65 years plus, the total number of people aged 40 years plus and the prevalence weight in year i, respectively. The weighted proportion of people aged 65 years and above is given by,

$$\begin{aligned} prop_{65+} &= \sum_{i=1986}^{2016} w_i \cdot \frac{N_i^{65+}}{T^{40}+_i} \\ &= \sum_{i=1986}^{2016} \frac{N_i^{65+}}{\sum_{i=1986}^{2016} T_i^{40+}} \end{aligned} \tag{E.1}$$

where

$$w_i = \frac{T_i^{40+}}{\sum_{i=1986}^{2016} T_i^{40+}}.$$

The same method is used in the subsequent sections when comparing the study and the UK prevalence of various diseases and life styles.

Appendix F

Research Protocol

Ethics application – research protocol amendment

Title of the Study
Use of big health and actuarial data for understanding longevity and morbidity risks.

THIN Project Id
TL038

SRC Approval Number
16THIN095

Overview
The research programme aims to develop new methods for assessing risk and evaluating longevity based on THIN primary care database. The programme's objectives are:

- identification and quantification of the key factors affecting mortality/longevity such as lifestyle choices, medical conditions and/or interventions;
- modelling of temporal changes in the factors affecting morbidity and mortality
- evaluation of plausible scenarios in mortality trends due to particular medical advances or lifestyle changes on the population of insureds of relevance to actuarial community; and
- tools to forecast longevity risk based on realistic scenarios of uptake of various health behaviours and/or interventions, or of particular disruptions to population health.

Previous studies on acute myocardial infarction, diabetes mellitus, hip replacement, statin therapy and antihypertensive treatment have received ethics approval by the SRC. The medical conditions received approval on the 16th of June 2014 with SRC reference number 14-043 (previous THIN dataset), and the sub-studies 1 (statins) and 2 (intensive blood pressure control) received approval on the 14th of June 2017 with SRC reference number 16THIN095.

In this amendment of the above study protocol, we include the research protocol of three new sub-studies on diabetes mellitus type 2 (p3-4), hormone replacement therapy (p4-5), and stroke (p5-6). First, the common factors of these studies will be described in data source, safe storage of data, selection of data, statistical analyses at the first stage, and limitations of data. Sections.

Data source
Extract of The Health Improvement Network (THIN) data consisting of full patient records for patients born up to 1960 (inclusive) and followed up to 01.01.2017. Data are fully anonymised both in respect to patients and to practices; only year of birth is available for patients (THIN, 2013). THIN database was chosen as it is a rich data source fit for the study's purpose, with validated data quality, and it is widely used in medical and other research (Blak et al, 2011).

Safe storage of data
THIN data will be properly safeguarded to ensure confidentiality and to ensure that they cannot be accessed by unauthorised third parties. THIN data are stored on a new standalone server which can only be accessed by authorised persons using a username and password. This server and related software were specifically bought for this purpose. The hard disk and related encryption password are stored securely and separately at all times. The room where

the computer is held is locked and the building can only be accessed by coded entry (or lock and key).

Selection of data
General selection criteria include that the full medical record of a patient is available, up-to-date, and valid. This is ensured by selecting patients that (1) were registered at an active GP practice that coded death validly before the age requirement (age above 40 years for DM2 and stroke sub-studies, and above 46 for the HRT sub-study), (2) were not registered at a GP practice that is known for gaps in medical records or limited recording, (3) were registered for at least 12 months at a GP practice at the age requirement, and (4) their medical record had been accessed at least once within the last ten years before the age requirement.

Statistical analyses – first stage
Descriptive analyses are performed to check whether variables are coded correctly and assess their distribution. If needed, some categories are merged to improve statistical efficiency. Missing data are examined by checking whether subjects with missing values are different from subjects with complete medical records. If missing data is likely to be missing completely at random, unbiased estimates are obtained by doing complete cases analyses (Allison, 2001). In case of missing data being at random, multiple imputations will be applied in order to obtain unbiased estimates (idem). The models found with complete cases will function then as the basis for the imputation model. Variables that are related with the mechanism behind the missingness in the medical records will also be included in the imputation model.

Missing values in lifestyle factors are likely not to be completely at random, but associated with the health status of the subject, the frequency of visiting a GP, introduction of the Quality and Outcome Framework, and the computerisation date of the GP practice (Collins & Altman, 2010; Collins & Altman, 2012; Feary et al, 2010; Marston et al, 2010). THIN based studies executed multiple imputation when missing values were found not to be completely at random (idem).

Limitations
Large number of transfers to a different GP practice without linkage to subsequent health records of the transferred patients is a general characteristic of the THIN data. If these transfers are in any way informative for a particular condition (such as transfers to a nursing home for a degenerative disease), this may result in a biased estimate of survival. We are thoroughly investigating these issues and intend to apply competing risks models if the censoring is deemed informative to obtain robust results (Satagopan et al, 2004).

Another limitation is that there are only records on lifestyle factors if the GP deemed it important for the healthcare of the patient. Recording improved since the introduction of QOF (Langley et al, 2011; Taggar et al, 2012). Despite this, recordings on lifestyle factors are related to health and frequency of visits to the GP (Feary et al, 2010). This bias could be overcome by multiple imputation. The last limitation is that lab results before the computerisation of medical records can have abnormal readings due to typographical errors (Wijlaars, 2013b). Computerisation in each GP practice took place between 1988 and 2007.

Sub-study 3: Diabetes mellitus type 2

Background

The world has an ageing population where chronic medical conditions become increasingly more common. Deaths caused by non-communicable diseases (NCD) have been reported to be on the increase and NCDs in 2015 constituted 70% of the top ten cause of deaths worldwide (WHO, 2017). Of interest is diabetes mellitus II (DM-II), which constitute 90% of diabetic patients in the UK and was reported to be on the increase, ranking number 6 on the cause of death in 2015 and killing 1.6 million worldwide (WHO, 2017). In the UK, deaths caused by DM-II have been increasing on an exponential rate as from 2012 (ONS, 2017). According to the International Diabetes Federation (IDF) in 2015, the UK was the 9th country with the highest medical expenditure towards DM-II ($13 billion), though it was not among the top ten countries with the highest numbers of DM-II patients.

The two main streams of diabetes-related morbidities are macrovascular and microvascular complications. Macrovascular complications include: stroke (cerebrovascular disease CVD), coronary heart disease (CHD), and peripheral vascular disease (PVD). 50% of people with diabetes die of cardiovascular disease (primarily heart disease and stroke). Microvascular complications include: neuropathy (nerve damage, including diabetic foot), nephropathy (kidney disease), and retinopathy (eye disease). Other comorbidities are: impairment of immune system (T1DM), and periodontal and foot diseases.

Factors that influence the survival of diabetics are based on the studies listed below in the references list. General factors that influence survival of diabetics are: age, gender, socioeconomic status, smoking, alcohol use, and obesity. Comorbidities that have an effect on survival are: hypertension, dyslipidaemia, heart disease, chronic kidney disease, and cardiovascular disease. Complications of diabetes include retinopathy, neuropathy, nephropathy, amputations, and cancer. Treatments that modify survival are: insulin, metformin, sulfonylureas, GLP-1 agonists, bydureon, DPP-4 inhibitors, and thiazolidinediones, as well as treatments for hypertension and hypocholesteraemia, and adherence to treatments.

Research objective

The main objective of this study is to estimate associations between the first diagnosis of diabetes type 2 and mortality and morbidity in the general population treated in routine clinical practice in the UK, compared to no diagnosis.

Study design

This is a retrospective cohort study from 1986 to 2017. The start date of the study period is based on the first recorded diagnosis of diabetes type 2 in eligible cases with a good quality medical record (see ‘Selection of data’ on page 2). Patients who were diagnosed with diabetes for the first time and aged 40 or above at diagnosis, will be selected chronologically (i.e. starting from 1984) and matched to three controls (i.e. patients who were not diagnosed with diabetes by that date) on sex, year of birth, and general practice. Patients will be excluded if prior to the selection date they were diagnosed with cancer, chronic kidney disease (CKD) stages 3-5, dementia, heart failure, myocardial infarction, peripheral vascular diseases, or stroke.

Study variables
The primary outcome is time to all-cause mortality. The primary exposure is the first diagnosis of diabetes type 2. The secondary outcomes are the major comorbidities developed during the follow-up that may be associated with diabetes (CVD, CKD, cancer, cognitive impairment, dementia, retinopathy, neuropathy, nephropathy, and amputation). The secondary outcomes will be of interest for the group of matched cases and controls who did not have such an outcome prior to the start date. The full list of relevant confounding variables will be established from medical literature such as systematic reviews, and from expert knowledge within the team. The list will include co-morbidities, treatment, lifestyle factors, and demographical factors. Co-morbidities include: hypertension, dyslipidaemia, heart disease, chronic kidney disease, cardiovascular disease, coronary artery disease, cardiovascular system conditions, osteoporosis, and timing of presence of these medical conditions (including during follow-up, resulting in time-dependent variables). Treatments includes: anti-diabetic drugs, blood pressure and lipid-regulating drugs. Lifestyle factors include: BMI, alcohol consumption, and smoking. The demographical factors include: sex, year of birth, and socio-economic status as measured by Mosaic and by IMD deciles, and other postcode variables. All the confounding variables represent the latest reading before entering the study.

Statistical analysis
Cox's proportional hazards regression models for all-cause mortality and secondary outcomes for cases and controls will be fitted. The simple Cox's, competing risks, and landmark models are considered. A shared frailty term on GP practice will be specified to take into account the hierarchical structure of the data (Brown & Prescott, 2006). The final models are to be obtained through backward elimination, where the full models include order two interactions with the main exposure, sex, and age. The final models will then be assessed with regards to the assumptions, influential observations, goodness-of-fit, and accuracy of prediction by means of cross-validation.

Sub-study 4: A retrospective cohort study of hormone replacement therapy effect on longevity and morbidity.

Background
Most women suffer from menopausal symptoms when they reach their 50s because of the deficiency of female sex hormones, oestrogen and progesterone. Postmenopausal women are more prone than men to develop many life-threatening conditions, such as dementia, osteoporosis, cardiovascular disease, and gynaecological cancer (Lobo, R. A., et al.). These diseases are also the major cause of mortality in women in the United Kingdom (UK). In the UK, approximately 1.5 million women suffer from various menopausal symptoms each year (National Institute for health and care excellence, 2015). Hormone replacement therapy (HRT) has been used for more than sixty years to relieve women from troublesome menopausal symptoms (Anderson et al., 2007). However, the use of HRT to treat menopausal symptoms remains controversial as the outcomes from the therapy are not always positive. Numerous studies have been carried out to investigate the impact of HRT in women (Adami et al, 1989; Anderson et al., 1998; Ettinger et al., 1996; Emily et al., 2003; Freeman et al. 2007; Henderson et al., 2014). But the results are contradictory and varied widely depending on the region and population. Because of the fear of negative health effects, many postmenopausal women do not choose the therapy or stop treatment prematurely. If the pros and cons of HRT are clearly known, then it may be possible to provide a better treatment and

quality of life for women after menopause. Therefore, further research is necessary to understand the actual risk factors that affect the outcomes of the therapy.

Research objective
The main objective of this study is to estimate associations between the first prescription of HRT with mortality and morbidity in the general population treated in routine clinical practice in the UK, compared to no prescription.

Study design
This is a retrospective cohort study from 1986 to 2017. The start date of the study period is based on the first recorded start of HRT in eligible cases with a good quality medical record (see 'Selection of data' on page 2). Eligible cases are women aged 46 years or older, who received any kind of oral or transdermal HRT for the first time. Patients who started HRT during the study period will be selected chronologically (i.e. starting from 1984) and matched to three controls (i.e. patients who had not started HRT by that date) on year of birth and general practice. Women will be excluded if prior to the selection date they were diagnosed with acute myocardial infarction, cancer, chronic kidney disease stage 3-5, heart failure, oophorectomy before age 45, premature ovarian insufficiency, or premature menopause.

Study variables
The primary outcome is time to all-cause mortality. The primary exposure is the first prescription of any kind of HRT drugs containing oestrogen and/or progesterone except local HRT. The secondary outcomes are the major comorbidities developed during the follow-up that may be positively or negatively associated with HRT (CVD, CKD, cancer, diabetes, cognitive impairment, dementia, osteoporosis). The secondary outcomes will be of interest for the group of matched cases and controls who did not have such an outcome prior to the start date. The full list of relevant confounding variables will be established from medical literature such as systematic reviews, and from expert knowledge within the team. The list will include co-morbidities, treatment, lifestyle factors, and demographical factors. Comorbidities include osteoporosis, cardiovascular disease, diabetes, hypertension, hypercholesterolemia, and timing of presence of these medical conditions (including during follow-up, resulting in time-dependent variables). Treatments include HRT types/modes of administration, statins and antihypertensive drugs. Lifestyle factors include BMI, alcohol consumption, and smoking. The demographical factors include year of birth, socio-economic status as measured by Mosaic and by IMD deciles, and other postcode variables. All the confounding variables represent the latest reading before entering the study.

Statistical analysis
Cox's proportional hazards regression models for all-cause mortality and secondary outcomes for cases and controls will be fitted. The simple Cox's, competing risks, and landmark models are considered. A shared frailty term on GP practice will be specified to take into account the hierarchical structure of the data (Brown & Prescott, 2006). The final models are to be obtained through backward elimination, where the full models include order two interactions with the main exposure, and age. The final models will then be assessed with regards to the assumptions, influential observations, goodness-of-fit, and accuracy of prediction by means of cross-validation.

Sub-study 5: A retrospective cohort study of the effect of the first diagnosis of Ischemic stroke and TIA on longevity and morbidity.

Background

Stroke is a severe and wide-spread disease which is the second-most cause of death worldwide. Stroke can be subdivided into an ischaemic stroke, IS and haemorrhagic stroke, HS. About 85% of all strokes are ischaemic and 15% haemorrhagic. Transient Ischaemic Attack (TIA), often referred as "mini-stroke" is regarded as a warning sign for future strokes. For the purpose of this study, TIA and IS stroke types are considered.

A stroke can be life-threatening condition and may also cause long-term problems and disability which can necessitate rehabilitation and further care. Many of these stroke survivors experience significant and long-term physical and psychological impacts, repeat strokes, transient ischaemic attacks (TIAs) and/or death within a year of stroke. Stroke is the third leading causes of disability-adjusted life years lost, DALYS (Feigin et al., 2009). It is no longer regarded as the "disease of the old age" due to the worrying trend of younger people being affected by it. The trend is expected to double by 2030 unless proper strategies are devised. The economic burden of stroke in UK is estimated to be around £9 billion a year (Saka et al., 2009). The annual NHS costs of stroke are estimated to hit £10.2 billion in 2035. Despite significant progress in prevention, treatment and rehabilitation, there is still great capacity for further improvements, which in turn could reduce these large economic burdens. The risk of premature death and disability is quite high among stroke survivors. It is of utmost importance to gain insight on risk factors, trends in incidence and diagnosis after stroke in an effort to reduce the risk of mortality and morbidity and consequently, improve the outcomes.

Research objective

The main objective of this study is to estimate associations between the first diagnosis of TIA or IS with mortality and morbidity in the general population treated in routine clinical practice in the UK, compared to no diagnosis.

Study design

This is a retrospective cohort study from 1986 to 2017. The start date of the study period is based on the first recorded diagnosis of TIA or IS in eligible cases with a good quality medical record (see 'Selection of data' on page 2). Patients who were diagnosed with TIA or IS for the first time and aged 40 or above at diagnosis, will be selected chronologically (i.e. starting from 1984) and matched to three controls (i.e. patients who were not diagnosed with TIA or IS by that date) on sex, year of birth, and general practice. Patients with prior acute myocardial infraction, cancer, chronic kidney disease stages 3-5, dementia, heart failure, or any other types of stroke were excluded.

Study variables

The primary outcome is time to all-cause mortality. The primary exposure is the first diagnosis of TIA or IS. The secondary outcomes are the major comorbidities developed during the follow-up that may be associated with stroke/TIA (further strokes, cognitive impairment, dementia, CVD, pneumonias, venous thromboembolism, dysphagia, incontinence, and depression). The secondary outcomes will be of interest for the group of matched cases and controls who did not have such an outcome prior to the start date. The full

list of relevant confounding variables will be established from medical literature such as systematic reviews, and from expert knowledge within the team. The list will include co-morbidities, treatments, lifestyle factors, and demographical factors. Comorbidities include asthma, atrial fibrillation, congenital heart defects, chronic obstructive pulmonary disease, diabetes, hypothyroidism, peripheral arterial disease (PAD), hypertension, hypercholesterolemia, and timing of presence of these medical conditions (including during follow-up, resulting in time-dependent variables). Treatments include antiplatelet drugs, anticoagulants, statins and anti-hypertensive drugs. Lifestyle factors include BMI, alcohol consumption, and smoking. The demographical factors include sex, year of birth, socio-economic status as measured by Mosaic and by IMD deciles, and other postcode variables. All the confounding variables represent the latest reading before entering the study.

Statistical analysis

Cox's proportional hazards regression models for all-cause mortality and secondary outcomes for cases and controls will be fitted. The simple Cox's, competing risks, and landmark models are considered. A shared frailty term on GP practice will be specified to take into account the hierarchical structure of the data (Brown & Prescott, 2006). The final models are to be obtained through backward elimination, where the full models include order two interactions with the main exposure, sex, and age. The final models will then be assessed with regards to the assumptions, influential observations, goodness-of-fit, and accuracy of prediction by means of cross-validation.

References

Adami, Hans-Olov, et al. "Risk of cancer in women receiving hormone replacement therapy." International journal of cancer44.5 (1989): 833-839.

Afkarian M, Sachs MC, Kestenbaum B, Hirsch IB, Tuttle KR, Himmelfarb J, de Boer IH Kidney disease and increased mortality risk in type 2 diabetes. J Am Soc Nephrol. 2013 Feb;24(2):302-8. doi: 10.1681/ASN.2012070718. Epub 2013 Jan 29.

Anderson, G., Cummings, S., Freedman, L., Furberg, C., Henderson, M., Johnson, S., Kuller, L., Manson, J., Oberman, A., Prentice, R., et al. (1998). Design of the women's health initiative clinical trial and observational study. Controlled clinical trials, 19(1):61-109.

Cristian BAICUS Update in Internal Medicine: Intensive Glucose Lowering Does Not Reduce Mortality in Type 2 Diabetes Mellitus. Maedica (Buchar). 2012 January; 7(1): 87.

Barr, E.L. et.al. (2008). Mortality and life expectancy associated with diabetes. The Epidemiology of Diabetes Mellitus, Second Edition, pages 603-625.

Blak, B. et. al. (2011). Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Journal of Innovation in Health Informatics, 19(4):251-255.

Boussageon R, Bejan-Angoulvant T, Saadatian-Elahi M, et al. Effect of intensive glucose lowering treatment on all cause mortality, cardiovascular death, and microvascular events in type 2 diabetes: meta-analysis of randomised controlled trials. BMJ. 2011;343:d4169–d4169.

Clarke P., Gray A., Briggs A., Farmer A., Fenn P., Stevens R., Matthews D., Stratton I. and Holman R. (2004) **A model to estimate the lifetime health outcomes of patients with Type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68)**. *Diabetologia*. **47**. 1747-1759. http://dx.doi.org/10.1007/s00125-004-1527-z

Currie CJ, Poole CD, Evans M, Peters JR, Morgan CL. Mortality and other important diabetes-related outcomes with insulin vs other antihyperglycemic therapies in type 2 diabetes. J Clin Endocrinol Metab. 2013 Feb;98(2):668-77.

Currie, C. J. et. al. (2010). Survival as a function of hba1c in people with type 2 diabetes: a retrospective cohort study. The Lancet, 375(9713):481-489.

Emily et al., v. (2003). Breast cancer and hormone-replacement therapy in the million women study. The Lancet, 362(9382):419–427.

Ekström N., Miftaraj M., Svensson A. M., Andersson Sundell K., Cederholm J., Zethelius B., Gudbjörnsdottir S. and Eliasson B. (2012) **Glucose-lowering treatment and clinical results in 163 121 patients with type 2 diabetes: an observational study from the Swedish national diabetes register**. *Diabetes, Obesity and Metabolism*. 14. 717-726.

Ettinger, Bruce, et al. "Reduced mortality associated with long-term postmenopausal estrogen therapy." Obstetrics & Gynecology 87.1 (1996): 6-12.

Feary, J., Rodrigues, L., Smith, C., Hubbard, R.B., and Gibson, J.E. (2010). Prevalence of major comorbidities in subjects with COPD and incidence of myocardial infarction and stroke: a comprehensive analysis using data from primary care. *Thorax,* 65(11): 956-62.

Feigin, V.L., Lawes, C.M., Bennett, D.A., Barker-Collo, S.L. and Parag, V., 2009. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. The Lancet Neurology, 8(4), pp.355-369.

Freeman, E. and Sherif, K. (2007). Prevalence of hot flushes and night sweats around the world: a systematic review. Climacteric, 10(3):197–214.

Gaede Peter, Lund-Andersen Henrik, Parving Hans-Henrik and Pedersen Oluf (2008). **Effect of a Multifactorial Intervention on Mortality in Type 2 Diabetes**. *New England Journal of Medicine*. 358. 580-591. http://www.nejm.org/doi/full/10.1056/NEJMoa0706245

Grant RW, Wexler DJ Lower blood pressure associated with higher mortality in retrospective study of patients with newly diagnosed type 2 diabetes. Evid Based Med. 2012 Nov 20.

Hall, G. C. (2009). Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiology and drug safety, 18(2):120-131.

Hansen M., Jensen M. and Carstensen B. (2011). Causes of death among diabetic patients in Denmark. *Diabetologia*. 1-9.

Harjutsalo V, Forsblom C, Groop PH. Time trends in mortality in patients with type 1 diabetes: nationwide population based cohort study. BMJ. 2011 Sep 8;343:d5364. doi: 10.1136/bmj.d5364.

Hemmingsen Bianca, Lund Søren S, Gluud Christian, Vaag Allan, Almdal Thomas, Hemmingsen Christina and Wetterslev Jørn. (2011) Intensive glycaemic control for patients with type 2 diabetes: systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. ***British Medical Journal***. **343**.

Henderson, Victor W. "Alzheimer's disease: review of hormone therapy trials and implications for treatment and prevention after menopause." The Journal of steroid biochemistry and molecular biology 142 (2014): 99-106.

Hex N., Bartlett C., Wright D., Taylor M. and Varley D. 2012 Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabetic Medicine*. 29. 855-862.

Huang Elbert S., Liu Jennifer Y., Moffet Howard H., John Priya M. and Karter Andrew J. (2011) Glycemic Control, Complications, and Death in Older Diabetic Patients. ***Diabetes Care***. 34. 1329-1336. http://care.diabetesjournals.org/content/34/6/1329.abstract

Kontopantelis, E. et. al. (2015). Glucose, blood pressure and cholesterol levels and their relationships to clinical outcomes in type 2 diabetes: a retrospective cohort study. Diabetologia, 58(3):505-518.

Kramer Caroline K., Rodrigues Ticiana C., Canani Luis H., Gross Jorge L. and Azevedo Mirela J. (2011) **Diabetic Retinopathy Predicts All-Cause Mortality and Cardiovascular Events in Both Type 1 and 2 Diabetes.** *Diabetes Care*. **34**. 1238-1244.

Langley, et. al. (2011). Can primary care data be used to monitor regional smoking prevalence? An analysis of The Health Improvement Network primary care data. BMC public health, 11(1):773.

Landman GW, Kleefstra N, van Hateren KJ, Gans RO, Bilo HJ, Groenier KH Educational disparities in mortality among patients with type 2 diabetes in The Netherlands (ZODIAC-23). Neth J Med. 2013 Mar;71(2):76-80.

Lewis, J. D. et. al. (2007). Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiology and drug safety, 16(4):393-401.

Lobo, R. A., et al. "Prevention of diseases after menopause." Climacteric 17.5 (2014): 540-556.

McEwen, Kim, Karter, Haan, Ghosh, Lantz, Mangione, Thompson, Herman (2007). Risk factors for mortality among patients with diabetes. *Diabetes Care*, 30:1736-41
Rotenstein Lisa S., Kozak Benjamin M., Shivers Joseph P., Yarchoan Mark, Close John and Close Kelly L. (2012) **The Ideal Diabetes Therapy: What Will It Look Like? How Close Are We?** *Clinical Diabetes*. **30**. 44-53.

Mulnier, H. et. al. (2006). Mortality in people with type 2 diabetes in the UK. Diabetic Medicine, 23(5):516-521.

NICE (2015). Putting NICE guidence into practice. https://www.nice.org.uk/guidance/ng23/resources/costing-report-pdf-556376653. Accessed: 2018-03-06.

ONS (2017). Deaths registered in England and Wales (series DR). www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/. Accessed: 22/02/2018.

Rothwell, P.M., Coull, A.J., Giles, M.F., Howard, S.C., Silver, L.E., Bull, L.M., Gutnikov, S.A., Edwards, P., Mant, D., Sackley, C.M. and Farmer, A., 2004. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (Oxford Vascular Study). The Lancet, 363(9425), pp.1925-1933.

Saka, Ö., McGuire, A. and Wolfe, C., 2009. Cost of stroke in the United Kingdom. Age and ageing, 38(1), pp.27-32.

Satagopan, J.M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. and Auerbach, A.D., 2004. A note on competing risks in survival data analysis. British journal of cancer, 91(7), p.1229.

Sluik D, Boeing H, Bergmann MM, Schütze M, Teucher B, Kaaks R, Tjønneland A, Overvad K, Arriola L, Ardanaz E, Bendinelli B, Agnoli C, Tumino R, Ricceri F, Mattiello A, Spijkerman AM, Beulens JW, Grobbee DE, Nilsson PM, Melander O, Franks PW, Rolandsson O, Riboli E, Gallo V, Romaguera D, Nöthlings U.

Soriano, L. C. et. al. (2015). Cardiovascular events and all-cause mortality in a cohort of 57,946 patients with type 2 diabetes: associations with renal function and cardiovascular risk factors. Cardiovascular Diabetology, 14(1):38.

Taggar, J. et. al. (2012). The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database. BMC public health, 12(1): 329.

Taylor KS, Heneghan CJ, Farmer AJ, Fuller AM, Adler AI, Aronson JK, Stevens RJ All-Cause and Cardiovascular Mortality in Middle-Aged People With Type 2 Diabetes Compared With People Without Diabetes in a Large U.K. Primary Care Database. Diabetes Care. 2013 Feb 22.

The Health Improvement Network (2013). London: The health improvement network. Retrieved from http://www.thin-uk.com

BRIAN J. WELLS, ANIL JAIN, SUSANA ARRIGAIN, CHANGHONG YU, WAYNE A. ROSENKRANS, JR, MICHAEL W. KATTAN. Predicting 6-Year Mortality Risk in Patients With Type 2 Diabetes. DIABETES CARE, VOLUME 31, NUMBER 12, DECEMBER 2008

WHO (2017). Fact sheets 310. www.who.int/mediacentre/factsheets/fs310/en/. Accessed: 03/11/2017.

Wijlaars, L.P., Nazareth, I., Whitaker, H.J., Evans, S.J. and Petersen, I., 2013. Suicide-related events in young people following prescription of SSRIs and other antidepressants: a self-controlled case series analysis. *BMJ open*, *3*(9), p.e003247.