Cardiovascular disease and its impact on longevity and longevity improvement

Lisanne Andra Gitsels

Doctor of Philosophy

June 2017

Cardiovascular disease and its impact on longevity and longevity improvement

Lisanne Andra Gitsels

Doctor of Philosophy
University of East Anglia
School of Computing Sciences
June 2017

Abstract

An increased risk or a history of cardiovascular disease (CVD) is associated with worse survival prospects. Clinical guidelines recommend several treatments for primary and secondary prevention. These guidelines are mainly based on clinical trials and hospital data. Data from routine clinical practice could provide insights in longevity and longevity improvement in the general population as opposed to selected patients.

The primary objectives of this research were to investigate how a history of CVD affects longevity in residents of the United Kingdom at retirement age, and to investigate which treatments improve longevity.

Medical records from 1987 to 2011 from general practices contributing to The Health Improvement Network (THIN) database were used to develop two specific survival models: to estimate the hazards of all-cause mortality associated with a history of acute myocardial infarction (AMI) and related treatments, and to estimate the hazard of all-cause mortality associated with statins prescribed as primary prevention of CVD. The models were multilevel Cox's proportional hazards regressions that included comorbidities, treatments, lifestyle choices, and socio-demographic factors. The models were specified for ages 60, 65, 70, and 75. More accurate estimates of longevity at these key ages could inform future medical management by clinicians and financial planning for retirement by individuals, actuaries, and the government.

This research found that survival prospects after AMI were reduced by less than previous studies have reported. Furthermore, currently recommended treatments for CVD were associated with mixed survival prospects, in which coronary revascularisation and prescription of beta blockers and statins were associated with improved prospects and prescription of ACE inhibitors and aspirin were associated with worsened prospects.

Table of Contents

List of Tables

List of Figures

List of Publications

- Gitsels, L. A., Kulinskaya, E., and Steel, N. (2016). Survival benefits of statins for primary prevention: a cohort study. *PloS One*, 11(11):e0166847.
- Gitsels, L. A., Kulinskaya, E., and Steel, N. (2017). Survival prospects after acute myocardial infarction in the UK: a matched cohort study 1987-2011. *BMJ Open*, 6:e013570.
- Kulinskaya, E. and Gitsels, L. A. (2016). Use of big health and actuarial data for understanding longevity and morbidity risk. *Longevity Bulletin*, (9):1518.

Survival

Acknowledgements

I would like to thank my supervisors Prof Elena Kulinskaya, Prof Nicholas Steel, and Mr Nigel Wright for their time and support. In particular, I would like to thank Elena for guiding me to become an independent academic researcher. I appreciate her vast knowledge in a wide range of fields and her assistance in all forms of writing, from conference abstracts to academic papers to this thesis. I would like to thank Nicholas for sharing his medical expertise and for his guidance in presenting and communicating my work to a medical public in a clear, concise, and effective way. I would like to thank Nigel for sharing his actuarial expertise.

I would also like to thank my former epidemiology classmates Maria Tran and Nisha Rajendran for discussing and proofreading parts of my work throughout my PhD project. Their feedback helped me improve explaining my work.

Finally, I would like to thank my mother Dr Janneke Gitsels-van der Wal for introducing me to the academic life. Her prediction that I would become a statistician has hereby come true. It was a great, unique experience to work together on conference posters and an academic paper in midwifery science, which were my first peer-reviewed publications.

Chapter 1

Introduction

This Chapter starts by explaining the rationale for developing survival models to estimate longevity given a history of cardiovascular disease and related treatments in people at retirement age. Next, the research objectives and aims are listed. Then, the contributions of the research are presented. Finally, the thesis outline is provided.

1.1 Rationale

Cardiovascular disease (CVD), which is an umbrella term for diseases of the heart and circulation (Townsend et al., 2015), is one of the main causes of death in the world (Naghavi et al., 2015; Newton et al., 2015; WHO, 2015a). In the United Kingdom (UK), CVD is the number two cause of death for men and women, accounting for 28% and 26% of deaths, respectively (Townsend et al., 2015). Longevity prospects of a person can inter alia be explained by comorbidities, treatments, lifestyle choices, and socio-demographic factors (WHO, 2015a). Longevity can be estimated by a survival model, which ideally consists of all risk factors that can explain the variations in the outcome. Precise estimates of longevity prospects and understanding variations in longevity prospects are important to many parties.

1.1.1 Relevance of survival models in medicine

For instance, survival models are of interest to clinicians, because they can identify specific patient characteristics associated with different survival rates. These findings can be used to counteract the harmful effects and enhance the protective effects of modifiable risk factors. For example, cholesterol level and blood pressure can be targeted to improve survival prospects (NICE, 2011, 2015). This could be in the form of routine screening, early intervention, or patient education programmes that could help patients better understand the risks associated with certain lifestyles and how these risks can be lowered by changing their lifestyle.

With an ageing population and medical advances improving survival prospects, chronic medical conditions like CVD become increasingly prevalent (Naghavi et al., 2015; WHO, 2015c). With a higher prevalence of these medical conditions, survival variations can be analysed in greater detail. In other words, a higher degree of differentiation between patients is possible, in which interactions between medical conditions, treatments, lifestyle choices, and socio-demographic factors can also be studied. A survival model that is estimated on a heterogeneous sample of patients, could lead to pharmacosurveillance in which the safety and effectiveness of treatments in groups of patients can be assessed (Platt et al., 2008). The effect of treatments might differ by sex, age, or other clinically defined subpopulations (Hippisley-Cox and Coupland, 2010a,c). By detecting the differences in effectiveness, clinicians can provide custom tailored care for the patient that helps improve the respective survival prospects. Subsequently, survival models can provide risk thresholds for action and updates for clinical guidelines of prevention and risk management (Wright and Dent, 2014).

Survival models can not only be used for individual risk assessment but also to assess the well-being of an entire population. This in turn can inform resource allocation decisions for optimal positive net benefits (Hingorani et al., 2013). The increase in prevalence of chronic medical conditions can put pressure on resources such as general practitioners, specialised doctors, surgeons, caretakers, medical centres, medical equipment, drugs, money, and time. The higher prevalence of chronic medical conditions can also increase the experience of doctors and surgeons, and lead to a greater variation in performance of health care and greater difference in survival prospects by medical centre. With more precise estimates of survival prospects and a greater understanding of survival variations among patients and medical centres, resources can be allocated in a strategic way.

Thus, medical professionals and local health authorities can benefit from survival models because the results can inform the shape of future medical management and strategic resource allocation.

1.1.2 Relevance of survival models in retirement planning

Survival models are of interest in retirement planning, because they can inform individuals about how to spend their pension pot during retirement, inform actuaries about pricing of annuities and life insurance, and inform governments about taxation, national insurance rates, and pensions.

Individuals

It is recommended to plan one's finances for retirement to ensure there is enough income to live off during retirement. The key information in financial planning for retirement is one's life expectancy, because this will inform the individual how much income can be spent per year. In the UK, sources of income could be the state pension, occupational pension, personal pension, defined benefit (DB) pension, or defined contribution (DC) pension (Office for National Statistics, 2013). Access to the pension pot is granted when a person reaches the minimum pension age of 55,

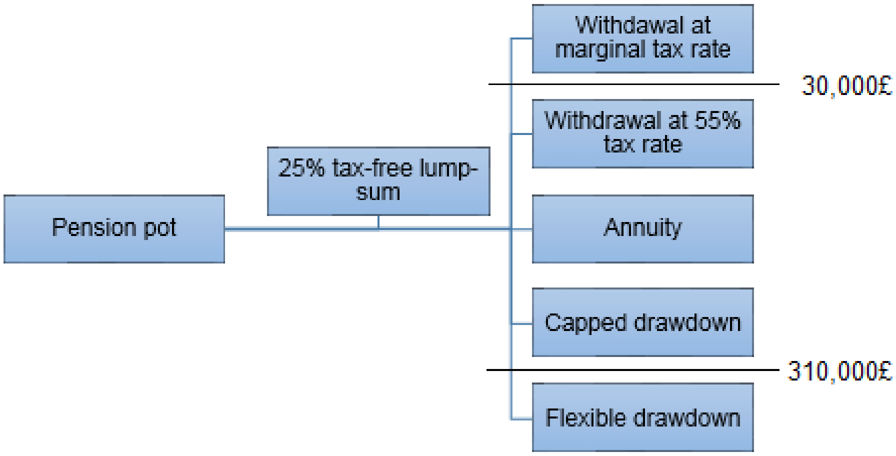


Figure 1.1: Former British pension system (2011-2015)
People who reached the minimum pension age of 55 could take out a 25% tax-free lump sum from their pension pot, after which they had the following options for spending their pension pot: withdrawal of first 30,000£ at marginal tax rate, withdrawal after first 30,000£ at 55% tax rate, purchase of annuity, capped drawdown, and flexible drawdown after first 310,000£ (Baxter, 2015a). This system pushed people in purchasing an annuity unless they had a very small or large pension pot.

which will rise to age 57 in 2018 (Baxter, 2015a). In April 2015, the laws regarding how and when the pension pot can be spent were reformed (Baxter, 2015b). Under the former pension system of the UK, which was active from April 2011 to April 2015, people were effectively encouraged to buy an annuity if their pension pot was worth between 30,000 and 310,000£, see Figure 1.1. As the majority of people had a pension pot of this size, 75% of people ended up with an annuity (HM Treasury, 2014).

In April 2015, the British government modified the rules on pensions to offer greater freedom to individuals in choosing how and when to access their pension pots during retirement. The reasoning behind this change was that annuities no longer suited everyone due to increasing life expectancy and diverse wishes for retirement (HM Treasury, 2014). Fifty years ago a 65-year old had a life expectancy of 12

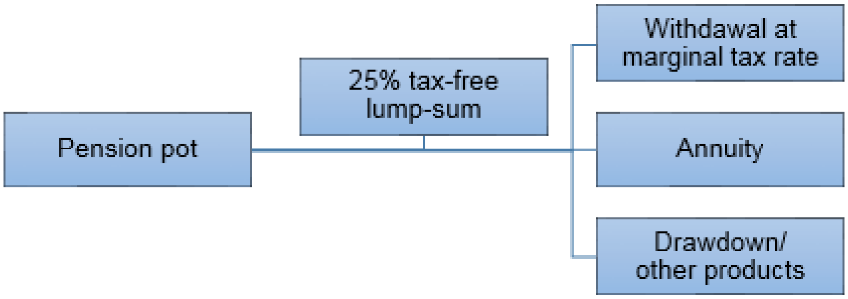


Figure 1.2: Current British pension system (since April 2015)
People who reached the minimum pension age of 55 can take out a 25% tax-free lump sum from their pension pot, after which they had the following options for spending their pension pot: withdrawal at marginal tax rate, purchase of annuity, drawdown, and purchase of other products created by providers (Baxter, 2015a,b).

years, whereas it is 21 years today. Furthermore, while previously retired people were relatively inactive, nowadays they are more active, often having part-time jobs at the start of their retirement. With increasing life expectancy and a varied lifestyle during retirement, purchasing a range of products instead of one product might be more suitable (Baxter, 2015a). With the current pension system, everyone has the option of a 25% tax-free lump-sum, withdrawal of money, drawdowns of money, purchase of an annuity, and the purchase of other products created by providers, see Figure 1.2.

Under the former and current pension system, buying an annuity was and is a fixed and one-time purchase. In April 2017, the pension system will be reformed to permit second-hand annuities (Baxter, 2015a). It is expected that only a minority of people would like to sell their annuity. An example of when it would be attractive to sell an annuity and receive a lump sum is when life expectancy has declined due to an unexpected unfavourable event that happened after the annuity was bought. Second-hand annuities will be subject to adverse selection in which sellers have the advantage over buyers. This is because sellers would have better knowledge about their own life expectancy due to complete information on known medical history, lifestyle choices, and socio-demographic factors (Baxter, 2015a).

Thus, the change in pension system and the future reforms of second-hand annuities requires more active decision-making by individuals regarding how to spend their pension pot (Baxter, 2016). Survival models can inform individuals how certain medical conditions, treatments, lifestyle choices, and socio-demographic factors affect their survival prospects at different retirement ages.

Actuaries

Survival models are of interest to actuaries because they can provide insights on survival variations and are thereby informative for the pricing of annuities and life insurance. In estimating life expectancy, actuaries deal with basis and longevity risk (Barrieu et al., 2012). The basis risk is here defined as life expectancy being incorrectly estimated, i.e. that there is a residual between estimation and observation. When life expectancy is over-predicted, the insurance company gains profit. This is because the annuity was specified to provide an income for more years than expected and when the client has passed away, the insurance company can keep the money that was left over. However, when life expectancy is under-predicted, the insurance company loses money. This is because the annuity was specified to provide an income for fewer years than expected and since an annuity is a guaranteed income for life, the insurance company has to continue providing an income until death. It is therefore of great importance to estimate life expectancy as accurately as possible to minimise the basis risk.

Longevity risk is defined as the risk of an unexpected increase in life expectancy, due perhaps to changing lifestyles in the population and to medical advances (Barrieu et al., 2012). Prior to modelling survival data, the baseline characteristics of the sample are examined. During this stage, time trends in incidence and prevalence of risk factors can be identified. Examples include an obesity epidemic and a rise in drug prescriptions (Hardoon et al., 2011; WHO, 2015b). Survival models in turn can test whether the hazardous or protective survival effects of lifestyle choices, treatments, or other factors change over time. The results can inform the insurance company which risk factors might be a longevity risk and should be taken into account when predicting life expectancy.

There are different types of annuities that focus on various details of the client. For example, enhanced annuities specialise in poor health status or lifestyle choices (Thurley, 2015). Survival models can provide estimates of the hazards associated with certain medical conditions, treatments, and lifestyle choices. By having more information about the health status and lifestyle of the client, life expectancy can be more accurately estimated. This is beneficial for the client because the client would receive a higher retirement income per year than when the information is not available and the life expectancy of the average, healthier person is used in calculating the income. Greater accuracy in estimating life expectancy is also beneficial for the insurance company because it minimises the basis risk.

With increasing life expectancy and increasing years of retirement, annuities need to cover more years (HM Treasury, 2014). This means that there is more uncertainty to take into account, and therefore the estimated life expectancy is less precise and the basis risk is greater. Survival models developed on long follow-up data of a heterogeneous sample can identify new risk factors or combinations of risk factors that explain survival variations to a higher degree and lead to better differentiation between groups of people and their respective survival prospects at the baseline.

Thus, actuaries can benefit from survival models because the results can provide insights into the basis and longevity risks of estimating life expectancy, and in turn can lead to better pricing of annuities and other insurance products.

Government

Survival models are also of interest to the government as the estimates can inform decisions related to the tax schemes, national insurance rates, and pension system. The UK has an ageing population, which means that there is an increase in the dependency of retired people on the workforce (UK Parliament, 2015). This puts pressure on welfare spending while less revenue is collected. It is therefore of importance to identify and forecast demographic and health trends in the population and to understand survival variations within the population in order to sustain the economy.

Survival models can identify age-specific risk factors of ill-health and mortality. The results can be informative for predicting healthy life expectancy and total life expectancy. Differentiating between the two can be indicative of when people are likely to retire and the length of their retirement. This in turn can inform the expected participation in the workforce at each age, what a reasonable minimum retirement age is for the population, and the expected dependency by retired people on the state.

There are great variations in healthy and total life expectancy. For example, between the least and most deprived areas in the UK, there is almost 17 years difference in healthy life expectancy for both men and women, and there is 8 and 6 years difference in total life expectancy for men and women, respectively (White and Butt, 2015). Survival models can identify specific profiles associated with different life expectancies. Furthermore, these profiles would help elucidate which modifiable risk factors to target in order to improve the well-being of the population. The results could give rise to or enhance the promotion of a healthy lifestyle, provision of preventative healthcare, provision of services to overcome addictions, taxation of unhealthy goods, and allocation of medical resources to the ones most in need or who would benefit the most. A healthier population means that more people can be part of the workforce and be part of it for a longer time period. Better distribution of governmental funds, guided by the estimates of quantitative methods such as survival models, would release pressure on welfare spending, increase tax revenues, and increase pension savings (UK Parliament, 2015).

Thus, the government can benefit from survival models, because the results can assess the well-being of the population and thereby inform the shape of future health policies, tax schemes, national insurance rates, and the pension system.

1.1.3 Existing cardiovascular disease survival models

Numerous survival models for CVD have been developed (NICE, 2013a). In the past, health scientists typically performed patient-level, incidence-based data analysis. In contrast, actuaries typically performed clustered, prevalence-based data analysis and used the results from clinical studies. As the insurance industry does not publish their survival models, this thesis reports survival models developed in clinical studies. The current subsection provides an overview of the different study designs, data sources, and data modelling techniques used in developing the survival models. These models are described in detail in Chapter 2.

CVD survival models were developed by either randomised control trials or cohort studies (CTTC, 2012; Hardoon et al., 2011; Luepker, 2011; Nakamura et al., 2006; NICE, 2013a; Ridker et al., 2008; Smolina et al., 2012a). The study populations consisted either of cases and controls or cases only. The studies that only included cases, could investigate survival variations after the diagnosis of the medical condition in great detail. However, these studies could not estimate the effect of the medical condition itself on survival time due to the lack of a control group. In contrast, studies that included both cases and controls, could not investigate survival variations given the medical condition in great detail due to limited medical information available for the entire sample, but could estimate the effect of the medical condition itself on survival time. The effect of the medical condition on survival time, however, was most likely overestimated as the estimate could not be adjusted for important risk factors. The cases may be more likely to have comorbidities and an unhealthy lifestyle, which are independent predictors of survival, and so adjustment for these risk factors is important.

Multiple data sources were used to develop CVD survival models, ranging from prospectively collected trial-cohort data to routine data from hospitals, primary care, or disease and mortality registers (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; CTTC, 2012; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; NICE, 2013a; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). The data source(s) used define the constraints of the study design, the inclusion and exclusion criteria of the sample, the sample size, the length of follow-up period, and the range of risk factors that can be adjusted for in the analysis. Thus, the data source determines what health outcomes can be studied and the generalisability of the results.

Of the different data sources, primary care data have rarely been used to develop CVD survival models. Primary care data could be an important new source of information on survival prospects associated with medical conditions treated in routine clinical practice. Approximately three decades ago, the migration from paper to electronic medical records began to take place, giving rise to electronic medical records databases (Shephard et al., 2011). This provides relatively easy access to data on the target population due to the number of medical records included in the database. Such a database is populated with medical records from multiple medical centres and is updated on a routine basis. The long follow-up of a large sample of patients from multiple medical centres can lead to greater confidence in the results due to more precise estimates and better coverage of the underlying population. Also, the high volume of person-years of data on a wide range of available risk factors, permits the development of more complex statistical models. Complex modelling can enable better understanding of the variations in the outcome of interest.

Various techniques of data modelling can be considered in developing survival models. This involves making assumptions about censoring, covariate selection including interaction effects, time dependency of survival prospects, survival variations by medical centres when applicable, and types of missing data when present (Allison, 2001; Therneau and Grambsch, 2000). Most of these assumptions were not explored by previous CVD survival models (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). In case of violated assumptions, the results might be biased or less precise.

Thus, previous CVD survival models have been developed using various study designs, data sources, and data modelling techniques. Studies including only cases were used to estimate survival variations given CVD while studies including both cases and controls were used to estimate the survival prospects of CVD. Combining these two study designs to create a new survival model of CVD should yield more accurate estimates of survival prospects of CVD where the risk factors explaining survival variations given CVD can be adjusted for. This new survival model can be achieved by making use of primary care data, which has information on general and CVD specific risk factors for both cases and controls. Due to the extensive content in primary care data, there is an opportunity to test new combinations of risk factors, including interaction effects and the time-dependency of effects, in explaining survival variations. Furthermore, there is an opportunity to take general practices into account in the survival model such that survival variations between practices can be explored.

Pursuing these options can provide insights on current treatments in routine clinical practice.

1.2 Research objectives and aims

The primary objectives of this research are to investigate how a history of a CVD, in particular acute myocardial infarction (AMI), affects longevity in residents of the UK at retirement age, and also to investigate which treatments improve longevity. Specifically, data from The Health Improvement Network (THIN) primary care database are used to develop survival models for longevity in the presence or absence of AMI and survival models for longevity in the presence or absence of statins prescribed as primary prevention of CVD. Snapshots of medical history are obtained at four different target ages, namely 60, 65, 70, and 75. The research focuses on survival prospects and variations given this medical history. These results can inform both the style of financial spending by an individual during retirement and the management of their basis and longevity risks by insurance companies. The results can also inform health care requirements and resource allocation in the population.

The main objectives are to develop population-based survival models addressing the following goals:

1. Determine the effects of AMI, statins prescribed as primary prevention of CVD, and other AMI related treatments on longevity at the four target ages.

2. Establish a list of additional risk factors affecting longevity by themselves or in interaction with the medical condition, treatments, and other risk factors.

3. Quantify the protective or harmful effects of these risk factors.

4. Establish the clinical and actuarial implications of found variations in longevity.

The aims of the research are:

1. Investigate how the presence and duration of comorbidities and treatments affect the hazard of mortality at each age and whether they can be related to age-specific medical management.

2. Investigate the survival benefits of statins prescribed as primary prevention of CVD for various CVD risk groups at each age and whether this can inform risk thresholds for action.

3. Investigate how modifiable risk factors such as cholesterol level, blood pressure, body mass index, alcohol consumption, and smoking affect the hazard of mortality at each age and whether they can inform public health measures.

4. Investigate the effect of general practice on the hazard of mortality at each age and whether this is a factor additional to the socio-demographic factors of a district to consider in resource allocation.

5. Estimate the years lost or gained in effective age for each of the medical conditions, treatments, lifestyle choices, and socio-demographic factors at each age, and investigate how this could inform individuals about financial planning for retirement.

6. Investigate which medical conditions, treatments, lifestyle factors, socio-demographic factors, and interactions of risk factors at each age do and do not contribute in explaining survival variations and therefore to minimise the basis risk of estimating life expectancy for the pricing of annuities.

7. Investigate whether the effects of treatments, lifestyle choices, or other risk factors on longevity change over time and might form longevity risks that should be taken into account with pricing of annuities.

1.3 Contributions

This research contributes to existing CVD research by developing survival models that estimate both the effect of AMI on survival time and the survival variations given a possible history of AMI at different retirement ages, and by developing survival models that estimate the effect of statins prescribed as primary prevention of CVD at different retirement ages.

The newly developed AMI survival models address the issue of lack of estimation of the hazardous effects of AMI by incidence study designs due to the exclusion of controls, as well as address the issue of overestimation of the hazardous effects of AMI by prevalence study designs due to the limited number of risk factors adjusted for. This is achieved by making use of primary care data, which have information on a wide range of risk factors for both cases and controls. Hence, many of the different groups of risk factors, which are classified in this research as comorbidities, treatments, lifestyle choices, and socio-demographic factors, can be represented and adjusted for in the analysis. In general, an epidemiological study would only test for interaction effects with age, sex, and the main exposure of interest. However, the newly developed AMI survival models test for interaction effects within and between all groups of risk factors. In turn, survival variations are examined in greater detail. The results could inform pharmacosurveillance, lead to improved resource allocation, be of guidance in strategic financial planning for retirement, and contribute to better pricing of annuities by minimising the basis risk and managing the longevity risk of life expectancy estimations.

The newly developed statins survival models address the issue of strict inclusion and exclusion criteria of clinical trials. Clinical trials typically perform analyses on ideal patients, making it difficult to generalise the results to the wider population

(Godlee, 2014). The newly developed statins survival models more accurately assess the potential survival benefits of statins prescribed in the general population by performing the analysis on an intention-to-treat basis on primary care data.

Under the National Health Service (NHS), 99% of British citizens are registered at a general practice (NHS, 2013). Survival models produced from a sample of primary care data can thus be representative of the whole of the UK. Furthermore, primary care has a better coverage of AMI cases compared to hospitals and disease registers, because it includes patients who were diagnosed immediately and patients who were not sent to the hospital but were diagnosed in routine practice later by blood test results (Herrett et al., 2013b). This means that the results of these newly developed survival models are representative of a wider range of AMI cases in the UK than previously. The survival models are in turn more applicable in the clinical and actuarial fields because the sample is more similar to the target population. Furthermore, risk management of patients will be relatively straightforward for clinicians, because the risk predictions are based on routinely measured factors. In addition, potential longevity risks are easier to identify for actuaries due to longer follow-up in primary care data compared to hospital data and disease registers, which do not routinely record death dates.

The newly developed survival models address the issue of interdependence of patients from the same general practice and the issue of missing data. Most previous models failed to address both issues appropriately when present and the interaction of these issues. General practices vary in health outcomes and survival rates due to differences in their patient populations and provision of patient care (Rasbash et al., 2012). As it is impossible to adjust for all important risk factors on the individual level in the analysis, taking clustering by practice into account could lead to increased explanation of survival variations and more accurate estimates of survival prospects.

Missing data is also a common issue with observational data. There are several methods available to deal with missing data in order to obtain unbiased estimates. The developed survival models provide more accurate and unbiased estimates by addressing clustering by general practice and dealing with missing data appropriately.

The newly developed survival models are estimated at four different target ages, namely 60, 65, 70, and 75. These are ages when people would typically retire from work, and therefore it would inform financial planning of retirement for individuals and pricing of annuities for actuaries. Furthermore, CVD becomes more prevalent from age 60 onwards (Townsend et al., 2015), making primary prevention of CVD by administration of statins more relevant to individuals aged 60 and older. The results can facilitate individuals and general practitioners to make a decision about statins use at key ages. With CVD being more common from age 60 onwards, these age-specific results can also inform clinicians about ongoing medical management of AMI and can inform local authorities about resource allocation.

Thus, this research contributes to existing CVD survival research by assessing and quantifying effects of various risk factors by making use of primary care data, addressing assumptions of homogeneous population and complete survival data, and analysing age-specific data of people at retirement age. As noted above, these results are relevant to several interested parties in the medicine and retirement planning fields.

The findings of the survival models were published in the peer-reviewed journals PLoS One (Gitsels et al., 2016) and BMJ Open (Gitsels et al., 2017), and in the Longevity Bulletin of the Institute and Faculty of Actuaries (Kulinskaya and Gitsels, 2016).

1.4 Thesis outline

This Section provides the outline of the following chapters of the thesis.

Chapter 2 is a literature review of CVD. First, the risk assessment of a first cardiovascular event and risk management by statin prescription in the UK are discussed. Second, the CVD subtype AMI is defined and the risk management for secondary prevention of AMI in the UK is described. Third, existing AMI survival models are discussed in detail.

Chapter 3 is a review of primary care data and its use. First, routine data are compared and contrasted with prospectively collected trial-cohort data. Second, the availability and validity of primary care data in the UK are discussed, in particular The Health Improvement Network (THIN) database that is used for this research. Third, the inclusion and exclusion criteria of the studied age cohorts and the recorded characteristics on these cohorts are described.

Chapter 4 is a review of statistical methods. First, the choice and assumptions of the specified survival model are explained. Second, the process of model development with regard to the study design and the selection of covariates is described. Third, the way missing data were dealt with is discussed. Fourth, the assessment of the final models is explained.

Chapter 5 presents the survival models that estimated the hazard of all-cause mortality associated with a history of AMI and estimated the survival variations given a possible history of AMI. The analysis procedure is explained and the studied cohorts are described. The effects of a history of AMI, treatments, and general practice on survival time at different retirement ages are presented. The survival models and the estimated effects are assessed and compared with the results of previous studies.

Chapter 6 presents the survival models that estimate the hazard of all-cause mortality associated with statins prescribed as primary prevention of CVD. The analysis procedure is explained and the studied cohorts are described. The potential survival benefits of statins in various cardiovascular risk groups at different retirement ages are presented. The survival models and the estimated effects are assessed and compared with the results of previous studies.

Chapter 7 discusses the research' findings. First, the main results are summarised and the contributions to the existing evidence are presented. Second, the strengths and limitations of the research are discussed. Third, the implications in medical management and retirement planning are discussed by addressing the research' aims. Fourth, an overall conclusion is provided.

Chapter 2

Review of cardiovascular disease

This Chapter is a literature review of cardiovascular disease (CVD) and its subtype acute myocardial infarction (AMI), in which the respective survival models are discussed. The objective of reviewing existing survival models is to survey the current ‘state of the art’ and to identify gaps in CVD research.

The first Section of this chapter defines the medical class CVD and describes the risk assessment and management of a first cardiovascular event in the United Kingdom (UK). The second Section defines the CVD subtype AMI, describes the risk management for secondary prevention of AMI in the UK, and discusses existing survival models that estimated survival prospects and variations after AMI.

2.1 Cardiovascular disease

CVD is the medical classification for diseases of the circulatory system, and includes: acute rheumatic fever; chronic rheumatic heart diseases; hypertensive diseases; ischaemic heart diseases; pulmonary heart disease and diseases of pulmonary circulation; other forms of heart disease; cerebrovascular diseases; diseases of arteries, arterioles and capillaries; diseases of veins, lymphatic vessels and lymph nodes; and other and unspecified disorders of the circulatory system (WHO, 2010). The underlying cause of CVD is in most cases atherosclerosis, which is a build-up of plaque on the walls of the circulatory system caused by excess cholesterol (Townsend et al., 2015). As stated in the previous Chapter, CVD is the number two cause of death for men and women in the UK, accounting for 154,639 deaths in total in 2014 (Townsend et al., 2015). The greatest contributor to these deaths are ischaemic heart diseases, such as angina pectoris and AMI, which account for 69,163 deaths (Townsend et al., 2015).

Most of these cardiovascular events could potentially be prevented by pursuing a healthy lifestyle including being physically active, having a healthy varied diet, having a healthy body mass index, drinking alcohol in moderation, and abstaining from smoking (WHO, 2015a). These healthier lifestyle choices are promoted using national strategies. Besides lifestyle choices, other modifiable risk factors of CVD are hypertension and hypercholesterolaemia. These two risk factors are addressed at the individual level by offering antihypertensive or lipid-lowering drug therapies such as beta blockers or statins, respectively (NICE, 2011, 2015).

For secondary prevention of CVD, all patients should be offered the following drug therapy: angiotensin-converting enzyme (ACE) inhibitors, beta blockers, dual antiplatelet agents of which one is aspirin, and statins (NICE, 2013b). Up to 75% of recurrent events may be prevented when all these drugs are prescribed in combination with smoking cessation (WHO, 2015a). For some patients, it is beneficial to have heart surgery, which includes coronary artery bypass grafts, coronary angioplasty, valve repair and replacement, heart transplantation, and artificial heart operations (WHO, 2015a).

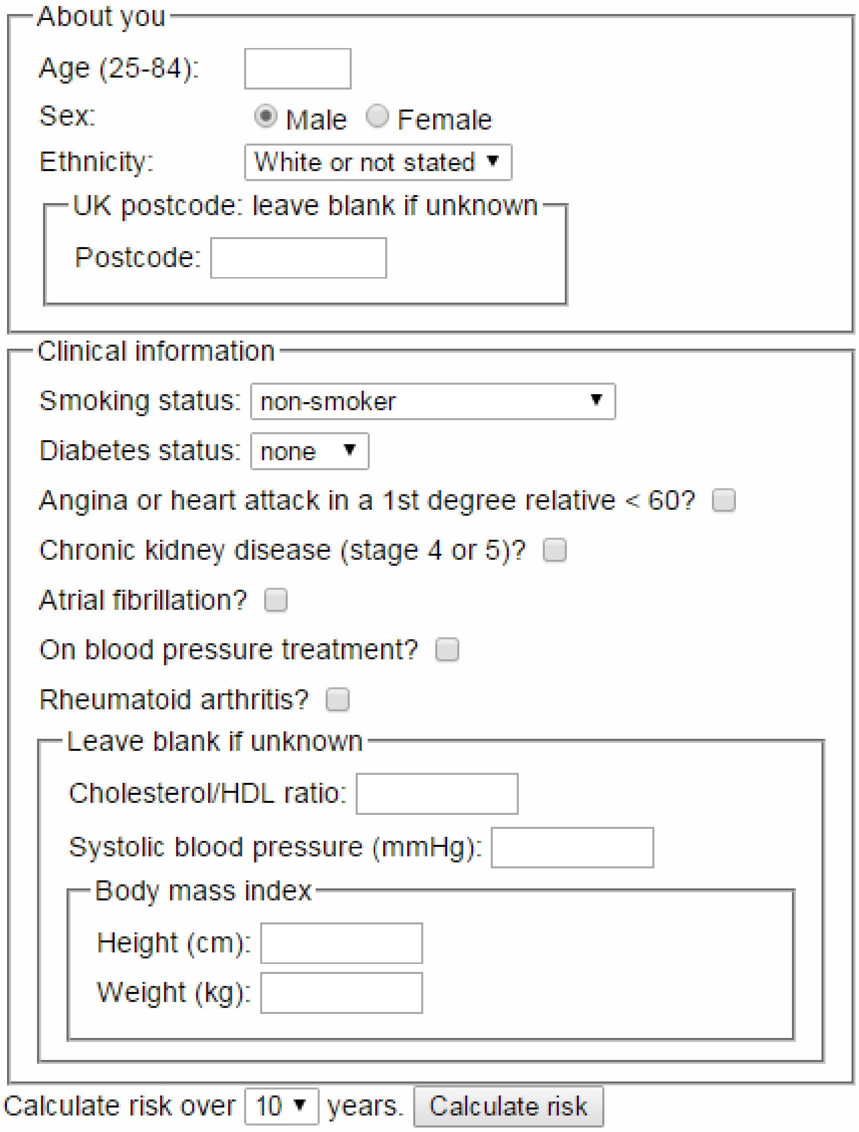
2.1.1 Primary risk assessment

The National Institute of Health and Clinical Excellence (NICE), which is a UK national body providing guidelines on health and social care, recommends the QRISK2 assessment to calculate the risk of developing a first cardiovascular event in the next ten years (NICE, 2015). This risk assessment incorporates information on multiple demographic, medical, and lifestyle factors, see Figure 2.1.

QRISK2 was developed in 2008, using two million UK patient records from 550 general practices that contributed to the QResearch primary care database (Hippisley-Cox et al., 2008). Using the QResearch and The Health Improvement Network (THIN) primary care database, the QRISK2 scores were validated against the Framingham scores, which was the recommended cardiovascular risk assessment at that time (Collins and Altman, 2009; Hippisley-Cox et al., 2008). The results showed that the QRISK2 scores estimated cardiovascular risk more accurately than the Framingham scores. As a result, since 2010 QRISK2 is the recommended tool to assess cardiovascular risk (NICE, 2015). QRISK2 is updated annually, including the set of risk factors and the coefficients of the risk factors (Ltd, 2015). The updated version is in turn externally validated (Collins and Altman, 2012).

The QRISK2 risk assessment shows that men are at a higher risk of developing CVD than women. Compared to people with a Caucasian background, people with an Indian, Pakistani, or Bangladeshi background, have higher cardiovascular risk. In contrast, people with a black Caribbean background or men with a black African or Chinese background, have lower cardiovascular risk. The risk of CVD increases with level of deprivation, which has a greater hazardous effect in women than in men. Although age is the main driver behind cardiovascular risk, the risk factors with the greatest hazardous effect are, listed in descending order: atrial fibrillation, type 2 diabetes, and family history of ischaemic heart disease.

QRISK2 is used to identify people aged between 25 and 85 who are likely to be at high risk of developing a first cardiovascular event. These age boundaries are specified because people younger than 25 have practically no cardiovascular risk, while



About you

Age (25-84):

Sex: ● Male ○ Female

Ethnicity: White or not stated ▾

UK postcode: leave blank if unknown

Postcode:

Clinical information

Smoking status: non-smoker ▾

Diabetes status: none ▾

Angina or heart attack in a 1st degree relative < 60? □

Chronic kidney disease (stage 4 or 5)? □

Atrial fibrillation? □

On blood pressure treatment? □

Rheumatoid arthritis? □

Leave blank if unknown

Cholesterol/HDL ratio:

Systolic blood pressure (mmHg):

Body mass index

Height (cm):

Weight (kg):

Calculate risk over 10 ▾ years. Calculate risk

Figure 2.1: QRISK®2 cardiovascular disease calculator

This tool calculates the risk of developing a first cardiovascular event in the next ten years for an individual aged between 25 and 85 (ClinRisk Ltd, 2015).

people aged 85 and older are at high risk no matter their demographic, medical, or lifestyle background. People who have a QRISK2 score above a certain threshold, and are thereby classified as being at high risk, are offered statin therapy for primary prevention of CVD (NICE, 2015). The set risk threshold is based on results from clinical trials that estimated the effectiveness of the drug in specific risk groups.

2.1.2 Primary risk management

Statins have been widely prescribed for primary and secondary prevention of CVD since the Scandinavian Simvastatin Survival Study in the 1990s demonstrated benefits of statin therapy in patients with established CVD (4S, 1994). Since then, the results of many statin trials have been combined into an individual patient-based meta-analysis of 27 randomised control trials and over 90,000 patients by the Cholesterol Treatment Trialists' Collaboration (CTTC) (CTTC, 2012). This meta-analysis reported that in participants without a history of vascular disease, statins reduced the overall risk of all-cause mortality by 9% per 1.0 mmol/L reduction in low-density lipoprotein (LDL). The study, however, could not conclude survival benefits of statins for the individual risk groups due to the small number of deaths.

Based on the CTTC findings published in 2012, NICE lowered the risk threshold at which statins should be prescribed from 20% to 10% in July 2014 (NICE, 2015). This caused a 'storm of controversy' about the benefits to people at low risk of CVD (Parish et al., 2015). The lowered risk threshold translated to an increasing number of people being eligible for the drugs; that is an additional 4.5 million UK residents (NICE, 2014b). The risk threshold of 10% recommended by NICE identified similar numbers of patients as the 2013 American College of Cardiology/American Heart Association (ACC/AHA) guideline, which recommends statin prescription when the Pooled Cohort Equations (PCE) estimated 10-year risk of a cardiovascular event is

$\geq$7.5% (Mortensen and Falk, 2014; Stone et al., 2014). The 2012 European Society of Cardiology (ESC) guideline recommends considering statins when the Systematic COronary Risk Evaluation (SCORE) estimated 10-year risk of cardiovascular mortality is $\geq$5%, but this identifies much fewer patients than the NICE and ACC/AHA guidelines, because it focuses on mortality rather than events (Mortensen and Falk, 2014; Perk et al., 2012).

The CTTC meta-analysis was one of the most comprehensive sources of evidence assembled for any medical condition, but still left some major uncertainties about the survival benefits of statins for those without a history of vascular disease. First, the strict inclusion criteria of most of the included clinical trials make it difficult to apply the findings to patients in routine clinical practice, most of whom would not have been eligible for the trials on the grounds of age or morbidity (Downs et al., 1998; Nakamura et al., 2006; Ridker et al., 2008; Shepherd et al., 1995). Second, the risk groups were based on the study's own prediction of the 5-year risk of a major vascular event, which makes comparison with the QRISK2, SCORE, or PCE risk over 10 years, as widely used and recommended in European or American clinical practice, difficult and uncertain. Third, the average age of a trial participant was 63 years and the trials only included a small number of older participants, making estimates of effectiveness in different age groups difficult. Fourthly, the follow-up time of each trial was at most five years, which is much shorter than the monitoring of many patients in routine clinical practice. The short follow-up time resulted in a small number of deaths observed, leading to uncertain results for the individual risk groups; only between 300 and 1,500 deaths were observed in the individual risk groups of patients with no history of vascular disease. There are also concerns that anonymised individual patient data from statin trials have not been made available for independent scrutiny, particularly as statins are among the most widely prescribed drugs globally (Parish et al., 2015).

NICE identified several gaps in the research evidence of risk management with statins when the guidance was updated in 2014, and recommended further research into the effectiveness of age alone and other routinely available risk factors compared with formal structured multi-factorial risk assessment to identify people at high risk of developing CVD, as well as into the effectiveness of statin therapy in older people in general (NICE, 2015). These identified gaps and the uncertain results of survival benefits of statins in the individual risk groups led to the current research objective to estimate the long-term survival benefits of statin prescribed in the general population with no previous history of CVD, stratified by age and QRISK2 groups. This is pursued in Chapter 6.

2.2 Acute myocardial infarction

AMI is pathologically defined as myocardial cell death due to prolonged ischaemia (Swanton and Banerjee, 2009). The risk factors of AMI were established by the INTERHEART study that took place in 52 countries from 1999 to 2002 and included roughly 30,000 participants (Yusuf et al., 2004). The study found that the population attributable risk (PAR) in men and women could be explained up to 90% and 94%, respectively, by the following risk factors: smoking, alcohol consumption, abdominal obesity, hypertension, hypercholesterolaemia, diabetes, psychosocial factors (an index score that combines depression, stress at work/home, financial stress, life events, and locus of control factors), consumption of fruits and vegetables, and regular physical activity. This means that if exposure to these risk factors were removed, the incidence would be reduced by 90% in men and 94% in women. Of the nine risk factors, hypercholesterolaemia was the most hazardous.

In 2012 in the UK, the average age for men and women to have their first AMI episode was 65 and 73 years, respectively, and the case-fatality (here death within first 30 days) was 8% (NICE, 2013a). Mortality ratios reduce markedly over the first year following AMI, but start to level off thereafter. The latest population-based cohort study in England with data from 2004 to 2010, concluded that after seven years people with a first or recurrent AMI have double or triple the risk of mortality compared to the general population of equivalent age (Smolina et al., 2012b).

Incidence and mortality rates have declined considerably over the past few decades in developed countries including the UK (Briffa et al., 2009; Capewell et al., 2000; Hardoon et al., 2011; Luepker, 2011; Smolina et al., 2012a). The Multinational Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA) project, which was set up by the World Health Organization, collected data from 38 medical centres in 21 developed countries from 1985-87 to 1995-97 (Luepker, 2011). The sample consisted of approximately 10 million subjects aged 25-64. Two thirds of the decline in mortality after ischaemic heart disease was explained by a decrease in the incidence rate and a third by a decrease in the case-fatality rate (here death within first 28 days). A more recent study in England made use of data from the Hospital Episode Statistics and Mortality Statistics from 2002 to 2010 (Smolina et al., 2012a). Over that period of time, the incidence and case-fatality rate (here death within first 30 days) of AMI fell both by a third. Both declines contributed approximately the same to the halved one-year mortality rate.

The European Society of Cardiology and the American College of Cardiology introduced a new diagnostic criterion of AMI in 2000 (Smolina et al., 2012a). The new criterion measures the amount of troponin I or T in a blood sample. These proteins are released when there is heart damage; the more damage, the more of these proteins can be found (Antman et al., 2000). The new criterion led to more diagnoses of AMI and thus also to more reported incidence of milder cases. It takes time before a new criterion is standardised in the diagnosis of a disease. Studies in Denmark (Abildstrom et al., 2005), Finland (Salomaa et al., 2006), Australia (Sanfilippo et al., 2008), and the United States (Roger et al., 2010) showed that the new criterion affects the incidence rate but not the mortality rate. This is because the new criterion increased the incidence of AMI in patients aged 70 or older, who have a worse survival rate than younger patients.

The improved incidence and mortality rates over the past few decades in developed countries can partly be explained by an increase in coronary revascularisation, more effective drug therapy, and healthier lifestyles (Bata et al., 2006; Briffa et al., 2009; Capewell et al., 2000; Hardoon et al., 2011; Smolina et al., 2012a). With respect to the lifestyles, there was a decrease in smoking, sedentary lifestyle, hypertension, and hypercholesterolemia. Even though the incidence and mortality rates have improved, a considerable number of people are still affected by AMI and continue to have worse survival prospects than people without AMI. In 2012 in the UK, there were approximately one million men and almost half a million women with a history of AMI (NICE, 2013a).

2.2.1 Secondary risk management

After an AMI, all patients are encouraged to attend a cardiac rehabilitation programme (NICE, 2013b). This programme includes exercise plans, health education, and stress management to reduce their risk of a next cardiovascular event. Patients are advised to be physically active, stop smoking, regularly consume a moderate amount of alcohol, eat a Mediterranean-style diet, and manage their weight. All patients should be offered ACE inhibitors, beta blockers, dual antiplatelet agents of which one is aspirin, and statins, and be considered for coronary revascularisation.

Survival prospects after an AMI vary by a number of factors, here grouped by socio-demographic factors, lifestyle choices, comorbidities, and treatments. Examples of survival variations by socio-demographic factors are age, sex, socioeconomic status, and psychosocial factors. Older patients have a higher case-fatality rate and are at higher risk of a recurrent cardiovascular event (Capewell et al., 2000; Smolina et al., 2012b). Women tend to have a worse survival rate of AMI in the short-term but have the same long-term survival prospects as men (Capewell et al., 2000; Chang et al., 2003; Gottlieb et al., 2000; Koek et al., 2007; Rosengren et al., 2001; Smolina et al., 2012b; Vaccarino et al., 1999). People with lower socioeconomic status measured at the individual or neighbourhood level have worse survival prospects after an AMI. Neighbourhood socioeconomic status possibly captures the residual confounding factors of unequal hospital resources and social characteristics of an area such as social cohesion and attitudes towards health (Capewell et al., 2000; Gerber et al., 2010; Smolina et al., 2012b). Compared to patients with ischaemic heart disease and no psychosocial factors, patients with ischaemic heart disease suffering from depression, anxiety, job strain, or lack of social support have a worse survival rate (Hemingway and Marmot, 1999). Therefore, offering stress management and psychosocial support to patients who had an AMI could help improve their survival prospects.

Lifestyle choices that affect survival prospects after an AMI are: smoking, body mass index, alcohol consumption, and physical activity. There is a smoker's paradox in which smokers have a better short-term survival rate than non-smokers (Gerber et al., 2009; Gourlay et al., 2002). This is partly explained by the fact that smokers tend to have an AMI at a younger age and therefore fewer additional risk factors. In the long-term, non-smokers have a better survival rate than ex- and current-smokers. Smoking cessation either before or after an AMI is associated with improved short- and long-term survival rates (Gerber et al., 2009). Similarly, obese AMI patients have a better survival rate in the first six months compared to AMI patients with a healthy weight (Nigam et al., 2006). After six months, patients with a healthy weight have a lower risk of a recurrent event and mortality. The reason for this paradox is not fully explained, but could be due to the fact that obese patients typically have an AMI at a younger age and receive more aggressive treatment than patients with healthy weights. At an older age, overweight patients have a better survival rate than healthy weight patients even though overweight people are more likely to have cardiovascular disease (Chapman, 2010). From the age of 65 onwards, the optimal body mass index was found to be between 27 and 30, and from the age of 75 onwards, obesity has little to no harmful effects on survival rate.

Comorbidities that affect survival prospects after an AMI are: previous AMI, angina pectoris, cerebrovascular disease, peripheral vascular disease, heart failure, cancer, diabetes, renal disease, respiratory disease, hypertension, and hypercholesterolaemia. Of these comorbidities, diabetes is the most hazardous. Co-occurrence of these medical conditions increases with age (van Baal et al., 2011). Considering AMI, diabetes, cerebrovascular disease, and cancer, the two conditions that occur most often together in absolute numbers are AMI with diabetes and in relative numbers AMI with cerebrovascular disease. The Charlson comorbidity index measured five years prior to the AMI event is also a strong predictor of survival prospects (Schmidt et al., 2012). The Charlson comorbidity score is calculated as follows: one point for AMI, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, ulcer disease, mild liver disease, and diabetes without end organ damage; two points for diabetes with end organ damage, hemiplegia, moderate to severe renal disease, non-metastatic solid tumour, leukaemia, and lymphoma; three points for moderate to severe liver disease; and six points for metastatic cancer and AIDS (Charlson et al., 1987). This score has been extensively validated (O'Connell and Lim, 2000; Schmidt et al., 2012). Treatment of these conditions should be in line with the respective clinical guidelines (NICE, 2013b).

As stated above, all patients should be offered drug therapy to reduce the risk of next cardiovascular event. Non-compliance with the drug therapy could result in a higher risk of adverse outcomes. Approximately half of patients are non-compliant in taking aspirin after several years (Graham et al., 2007). A systematic review based on approximately 50,000 patients showed that this can cause a threefold increased risk of another cardiovascular event (Biondi-Zoccai et al., 2006). Another study found that people who stopped taking aspirin within the last six months were worse off compared to current users with regards to risk of non-fatal AMI or fatal ischaemic heart events (García Rodríguez et al., 2011). However, people who stopped taking the drug for more than six months were not significantly better or worse off than current users. People who stopped because of safety concerns or used over-the-counter aspirin were also not significantly better or worse off than current users. NICE reported mixed clinical evidence of the effectiveness of drug therapy versus placebo with regards to long-term all-cause mortality. Aspirin had inconclusive benefits (CDP, 1976; NICE, 2013a). ACE inhibitors seem to be effective in AMI patients with left ventricular systolic dysfunction (LVSD; relative risk (RR) of 0.84 (95% confidence interval 0.78-0.91)) (NICE, 2013a). This evidence is of moderate quality with no serious inconsistency, indirectness, or imprecision. ACE inhibitors, however, seem to be ineffective in AMI patients with unselected LVSD (RR=1.02 (0.57-1.84)) (NICE, 2013a). This evidence is of low quality due to its imprecision. Patients receiving beta blockers in the first 72 hours after the onset of AMI or after 72 hours to a year have a lower hazard of mortality (RR=0.87 (0.67-1.20) and RR=0.76 (0.49-1.16), respectively) (NICE, 2013a). This evidence is of low quality due to its imprecision. Statins in any patient or specifically in CVD patients reduced the hazard of all-cause mortality (RR=0.87 (0.84-0.91) and RR=0.87 (0.83-0.91), respectively) (NICE, 2015). This evidence is of high quality but not clinically important due to the low effect size. The NICE-recommended drug therapy's primary objective is to improve survival prospects by reducing the risk of next cardiovascular event and not per se by reducing the risk of mortality. If the benefits of a drug in reducing the risk of a next cardiovascular event outweighs the adverse effects and is not harmful for life expectancy, the drug could be included in the clinical guideline. The mixed clinical evidence of reduction in the risk of mortality associated with drug therapy led to the research objective to estimate the long-term survival benefits of treatments. This is pursued in Chapter 5.

2.2.2 Existing survival models

There are numerous studies that have examined survival prospects and their variations after AMI by estimating case-fatality, one-year mortality, and long-term mortality. These studies either estimated mortality rates of AMI standardised for age, sex, deprivation or region (Capewell et al., 2000; Hardoon et al., 2011; Luepker, 2011; Smolina et al., 2012a,b) or examined survival variations among AMI patients by a range of risk factors including socio-demographic factors, lifestyle choices, comorbidities, and treatments (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). The first type of study is likely to overestimate the hazardous effect of AMI on mortality, because it was limited in the number of risk factors to adjust for between patients who had an AMI and who did not have an AMI; while the second type of study cannot estimate the hazardous effect of AMI on mortality due to the lack of a control group. Thus, there has not been a study that estimated long-term survival prospects after AMI compared to no AMI while adjusting for a range of risk factors. To inform the choice of risk factors in the survival models developed for this research, existing survival models that estimated long-term all-cause mortality in AMI patients were reviewed. The review also included listing the study designs, data sources, and data modelling techniques used. The survival models reviewed are presented in Table 2.1.

The studies took place in various developed countries: Australia, Canada, England, Germany, Israel, the Netherlands, Scotland, and the United States. Either a city, region, county, or whole country was eligible for the study. All studies had as inclusion criteria that patients had to be hospitalised for an AMI. Additional inclusion criteria were that patients had to survive for a specific period of time (8/11 studies), the AMI had to be the first one in the medical history (7/11 studies), patients had to be of a certain age (6/11 studies), and patients had to have an additional medical condition (2/11 studies). The recruitment periods ranged from 1 to 18 years. Six studies followed up the patients for longer than that period, resulting in study periods ranging from 6 to 21 years. Together the studies analysed data from 1977 to 2011. The sample size varied greatly, from less than 1,000 to almost 400,000 patients.

All studies used clinical data from either hospitals or register databases. Five studies made use of the extra information available in hospital data regarding the severity of AMI and the treatment given at admission and/or discharge (Briffa et al., 2009; Chang et al., 2003; Gerber et al., 2010, 2009; Kirchberger et al., 2014). Three studies obtained additional data from questionnaires or interviews, in which information regarding the patient's socioeconomic status (education, income, employment, and living with a steady partner), lifestyle (smoking and physical activity), and health status (self-rated) were asked (Gerber et al., 2010, 2009; Kirchberger et al., 2014). One study also made use of primary care data, although AMI patients were not selected for the study using that source of data (Quint et al., 2013). A study population and survival model based on primary care data could lead to different results due to the extensive information available on this slightly different population. Compared to hospital and register data, primary care data has more information on socio-demographic and lifestyle factors and has a greater coverage of AMI cases in the UK (Herrett et al., 2013b). Therefore, primary care data could be an important new source of information on survival prospects of patients who have had an AMI.

All but one study included patients from multiple hospitals. Of these studies, only one included a random effect of hospitals in the survival model, but found no significant survival variations between the eight hospitals (Gerber et al., 2009). The other studies assumed that there were no survival variations by hospital. In case that this assumption is violated, failing to adjust for clustering by hospital can lead to false precision (Therneau and Grambsch, 2000).

All studies estimated all-cause mortality from the moment the AMI was diagnosed by means of a Cox's proportional hazards regression. Five studies reported checking the assumption of proportional hazards between the baseline category and the other categories of a risk factor (Gerber et al., 2010, 2009; Kirchberger et al., 2014; Koek et al., 2007; Smolina et al., 2012b). In case of violation, the estimated hazard of mortality associated with a risk factor would be imprecise at specific points in time as the hazard is an average over the whole study period (Kleinbaum and Klein, 2011). Moreover, there could be a type-II error, i.e. a false negative, where the risk factor is protective in one period of time and hazardous in another period of time, like the smoking paradox in one-year mortality rates as explained above (Gerber et al., 2009; Gourlay et al., 2002). Therefore, it is important to check the proportional hazards assumption. All studies assumed uninformative censoring, where patients lost to follow-up were assumed to have the same mortality rate as patients observed until the end of the study (Hosmer et al., 2008). Two studies reported the number of patients lost to follow-up, which was less than 3% (Herzog et al., 1998; Kirchberger et al., 2014). The other studies probably had a similar percentage as they made use of death registers and thus loss to follow-up would only happen in case of emigration. These low percentages of loss to follow-up were unlikely to affect the results.

Seven studies had missing data in socioeconomic status and lifestyle factors and one study had missing data in drug prescriptions. One study imputed missing values in income with the lower of the two categories, which affected only 2% of the sample (Chang et al., 2003). Five studies performed a complete case analysis thereby excluding between 4 and 24% of the sample (Capewell et al., 2000; Gerber et al., 2010, 2009; Kirchberger et al., 2014; Quint et al., 2013). In one study it is unknown how missing data were addressed (Briffa et al., 2009). When data are missing completely at random, meaning that there is no systematic difference between observed and unobserved data, unbiased results are obtained when complete case analysis is performed (Allison, 2001). It is, however, highly likely that patients with missing data differ from patients with complete data, as other studies found that recording of lifestyle factors is associated with health and thereby indirectly associated with survival rate (Marston et al., 2010; Shephard et al., 2011; Szatkowski et al., 2012). Thus, the results of the complete case analyses of the reviewed studies could potentially be biased.

Half of the survival models were built in blocks of risk factors; first sex and age were included in the model, followed by a group of risk factors, and ending with a full model (Gerber et al., 2010, 2009; Kirchberger et al., 2014; Koek et al., 2007; Quint et al., 2013). One study also reported the leanest model possible found by backward elimination (Kirchberger et al., 2014). The groups of risk factors used in the survival models were socio-demographic factors, lifestyle choices, severity of

AMI, comorbidities, treatments, and hospital details. All studies adjusted for age and sex, and most studies adjusted for socioeconomic status, types of CVD, diabetes, hypertension, and coronary revascularisation procedures or drug therapy. Two studies estimated survival variations separately by sex and/or age group (Capewell et al., 2000; Smolina et al., 2012b), and another two studies tested for interaction effects with sex, age, and the exposure of interest (diabetes or education level) (Kirchberger et al., 2014; Koek et al., 2007). Testing interactions between all risk factors could elucidate different effects in subpopulations and thereby explain survival variations to a greater extent. Furthermore, given that it is impossible to include all risk factors in the survival model due to limited recording, all groups of risk factors should be represented to obtain a holistic estimation of survival prospects.

The estimated hazards of mortality associated with different risk factors are study specific; the hazards depend on the sample studied, the time-period studied, the length of the study period, and the adjustment of other risk factors. Compared to a first AMI, the hazard ratio of mortality associated with a recurrent AMI was estimated to be 1.2 to 1.8 (Chang et al., 2003; Nigam et al., 2006; Smolina et al., 2012b). This hazard was greater at a younger age; in people aged 30 to 54 the hazard ratio was 2.3 while in people aged 85 or older the hazard ratio was 1.4 (Smolina et al., 2012b). The hazard of mortality associated with diagnosis of AMI decreased with calendar year; with every 4 to 8 years the hazard of mortality was significantly lower (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Herzog et al., 1998). Of the comorbidities, heart failure, diabetes, and chronic kidney disease, were with hazard ratios of 1.4 to 2.5, the most hazardous for the survival rate (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Herzog et al., 1998; Nigam et al., 2006; Smolina et al., 2012b). Again, these hazards were greater in younger patients than in older patients (Smolina et al., 2012b).

Coronary revascularisation had a survival benefit with a hazard ratio of 0.4 to 0.5 when it was undertaken during hospitalisation (Nigam et al., 2006; Smolina et al., 2012b) and of 0.8 to 0.9 when it was undertaken prior to the AMI or within the six months after the AMI (Chang et al., 2003; Herzog et al., 1998). Drug prescriptions had mixed effects on survival prospects. The greatest survival benefit was by prescription of antiplatelet drugs, like aspirin, and beta blockers with a hazard ratio of 0.5 to 0.8 (Briffa et al., 2009; Nigam et al., 2006; Quint et al., 2013). In contrast, lipid-lowering drugs, like statins, and ACE inhibitors were associated with a hazard ratio of 1.4 to 2.2 (Briffa et al., 2009; Nigam et al., 2006). Survival prospects were better when patients were prescribed multiple drugs during hospitalisation, but this was not necessarily true at discharge. The optimal number of prescriptions was three during hospitalisation and two at discharge (Briffa et al., 2009). Prescription of only one drug during hospitalisation or more than two different drugs at discharge was not associated with a significant survival benefit (Briffa et al., 2009).

Men and women did not vary in long-term survival prospects after an AMI, but experienced different effects of age and deprivation on their survival prospects (Capewell et al., 2000; Smolina et al., 2012b). Compared to men and women aged under 55, men and women aged 85 and older had a hazard ratio of 12 to 20 and 11 to 16, respectively (Capewell et al., 2000; Smolina et al., 2012b). Compared to men and women living in the least deprived area measured by the index of multiple deprivation (IMD), men and women living in the most deprived areas had a hazard ratio of 1.3 to 1.4 and 1.2, respectively (Capewell et al., 2000; Smolina et al., 2012b). Lastly, the hazard of mortality associated with smoking was 1.3 to 1.4 (Briffa et al., 2009; Chang et al., 2003), and a body mass index of greater than 25 was 0.7 (Nigam et al., 2006).

The identified gaps in the research evidence around survival prospects of AMI patients led to the research objective to estimate the hazard of mortality associated with a history of a single or multiple AMIs at key ages in UK residents based on primary care records while adjusting for a wide range of risk factors. Additionally, the widespread model assumptions in the research evidence led to the aim to investigate possible survival variations by general practice. The research objectives and aims are pursued in Chapter 5.

Table 2.1: Existing survival models of all-cause mortality after acute myocardial infarction (AMI)

Lead author (year published)	Study period	Study population	Sample size (deaths)	Survival model (assumptions)	Approach to missing data (% missing)	Risk factors in model
Briffa (2009)	1984-2005	Patients aged 35-64 admitted to hospitals in Perth (Australia) in 1984-1993 with first AMI and survived first 28 days	4,451 (1,182)	Cox's regression (proportional hazards assumed)	Imputed null values (unknown)	Sex, age, subcohort, diabetes, hypertension, smoking, ECG PREDICT score, heart failure, cardiogenic shock, tachycardia, systolic blood pressure, thrombolyisis, antiplatelet drugs, beta blockers, angiotensin converting enzyme inhibitors, lipid-lowering drugs, and coronary artery bypass graft
Capewell (2000)	1986-1996	Patients admitted to hospitals in Scotland in 1986-1995 with first AMI or angina and survived first 30 days	96,026 (39,449) AMI and 37,403 (8,153) angina	Cox's regression by sex (proportional hazards assumed)	Complete case analysis (4%)	Age group, deprivation, cerebrovascular disease, cancer, ischaemic heart disease, diabetes, heart failure, peripheral vascular vascular disease, and respiratory disease
Chang (2003)	1993-2000	Patients admitted to hospitals in Alberta (Canada) in 1993-2000 with AMI or unstable angina diagnosis	22,967 (7,014) AMI and 8,441 (1,718) unstable angina	Cox's regression (proportional hazards assumed)	Missing values of income imputed with the lower category (2%)	Sex, age group, year of admission, diabetes, previous congestive heart failure, chronic obstructive pulmonary disease, previous AMI, previous angioplasty, chronic renal disease, anaemia, cardiovascular specialist, coronary artery bypass graft capable hospital, health region, income, revascularisation

Continued on next page

Table 2.1 – *Continued from previous page*

Lead author (year published)	Study period	Study population	Sample size (deaths)	Survival model (assumptions)	Approach to missing data (% missing)	Risk factors in model
Gerber (2009)	1992-2005	Patients aged ≤65 admitted to 8 hospitals in central Israel in 1992-1993 with first AMI and survived hospital admission	1,521 (427)	Cox's regression (proportional hazards checked)	Complete case analysis (24%)	Sex, age, ethnic origin, education, income, pre-AMI employment, smoking status, hypertension, dyslipidemia, diabetes, obesity, physical activity, Q-wave AMI, anterior AMI, Killip class, comorbidity, thrombolysis, coronary artery bypass graft, and angioplasty (note: random effect on hospital was insignificant)
Gerber (2010)	1992-2005	Patients aged ≤65 admitted to 8 hospitals in central Israel in 1992-1993 with first AMI and survived hospital admission	1,521 (427)	Cox's regression (proportional hazards checked)	Complete case analysis (24%)	Sex, age, ethnic origin, hypertension, diabetes, dyslipidemia, smoking, physical activity, admission to ICU, anterior AMI, comorbidity index, Killip class, coronary artery bypass graft, angioplasty, self-rated health, education, income, pre-AMI employment, living with a steady partner, and neighbourhood socioeconomic status

Continued on next page

Table 2.1 – *Continued from previous page*

Lead author (year published)	Study period	Study population	Sample size (deaths)	Survival model (assumptions)	Approach to missing data (% missing)	Risk factors in model
Herzog (1998)	1977-1995	Patients recorded in United States Renal Data System with first AMI in 1977-1995 after renal-replacement therapy for ≥90 days and dialysis for ≥60 days	34,189 (approximately 32,000)	Cox's regression (non-informative censoring (858/2.5% censored) and proportional hazards assumed)	Not applicable	Sex, age, ethnicity, AMI calendar year, cause of end-stage renal diagnosis, congestive heart failure, other cardiovascular conditions, cancers other than skin cancer, chronic obstructive pulmonary disease, cerebrovascular ischaemia, peripheral vascular disease, gastrointestinal disease, gallbladder disease, liver disease, duration of end-stage renal disease, and revascularisation
Kirchberger (2014)	2000-2011	Patients aged 28-74 admitted to 10 hospitals in Augsburg region (Germany) in 2000-2008 with first AMI and survived first 28 days	4,405 (471)	Cox's regression (non-informative censoring (5/0.1% censored) assumed and proportional hazards checked)	Complete case analysis (22%)	Sex, age group, education level, living alone, diabetes, angina pectoris, hypertension, hyperlipidaemia, stroke, any reperfusion therapy, AMI type, left ventricular ejection fraction, any in-hospital complications, age group*sex, and age group*education level (note: sex*educational level was insignificant)
Koek (2007)	1995-2000	Patients admitted to hospitals in the Netherlands in 1995 with first AMI	21,565 (3,149)	Cox's regression (proportional hazards checked)	Not applicable	Sex, age, diabetes, previous cardiovascular disease, ethnic origin, age*diabetes, and age*sex (note: sex*diabetes was insignificant)

Continued on next page

Table 2.1 – *Continued from previous page*

Lead author (year published)	Study period	Study population	Sample size (deaths)	Survival model (assumptions)	Approach to missing data (% missing)	Risk factors in model
Nigam (2006)	1988-2001	Patients aged <80 residential in Olmsted County (United States) admitted to Mayo Clinic Coronary Care Unit in 1988-2001 with AMI and survived the first six months	894 (233)	Cox's regression (proportional hazards assumed)	Not applicable	Sex, age, body mass index, diabetes, systolic blood pressure, diastolic blood pressure, smoking status, family history of ischaemic heart disease, lipid lowering drugs, beta blockers, aspirin, and angiotensin converting enzyme inhibitors
Quint (2013)	2003-2008	Patients aged $\geq$18 with chronic obstructive pulmonary disease (COPD) and a first AMI recorded in MINAP (Myocardial Ischaemia National Audit Project, England) and registered at a General Practice Research Database (GPRD) practice in 2003-2008 and survived first year	1063 (447)	Cox's regression (proportional hazards assumed)	Complete case analysis (24%)	Sex, age, smoking, family history of cardiovascular disease, angina, hypertension, dyslipidaemia, peripheral arterial disease, cerebrovascular disease, heart failure, diabetes, frequency of exacerbations of COPD, type of myocardial infarction, diuretics, anti-arrhythmia drugs, antiplatelet agents, angiotensin converting enzyme inhibitors, beta blockers, statins, nitrates, and calcium-channel blockers
Smolina (2012)	2004-2011	Patients admitted to hospitals in England in 2004-2010 with AMI and survived first 30 days	387,452 (100,442)	Cox's regressions by sex and by age group (proportional hazards checked)	Not applicable	Sex/age group, deprivation, coronary artery bypass graft, angioplasty, cancer, cardiovascular disease, diabetes, respiratory disease, and renal disease

Chapter 3

Review of primary care data

This Chapter is a review of primary care data, in particular The Health Improvement Network (THIN) database, used for this research. The previous Chapter reviewed survival models of cardiovascular disease (CVD) and revealed that primary care data were rarely used for model development. This Chapter starts by discussing in what aspects primary care data could be valuable in developing survival models by comparing prospectively collected trial-cohort data with routine data. Next, the main primary care databases in the United Kingdom (UK) are introduced and the availability and validity of the data are discussed. This is followed by an overview of the THIN database. Then, the inclusion and exclusion criteria used to identify eligible practices and patients for the age-specific cohorts are discussed. Finally, the baseline measures and the outcomes during follow-up of the cohorts are defined.

3.1 Medical data sources

Important features regarding the usability of a model are that the sample is representative of the target population, there is a high volume of person-years of data to obtain precise estimates, there is high statistical power to test for risk factors, the most important risk factors are adjusted for to minimise residual confounding, the appropriate model is chosen to estimate the outcome and the respective model assumptions are addressed, and the model is validated. The data source used for model development determines which outcomes can be studied and the generalisability of the results. The data sources used for the CVD survival models can be divided into prospectively collected trial-cohort data and routine data from primary care, secondary care, disease registers, or mortality registers (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; CTTC, 2012; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; NICE, 2013a; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). These two types of data sources are compared with respect to the outcome ascertainment, the completeness and accuracy of risk factors, the representativeness of the sample, and the cost of collecting data.

Randomised control trials (RCTs) and cohort studies that prospectively collect data are designed to measure specific outcomes, exposures, and risk factors. Study protocols usually ensure that the data collection is done in a consistent and complete matter at specific points in time. The data collection, however, might be subject to recall or interviewer bias. Furthermore, there might not be data collected on certain risk factors, because at the start of the study these risk factors were deemed not to be relevant in predicting the outcome.

With routine data, information is recorded at the time the patient enters primary or secondary care. Therefore, the amount of data available is dependent on how frequently the patient enters primary or secondary care and what the clinicians find relevant to record for the patient's healthcare (MacDonald and Morant, 2008; Shephard et al., 2011; Wijlaars, 2013a). This could result in a substantial amount of missing or sporadic entries for the risk factors and outcomes the researcher is interested in. Furthermore clinicians can code medical conditions and treatments differently from each other and can have trouble classifying them or even misclassify them (Hippisley-Cox and Coupland, 2010b; MacDonald and Morant, 2008; Shephard et al., 2011). With the introduction of the Quality and Outcomes Framework (QOF) in 2004, a pay scheme to improve the quality of health care provided by general practitioners, recording has greatly improved in primary care (Langley et al., 2011; NICE, 2014a; Szatkowski et al., 2012). There are also lists of clinical codes published to ensure that different researchers identify the same cases for a medical condition or treatment, thereby improving reproducibility of results (ClinicalCodes.org, 2016). Moreover, there are widespread accepted methods to deal with missing data (Allison, 2001; van Buuren and Groothuis-Oudshoorn, 2011). All-cause mortality might be best recorded in primary care after the mortality register, because general practitioners must be informed that their patients have died (HSCIC, 2016a). Furthermore, because primary care data hold information on the comprehensive medical history and are not specific to a condition, new factors can be tested for that are not traditionally collected in secondary care, disease registers, or prospective trial-cohort studies. As these potential risk factors are already routinely recorded in primary care, there is greater usability of the developed risk models in routine clinical practice.

RCTs and cohort studies that prospectively collect data might be restricted in the generalisability of the results due to strict inclusion/exclusion criteria and the relatively small sample of medical centres participating in the study. Secondary data and disease registers might only be representative of severe cases. Primary care data might be the most representative of the UK population, because under the National Health Service (NHS), 99% of the UK population is registered at a general practice (NHS, 2013). General practitioners are informed when the patients are picked up by ambulances and/or enter secondary care (Hall, 2009). This means that primary care data would be representative of both mild and severe cases. Regarding CVD, after initial treatment in secondary care, the cardiac rehabilitation programme and follow-up will take place in primary care (Dalal et al., 2015). This means that the risk factors affecting survival prospects of CVD patients are recorded in primary care medical records. Primary care data thus ideally reflects current clinical practice by providing an almost complete picture of the medical history of all patients (Wijlaars, 2013a). The medical history is not 100% complete, as for example, the intake of over the counter medication is unknown (Wijlaars, 2013a). Although primary care medical records might be representative of the national population, the existing databases consist of medical records from approximately 6 to 10% of all general practices. Hence its representativeness may vary by clinical system or region, and therefore it is important to validate risk models on data from different general practice groups (Collins and Altman, 2012).

Data collection is costly for prospective studies because of the ongoing cost during the study period. The costly data collection can limit the sample size, the amount of data that is collected, and the length of the study period. Collecting routine data, on the other hand, has only a high initial cost of setting up the software and training clinicians to use the software, followed by low maintenance cost of validity checks and providing feedback on data quality to clinicians. Due to the lower cost of data collection, large study samples with rich, long follow-up data can be obtained relatively easily (Wijlaars, 2013a). Databases with routine data are frequently updated and therefore lend themselves to obtaining the most recent statistics (Wijlaars, 2013a). With an almost complete picture of the medical history of a patient in primary care data, new combinations of risk factors can be examined on the historic data. Furthermore, primary care data allow the selection of cases and controls from the same source population, whereas this tends to be a challenge for other data sources (Wijlaars, 2013a).

Thus, while each source of data has its strengths and limitations, primary care data are suitable for addressing the research objectives of estimating the effect of a history of acute myocardial infarction (AMI) on the hazard of mortality in UK residents and also estimating the potential survival benefits of primary and secondary treatments of CVD.

3.2 Primary care databases in the United Kingdom

There are three large primary care databases in the UK: QResearch; Clinical Practice Research Datalink (CPRD), which was previously called General Practice Research Database (GPRD); and The Health Improvement Network (THIN) (CPRD, 2016; IMS Health Incorporated, 2015b; QResearch, 2016), see Table 3.1. QResearch includes medical records from approximately 1,000 practices that use the EMIS clinical system, whereas CPRD and THIN include records from approximately 600 practices that use the Vision clinical system (Reeves et al., 2014). The records are from patients who have at some point been registered at the contributing general practice, and thus also include non-active patients who have transferred to another practice or died.

The clinical system in operation allows staff of a general practice to store an electronic record on the details of a patient's medical history and the care provided and planned, including appointments, symptoms, test results, diagnoses, and prescriptions (Department of Health, 2011). Clinical systems can provide templates for specific medical conditions or for certain risk groups as reminders of what information needs to be recorded during a consultation (Department of Health, 2011). For example in people aged 40 or older, the risk of a first cardiovascular event should be reassessed on a regular basis (NICE, 2015).

Table 3.1: Primary care databases and national surveys in the United Kingdom

Sources: CPRD, 2016; HSCIC, 2006, 2013, 2015, 2016b; Herrett et al., 2015; IMS Health Incorporated, 2015b,d; ONS, 2013, 2014; QResearch, 2016. [1]Previously known as General Practice Research Database (GPRD). [2]Previously known as General Lifestyle Survey (GLS).

	Qresearch	CPRD	THIN	QOF	GHS	HSE
Full name	Qresearch	Clinical Practice Research Datalink[1]	The Health Improvement Network	Quality and Outcomes Framework	General Household Survey[2]	Health Survey England
Type	Primary care database	Primary care database	Primary care database	Incentive payment scheme in primary care resulting in electronic medical records database	Repeated cross-sectional survey	Repeated cross-sectional survey
Study period	1993-ongoing	1987-ongoing	1987-ongoing	2004-ongoing	1971-2011	1991-ongoing
Study population	Volunteering general practices using EMIS clinical system in the United Kingdom	Volunteering general practices using Vision clinical system in the United Kingdom	Volunteering general practices using Vision clinical system in the United Kingdom	Volunteering general practices in England	Probability, stratified two-stage sample design in Great Britain	Multi-stage stratified random sample in England

Continued on next page

Table 3.1 – *Continued from previous page*

	Qresearch	CPRD	THIN	QOF	GHS	HSE
Sample size	Over 18 million patients from over 1,000 general practices (in 2015)	Over 11 million patients from 674 general practices (in 2015)	Over 12 million patients from 587 general practices (in 2015)	Over 56 million patients from nearly 8,000 general practices (in 2015)	18,367 individuals aged 16+ from 7,937 households (in 2011)	10,080 individuals (in 2014)
Data collection	Electronic medical records	Electronic medical records	Electronic medical records	Electronic medical records	Annual telephone and face-to-face interviews collecting data regarding education, employment and labour, health, housing, social indicators and quality of life, use and provision of specific social services	Annual face-to-face interviews collecting data on physical health, mental health and well-being, social care, lifestyle behaviours, and physical measures

The data from these databases made available to researchers are anonymised records on the patient's demographics, consultations, diagnoses, treatments offered, and lifestyle choices (Wijlaars, 2013b). The anonymisation inter alia includes making available only part of the date of birth, and postcode of a patient and the general practice at which the patient is registered. Historical paper records, which are typically scanned in, are not available to researchers. Free-text comments are sometimes made available to researchers (Wijlaars, 2013b). Free-text could include extra information that are not recorded using the structured Read codes, which could result in underestimation of diagnoses or treatments when free-text comments are not used by researchers (Nicholson et al., 2011). With the introduction of the QOF in 2004, the use of structured Read codes has increased (NICE, 2014a).

When medical records are updated or new ones added to the primary care database, non-clinical information such as inputs on demographics and registration are checked (IMS Health Incorporated, 2015c). These validity checks are reported to clinicians, so that they can improve recording, and to researchers so that they can adjust the analyses when needed. The validity of clinical information was investigated by systematic reviews and external validations of primary care research using data from QResearch, CPRD, THIN, General Household Survey (GHS), and Health Survey England (HSE) (Blak et al., 2011; Collins and Altman, 2009, 2012; Hall, 2009; Herrett et al., 2010; Hippisley-Cox et al., 2008; Jordan et al., 2004; Khan et al., 2010), see Table 3.1. Even though these three primary care databases include different general practices and patients, the estimated incidence, prevalence, and mortality rates are similar across the databases and with national surveys when adjusted for sex, age, and deprivation. Family history of medical conditions, lifestyle choices, and demographics are, however, not systematically recorded or reassessed in primary care (Hippisley-Cox and Coupland, 2010b; Hippisley-Cox et al., 2008; Marston et al., 2010; NICE, 2014a; Szatkowski et al., 2012). Recording of these factors is related to ill health; people who are ill, visit a general practice more often and general practitioners are more likely to ask these patients about their lifestyle compared to healthier patients (MacDonald and Morant, 2008; Shephard et al., 2011). Recording of lifestyle choices improved greatly after the introduction of the QOF in 2004. For example, from 2006 onwards the prevalence of smokers in THIN is comparable to that of the GHS, which is considered the gold standard with regards to prevalence of smoking (Langley et al., 2011; Szatkowski et al., 2012). Records on smoking status and alcohol consumption are self-reported in primary care databases. As these are not checked by biochemical data, there could be reporting bias towards the healthier choice (Langley et al., 2011; Szatkowski et al., 2012). Primary care does also not hold information on intake of drugs. This means that actual treatment levels can be lower than prescription records indicate, and hence the estimated effect of prescribed treatment on an outcome could be smaller than the actual effect of treatment itself on an outcome (MacDonald and Morant, 2008).

This research used THIN data, therefore it is useful to check the prevalence of medical conditions related to CVD in THIN against the other data sources. THIN had a slightly higher crude prevalence of medical conditions compared to QOF data from 2006/2007: 0.2% for ischaemic heart disease, 0.2% for chronic kidney disease, 0.1% for hypertension, and 0.8% for obesity (Blak et al., 2011). The prevalence of diabetes in THIN was 0.2% lower than in QOF in the same year (Blak et al., 2011). The differences in crude prevalence were likely due to the lack of adjustment for sex, age, and deprivation (Blak et al., 2011). Compared to HSE, the sex- and age-adjusted prevalence of diabetes in THIN for 1996-2005 was by 0.2% higher (Massó González et al., 2009). Another study compared the sex- and age-adjusted incidence of chronic kidney disease in THIN with that of QResearch for 2002-2008 and found similar results

(Hippisley-Cox and Coupland, 2010b). Prevalence of hypertension and hypercholesterolaemia were lower in THIN compared to the national rates of 1998, 2003 and 2006 (MacDonald and Morant, 2008). The rates from the different data sources converged over time and with increasing age (MacDonald and Morant, 2008). Studies report terminal digit bias towards zero in blood pressure records from general practices, hypertension clinics, clinical trials, and hypertension screenings (Ali and Rouse, 2002; Nietert et al., 2006; Thavarajah et al., 2003; Wingfield et al., 2002a,b). In THIN, the digit bias declined from 71% to 37% in systolic blood pressure and from 64% to 36% in diastolic blood pressure from 1996 to 2006, respectively (Harrison et al., 2007). This decline can partly be explained by the shift from manual mercury to digital sphygmomanometers. The digit bias seems to be associated with lower average blood pressure records, which could mean underdiagnosis of hypertension when it is solely based on blood pressure records.

Thus, although researchers do not have access to the complete medical history of a patient, the available data from primary care databases appear to be valid and accurate enough to use for model development.

3.3 The Health Improvement Network database

This research made use of data from THIN primary care database. The database was set up in 2003 by In Practice Systems Ltd (INPS) and Epidemiology and Pharmacology Information Core (EPIC) (IMS Health Incorporated, 2015b). INPS developed the clinical system Vision to store and maintain electronic medical records. The way information of a patient is recorded, is practically the same across the different operating clinical systems in the UK (Department of Health, 2011). EPIC collects the electronic medical records, checks the non-clinical information, anonymises the records, and adds the records to the database with postcode related indicators. The postcode-related indicators are ethnicity, socioeconomic factors such as deprivation, and environmental factors such as air quality (IMS Health Incorporated, 2015c). This information is derived from census data of England and Wales, and is thus not added to the medical records of patients living in Scotland or Northern Ireland. THIN data is structured by seven ASCII (American Standard Code for Information Interchange) standardised files, which are (Wijlaars, 2013b):

- Patient: data on demographics.
- Therapy: data on prescriptions.
- Medical: data on medical events.
- Additional Health Data (AHD): data on prevention, lifestyle and diagnostics.
- Postcode Variable Indicators (PVI): data on socioeconomic, ethnicity and environmental factors
- Consultation: data on location, time, and length of consultation.
- Staff: data on staff who entered the data.

Approximately 1,800 general practices make use of Vision, of which 587 practices contribute to THIN (IMS Health Incorporated, 2015d). This corresponds to almost 6% of all general practices in the UK contributing to THIN. When a practice signs up with THIN, previous and current medical records are imported to the database (IMS Health Incorporated, 2015b). After that, the already imported medical records are updated and new medical records are imported on a monthly basis (IMS Health Incorporated, 2015b). THIN consists of over 12 million patient records, of which almost 4 million are active, i.e. these patients are alive and still registered at the contributing general practices (IMS Health Incorporated, 2015d). There are roughly

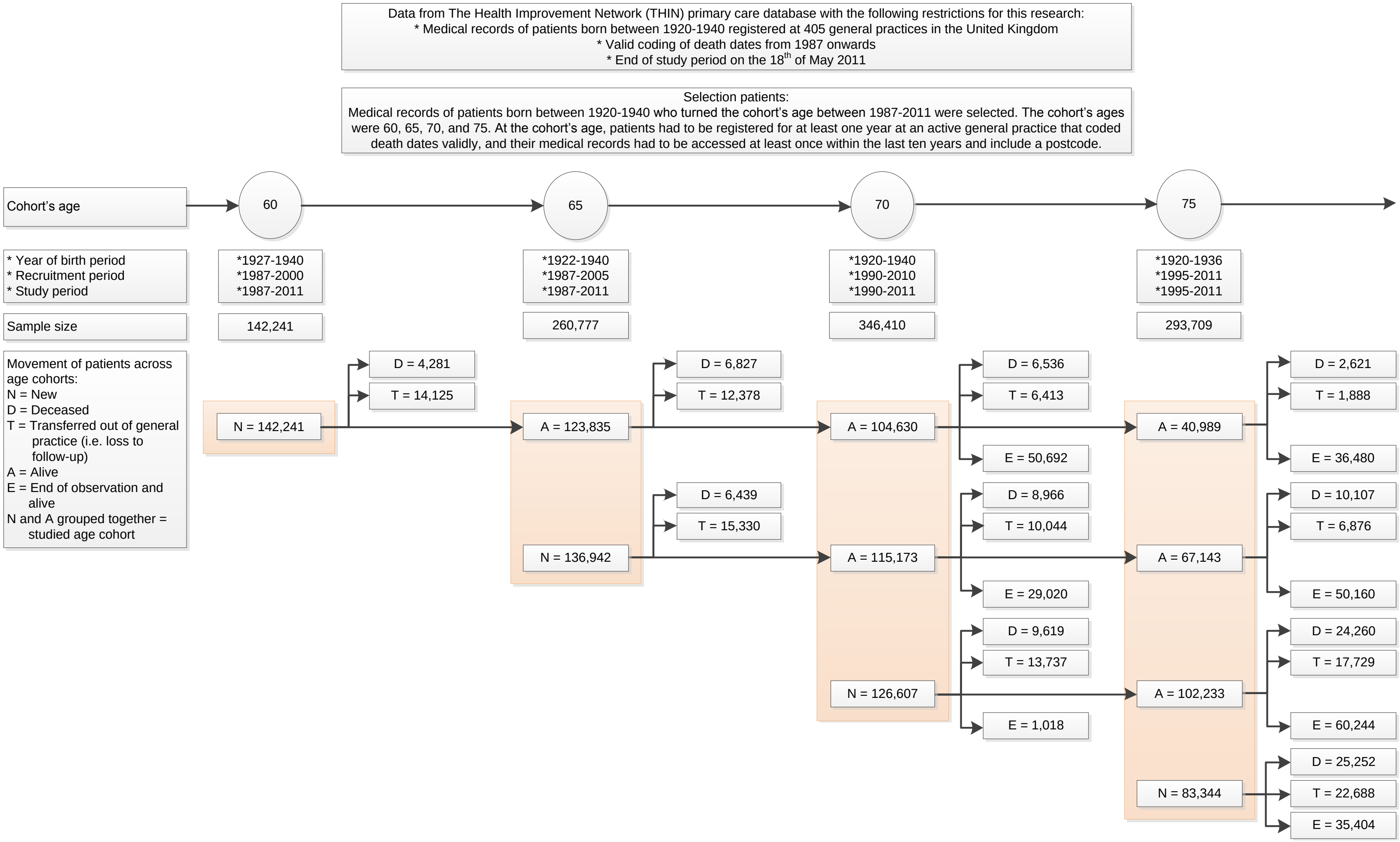
86 million patient-years of data, of which 73 million are after the date that general practices started reporting valid mortality rates (IMS Health Incorporated, 2015d). THIN holds an approximately equal amount of medical records from men and women, 48 and 52%, respectively (IMS Health Incorporated, 2015d).

The patients included in THIN are representative of the UK with regards to demographics, prevalence of medical conditions, and mortality rates (Blak et al., 2011; Hall, 2009). THIN database follows the sex and age distribution of the UK population, although it includes slightly fewer people aged younger than 25 (Blak et al., 2011), which is not a problem for this research as it focusses on people around retirement age. THIN database also includes more people from affluent areas than deprived ones (Blak et al., 2011). It is therefore important to include sex, age, and deprivation in model development to obtain representative estimates of the UK population (Blak et al., 2011; Hippisley-Cox and Coupland, 2010b; MacDonald and Morant, 2008; Massó González et al., 2009).

3.4 Selected age cohorts

Access to THIN data can be through sub-licence, data extracts, and summaries and reports (IMS Health Incorporated, 2015a). The data cut for this research was restricted to medical records up to the 18th of May 2011 of patients born between 1920 and 1940 from 405 general practices in the UK. Four age cohorts of patients were selected who turned the target ages (60, 65, 70, and 75) by 2011, see Figure 3.1. Patients were selected when at the cohort's age they were registered for at least one year at an active general practice that coded death dates validly and had no data issues. Furthermore, the medical records had to include a postcode and be accessed at least once within the last ten years. These inclusion criteria were specified to ensure that the patient's full medical record was available, up-to-date, and valid (Ashman et al., 2012; Collins and Altman, 2012). EPIC provided flags on practice level to report issues in data recording and the year when acceptable mortality rate (AMR) recording was achieved (Research, 2011). Data issues could be due to gaps or limited recording before conversion to Vision, missing consultations in the medical or additional health data files, or that the general practice split, merged, or changed user number. The AMR date ensures that general practices reported mortality as expected for the demographic structure of their patient population (Blak et al., 2011; Maguire et al., 2009). Between 1987 and 2000, all general practices achieved validated death dates. Thus, the data available were medical records from 1987 to 2011 of patients born between 1920 and 1940. These constraints resulted in a recruitment period of 13, 18, 20, and 16 years, and a study period of 24, 24, 21, and 16 years for the 60-, 65-, 70-, and 75-year old cohorts, respectively. The length of the recruitment period and the calendar years included affected the sample size of the age cohort. More medical records were available at later years because of AMR dates and inclusion of new general practices. The sample sizes ranged between 140,241 and 346,410 patients per age cohort. The selection of medical records was performed in SQL 2012.

Patients observed for more than five years were part of multiple cohorts, see Figure 3.1. Patients who transferred out of a general practice were lost to follow-up. When a patient transferred from one general practice to another that contributed to THIN database, two medical records of the same patient would exist. These medical records could not be linked due to THIN's privacy preservation policy. The inclusion of new patients at the older age cohorts was due to eligibility of general practices and patients, more specifically due to the AMR dates, the year of birth period, and the recruitment period.



Data from The Health Improvement Network (THIN) primary care database with the following restrictions for this research:
* Medical records of patients born between 1920-1940 registered at 405 general practices in the United Kingdom
* Valid coding of death dates from 1987 onwards
* End of study period on the 18th of May 2011

Selection patients:
Medical records of patients born between 1920-1940 who turned the cohort's age between 1987-2011 were selected. The cohort's ages were 60, 65, 70, and 75. At the cohort's age, patients had to be registered for at least one year at an active general practice that coded death dates validly, and their medical records had to be accessed at least once within the last ten years and include a postcode.

Movement of patients across age cohorts:
N = New
D = Deceased
T = Transferred out of general practice (i.e. loss to follow-up)
A = Alive
E = End of observation and alive
N and A grouped together = studied age cohort

Figure 3.1: Selected age cohorts

3.5 Selected medical history

At the cohort's age, a snapshot of the patient's medical history was obtained, after which it was observed whether and when the patient died from any cause during follow-up. Patients were censored when they transferred out of their general practice, as the death date could not be observed. Due to THIN's privacy preservation policy, only the year of birth of the patient was known. Therefore, the snapshot of the patient's medical history of relevant medical, lifestyle, and socio-demographic information was taken at the first of January of the year the patient turned the cohort's age. The selected information are listed in Table 3.2. Selection was based on the literature review discussed in the previous Chapter.

Table 3.2: Information selected from medical records contributing to The Health Improvement Network (THIN) primary care database

[1]The read codes in The Health Improvement Network (THIN) primary care database are comparable to the ICD-10 (International Classification of Disease 10th revision) codes. [2]BNF=British National Formulary. [3]All values represented the latest reading before entering the study, which is the first of January of the year the patient turned the cohort's age of 60, 65, 70, or 75.

Category	Covariate[3]
Medical conditions[1]	Acute myocardial infarction diagnosis (ICD-10 I21)
	Angina pectoris diagnosis (ICD-10 I20)
	Cerebrovascular disease diagnosis (ICD-10 I60-69)
	Chronic kidney disease end stage (ICD-10 N18.5)
	Diabetes mellitus diagnosis (ICD-10 E10-14)
	Family history of acute myocardial infarction (ICD-10 Z82.4)
	Family history of cardiovascular disease (ICD-10 Z82.3 or Z82.4)
	Heart failure diagnosis (ICD-10 I50)
	Other cardiovascular system disorders diagnosis (ICD-10 I95-99)
	Peripheral vascular disease diagnosis (ICD-10 I73)
	Valvular heart disease diagnosis (ICD-10 I34-39)
Treatments[2]	ACE inhibitors prescription (BNF Chapter 2.5.5.1 and 2.5.5.2), which include: angiotensin-converting enzyme inhibitors and angiotensin-II receptor antagonists
	Aspirin prescription (BNF Chapter 2.9.1)
	Beta adrenoceptor blockers prescription (BNF Chapter 2.4)

Continued on next page

Table 3.2 – *Continued from previous page*

Category	Covariate[3]
	Calcium-channel blockers prescription (BNF Chapter 2.6.2)
	Coronary revascularisation, which include coronary artery bypass graft and coronary angioplasty
	Statins prescription (BNF Chapter 2.12), which include: atorvastatin, cerivastatin, fluvastatin, pravastatin, rosuvastatin, and simvastatin
Clinical measurements & lifestyle choices	Alcohol consumption status, measured as non/ex/current
	Diastolic blood pressure in mmHg
	Height in meters
	Hypercholesterolaemia diagnosis
	Hypertension diagnosis
	Smoking status, measured as non/ex/current
	Systolic blood pressure in mmHg
	Total cholesterol reading in mmol/L
	Weight in kilograms
Demographics - patient level	Sex
	Year of birth
Demographics - postcode level	Air pollution, which includes separate covariates for 2001 estimated level of nitrogen dioxide, nitrogen oxide, sulphur dioxide, and particulate matter, measured in quintiles
	Ethnicity, which includes separate covariates for proportion of district population defining themselves as white, mixed, Asian or Asian British, black or black British, and other, measured in quintiles
	Index of multiple deprivation (IMD), measured in quintiles
	Limiting long-term illness, health problem, or disability of which proportion of district population defines as limiting daily activities, measured in quintiles
	Mosaic, which is a consumer classification that captures demographics, lifestyles and behaviour of a person, see Table 3.3 for classification
	Townsend deprivation score, measured in quintiles
	Urbanisation: (1) urban, (2) town and fringe, or (3) village, hamlet and isolated dwelling

The selected medical conditions were: type of CVD (AMI, angina, cerebrovascular disease, heart failure, peripheral vascular disease, and other cardiovascular system disorders), family history of CVD and specifically of AMI, chronic kidney disease, and diabetes. As a history of AMI is of specific interest to this research, the number of events was recorded. To separate multiple events, it was specified that the interval between events should be at least 30 days.

The treatments investigated were based on the UK National Institute of Health and Care Excellence (NICE) and the British National Formulary (BNF) recommended treatments for primary and secondary prevention of CVD during the study period, which include: coronary revascularisation and prescription of ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, and statins (Joint Formulary Committee, 2016b; NICE, 2013b). Coronary revascularisation consisted of coronary artery bypass graft and coronary angioplasty.

The selected clinical measurements and lifestyle choices were: systolic and diastolic blood pressure, hypertension, total cholesterol reading, hypercholesterolaemia, alcohol consumption status, height, weight, and smoking status. The ideal blood pressure is 120/80 mmHg, i.e. a systolic blood pressure (SBP) of 120 mmHg and a diastolic blood pressure (DBP) of 80 mmHg (NHS, 2014b). Hypertension is diagnosed when multiple blood pressure readings are 140/90 mmHg or higher, more specifically when SBP is $\geq$140 or DBP is $\geq$90 (NICE, 2011). The ideal total cholesterol level is below 5 mmol/L and is recommended to be below 4 mmol/L in patients with CVD (Thompson et al., 2008). Hypercholesterolaemia is diagnosed when the cholesterol reading is above 7.5 mmol/L, although treatment is initiated when total cholesterol is above 5 mmol/L (NICE, 2015; Thompson et al., 2008). Body mass index is a weight to height measurement and is calculated as weight(kg)/height(m)^2 (WHO, 2006).

The selected socio-demographic information at patient level were: sex and year of birth. The selected socio-demographic information at the postcode level were: air pollution, ethnicity, index of multiple deprivation, Mosaic, limiting long-term illness, and urbanisation. Apart from Mosaic, the postcode indicators were derived from census data of England and Wales, and therefore not available for patients living in Scotland or Northern Ireland. The postcode indicators were selected to examine possible survival variations by general practice in more detail for the respective patients.

Mosaic was selected as a measure of socioeconomic status to adjust for in estimating survival prospects of patients with a possible history of CVD. Mosaic is a consumer classification that captures demographics, lifestyles and behaviour of a person (Experian Ltd., 2009). It is developed by Experian, which is an information services company (Experian Ltd., 2009). Mosaic consists of 67 categories that are classified in 15 groups, see Table 3.3. Mosaic codes are assigned to each UK postcode, which entails circa fifteen households (Experian Ltd., 2009). Mosaic classification is based on information from 440 covariates. The information comes from multiple sources: 38% is from the UK decennial census and the remaining 62% is from the Experian's UK Consumer Dynamics Database (Experian Ltd., 2009). The latter is built on a number of data sources such as the Electoral Roll, Council Tax property values, house sale prices, and self-reported lifestyle surveys. Not all sources are known nor the exact calculation of Mosaic classifications due to commercial considerations. Although this makes it hard to verify the validity of Mosaic, studies have shown that Mosaic performs well as a measure of socioeconomic status (Douglas and Szatkowski, 2013; Sharma et al., 2010; Szatkowski et al., 2012). Other measures of socioeconomic status used in the UK are index of multiple deprivation (IMD), Townsend, and Carstairs (UKDS, 2006). These measures are country specific whereas Mosaic scores are available for the whole of the UK. Another advantage of Mosaic over the other measures of socioeconomic status is that it is based on a greater variety and amount of data and distinguishes between people to a greater extent, resulting in a higher accuracy in capturing social disparities (Douglas and Szatkowski, 2013; Sharma et al., 2010). Finally, actuaries make use of Mosaic in setting out pensions and annuity portfolios as it was found to be one of the main contributors in explaining longevity variations in pensioner mortality data next to sex, age, and policy size (Richards, 2008; Ridsdale

Table 3.3: Mosaic classification

A consumer classification that captures demographics, lifestyles and behaviours of a person (Experian Ltd., 2009). [1]Due to the low prevalence of these categories in the age cohorts, they were merged together to increase precision in the estimates of the survival models.

Category	Description
Alpha territory	Most wealthy and influential individuals
Professional rewards	Executive and managerial classes
Rural solitude	People who live in small villages
Small town diversity	People who live in medium sized and smaller towns
Careers and kids	Young couples, married or living with their partner
New homemakers	Neighbourhoods containing mostly houses that were built in the last five years
Ex-council community	People who are practical and enterprising
Claimant cultures	Some of the most disadvantaged people
Upper floor living	People who are on limited incomes
Active retirement[1]	People aged over 65
Suburban mindsets[1]	People of middle age living together with their children in family houses
Elderly needs[1]	Older pensioners who can no longer easily manage household responsibilities
Industrial heritage[1]	People who are traditional and conservative, approaching retirement age
Terraced melting pot[1]	People who work in relatively menial, routine occupations and are poorly educated
Liberal opinions[1]	Young, professional, well educated people

and Gallop, 2010).

Indicators of psychosocial factors such as job strain and lack of social support, fruit and vegetable intake, and physical activity were not included in the analysis because THIN does not provide this information.

3.6 Coding of covariates

The prevalence of each medical, lifestyle, and socio-demographic covariate was calculated. Where appropriate, covariates or categories of covariates that were relatively rare, approximately a prevalence of less than 5%, were merged together. This was done to increase statistical power and precision in the estimates (Greenland and Morgenstern, 1990; Hennekens et al., 1987).

With respect to the medical covariates, subtypes of CVD were transformed. A history of AMI was categorised as having had no, single, or multiple events. This categorisation was chosen because previous studies found different survival prospects for patients who had single compared to multiple events, or adjusted for recurrent events when estimating survival variations in AMI patients (Gerber et al., 2009; Gottlieb et al., 2000; Kirchberger et al., 2014; Nigam et al., 2006; Smolina et al., 2012b; Vaccarino et al., 1999). The exact number of events was not used, because only 5% of the patients who had a history of AMI had more than two events. Different subtypes of CVD were merged together due to the low prevalence in the youngest cohorts, which were: diagnosis of cerebrovascular disease, peripheral vascular disease, valvular heart disease, and other cardiovascular system conditions. The prevalence of these types were 3.5, 2.4, 0.9, and 4.6% at age 60, respectively, and 10.0, 6.7, 2.8, and 13.2% at age 75, respectively. The prevalence of the medical conditions combined as 'cardiovascular system conditions' was 9.9% at age 60 and 26.0% at age 75. The CVD subtype heart failure was not included in 'cardiovascular system conditions' even though its prevalence was less than 5% in the youngest cohorts. This was because heart failure is directly related to prescription of ACE inhibitors, which is a first line treatment for AMI patients and was studied for this research (NICE, 2013a). ACE inhibitors are prescribed to reduce the risk of a next cardiac event, in particular heart failure.

With respect to the socio-demographic covariates, year of birth and Mosaic were transformed. Year of birth was categorised in five year intervals for the matching procedure, which is explained in Section 4.2.1 in Chapter 4. The Mosaic groups 'active retirement', 'suburban mindsets', 'elderly needs', 'industrial heritage', 'terraced melting pot', and 'liberal opinions' were combined as 'other', see Table 3.3. These categories were merged together because the prevalence of the former two categories was less than 5% and the latter four categories less than 1% in the age cohorts.

With respect to the lifestyle covariates, body mass index (BMI) and alcohol consumption status were transformed. BMI values smaller than 15 and greater than 50 were removed to deal with extreme values that were likely due to typographical errors. BMI was categorised to reflects its non-linear effect on the hazard of mortality in older patients more accurately; overweight patients have a better survival rate compared to healthy weight or obese patients (Chapman, 2010; Gerber et al., 2009; Nigam et al., 2006; Yusuf et al., 2004). BMI was categorised into underweight/normal weight (BMI<25), overweight (25≤BMI<30), and obese (30≤BMI) (WHO, 2006). Underweight was grouped with normal weight because its prevalence was less than 1% in the age cohorts. With alcohol consumption status, non- and ex-consumers were combined as 'not current consumer', because the prevalence of ex-consumers was less than 2% in the age cohorts.

Chapter 4

Review of statistical methods

This Chapter is a review of the statistical methods concerning modelling incomplete survival data from primary care medical records. Chapter 2 reviewed survival models of cardiovascular disease (CVD) and revealed some gaps in model development with respect to addressing the model assumptions and selecting covariates to adjust for in the model. Chapter 3 reviewed primary care data and revealed that not all medical records are complete with respect to the covariates of interest for this research. This Chapter starts with discussing survival analysis in general and the type of survival model used for this research. Next, the model building is explained that involves the study design and selection of covariates in the final model. Then, methods to deal with missing data are described. Finally, the assessment of the final models is discussed.

4.1 Survival analysis

Survival analysis is the field of statistics concerned with analysing the length of time until occurrence of an event (Kleinbaum and Klein, 2011). For this research, the event of interest is death. Although the survival analysis techniques were primarily developed in the medical and biological sciences, they are now widespread across the sciences. Only the techniques applicable for this research are described in this

Section. First survival time is introduced, second descriptive analysis techniques are reviewed, and third regression analysis techniques are discussed. The main focus is on the regression analysis techniques.

4.1.1 Survival time

Survival time is the time until the occurrence of an event (Kleinbaum and Klein, 2011). Let T be a non-negative continuous random variable representing survival time and t be a specific value for T. The probability function of survival time T can be stated in terms of the survival function, the hazard function, and the cumulative hazard function (Kleinbaum and Klein, 2011).

The survival function is the probability that the survival time T is greater than an arbitrary point in time t. This is denoted as (Kleinbaum and Klein, 2011):

$$S(t) = P(T > t). \tag{4.1}$$

The survival function is a non-increasing function with $S(0)$=1 since at the start of survival time there has not been an event.

The hazard function, also known as the conditional hazard rate, is the probability that an event occurs within the time interval of $t + \Delta t$, given survival to time t. This is denoted as (Kleinbaum and Klein, 2011):

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t}. \tag{4.2}$$

The hazard function is a non-negative function with no upper bound. In terms of the survival function, the hazard function can be written as the probability density function divided by the probability that an event has not occurred before time t (Kleinbaum and Klein, 2011):

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-\frac{dS(t)}{dt}}{S(t)} = \frac{-d\log(S(t))}{dt}. \tag{4.3}$$

The cumulative hazard function $\Lambda(t)$ is the integral of the hazard function, i.e. the 'sum total' of the hazards up to time t (Kleinbaum and Klein, 2011):

$$\Lambda(t) = \int_0^t \lambda(t)dt = -\log(S(t)). \tag{4.4}$$

Survival time could be incomplete, where the exact timing of the event is unknown. Presence of incomplete survival times is called censoring. There is left, interval, and right censoring (Hosmer et al., 2008). With left censoring, the event happened before the study started and thus the actual survival time is shorter than the observed one. With interval censoring, the event happened within a time interval with the exact time unknown. With right censoring, the event did not happen before or during the study period, and thus the actual survival time is longer than the observed one. Right-censoring is the most common form of censoring. The main causes of right-censoring are that the subject was no longer part of the study or the study ended before the subject could have the event (Kleinbaum and Klein, 2011). In this research, there was right-censoring data, where the censoring was due to patients transferring to another general practice during the study period or remaining alive at the end of the study period.

4.1.2 Descriptive analysis

In the descriptive analysis, survival variations are assessed between subgroups as specified by the covariate of interest. The most popular descriptive analysis techniques include the Kaplan-Meier estimator, median survival time, and comparison of survival functions (Kleinbaum and Klein, 2011). For this research, the goal of the descriptive analysis was to obtain an overview of the data, to check the coding of the covariates, and to be informed about the possible directions and effect sizes associated with the covariates estimated in the regression analysis.

Let t stand for the observed death times that are sorted in ascending order; d_i for the number of deaths at t_i; and n_i for the number at risk of dying at t_i. The Kaplan-Meier estimator estimates the survival function as (Kaplan and Meier, 1958):

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{4.5}$$

The Kaplan-Meier estimator $\hat{S}(t)$ is a right-continuous step function with jumps at death times. Not all subjects' deaths are observed during the study period, which means they are censored at the last observation. These subjects are included in the number at risk of dying until censoring, after which they are no longer included in the survival function. The variance of the Kaplan-Meier estimator $\hat{S}(t)$ is estimated by Greenwood's formula (Greenwood, 1926):

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{4.6}$$

The Kaplan-Meier estimator $\hat{S}(t)$ is asymptotically normally distributed. Let z_α stand for the critical value at α level from the standard normal distribution, then the confidence intervals for estimated survival times are calculated as (Kleinbaum and Klein, 2011):

$$\hat{S}(t) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))}. \tag{4.7}$$

The estimated median survival time is the estimated survival time at which the cumulative survival function falls to 0.5 (Kleinbaum and Klein, 2011). The median rather than the mean survival time is estimated because the distribution is often highly skewed.

With the comparison of survival functions, the null hypothesis that the hazard rates are the same across groups is tested. This involves comparing the observed and expected number of deaths at each time a death is observed in a subgroup of subjects as specified by the covariate of interest (Kleinbaum and Klein, 2011). With the log-rank test, equal weights are given to each time a death is observed, i.e. all deaths during the study period are of equal importance.

Let i stand for a group of subjects, j for the observed death times that are sorted in ascending order, O_j for the observed number of deaths across groups, N_j for the number of subjects at risk of death across groups, and N_{ij} for the number of subjects at risk of death in group i at time j. Then, the expected number of deaths for group i at ordered observed death time j is calculated as $E_{ij}{=}O_jN_{ij}/N_j$. For a binary covariate, the log-rank test Z is defined as (Peto et al., 1977):

$$Z = \frac{\sum\limits_{j=1}^{J}(O_{1j} - E_{1j})}{\sqrt{\sum\limits_{j=1}^{J} V_j}}, \tag{4.8}$$

where

$$V_j = \frac{O_j(N_{1j}/N_j)(1 - N_{1j}/N_j)(N_j - O_j)}{N_j - 1}. \tag{4.9}$$

For a covariate with $K > 2$ categories, the test statistic is calculated as $u'V^{-1}u$, where u is the vector of observed minus expected deaths for each category, defined as:

$$u = (u_1, ..., u_K)^T, \tag{4.10}$$

where

$$u_k = \sum_{j=1}^{J}(O_{kj} - E_{kj}), k = 1, ..., K. \tag{4.11}$$

And V is the covariance matrix, defined as:

$$V_{il} = \sum_{j=1}^{J} \frac{O_j(N_{1j}/N_j)(N_j - O_j)}{N_j - 1} \left(\delta_{il} - \frac{N_{ij}}{N_j} \right), \tag{4.12}$$

with δ_{il}=1 for i=l and δ_{il}=0 otherwise, and i and l goes from 1 to k-1.

The log-rank test follows a χ^2-distribution with k-1 degrees of freedom. The log-rank test might not be informative when the survival functions of subgroups of subjects are not proportional to each other over time. The proportionality of the survival functions can visually be assessed by plotting the Kaplan-Meier estimators.

4.1.3 Regression analysis

Unlike the descriptive analysis, the regression analysis can assess the effects of the covariates associated with survival time simultaneously (Hosmer et al., 2008). This means that adjusted effects can be obtained, which may result in more precise estimation of the true effects. For this research a Cox's proportional hazards regression model was used, which regresses the hazard function on the covariates. The Cox's model estimates the hazard $\lambda_i(t)$ for subject i by multiplying the baseline hazard function $\lambda_0(t)$ by the subject's risk score r_i (Cox, 1972):

$$\lambda_i(t, \beta, X_i) = \lambda_0(t)\, r_i(\beta, X_i) = \lambda_0(t)\, e^{\beta X_i}. \tag{4.13}$$

The risk score is dependent on the values for the multiple covariates X and their coefficients β. Taking a ratio of the hazard functions for two subjects i and j who differ in one covariate x and not in the other covariates, the coefficient β_x or the hazard ratio (HR) e^{β_x} of that particular covariate x can be calculated (Hosmer et al., 2008):

$$\lambda(t, \beta, X) = \frac{\lambda_i(t, \beta, X)}{\lambda_j(t, \beta, X)} = \frac{\lambda_0(t) e^{\beta X_1}}{\lambda_0(t) e^{\beta X_0}} = \frac{e^{\beta_x x_1}}{e^{\beta_x x_0}} = e^{\beta_x (x_1 - x_0)}. \tag{4.14}$$

This means that the baseline hazard $\lambda_0(t)$ does not have to be specified and that the HR is constant with respect to time t. In other words, the Cox's model does not make any assumptions about the shape of the baseline hazard function, but does assume proportional hazards for the covariates over time t. Evaluation of the proportional hazards assumption is described later in this subsection.

An adjusted HR is interpreted as the average increased or decreased risk of mortality when belonging to one category compared to the baseline category of a covariate during the length of the study period while adjusting for the other covariates (Kleinbaum and Klein, 2011). If the HR is smaller than one, thus the estimated coefficient $\hat{\beta}$ is negative, then there is a decreased hazard of mortality and a longer survival time. If the HR is greater than one, thus the estimated coefficient $\hat{\beta}$ is positive, then there is an increased hazard of mortality and a shorter survival time. If the HR is equal to one, thus the estimated coefficient $\hat{\beta}$ is zero, then the covariate is not associated with the hazard of mortality and survival time.

An adjusted HR can be translated to the numbers of years lost or gained in effective age by dividing the log HR by the log increase in annual hazard of mortality associated with ageing one year in the population (Brenner et al., 1993). As showed by Gompertz model applied to numerous populations over time, the increase in annual hazard of mortality associated with ageing one year is approximately constant between ages 30 and 95 (Brenner et al., 1993; Vaupel, 2010). For England and Wales in 2010-2012, the increase in the hazard between those ages was approximately 1.1 (Spiegelhalter, 2016). Subsequently, the number of years lost or gained in effective age δt is the log HR divided by log 1.1:

$$\delta t \approx \frac{\log HR}{\log 1.1}. \tag{4.15}$$

The underlying assumptions of this calculation are that the increase in annual hazard of mortality associated with ageing one year is constant with age and that there are proportional hazards for the covariates over time (Brenner et al., 1993). The first assumption will most likely hold for this research, because the cohorts studied consisted of people aged 60, 65, 70 or 75. Evaluation of the proportional hazards assumption is described later in this subsection.

Using the UK life tables of 2010-2012 (ONS, 2016), the number of years lost or gained in effective age associated with the covariates can be translated into the average period expectation of life at the effective age. The average period expectation of life at the cohort's ages of 60, 65, 70, and 75 for men are 22, 18, 14, and 11 years, respectively, and for women are 25, 21, 17, and 13 years, respectively. Covariates that are associated with an increased hazard of mortality translate in a higher effective age and therefore a lower period expectation of life. In contrast, covariates that are associated with a decreased hazard of mortality translate to a lower effective age and therefore in a higher period expectation of life.

The coefficients β are estimated by maximising the partial log likelihood of the Cox's model (Hosmer et al., 2008). The likelihood is called partial, as it is only based on the m subjects who had the event of interest d during the study period instead of all n subjects who might or might not have had the event of interest d. The partial likelihood function is the product of the i^{th} subject's probability of dying at time t instead of the other subjects j in the risk set R that comprises of subjects who die at the specified time or later during the study period. With distinct survival times sorted in ascending order, the partial likelihood function is given by (Hosmer et al., 2008):

$$PL(\beta) = \prod_{i=1}^{m} \frac{e^{\beta X_i}}{\sum_{j \in R(t_i)} e^{\beta X_j}}, \tag{4.16}$$

The partial log likelihood is (Hosmer et al., 2008):

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} d_i \left(\beta X_i - \log \left(\sum_{j \in R(t_i)} e^{\beta X_j} \right) \right) \\ &= \sum_{i=1}^{m} \left(\beta X_i - \log \left(\sum_{j \in R(t_i)} e^{\beta X_j} \right) \right). \end{aligned} \tag{4.17}$$

The partial log likelihood is maximised by setting its derivative to zero and solving it for the unknown coefficients β (Hosmer et al., 2008). The derivative is also called the score vector, which is calculated as (Hosmer et al., 2008):

$$\begin{aligned} U(\beta) = \frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{m} \left(X_i - \frac{\sum_{j \in R(t_i)} X_j e^{\beta X_j}}{\sum_{j \in R(t_i)} e^{\beta X_j}} \right) \\ &= \sum_{i=1}^{m} \left(X_i - \sum_{j \in R(t_i)} w_{ij}(\beta) X_j \right) \\ &= \sum_{i=1}^{m} (X_i - \bar{X}_{w_i}). \end{aligned} \tag{4.18}$$

The variance of the estimated coefficients is calculated by the inverse of the negative second derivative of the partial log likelihood function. The negative second derivative is called the Fisher's Information Matrix, which is denoted as (Hosmer et al., 2008):

$$\begin{aligned} I(\beta) = -\frac{\partial l^2(\beta)}{\partial \beta^2} &= \sum_{i=1}^{m} \left(\frac{\left(\sum_{j \in R(t_i)} e^{\beta X_j} \right) \left(\sum_{j \in R(t_i)} X_j^2 e^{\beta X_j} \right) - \left(\sum_{j \in R(t_i)} X_j e^{\beta X_j} \right)^2}{\left(\sum_{j \in R(t_i)} e^{\beta X_j} \right)^2} \right) \\ &= \sum_{i=1}^{m} \sum_{j \in R(t_i)} w_{ij}(\beta)(X_j - \bar{X}_{w_{ij}})^2. \end{aligned} \tag{4.19}$$

By taking the inverse of the Fisher's Information Matrix, the variance of the estimated coefficients is obtained:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = I(\hat{\beta})^{-1}. \tag{4.20}$$

By means of the Newton-Raphson algorithm the coefficients are obtained through n iterations as (Therneau and Grambsch, 2000):

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + \widehat{\mathrm{Var}}(\hat{\beta}^{(n)})U(\hat{\beta}^{(n)}). \tag{4.21}$$

This iterative process starts with the estimated coefficients set to zero; $\hat{\beta}^{(0)} = 0$. The end point is convergence in the estimated coefficients, which is when there is stability in the partial log likelihood; $l(\hat{\beta}^{(n+1)}) \approx l(\hat{\beta}^{(n)})$.

To test the significance of the model, i.e. whether the model with the included covariates can estimate a subject's survival time more accurately than taking the average survival time of the whole sample, the likelihood ratio test statistic is calculated. This statistic is calculated as (Hosmer et al., 2008):

$$LRT_{\beta} = 2\left(l(\hat{\beta}) - l(\beta^{(0)})\right), \tag{4.22}$$

where $\beta^{(0)}$ stands for the initial values of the coefficients, and $\hat{\beta}$ for the estimated coefficients. The initial values of the coefficients are set to zero as the null hypothesis is that the covariates are not associated with the hazard of mortality. Therefore, the likelihood ratio test statistic is essentially twice the difference in the partial log likelihood of a full and an empty model. The likelihood ratio test statistic follows a χ^2-distribution with the degrees of freedom equal to the difference in the number of estimated coefficients by the two models (Hosmer et al., 2008).

To test the significance of a specific covariate x, i.e. whether the covariate is significantly associated with survival time, the Wald statistic z is calculated. With this statistic, it is tested whether the covariate's estimated coefficient $\hat{\beta}_x$ is significantly different from zero. The statistic z is calculated as (Hosmer et al., 2008):

$$z = \frac{\hat{\beta}_x}{\widehat{\mathrm{se}}(\hat{\beta}_x)}, \tag{4.23}$$

where se stands for the standard error and is calculated as the square root of the variance of the estimated coefficient x. The Wald statistic z follows a standard normal distribution (Hosmer et al., 2008). The confidence interval (CI) of the estimated coefficient $\hat{\beta}_x$ is calculated as (Hosmer et al., 2008):

$$\mathrm{CI}(\hat{\beta}_x) = \hat{\beta}_x \pm z_{1-\alpha/2}\widehat{\mathrm{se}}(\hat{\beta}_x), \tag{4.24}$$

where z_α stands for the critical value at a specific α level from the normal distribution. If the confidence interval for the estimated coefficient does not include zero, then the covariate is associated with a significant increased or decreased hazard of mortality.

The regression techniques of the basic Cox's model, as described above, can be extended to fit various scenarios. For this research, the assumptions of no time-tied events, uninformative censoring, homogeneous population, and proportional hazards were addressed.

Assumption of no time-tied events

The partial log likelihood of the basic Cox's model, as described above, assumes that survival time is continuous and thus that there are no time-tied events (Therneau, 2014). The data used for this research included death dates, making survival time discrete. Multiple deaths could be observed on the same day in which the order of the deaths is unknown, resulting in time-tied events.

There are three main methods to deal with time-tied events in the Cox's model: Breslow's or Efron's approximation of the partial log likelihood or the exact partial log likelihood (Therneau, 2014). The three methods are explained by means of an example.

Let there be three subjects in the risk set R and each has the event at a different time. Then based on Equation (4.16), the partial log likelihood for the subject who first had the event would be:

$$PL_1 = \frac{e^{\beta X_1}}{\sum_{j \in R(t_i)} e^{\beta X_j}} = \frac{r_1}{r_1 + r_2 + r_3}. \tag{4.25}$$

Now let the first time of an event be a time-tied event for two subjects. Then the partial log likelihood for the first two subjects with the events could be one of the following if time was continuous:

$$\begin{aligned} PL_{1\&2} &= \left(\frac{r_1}{r_1 + r_2 + r_3}\right)\left(\frac{r_2}{r_2 + r_3}\right), \\ PL_{1\&2} &= \left(\frac{r_2}{r_1 + r_2 + r_3}\right)\left(\frac{r_1}{r_1 + r_3}\right). \end{aligned} \tag{4.26}$$

It is unknown which partial log likelihood is the correct one. Breslow's and Efron's approximations estimate the average of the two possible partial log likelihoods. With Breslow's approximation, the fractions with the largest risk pool for each time-tied event are used (Breslow, 1972). The partial log likelihood for the two subjects would thus be:

$$PL_{1\&2-Breslow} = \left(\frac{r_1}{r_1 + r_2 + r_3}\right)\left(\frac{r_2}{r_1 + r_2 + r_3}\right). \tag{4.27}$$

Generalising this to the whole sample, Breslow's approximation can be written as (Hosmer et al., 2008):

$$PL(\beta)_{Breslow} = \prod_{i=1}^{m} \frac{e^{\beta X_{i+}}}{\sum_{j \in R(t_i)} (e^{\beta X_j})^{d_i}}, \tag{4.28}$$

where d_i is the number of events at time t_i, $D(t_i)$ is the set of subjects who have the event at time t_i, and X_{i+} is the summed values for the covariates, $X_{i+} = \sum_{j \in D(t_i)} X_j$.

The problem with Breslow's approximation is that d_i-1 subjects are included in the denominator too many times. The more subjects have the event at the same time, the less accurate Breslow's approximation is. This leads to a conservative bias and thus underestimation of the coefficients (Cox and Oakes, 1984).

A more accurate approximation of the partial log likelihood was proposed by Efron (1977). With his approximation, the risk scores of the subjects with time-tied events in subsequent risk sets are multiplied by the probability they would be in the subsequent risk set (Hosmer et al., 2008). Coming back to the previous example, with Efron's approximation, the risks of the two subjects with time-tied events are multiplied by 50% in the second risk set. The idea behind this is, that subject 1 and 2 are for sure in the first risk set, but have 50% chance each to be in the second risk set. Thus, the partial log likelihood for the two subjects would be:

$$PL_{1\&2-Efron} = \left(\frac{r_1}{r_1 + r_2 + r_3}\right)\left(\frac{r_2}{(0.5)r_1 + (0.5)r_2 + r_3}\right). \tag{4.29}$$

Generalising this to the whole sample, Efron's approximation can be written as (Hosmer et al., 2008):

$$PL(\beta)_{Efron} = \prod_{i=1}^{m} \frac{e^{\beta X_{(i)+}}}{\prod\limits_{k=1}^{d_i}\left(\sum_{j\in R(t_i)} e^{\beta X_j} - \frac{k-1}{d_i}\sum_{j\in D(t_i)} e^{\beta X_j}\right)}. \tag{4.30}$$

As stated above, the exact partial log likelihood for time-tied events assumes time to be discrete and thus expects time-tied events. A Cox's model with discrete-time is based on the logistic model, where the risk score is multiplied with the odds ratio of the baseline hazard instead of the baseline hazard itself (Therneau and Grambsch, 2000):

$$\frac{\hat{h}_i(t)}{1 + \hat{h}_i(t)} = \left(\frac{\hat{h}_0(t)}{1 + \hat{h}_0(t)}\right) e^{\beta X_i}. \tag{4.31}$$

Coming back to the previous example, the exact partial log likelihood calculates the probability that subject 1 and 2 are part of the group with time-tied events instead of any other selection of the subjects in the risk set. Thus, the partial log likelihood for the two subjects would be:

$$PL_{1\&2-Exact} = \frac{r_1 r_2}{r_1 r_2 + r_1 r_3 + r_2 r_3}. \tag{4.32}$$

Generalising this to the whole sample, the exact partial log likelihood is (Klein and Moeschberger, 2003):

$$PL(\beta)_{Exact} = \prod_{i=1}^{m} \frac{e^{\beta X_{i+}}}{\sum_{q \in Q(t_i)} e^{\beta X_q}}, \tag{4.33}$$

where Q are all the combinations of selected subjects for the group with time-tied events, q is one combination of selected subjects, d_i is the number of events at time t_i, and X_q is the summed values for the covariates, $X_q = \sum_{j=1}^{d_i} X_j$. The exact partial log likelihood is accurate but time-consuming in its computation.

For each method, the same steps are taken to estimate the coefficients β as with the partial log likelihood that assumed no time-tied events (Hosmer et al., 2008). In case of no time-tied events, all three methods provide the same results. In case of time-tied events, the default setting is to use Efron's approximation as it is more accurate than Breslow's approximation and more efficient than the exact partial log likelihood (Hosmer et al., 2008). For this research, Efron's approximation was used.

Assumption of uninformative censoring

Survival models assume uninformative censoring, which means that the distribution of time-to-event and the distribution of time-to-censorship do not inform each other (Hosmer et al., 2008). As stated above, this research had right-censoring data, where the censoring was due to patients transferring to another general practice during the study period or remaining alive at the end of the study period. Censoring due to end of study period is believed to be random and thus uninformative of time-to-event distribution. Censoring due to transfer to another general practice could be non-random and affect the distribution of time-to-event.

Previous studies identified two groups of people who move at an older age in the United Kingdom (Pennington, 2013; Uren and Goldring, 2007). The first group moves to match the desired lifestyle during retirement. For example, people move to an area more suitable for active retirement or move to a more secure area with respect to proximity to health and social care, and support from nearest family. This move tends to take place between ages 60 to 70, and is mostly done by households from affluent areas. The second group moves due to ill-health. This move tends to take place after age 70, and is mostly done by households from deprived areas.

If these trends are observed in the studied age cohorts for this research, the hazard of mortality could be overestimated before age 70, when transferring to another general practice is associated with excellent health. Similarly, the hazard of mortality could be underestimated after age 70, when transferring to another general practice is associated with ill-health. For this research, patients who did not transfer from their general practice were compared with patients who transferred before or after age 70 with respect to health status, lifestyle choices, and socio-demographic factors.

Assumption of homogeneous population

The basic Cox's model assumes that the sample comes from a homogeneous population. The data used for this research are medical records of patients from multiple general practices. Practices vary in health outcomes due to differences in their patient population and provision of patient care (Rasbash et al., 2012). Patients from one practice are thus not independent from each other, instead they have a shared risk (Hosmer et al., 2008). The shared risk is also called shared frailty, where patients of one practice are more frail than patients from another practice, and this excessive risk translates in worse health outcomes (Therneau and Grambsch, 2000). The differences between shared frailties result primarily from the fact that it is impossible to adjust for all covariates on subject level in the analysis. This could be due to financial, legal, or ethical restrictions in identifying and measuring the covariates. Even if all covariates were to be measured, there could be too many of them to adjust for in the model due to limited computational or statistical power.

To adjust for the cluster effect of general practices due to unmeasured covariates, a shared frailty term is specified in the Cox's model. This means that the Cox's model is multilevel, modelling data on patient and general practice level. The unmeasured covariates are modelled by a random effect in the analysis. The multilevel Cox's model estimates the hazard λ_{ij} for patient j in general practice i by multiplying the shared frailty term of the general practice Z_i by the baseline hazard $\lambda_0(t)$ and the patient's risk score r_{ij} (Hosmer et al., 2008):

$$\lambda_{ij}(t, Z_i, \beta, X_{ij}) = \lambda_0(t)\, Z_i\, r_{ij}(\beta X_{ij}) = \lambda_0(t)\, Z_i\, e^{\beta X_{ij}}, \tag{4.34}$$

where

$$Z_i \sim i.i.d.\ \Gamma(1, \theta). \tag{4.35}$$

The shared frailties Z_i's are assumed to be identical and independently distributed random factors that follow a Gamma distribution with mean 1 and variance θ. A Gamma distribution is chosen instead of a Gaussian distribution, because the frailty terms should only be able to take positive values since the hazard function is a non-negative function (Hosmer et al., 2008). The variance θ quantifies the degree of variability among the clusters and the degree of correlation within a cluster. In other words, a greater variance means that the cluster effects are more dispersed and the correlation within clusters is stronger. Inter-cluster correlation, calculated by

Kendall's τ, is $\theta/(2+\theta)$ in the model (Therneau, 2014).

In case of no cluster effects ($\theta = 0$ and $Z = 1$), i.e. patients from the same general practice are independent from each other, the multilevel Cox's model reduces to the standard Cox's model specified in Equation (4.13). In case of cluster effects ($\theta \neq 0$), a shared frailty of a general practice greater than one means that patients from that practice are more frail and thus have a shorter than average survival time. Similarly, a shared frailty of a general practice smaller than one, means that the patients registered at that practice are less frail and thus have a longer than average survival time.

The multilevel Cox's model is estimated by maximising the partial log likelihood function. Let there be G general practices where the i^{th} general practice has n_i patients. Let $D_i = \sum_{j=1}^{n_i} d_{ij}$ stand for the number of events in the i^{th} general practice and $\Lambda_0(t)$ stand for the cumulative baseline hazard. The partial log likelihood is formulated as (Klein and Moeschberger, 2003):

$$\begin{aligned} l(\theta,\beta) = \sum_{i=1}^{G} & [D_i \ln(\theta) - \ln\left[\Gamma\left(1/\theta\right)\right] + \ln\left[\Gamma\left(1/\theta + D_i\right)\right] \\ & - (1/\theta + D_i)\ln\left[1 + \theta \sum_{j=1}^{n_i} \Lambda_0(t) e^{\beta X_{ij}}\right] \\ & + \sum_{j=1}^{n_i} d_{ij}\left\{\beta X_{ij} + \ln\left[\lambda_0(t)\right]\right\}]. \end{aligned} \tag{4.36}$$

The partial log likelihood is maximised by applying the Estimation-Maximization (EM) method, which entails the following steps (Klein and Moeschberger, 2003):

1. Fit the basic Cox's model and obtain the estimated coefficients $\hat{\beta}$.

2. (E step) Estimate the shared frailty terms $\hat{Z}_i$.

3. (M step) Update the estimated coefficients $\hat{\beta}$ by fitting the Cox's model with the estimated shared frailty terms $\hat{Z}_i$.
4. Iterate between E and M steps until convergence in the estimates $\hat{Z}_i$ and $\hat{\beta}$.

To assess whether there is a cluster effect of general practices, it is tested whether the variance of the shared frailty terms θ is significantly different from zero. This is done by calculating the likelihood ratio test (LRT) statistic (Hosmer et al., 2008):

$$LRT_{\theta} = 2\left(l(\hat{\theta}, \hat{\beta}) - l(0, \hat{\beta})\right), \tag{4.37}$$

where

$$l(\theta = 0, \hat{\beta}) = \sum_{i=1}^{G} \sum_{j=1}^{n_i} d_{ij} \left\{\beta X_{ij} + \ln\left[\lambda_0(t)\right]\right\} - \sum_{i=1}^{G} D_i. \tag{4.38}$$

The likelihood ratio test statistic is thus twice the difference between the partial log likelihood of the multilevel Cox's model and the partial log likelihood of the basic Cox's model. The test statistic follows a χ^2-distribution with one degree of freedom (Hosmer et al., 2008).

Assumption of proportional hazards

The underlying assumption of the Cox's model is that the hazards are proportional over time between subgroups of subjects (Therneau, 2014). The model is adequate when this assumption holds. Popular techniques used to assess the proportional hazards assumption are visual assessment of Kaplan-Meier plots or numerically by Grambsch and Therneau's test (Persson, 2002). Graphs are not as objective as statistical tests, and are therefore not recommended to rely solely on (Persson, 2002).

The Grambsch and Therneau's test is based on the scaled Schoenfeld residuals. Schoenfeld residuals are obtained from the first derivative of the partial likelihood as specified in Equation (4.18). Recall that m is the number of subjects who had the event during the study period, X_i stands for the values of the covariates for subject i with ordered survival time t_i, and $\bar{X}_{w_i}$ stands for the risk set conditional means of the covariates for subject i with ordered survival time t_i. The Schoenfeld residual for the k^{th} covariate equals (Hosmer et al., 2008):

$$\hat{r}_k = \sum_{i=1}^{m}(X_{ik} - \bar{X}_{w_{ik}}). \tag{4.39}$$

From this, the Schoenfeld residual for the k^{th} covariate for the i^{th} subject who had the event of interest during the study period equals (Hosmer et al., 2008):

$$\hat{r}_{ik} = X_{ik} - \bar{X}_{w_{ik}}. \tag{4.40}$$

The sum of the Schoenfeld residuals should equal zero, as the coefficients β were obtained by setting the derivative of the partial log likelihood to zero. As the partial log likelihood is only based on the subjects who had the event of interest during the study period, the Schoenfeld residuals for censored subjects will be missing and will thus not be informative for the fit of the model. Grambsch and Therneau proposed to scale the Schoenfeld residuals by its estimated variance to obtain a more informative measure (Therneau, 2014). Let $\hat{r}_i'$ be a vector of Schoenfeld residuals for the i^{th} subject who had the event of interest during the study period, as (Hosmer et al., 2008):

$$\hat{r}_i' = (\hat{r}_{i1}, \hat{r}_{i2}, ..., \hat{r}_{ik}). \tag{4.41}$$

The variance of the Schoenfeld residuals can be approximated by the covariance matrix of the estimated coefficients $\widehat{\text{Var}}(\hat{\beta})$ multiplied by the number of events m, so that the scaled Schoenfeld residuals for the i^{th} subject are computed as (Hosmer et al., 2008):

$$r_i^* = \frac{r_i'}{m\widehat{\text{Var}}(\hat{\beta})}. \tag{4.42}$$

Let transformed survival time be a vector ranking subjects in the order of events, such that $t^* = (1, 2, ..., m)$ (Therneau, 2014). The Grambsch and Therneau's test estimates the correlation between the scaled Schoenfeld residual of a covariate and the transformed survival time (Therneau, 2014). The correlation statistic follows a χ^2-distribution with one degree of freedom. A correlation statistic of a covariate that is significantly different from zero indicates that the proportional hazards assumption is violated for that covariate.

If the proportional hazard assumption is violated, the covariate should be specified differently. For this research, such covariates were assumed to be time-dependent. Follow-up time was split in intervals in which the assumption was no longer violated. The time intervals of 10, 5, or 1 year were tested. When the coefficients of a covariate were not significantly different in multiple time intervals, the time intervals were merged together.

4.2 Missing data

Most studies that make use of observational data, have a proportion of the data missing. There are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Graham, 2009). With MCAR data, there is no systematic difference between observed and unobserved data (Graham, 2009). For instance, cholesterol readings might be missing due to breakdown of equipment. With MAR data, there is a systematic difference between observed and unobserved data and this difference can completely be explained by the observed data (Graham, 2009). For instance, given that the data consist of cholesterol readings and age, recorded cholesterol readings might be higher than missing cholesterol readings only because older patients are more likely to have a record of cholesterol readings. With MNAR data, there is a systematic difference between observed and unobserved data and this difference is at least partly explained by unobserved data (Graham, 2009). For instance, given that the data consist of cholesterol readings and drug therapy, patients with high cholesterol readings who do not adhere to drug therapy might be less likely to go to doctor appointments. This could mean that missing cholesterol readings are higher than recorded cholesterol readings of patients who do adhere to drug therapy.

The next subsections discuss the type of missing data present in primary care records, methods to deal with this type of missing data to obtain unbiased results, and the specification of the multiple imputation process that was carried out for this research to deal with the missing data.

4.2.1 Missing data in primary care records

With primary care data, it is established that there is a systematic difference between observed and unobserved data (Hippisley-Cox and Coupland, 2010b; Hippisley-Cox et al., 2008; Marston et al., 2010; NICE, 2014a; Szatkowski et al., 2012). Recording of medical, lifestyle, and socio-demographic information are related to ill health; people who are ill, visit the general practice more often and general practitioners are more likely to record background information of these patients compared to healthier patients (MacDonald and Morant, 2008; Shephard et al., 2011). Women are more likely to have a complete medical record than men, as they tend to be sicker and visit the general practice more often (Bartley, 2004). Information on lifestyle is predominantly recorded because of its association with medical conditions (Marston et al., 2010).

Over time, the incentives and methods for recording information other than the medical condition or treatment by general practitioners have changed (Marston et al., 2010). For example, with the introduction of the Quality and Outcomes Framework (QOF) in 2004, a pay scheme to improve the quality of the health care provided by general practitioners, recording has greatly improved in primary care (Langley et al., 2011; NICE, 2014a; Szatkowski et al., 2012).

For this research, patients with and without complete medical records were compared to each other in respect to the medical history of relevant medical, lifestyle, and socio-demographic covariates as described in Chapter 3 Section 5. It was assumed that patients who had no record of a medical diagnosis or treatment, did not have the medical condition or receive the treatment. This means that there were only missing records in the lifestyle covariates, i.e. in cholesterol level, blood pressure, body mass index, alcohol consumption status, and smoking status. It was assessed whether there was a significant difference in the proportion of missing records in a lifestyle covariate by the medical conditions, treatments, socio-demographic factors, and other lifestyle covariates. This was assessed by the χ^2-test of independence, of which the statistic is calculated as (Kleinbaum and Klein, 2011):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.43}$$

where r and c are the numbers of rows and columns in the contingency table, respectively, O_{ij} stands for observed count in row i column j, and E_{ij} for expected count in row i column j. The expected count E_{ij} is calculated as $O_i O_j / N$, where N is the sample size. This test follows a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom. The underlying assumptions of the test are that the sample size is sufficiently large and the smallest expected count is greater than five (Kleinbaum and Klein, 2011). These assumptions were met.

In the studied age cohorts for this research, the proportion of missing records in lifestyle covariates decreased with age, see Tables A.1 to A.4 in the Appendix. The proportion of missing records was the greatest in cholesterol readings with 83% at age 60 and 50% at age 75. Due to the high proportion of missing records, cholesterol reading was substituted by the diagnosis of hypercholesterolaemia in the analysis. The proportion of missing records was the lowest in blood pressure and smoking status with 28% at age 60 and 13% at age 75. The proportion of missing records in body mass index and alcohol consumption status was 37% at age 60 and 20% at age 75. Excluding cholesterol readings, 49% of the youngest cohort and 29% of the oldest cohort had incomplete medical records. A study using QResearch primary care database reported similar proportions of missing records (Hippisley-Cox et al., 2008).

The results of the χ^2-test of independence confirm that patients with and without complete medical records systematically differ from each other, see Tables A.1 to A.4 in the Appendix. There were significantly more incomplete records among men, patients born in earlier years, and patients without a medical diagnosis or treatment ($p<.001$). The exceptions to these trends were that there was no significant difference in the proportion of missing smoking status across sexes ($\chi^2(1)<.01$, p=.97) and body mass index categories ($\chi^2(2)=1.55$, p=.46) in the 70-year old cohort, and in blood pressure across sexes ($\chi^2(1)=1.76$, p=.18) in the 75-year old cohort.

4.2.2 Methods to deal with missing data

Over time, the approach to handling missing data has changed; when a significant proportion of data is missing (>5%), it is widely acknowledged that it is not acceptable to perform: complete case analysis, exclude covariates with missing data from the analysis, create a category for missing observations, or substitute a missing observation with a reasonable guess (Allison, 2001; Marshall et al., 2010; Sterne et al., 2009;

van Buuren and Groothuis-Oudshoorn, 2011). With complete case analysis, subjects with incomplete information are excluded from the analysis. This means that the sample size and thereby the statistical power of tests are reduced. In addition, if the reason for missing data is associated with the outcome of interest, the analysis would provide biased estimates because the estimates of coefficients would be biased towards the subset of sample that is observed instead of the entire sample. Excluding covariates with missing data from the analysis could mean that important covariates in predicting the outcome are not adjusted for, resulting in biased estimates. Creating a category for missing observations would also result in biased estimates because it would distort correlations with the other covariates.

Substituting a missing observation with a reasonable guess is called single imputation. Typically, missing observations of a given covariate are substituted with the mean of the observed values for that covariate. As with creating a category for missing observations, imputing missing observations with the mean of observed observations would distort correlations with the other covariates and thus provide biased estimates of coefficients. A more appropriate imputation is to substitute missing observations of a given covariate with the predicted values of that covariate given by a regression model that regresses that covariate on the other covariates that are part of the analysis. Imputation by a regression model provides unbiased estimates given that the covariates associated with the reason for having missing data are included in the regression model. The imputation, however, underestimates the standard errors, leading to false precision of the results.

When there is up to 50% of the data missing, the widely accepted method to deal with that is to perform multiple imputation (Allison, 2001; Marshall et al., 2010; Sterne et al., 2009; van Buuren and Groothuis-Oudshoorn, 2011). With this method, values are imputed for missing observations multiple times, creating multiple datasets with varying values. The results estimated on each of the datasets are then pooled using Rubin's rules, which take into account the variance of the coefficient within and between imputed datasets (Rubin, 1987). The goal of multiple imputation is to make valid statistical inferences by reflecting the uncertainty in missing data, and not to impute the true values (Rubin, 1987). The next subsection describes the process of multiple imputation in greater detail.

Multiple imputation involves several stages of model development and respective assumptions about the missing data, where any incorrect decisions could lead to biased estimates. The most important assumption is that the data are missing (completely) at random (Allison, 2001), as defined in the introduction to this Section. The type of missing data cannot be directly tested, however understanding why the data might be missing may support or oppose the use of multiple imputation.

Multiple imputation carried out on primary care data was shown to be an effective method of obtaining unbiased estimates in the presence of missing data (Carlin et al., 2007; Marston et al., 2010). For this research it was, therefore, assumed that the reason for missing records could fully be explained by the observed data on medical conditions, treatments, socio-demographic factors, and lifestyle factors, and thus that the missing data were missing at random. There was more than 5% and less than 50% missing data in the age cohorts, hence multiple imputation was chosen to deal with bias and imprecision caused by the presence of missing data.

4.2.3 Multiple imputation

The imputation model should consist of all covariates that are included in the analysis model and covariates that are associated with missing data (van Buuren, 2012). It might, however, not be possible to include all covariates in the imputation model because of multicollinearity or limited computational power to handle the complexity of the imputation model. The optimal number of included covariates is up to 30, although more than 15 hardly increases the explained variance in the imputed covariate (van Buuren, 2012). When the imputation model includes more covariates than the analysis model, the results of the analysis model could be biased if the imputation model is incorrect (Schafer and Graham, 2002). When the analysis model includes more covariates than the imputation model, the results of the analysis model are valid but could underestimate the effects of the covariates that were excluded from the imputation model, given that the analysis model is correct (Schafer and Graham, 2002).

The imputation model for each covariate with missing data is depended on its measurement scale. Continuous covariates are imputed using a linear regression. The imputed covariate does not have to be normally distributed because the aim of the imputation model is to provide a range of plausible values and not to impute the values that would have been observed if the data were not missing (van Buuren, 2012). Binary covariates are imputed using a logistic regression. Covariates with more than two categories are imputed using a multinomial regression. Covariates should be imputed on their original measurement scale and transformed after imputation to obtain unbiased estimates (von Hippel, 2009). Any covariates derived from covariates with missing records should be imputed as well instead of calculated based on the imputed values (i.e. passive imputed) (van Buuren, 2012). This is to ensure that the imputed values are consistent with the analysis model.

As described above, the studied age cohorts for this research had multiple covariates with missing values. These covariates were measured on different scales, where blood pressure and body mass index were continuous, and alcohol consumption and smoking status were categorical. There was a general missing pattern, meaning that the data could not be ordered in such a way that if there were missing values for one covariate then there were also missing values for another covariate (van Buuren and Groothuis-Oudshoorn, 2011). The data had a hierarchical structure as patients' medical records were clustered by general practice. Until recently, the only software that could specify an imputation model for mixed continuous-categorical data with a hierarchical structure and a general missing pattern was REALCOM-Imputation (Centre for Multilevel Modelling, 2016). This software imputes missing data by joint modelling.

Joint modelling assumes that the data come from a multivariate distribution, most commonly being a Gaussian distribution, which is what the software REALCOM-Imputation specifies (Allison, 2001; Centre for Multilevel Modelling, 2016). Imputations based on this distribution have been shown to be robust to non-normal data (van Buuren and Groothuis-Oudshoorn, 2011). Joint modelling imputes per pattern of missing data on a row to row basis (Allison, 2001). For example if there are two covariates with missing observations and the missing is at random, then the patterns of missing data are: subjects/rows with no missing data, subjects/rows with missing observations in the first covariate and no missing observations in the second covariate, subjects/rows with no missing observations in the first covariate and missing observations in the second covariate, and subjects with missing observations in the first and second covariates.

Multiple imputation is an iterative process that consists of the following steps (Allison, 2001):

1. Obtain the observed sample means and covariance matrix.

2. For each pattern of missing data, use the means and covariance matrix to estimate the regression coefficients for equations in which each covariate with missing data is regressed on all other covariates of interest.

3. Based on the estimated regression coefficients, predict the values for the subjects with missing data and add a residual term as specified by the Gaussian distribution. Without adding a residual term, the imputation process would be deterministic, resulting in underestimated variances of the covariates with missing data.

4. With the new dataset of observed and imputed values, recalculate the sample means and covariance matrix.

5. Use the recalculated means and covariance matrix to obtain a random draw from the posterior distribution of means and covariances.

6. Iterate by going through steps 2 to 5 continuously until convergence of the estimated regression coefficients.

The iteration process until convergence in the estimated regression coefficients is called the 'burn-in length'. Convergence of the coefficients can be checked by plotting the coefficients of the imputation model against the iterations. Each iteration results in a new dataset of observed and imputed values. To obtain independent imputed datasets, the datasets should be separated by a number of iterations. The number of iterations for the burn-in length and between imputed datasets can be as little as five iterations (van Buuren, 2012). The default setting in REALCOM-impute is a burn-in length of 100 iterations and 100 iterations between imputed datasets (Centre for Multilevel Modelling, 2016). Having more iterations than necessary, does not affect the results (van Buuren, 2012).

The number of imputed datasets is recommended to be roughly the percentage of missing records in the dataset (von Hippel, 2009). Although imputing more than five times will most likely not change the inferences made on five imputed datasets (van Buuren, 2012). For this research, the default setting of REALCOM-impute was used, which is 10 imputed datasets (Centre for Multilevel Modelling, 2016). The imputed values were checked by bar and density plots. Assuming that the missing data were missing at random, the imputed values should have a similar distributions as the recorded observations (van Buuren, 2012).

The imputed datasets were analysed separately, i.e. the covariate selection was done on each imputed dataset. As described in subsection 4.2.2, a general, non-age specific model was selected to have the same interpretation of the effects at each age. If an interaction effect was found in the majority of the models (≥50%), it was included in the final model. Given age cohort, this final model was then estimated on each imputed dataset and combined using Rubin's rules. Let there be m imputed datasets, where the j^{th} imputed dataset has estimated coefficients $\hat{\beta}_j$. Let $\widehat{W_j}$ be the estimated variances of estimated coefficients $\hat{\beta}_j$ of the j^{th} imputed dataset, $\widehat{W}$ be the average within imputation variance of estimated coefficients $\hat{\beta}$ across m imputed datasets, and $\widehat{B}$ be the between variance of estimated coefficients $\hat{\beta}$. Using Rubin's rules, the coefficients $\hat{\beta}$ and variances $\widehat{\text{Var}}(\hat{\beta})$ are calculated as (Rubin, 1987):

$$\hat{\beta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j, \tag{4.44}$$

$$\begin{aligned}\widehat{\text{Var}}(\hat{\beta}) &= \widehat{W} + \left(1 + \frac{1}{m}\right)\widehat{B}, \text{where}\\ \widehat{W} &= \frac{1}{m}\sum_{j=1}^{m}\widehat{W}_j,\\ \widehat{B} &= \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\beta}_j - \hat{\beta})^2.\end{aligned} \tag{4.45}$$

If one dataset is analysed, the variances of the coefficients $\widehat{\text{Var}}(\hat{\beta})$ would be equal to the within imputation variance $\widehat{W}$. The between imputation variance $\widehat{B}$ is the result of having missing data. The addition of the term $\widehat{B}/m$ is the result of estimating the coefficients on a limited number of imputed datasets.

To test whether the coefficient $\hat{\beta}$ is significantly different from zero, the t-statistic is calculated as (Rubin, 1987):

$$t = \frac{\hat{\beta}}{\widehat{\text{se}}(\hat{\beta})}, \tag{4.46}$$

where se stands for the standard error of the estimated coefficient $\hat{\beta}$. The t-statistic follows a t-distribution with degrees of freedom equal to (Rubin, 1987):

$$\begin{aligned} df &= (m-1)\left(1+\frac{1}{r}\right)^2, \text{where} \\ r &= \frac{\widehat{B}(1+1/m)}{\widehat{W}}. \end{aligned} \tag{4.47}$$

4.3 Model building

4.3.1 Study design

For this research, two survival models were developed. The first model was developed to estimate the hazards of all-cause mortality associated with a history of AMI and related treatments while adjusting for other risk factors. The second model was developed to estimate the hazard of all-cause mortality associated with statins prescribed as primary prevention of CVD while adjusting for other risk factors. The models were specified on slightly different age cohorts. For the first model all patients in the cohorts were eligible for participation while for the second model patients with a history of CVD were excluded from the cohorts.

The prevalence of AMI was relatively rare, especially in the youngest cohort. At ages 60, 65, 70, and 75, the prevalence of AMI in women was 1.1, 1.9, 2.7, and 3.7%, respectively, and in men 4.8, 6.6, 8.3, and 10.0%, respectively. To create more balanced cohorts with respect to the exposure of interest, patients with a history of

AMI were selected and each matched to three controls without this history on sex, year of birth category, and general practice.

A balanced dataset means that the groups being compared are of the same size and have similar characteristics with the exception of the exposure and outcome of interest (Greenland and Morgenstern, 1990; Hennekens et al., 1987). A more balanced cohort created by matching, translates in increased statistical power and efficiency (Greenland and Morgenstern, 1990; Hennekens et al., 1987). The statistical power of a test is the probability that the null hypothesis is correctly rejected, i.e. the power to identify an effect that truly exists. This is maximised when the subgroups of patients are of the same size. Statistical efficiency refers to the optimal use of the data, where a more efficient test or estimator needs fewer patients to reject the null hypothesis. Efficiency is maximised when the variance is minimised.

To improve statistical power and efficiency, the matched factor must be a confounder, which is a covariate that is associated with both the exposure and outcome and does not lay on the causal pathway (Hennekens et al., 1987; Rothman et al., 2008). If the matched factor is not a confounder, overmatching can take place, which in turn could harm the statistical efficiency and the validity of the results. Matched factors that are typically used are sex and age (Hennekens et al., 1987; Rothman et al., 2008). When a study includes multiple medical centres, the medical centres are also matched on. This is because patients of one centre are more alike compared to patients from another centre (Rasbash et al., 2012). By matching on medical centre, cases and controls are similar in unmeasured factors that are clustered in medical centres. For this research, cases and controls in the age cohorts were matched on sex, year of birth category, and general practice. Matching on year of birth category takes into account possible advances in medical management over time (Kleinbaum and Klein, 2011). Year of birth was categorised with categories being 1920-25, 1926-30, 1931-35, and 1936-40, because not all cases could be matched to a control when a specific year was used as a matching factor.

The traditional matching methods are individual or frequency matching (Rothman et al., 2008). With individual matching, matching is carried out case by case, whereas with frequency matching, it is carried out per stratum of cases. In a situation when not enough controls per case can be found, alternative matching methods are carried out such as partial, marginal, counter, probability, and propensity score matching (Caliendo and Kopeinig, 2008; Cologne et al., 2004; Langholz and Clayton, 1994). An optimal matching ratio is to match the cases with one to ten controls, although five or more controls hardly increases the statistical efficiency (Raboud and Breslow, 1989). The majority of studies use less than five controls per case due to feasibility and the limited increased statistical efficiency when there would be more controls (Cepeda et al., 2003). For this research, the matching ratio was set to the maximum number of controls that could be matched per case in each age cohort with the limit being five controls per case. This resulted in a matching ratio of 1 case to 3 controls. Frequency matching was carried out as this is more efficient than individual matching when there are multiple cases with the same values for the matched factors (Rothman et al., 2008).

4.3.2 Selection of covariates

The selection of medical, lifestyle, and socio-demographic covariates for this research was based on the literature review discussed in Chapter 2. This means that all covariates were previously shown to be risk factors that contributed in explaining survival variations in AMI patients or primary cardiovascular risk groups. Interaction effects within and between all groups of risk factors were not studied before, except for limited interactions with the main exposure, sex, and age. For this research, all second-order interactions were tested to examine survival variations in greater detail. To obtain the leanest model possible where the prediction error is minimised, a selection procedure was carried out to select the most important interaction effects.

The prediction error is the difference in observed and predicted outcome values (Harrell, 2001; Hosmer et al., 2008). The prediction error includes the bias between observed and predicted outcome values and the variance of the predicted outcome values. The minimal prediction error is observed with the true underlying model of the data. A model that includes less covariates than the true model ‘underfits’ the data by not capturing the underlying trend of the data. In other words, this model excludes important covariates in predicting the outcome and therefore has high bias and low variance of the predicted outcome values. A model that includes more covariates than the true model ‘overfits’ the data by not only capturing the underlying trend but also the noise of the data. In other words, this model includes important and unimportant covariates in predicting the outcome, and therefore has low bias and high variance of the predicted outcome values. To find the true model, several stepwise methods were proposed.

Stepwise methods select the most important covariates to include in a model solely based on their significance that is specified by some mathematical criterion (Harrell, 2001; Hosmer et al., 2008). Forward selection starts with an empty model and adds one covariate at the time where the covariate that is added first explains the greatest percentage of (the remaining unexplained) variation in the outcome. This process is continued until no covariate significantly contributes to the model in explaining the variation in the outcome. Backward elimination starts with a full model that includes all covariates and removes one covariate at the time where the covariate that is removed first contributes the least to the model. This process is continued until all the covariates in the model contribute significantly in explaining variations in the outcome. Bidirectional elimination starts with an arbitrary model and considers adding, removing, or swapping a covariate with each step until the model does not change. This stepwise method is thus a combination of forward selection and backward elimination. All methods have the risk of excluding important and/or including unimportant covariates (Harrell, 2001; Hosmer et al., 2008). Backward elimination is preferred over the other stepwise methods because it is the least likely to exclude important covariates that are only significant in relation with other covariates (Harrell, 2001; Hosmer et al., 2008).

Stepwise methods are based on some mathematical information criterion. This information criterion is an estimate of the prediction error. With the model selection process, the model with the lowest information criterion would be selected. There are numerous information criteria of which Akaike information criterion (AIC) and Bayesian information criterion (BIC) are most commonly used, and defined as (Hastie et al., 2009):

$$AIC = -2LL + 2k, \tag{4.48}$$

$$BIC = -2LL + k\ln(m), \tag{4.49}$$

where LL is the partial log likelihood of the Cox's model, k is the number of coefficients estimated by the model, and m is the limiting sample size, which is the number of events in survival analysis (Harrell, 2001; Hastie et al., 2009). As the sample size increases to infinity, model selection based on AIC has a non-zero and model selection based on BIC has a zero probability of overfitting the data, i.e. selecting unimportant covariates thereby making the model too complex (Hastie et al., 2009). However, with a finite limiting sample size, model selection based on BIC could select too few covariates, making the model too simple. As the stepwise method based on some information criterion is an automated process, it is important to check whether the resulting survival model is biologically plausible (Hosmer et al., 2008).

For this research, backward elimination of second-order interaction effects was carried out. For the automated process, BIC was used as information criterion. It was assumed that only interaction effects would be removed from the model and not main effects as their significance in explaining survival prospects were established in previous studies. The automated backward elimination process where BIC was minimised, resulted per age cohort in a survival model with 20 to 25 interaction effects with p-values ranging from $<.001$ to .500. Backward elimination was continued until all interaction effects were significant at 1% level in the decomposition of Cox's model by ANOVA (analysis of variance). Due to large sample sizes, the significance level was set at 1% to obtain only the interaction effects that contribute the most to the model in explaining survival variations. The low significance level comes with the cost that interaction effects between covariates that have relatively rare categories will most likely be excluded. This manual backward elimination process resulted per age cohort in a survival model with three to eight interaction effects. A common set of effects in the survival models was chosen to have the same interpretation of the effects at each age cohort. If an interaction effect was found in the majority of the models ($\geq$50%), it was included in all models. For the analysis that makes use of multiple imputation, the variable selection had a two-step approach. At the first step, interaction effects found in the majority of the models on the imputed datasets of an age cohort, were included in the age-specific model. At the second step, interactions found in the majority of the age-specific models, were included in the final overall model. It was checked whether the interaction effects were biologically plausible by comparing them with previous studies and clinical guidelines.

4.4 Model assessment

The final survival models were assessed in respect to overall performance, discrimination, and external validation, using Royston's R-square (Royston, 2006), Harrell's concordance (Harrell et al., 1996), and the shrinkage slope (Steyerberg et al., 2010), respectively. Furthermore, the results were compared internally with the complete case analysis, and externally with previous studies.

Royston's R-square

Overall model performance was assessed by calculating Royston's R-square, which is the percentage of variation in the outcome explained by the survival model (Royston, 2006). This is based on the goodness-of-fit measures proposed by Nagelkerke (1991) and O'Quigley et al. (2005).

Nagelkerke's ρ_n^2 is the most commonly calculated goodness-of-fit measure of a Cox's model (Steyerberg et al., 2010). The statistic ρ_n^2 is calculated by dividing the likelihood ratio test statistic of the empty versus full model, as specified in Equation (4.22), by the sample size N (Nagelkerke, 1991):

$$\rho_n^2 = 1 - \exp\left(\frac{LRT_\beta}{N}\right). \tag{4.50}$$

This statistic ranges from zero to one, where zero means that the survival model could not determine survival time for any subject and one means that the model could perfectly determine survival time for each subject. Nagelkerke's ρ_n^2 is biased towards zero when there is censoring during the study period (Royston, 2006). O'Quigley, Xu, and Stare (2005) therefore suggested to replace the sample size N in the denominator by the number of events M:

$$\rho_k^2 = 1 - \exp\left(\frac{LRT_\beta}{M}\right). \tag{4.51}$$

This statistic is a measure of explained variation, where zero means that the survival model could not explain the variation in survival time for any subject and one means that the model could perfectly explain the variation in survival time for every subject. Royston (2006) suggested to rewrite the statistic ρ_k^2 as a measure of explained variation. Let the model variance Var_m be $\rho_k^2/(1-\rho_k^2)$, the residual variance Var_r be $\pi^2/6$, and the total variance Var_t be the sum of the model variance Var_m and the residual variance Var_t (Royston, 2006). Then the percentage of variation in survival time explained by the model is the model variance Var_m divided by the total variance Var_t (Royston, 2006):

$$\rho_r^2 = \frac{V}{\pi^2/6 + V} = \frac{\rho_k^2}{\rho_k^2 + (\pi^2/6)(1-\rho_k^2)}. \tag{4.52}$$

This statistic ranges from zero to one, where zero means that the survival model could not explain any of the survival variation between subjects and where one means that the survival model could perfectly explained survival variation between subjects.

Harrell's concordance

The degree of discrimination between subjects by the model was assessed by calculating Harrell's concordance, which is the percentage of correspondence between the estimated hazard score and observed survival time for all combinations of two selected subjects (Steyerberg et al., 2010). Harrell's concordance C_H is calculated as (Steyerberg et al., 2010):

$$C_H = \frac{C + T/2}{C + D + T}, \tag{4.53}$$

where C stands for concordant pairs, D for discordant pairs, and T for tied pairs. A concordant pair is when the subject with the lower risk score has the longer survival time. A discordant pair is when the subject with the lower risk score has the shorter survival time. A tied pair is when it is unknown which of two subjects had the event first. This happens when two subjects have time-tied events, when both are censored, or when one subject is censored before the other subject had the event.

The concordance statistic ranges between zero and one, where one stands for perfect discrimination between two randomly selected subjects (Therneau, 2014). A concordance of .5 means that the model is as good as flipping a fair coin in its discrimination. In survival analysis, the concordance measure is typically between .6 and .7 (Therneau, 2014).

Shrinkage slope

External model validation was assessed by calculating the shrinkage slope, which assesses the agreement between observed and expected outcomes. Predictions could be systematically too low or too high, reflecting overfitting or underfitting of a model (Harrell et al., 1996). The shrinkage slope is the factor by which the coefficients may need to be shrunk due to overfitting or rise due to underfitting of the model (Harrell et al., 1996). In other words, this factor indicates how well the model would perform on new data and how well the results of the sample are generalisable to a wider population.

For this research, the shrinkage slope was calculated by a ten-fold cross-validation (Refaeilzadeh et al., 2009). Cross-validation estimates the expected extra prediction error due to sampling, where the prediction error is the difference between observed and predicted outcome values. The prediction error ϵ is calculated as (Hastie et al., 2009):

$$\epsilon = E[Y - \hat{Y}] = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i), \tag{4.54}$$

where subject i has observed outcome y_i and predicted outcome $\hat{y}_i$, and N is the total number of subjects. With K-fold cross-validation, the sample is randomly split in K folds of equal sizes. The model is fitted on K-1 folds and the prediction error of the fitted model on the excluded fold is calculated. In other words, the model is trained on K-1 folds and validated on the excluded fold. This process is repeated K times, so that each fold has been validated once. The cross-validation estimate of the prediction error $CV(\hat{\epsilon})$ is the average of the K prediction errors (Hastie et al., 2009):

$$\begin{aligned} CV(\hat{\epsilon}) &= \frac{1}{K}\frac{1}{n}\sum_{i=1}^{K}\sum_{i=1}^{n}(y_i^{-k(i)} - \hat{y}_i^{-k(i)}), \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^{-k(i)}), \end{aligned} \tag{4.55}$$

where $\hat{y}^{-k}$ is the predicted outcome for the k^{th} excluded fold and n the number of subjects in the k^{th} fold. A ten-fold cross-validation is typically chosen because there are multiple performance statistics and the training fold is large enough to generalise over the whole sample while the validation fold is not too small to obtain an accurate performance statistic (Refaeilzadeh et al., 2009).

The shrinkage slope is one minus the prediction error (Hastie et al., 2009). A shrinkage slope of one means that the coefficients were correctly estimated by the model (Harrell et al., 1996). In most cases, the shrinkage will be smaller than one, which means that the coefficients were overestimated by the model and that the true coefficients are likely to be smaller than estimated. In rare cases, a shrinkage slope will be greater than one, which means that the coefficients were underestimated by the model and that the true coefficients are likely to be greater than estimated. The closer the shrinkage slope is to one, the more accurate the survival model is predicting survival time.

Comparing results

The final models were compared to the models obtained from the complete case analysis. The comparison included the covariates and interactions adjusted for, the estimated hazard ratios, and the model diagnostics with respect to overall performance, discrimination, and external validation. This comparison within the study would inform how missing data affected the results and the validity of the multiple imputation.

The estimated hazard ratios from the final models were also compared with the ones estimated by models developed by previous studies. These studies were summarised in Chapter 2 in Table 2.1, which included what covariates and interactions were adjusted for. The previous studies did not report on model diagnostics, and thus these cannot be compared.

Chapter 5

Survival models for acute myocardial infarction

This Chapter presents the development of the survival models for acute myocardial infarction (AMI). The survival models estimated the adjusted hazards of all-cause mortality associated with a history of AMI and related treatments, comorbidities, lifestyle factors, and demographics in residents of the United Kingdom (UK) at retirement age. Survival variations by general practice were also examined. The findings were published in BMJ Open (Gitsels et al., 2017).

This Chapter starts with explaining the analysis procedure including the study design, selected medical history, and model development. This is followed by a description of the age cohorts. Next, the survival models are presented. Then, the survival models are evaluated with respect to model performance, internal validation, and external validation. Finally, recommendations based on the findings are provided.

5.1 Analysis procedure

5.1.1 Study design

The research objectives of investigating how a history of AMI affects longevity at retirement age and which treatments improve longevity, were addressed by a retrospective cohort design. Four cohorts of patients aged 60, 65, 70, and 75 were followed up. The data restrictions and the identification of eligible general practices and patients from The Health Improvement Network (THIN) primary care database, were discussed in Chapter 3 Section 4. The prevalence of AMI was relatively rare, especially in the youngest cohort. At ages 60, 65, 70, and 75, the prevalence of AMI in women was 1.1, 1.9, 2.7, and 3.7%, respectively, and in men 4.8, 6.6, 8.3, and 10.0%, respectively. To create more balanced cohorts with respect to the exposure of interest, patients with a history of AMI were selected and each matched to three controls without this history on sex, year of birth category, and general practice, see Figure 5.1.

The analysis of the age cohorts was performed in two stages, where the first stage of complete case analysis informed the second stage of full data analysis. Frequency matching was carried out once on complete medical records and once irrespective of completeness of records. The reason for this was to ensure that the balance in the cohorts created by matching would not be cancelled when excluding patients with missing data from the complete case analysis. Moreover, this led to optimal use of the available data by maximising the sample size of the complete case analysis and thereby increasing the statistical power. Patients could be part of multiple cohorts, where cases could be matched to different controls. Matching was done in Python version 3.4.2.

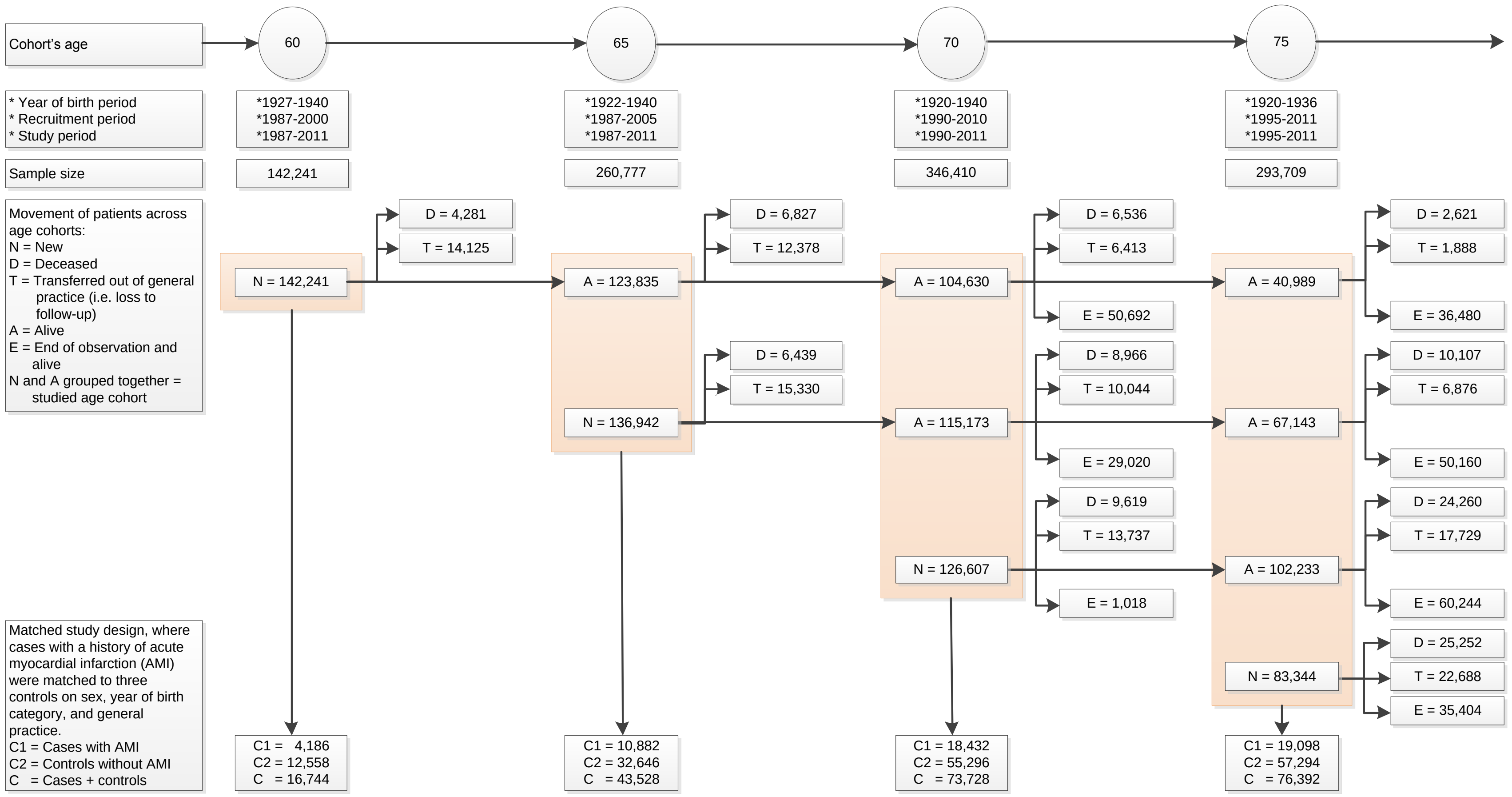


Figure 5.1: Selected age cohorts for acute myocardial infarction matched by sex, year of birth, and general practice

This Figure is an extended version of Figure 3.1 with information on the matched study design. Patients could be part of multiple cohorts, where cases could be matched to different controls.

5.1.2 Selected medical history

At the cohort's age, a snapshot of the patient's medical history was obtained. Additionally, the patient's survival was recorded during the follow-up. Chapter 3 Section 5 described the raw information selected from the medical records for the entire research. This Section describes the edited information selected for the survival models aimed at estimating the hazards of mortality associated with acute myocardial infarction and related treatments.

The primary exposure was AMI. Multiple events were required to be separated by 30 days. Information on the type of AMI was not available. However, a study that linked information from the Myocardial Ischaemia National Audit Project (MINAP) and the General Practice Research Database (GPRD), which has 60% of practices in overlap with THIN, found that 46% of AMIs were ST-elevated (ST segment elevation myocardial infarctions, STEMIs) in England and Wales in 2003-2008 (Herrett et al., 2013a). The selected variables of the medical records were based on the literature review, and consisted of: sex, year of birth, socioeconomic status, angina, heart failure, cardiovascular system conditions (cerebrovascular disease, peripheral vascular disease, valvular heart disease, and other cardiovascular system disorders), chronic kidney disease, diabetes, hypertension, hypercholesterolaemia, alcohol consumption status, body mass index (BMI), and smoking status, see Appendix B Table B.1. Socioeconomic status was measured by Mosaic, which is based on demographics, lifestyles, and behaviour of people at a postcode level (Experian Ltd., 2009).

The treatments investigated were based on the UK National Institute of Health and Care Excellence (NICE) recommended first line treatments to AMI patients during the study period, which includes: coronary revascularisation and prescription of ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, and statins (NICE, 2013b). Coronary revascularisation consisted of coronary artery bypass graft and coronary angioplasty. Since 2007, calcium-channel blockers are only recommended to treat hypertension or angina in AMI patients (NICE, 2013b). Since 2013, dual antiplatelet therapy (DAPT: aspirin plus another antiplatelet agent) is recommended to AMI patients (NICE, 2013b). Owing to the low prevalence of DAPT in the age cohorts, the survival effect of the therapy was not estimated, see Appendix B Table B.2.

The values of all variables represented the latest reading before entering the study, which was at the 1st of January of the year the patient turned the cohort's age. Family history of AMI and family history of cardiovascular disease (CVD) were not included in the analysis because of the very low rates of recording in primary care (Hippisley-Cox et al., 2008). Indicators of psychosocial factors such as job strain and lack of social support, fruit and vegetable intake, and physical activity were not included in the analysis because THIN does not hold information on them.

There were missing values in alcohol consumption status (proportion range across the four age cohorts 17-37%), BMI (18-37%), and smoking status (10-29%). The fraction of incomplete medical records decreased with age; 45% of the youngest cohort and 23% of the oldest cohort had incomplete records. Incomplete records were more common in patients without medical conditions or treatments and in patients born at an earlier year, see Appendix B Table B.3. This is in accordance with previous research that reported that recording is associated with sickness and has improved since the introduction of Quality and Outcomes Framework (QOF) in 2004 (Marston et al., 2010; Shephard et al., 2011; Taggar et al., 2012).

Missing values were dealt with by multiple imputation. The joint imputation model consisted of all factors from the snapshot of medical history including time to death and was two level to adjust for the correlation between patients from the same general practice. Imputations were done in REALCOM-Imputation software. The

Monte Carlo Markov Chain (MCMC) estimation had a burn-in length of 100 iterations and was in total 1,000 iterations long. The imputed values of every 100^{th} iteration were used. This resulted in ten imputed datasets. The distribution of recorded and imputed values of variables with missing data were similar, see Appendix B Table B.4.

Postcode classification indices other than Mosaic were only available for patients living in England or Wales. The following classification indices were selected to examine the hazard of mortality associated with general practice in more detail: index of multiple deprivation (IMD), level of urbanisation, quintile of limiting long-term illness, quintile of various ethnic groups, and quintile of various air pollution measures, see Appendix B Table B.1.

5.1.3 Model development

A Cox's proportional hazards regression model was fitted to estimate the effect of a history of AMI and respective treatments on the hazard of all-cause mortality at different ages. The outcome variable was time to death in days, that is, from the 1^{st} of January of the year the patient turned the cohort's age to the date of death. Starting from a model with second-order interaction effects of all variables with the main exposure AMI and the matching factors sex and year of birth category, backward elimination was performed to obtain the most parsimonious model possible. Interaction effects found in the complete case analysis, which were not restricted to the main exposure and matching factors, were also included in the backward elimination process. The extra interactions were hypercholesterolaemia with statin prescription, and BMI with smoking status. A unified model for all ages was chosen, to have the same interpretation of the hazards. This model consisted of the factors that were found significant in the majority of models with the alpha level set to 5% for fixed effects and 1% for interaction effects. The model also included a random effect of general practice to adjust for the correlation between patients from the same practice.

The Cox's regression assumes no time-tied deaths, uninformative censoring, and proportional hazards. Time-tied deaths were handled by Efron's approximation. Censoring due to loss of follow-up was examined by comparing patients who did and did not transfer out of their general practice with respect to health status and lifestyle factors. For the youngest two age cohorts, distinction was made between patients observed until age 70 or for longer. This was because previous studies indicated that there are two groups of people who move at an older age, where the younger group comes from more affluent areas and move in good health and the older group comes from more deprived areas and move in worse health (Pennington, 2013; Uren and Goldring, 2007). These trends were not observed in the age cohorts, see Appendix B Table B.5, and therefore uninformative censoring was assumed. The assumption of proportional hazards was checked by Grambsch and Therneau's test. When the assumption was violated for a variable, being significant at alpha level set to 1%, the variable's effect on survival time was made time-variant. Follow-up time was split in intervals in which the assumption was no longer violated, where the time intervals of 10, 5, or 1 year were tested.

The final model included sex, year of birth, socioeconomic status, AMI, angina, heart failure, cardiovascular system conditions, chronic kidney disease, diabetes, hypertension, hypercholesterolaemia, coronary revascularisation, ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, statins, alcohol consumption status, BMI, smoking status, and general practice. Chronic kidney disease was not adjusted for at ages 60 and 65 due to a low prevalence of less than 1%. There were five interactions: AMI with angina, AMI with beta blockers, AMI with calcium-channel blockers, hypercholesterolaemia with statins, and BMI with smoking status. Angina is, like AMI, a subtype of ischaemic heart disease (IHD) and has a similar treatment regime as AMI (NICE, 2013a,b). With the way these subtypes interact, a new factor representing IHD was generated that had the following levels: no history, angina only, single AMI with possibly angina, or multiple AMIs with possibly angina. Similarly, with the way hypercholesterolaemia and statins interact, a new factor representing cholesterol was generated that had the following levels: no diagnosis or drugs, hypercholesterolaemia only, and statins with possibly hypercholesterolaemia. Variables with time-varying hazards were cardiovascular system conditions and coronary revascularisation at all ages, and hypertension by ages 70 and 75. The contribution of each variable to the model in explaining survival variations was assessed by decomposing the Cox's regression by ANOVA (analysis of variance).

It was examined whether the adjusted hazard of mortality associated with general practice could be explained by postcode related indicators. These variables were not included in the original survival models because they were only available for patients living in England or Wales. Even though the additional variables explored were measured on district level, the relation between them and general practice was examined on patient level. This was because a practice could serve patients from a range of districts and an average calculated in the cohorts would not be representative of the practice as it would be biased to the older and sicker patients. High density scatterplots were made and Spearman's rank tests were performed to estimate the correlation of the postcode related indicators and the hazard of mortality associated with general practice. The correlation statistic ranges from -1 to 1, where 0 stands for no correlation and -1 or 1 for perfect correspondence between the two variables (Harrell et al., 1996). The probability values were not consulted because the correlation tests between aggregate measures were performed at patient level, thereby underestimating the variance and leading to false precision.

The number of years lost or gained in effective age associated with a history of AMI and related treatments, comorbidities, lifestyle factors, and demographics were calculated. The models were assessed on overall performance, discrimination, and external validation, using Royston's R^2, Harrell's concordance, and the shrinkage slope, respectively. Furthermore, the results were compared internally with the complete case analysis and externally with previous studies. The analyses were performed in R version 3.1.1, using the packages 'hmisc', 'rms', and 'survival'.

5.2 Description of cohorts

Four cohorts of patients who were either 60, 65, 70, or 75 years old at baseline were studied. The age cohorts included cases with history of AMI who were matched to three controls on sex, year of birth, and general practice. The profile of cases and controls with respect to comorbidities, treatments, and lifestyle and socio-demographic factors differed by age and changed over time, see Table 5.1, Figure 5.2 and Appendix B Figures B.1-B.2.

The prevalence of comorbidities was higher among AMI cases than controls, and increased with age, see Table 5.1. Angina was the most common comorbidity in AMI patients (prevalence range across the four age cohorts 46-48%), followed by cardiovascular system conditions (23-43%), diabetes (11-20%), heart failure (5-12%), and chronic kidney disease (0-10%). From 1995 to 2011, the prevalence of diabetes in AMI patients increased by 14-18%, while the prevalence of angina decreased by 5-10%, and the prevalence of the other comorbidities remained approximately the same over time, see Appendix B Figure B.1.

Table 5.1: Characteristics of cases and controls in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. [1]Missing values in alcohol consumption status, body mass index, and smoking status were dealt with by multiple imputation. The reported prevalences of these variables are the means across ten imputed datasets. The prevalences of comorbidities and lifestyle factors at the cohort's age were affected by calendar year, see Appendix B Figures B.1-B.2.

	Age 60		Age 65		Age 70		Age 75	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Number of patients	4,186	12,558	10,882	32,646	18,432	55,296	19,098	57,294
Total person-years of follow-up (mean)	46,686 (11.2)	150,471 (12.0)	93,056 (8.6)	299,841 (9.2)	114,700 (6.2)	370,006 (6.7)	91,884 (4.8)	298,140 (5.2)
Deaths during follow-up (%)	1,220 (29%)	2,008 (16%)	3,070 (28%)	5,782 (18%)	5,186 (28%)	10,557 (19%)	5895 (31%)	12,674 (22%)
Transfers during follow-up (%)	900 (22%)	3,035 (24%)	1,986 (18%)	6,597 (20%)	2,693 (15%)	8,781 (16%)	2,733 (14%)	8,971 (16%)
Male (%)	3,367 (80%)	10,101 (80%)	8,402 (77%)	25,206 (77%)	13,567 (74%)	40,701 (74%)	13163 (69%)	39,489 (69%)
Angina (%)	1,924 (46%)	594 (5%)	5,161 (47%)	2,445 (7%)	8,623 (47%)	5,528 (10%)	9,122 (48%)	7,472 (13%)
Heart failure (%)	205 (5%)	61 (0%)	676 (6%)	338 (1%)	1,568 (9%)	982 (2%)	2,198 (12%)	1,674 (3%)
Cardiovascular system conditions (%)	979 (23%)	681 (5%)	3,154 (29%)	2,941 (9%)	6,591 (36%)	7,672 (14%)	8,205 (43%)	11,674 (20%)
Chronic kidney disease (%)	1 (0%)	0 (0%)	4 (0%)	8 (0%)	965 (5%)	1,392 (3%)	1872 (10%)	3,039 (5%)
Diabetes (%)	449 (11%)	624 (5%)	1,622 (15%)	2,297 (7%)	3,398 (18%)	5,573 (10%)	3,726 (20%)	6,876 (12%)
Hypercholesterolaemia (%)	1,634 (39%)	1,907 (15%)	4,228 (39%)	7,423 (23%)	6,392 (35%)	14,936 (27%)	6,395 (33%)	15,814 (28%)

Continued on next page

Table 5.1 – *Continued from previous page*

	Age 60		Age 65		Age 70		Age 75	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Hypertension (%)	1,168 (28%)	1,991 (16%)	3,750 (34%)	7,608 (23%)	7,411 (40%)	17,955 (32%)	8,579 (45%)	22,330 (39%)
Alcohol consumption status recorded (%)	7,679 (61%)	2,874 (69%)	23,183 (71%)	8,663 (80%)	42,571 (77%)	15,746 (85%)	46,359 (81%)	16,894 (88%)
Alcohol consumer[1] (%)	3,385 (81%)	10,997 (88%)	8,780 (81%)	28,130 (86%)	14,494 (79%)	45,962 (83%)	14,293 (75%)	45,504 (79%)
Body mass index recorded (%)	7,664 (61%)	2,973 (71%)	23,023 (71%)	8,793 (81%)	42,655 (77%)	15,876 (86%)	46,048 (80%)	16,919 (89%)
Overweight[1] (%)	2,427 (58%)	7,239 (58%)	5,866 (54%)	17,609 (54%)	9,406 (51%)	28,253 (51%)	9,264 (49%)	28,030 (49%)
Obese[1] (%)	750 (18%)	1,418 (11%)	2,295 (21%)	4,687 (14%)	4,107 (22%)	9,180 (17%)	3,848 (20%)	9,365 (16%)
Smoking status recorded (%)	8,692 (69%)	3,244 (77%)	25,498 (78%)	9,493 (87%)	46,766 (85%)	16,817 (91%)	50,665 (88%)	17,884 (94%)
Ex-smoker[1] (%)	1,274 (30%)	2,398 (19%)	4,611 (42%)	10,903 (33%)	8,335 (45%)	19,305 (35%)	8695 (46%)	20,641 (36%)
Smoker[1] (%)	1,163 (28%)	3,507 (28%)	2,203 (20%)	6,544 (20%)	3,079 (17%)	8,973 (16%)	2,545 (13%)	7,660 (13%)

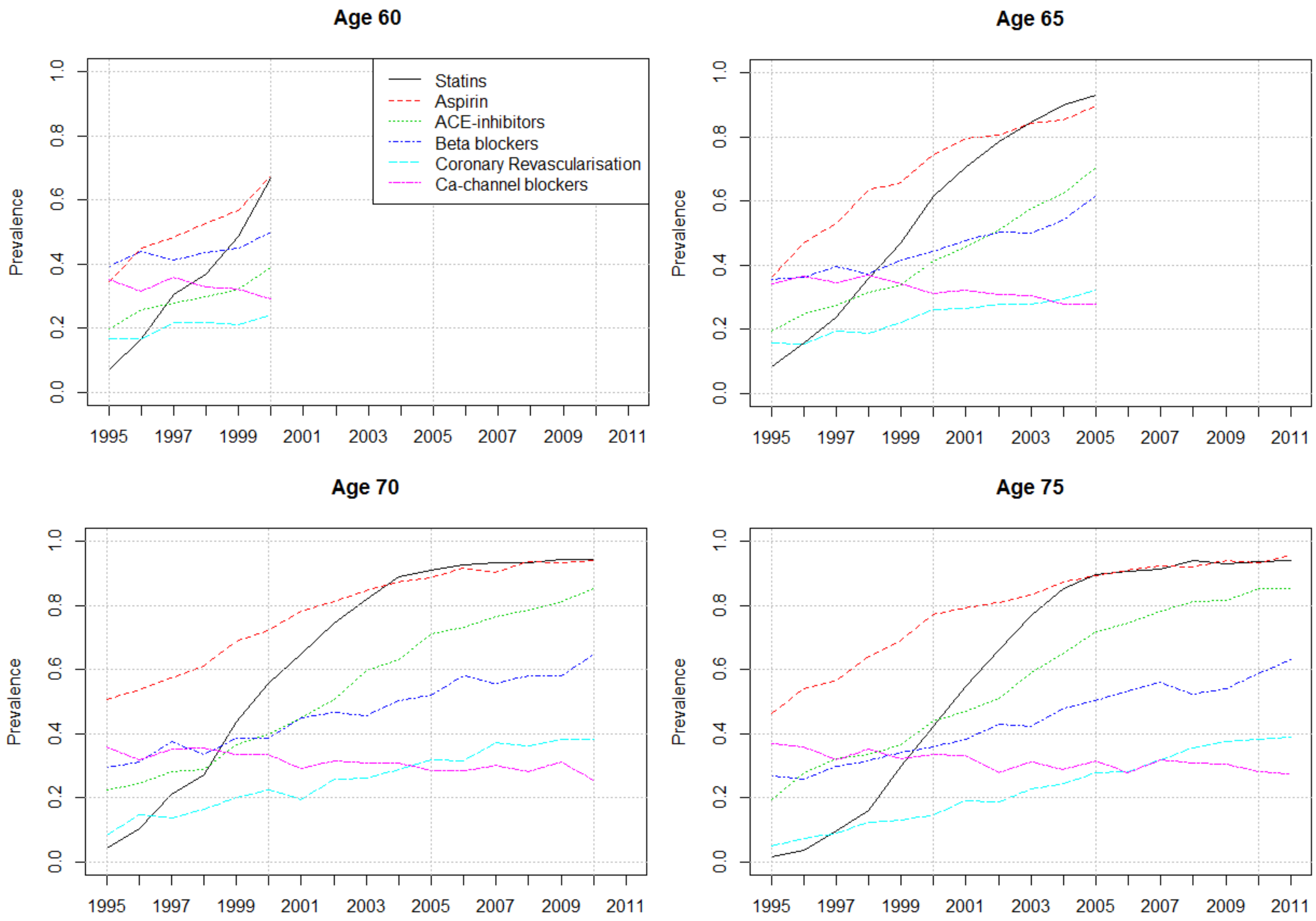


Figure 5.2: Prevalence of treatments by cohort's age in patients with acute myocardial infarction

The age cohorts differed in recruitment period. The prevalence prior to 1995 is not presented due to the small numbers of medical records available.

The clinical and lifestyle factors hypertension, hypercholesterolaemia, obesity (BMI$\geq$30 kg/m^2), and ex-smoking were more common among AMI cases than controls, see Table 5.1. Overweight (BMI 25-30 kg/m^2) and smoking was as common among cases as controls. Alcohol consumption was less common among cases than controls. Given calendar year, the prevalence of obesity and smoking in AMI patients decreased with increasing age, see Appendix B Figure B.2. Given calendar year, the prevalence of overweight and ex-smoking in AMI patients remained approximately the same with increasing age, while the prevalence of alcohol consumption decreased, and the prevalence of hypertension increased. The prevalence of hypercholesterolaemia was more common in younger AMI patients prior to 2000, and was approximately the same across age thereafter. From 1995 to 2011, the prevalence of alcohol consumption and overweight in AMI patients remained approximately the same; the prevalence of hypertension, ex-smoking, and obesity increased; and the prevalence of smoking decreased. Interestingly, the prevalence of hypercholesterolaemia increased from 1995 to 2000/2002 and decreased after that. This trend was the most pronounced in the oldest cohort, in which the prevalence increased from 15% in 1995 to 43% in 2002, after which it decreased to approximately 30% in 2011. The decrease in hypercholesterolaemia could reflect the cholesterol lowering effect by the more widespread prescription of statins. The prevalence order of the clinical and lifestyle factors in AMI patients changed over time and this varied by age, although smoking remained the least prevalent and alcohol consumption the most prevalent. From 2005 onwards, the prevalence of smoking was between 10-20%; obesity and hypercholesterolaemia between 20-40%; ex-smoking, overweight, and hypertension between 40-60%; and alcohol consumption between 60-80%.

The prevalence of coronary revascularisation and drug therapy was higher among patients who had multiple AMIs compared with patients who had a single AMI, see Table 5.2. The rates across the four age cohorts for coronary artery bypass graft (CABG) and percutaneous coronary intervention (PCI) were 16-19% and 3-8%, respectively, see Appendix B Table B.6. Men were approximately twice as likely to have had coronary revascularisation as women were, which could not be explained by age, deprivation, or diabetes, see Figure 5.3 and Appendix B Table B.7. Men and women were equally likely to be prescribed drugs. From 1995 to 2011, the prevalence of coronary revascularisation and drug therapy increased substantially, with the exception of prescription of calcium-channel blockers that decreased over the years, see Figure 5.2. The difference in treatment prevalence by the four initial ages converged over time. In 2010 the most widely prescribed drugs to AMI patients were statins and aspirin (both 94%) followed by ACE inhibitors (85%), beta blockers (65%), and calcium-channel blockers (25%). In the same year, 38% of the AMI patients have had coronary revascularisation; the prevalence was greater in patients living in the most affluent areas (IMD category 1: 45%) than in patients living in the most deprived areas (IMD category 5: 32%; trend $\chi^2(1)$=5.06, p=.20).

Table 5.2: Baseline treatments by ischaemic heart disease

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. History of ischaemic heart disease was a factor categorised as: no history, angina only, single AMI with possibly angina, or multiple AMIs with possibly angina. Thus, the former two categories consisted of the controls and the latter two categories consisted of the cases. The prevalences of treatments by the cohort's age were affected by calendar year, see Figure 5.2.

	Ischaemic heart disease	Size	Coronary revascularisation		Drug therapy				
			Men	Women	Aspirin	ACE inhibitors	Beta blockers	Statins	Ca-channel blockers
Age 60	No	11,964	0 (0%)	0 (0%)	271 (2%)	678 (6%)	1,156 (10%)	208 (2%)	615 (5%)
	Angina	594	97 (19%)	7 (8%)	211 (36%)	95 (16%)	238 (40%)	148 (25%)	264 (44%)
	Single AMI	3,465	486 (18%)	77 (11%)	1,467 (42%)	768 (22%)	1,482 (43%)	951 (27%)	1,080 (31%)
	Multiple AMIs	721	194 (32%)	20 (18%)	386 (54%)	256 (36%)	314 (44%)	247 (34%)	290 (40%)
Age 65	No	30,201	0 (0%)	0 (0%)	2,548 (8%)	3,299 (11%)	3,727 (12%)	2,194 (7%)	2,709 (9%)
	Angina	2,445	512 (25%)	30 (7%)	1,400 (57%)	701 (29%)	1,036 (42%)	1,164 (48%)	1,024 (42%)
	Single AMI	8,796	1532 (23%)	334 (16%)	5,751 (65%)	3,452 (39%)	4,011 (46%)	4,722 (54%)	2,762 (31%)
	Multiple AMIs	2,086	594 (35%)	67 (17%)	1,532 (73%)	1,072 (51%)	946 (45%)	1,272 (61%)	722 (35%)
Age 70	No	49,768	0 (0%)	0 (0%)	8,698 (17%)	9,756 (20%)	7,176 (14%)	8,863 (18%)	6,820 (14%)
	Angina	5,528	1,263 (28%)	125 (12%)	3,851 (70%)	2,204 (40%)	2,376 (43%)	3,335 (60%)	2,235 (40%)
	Single AMI	14,847	2,811 (26%)	730 (18%)	11,269 (76%)	7,770 (52%)	6,989 (47%)	9,638 (65%)	4,461 (30%)
	Multiple AMIs	3,585	1,012 (36%)	172 (22%)	2,918 (81%)	2,202 (61%)	1,721 (48%)	2,524 (70%)	1,219 (34%)
Age 75	No	49,822	0 (0%)	0 (0%)	12,592 (25%)	12,633 (25%)	7,945 (16%)	11,318 (23%)	8,574 (17%)
	Angina	7,472	1,652 (29%)	225 (13%)	5,642 (76%)	3,430 (46%)	3,188 (43%)	4,780 (64%)	2,952 (40%)
	Single AMI	15,319	2,705 (26%)	835 (17%)	12,487 (82%)	9,226 (60%)	7,036 (46%)	10,395 (68%)	4,676 (31%)
	Multiple AMIs	3,779	954 (35%)	230 (23%)	3,295 (87%)	2,574 (68%)	1,759 (47%)	2,767 (73%)	1,228 (32%)

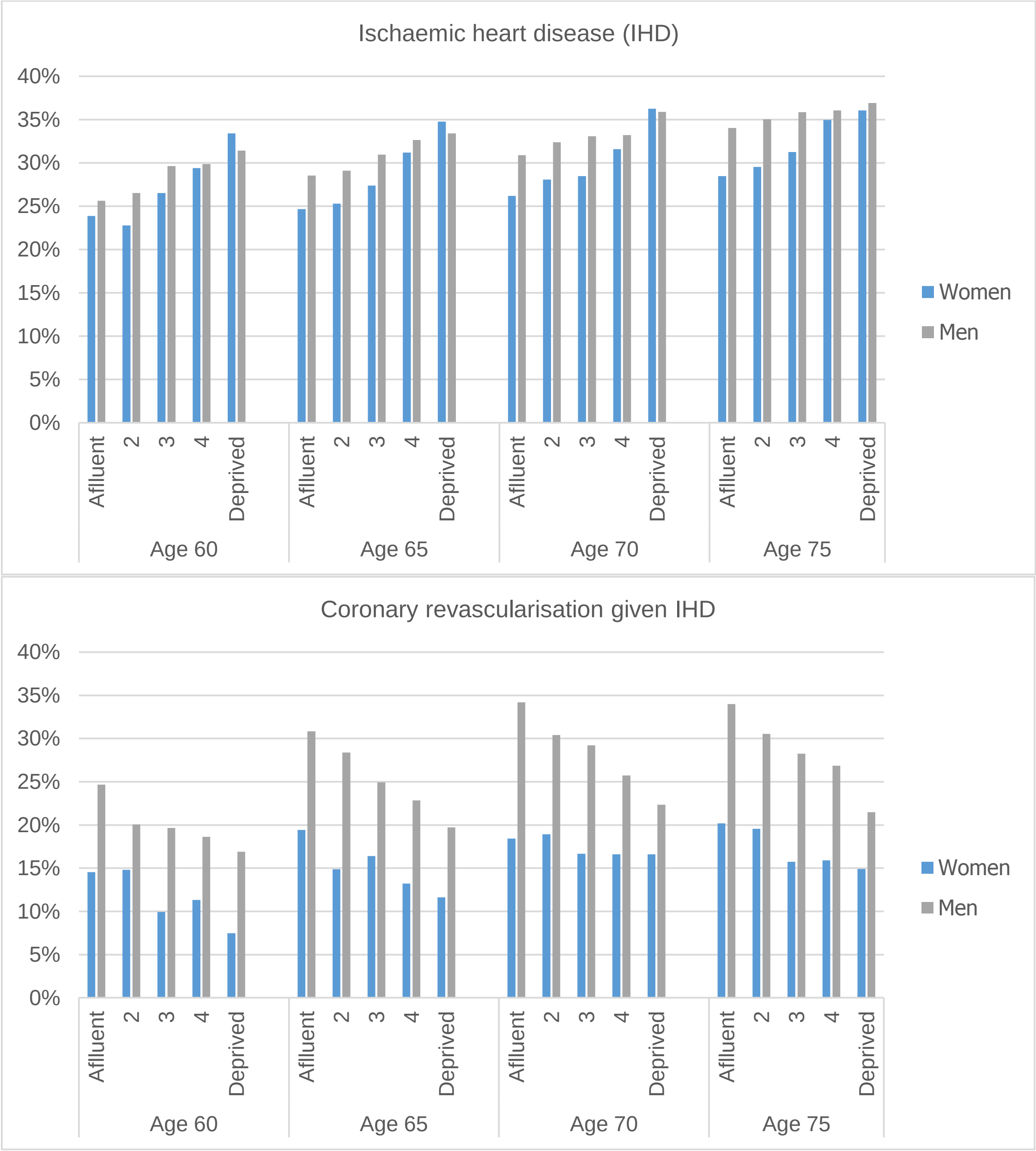


Figure 5.3: Prevalence of ischaemic heart disease (IHD) and coronary revascularisation given IHD, by deprivation

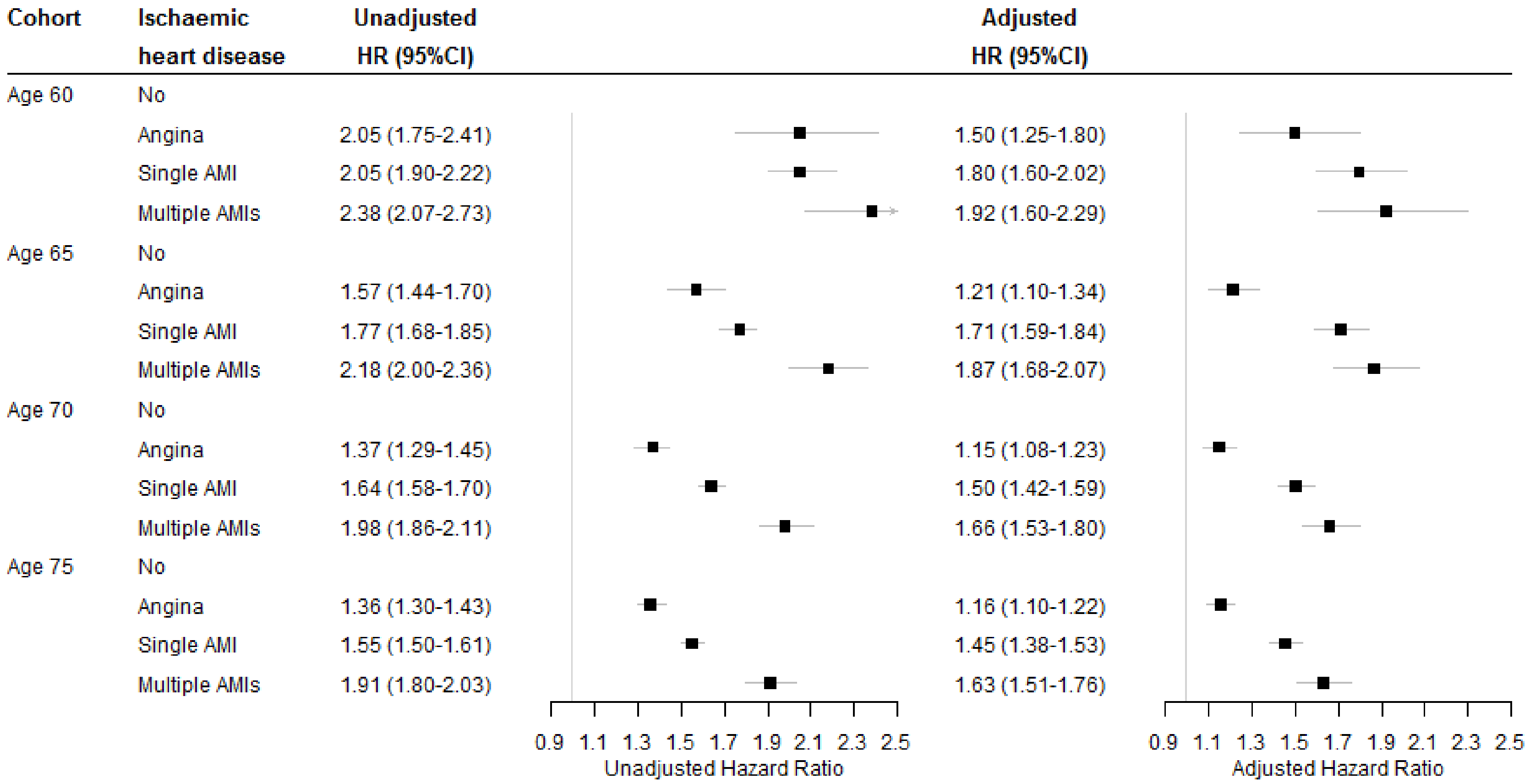
The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. History of IHD consisted of AMI and angina. Patients of both subtypes of IHD could be eligible for coronary revascularisation. Deprivation was measured by index of multiple deprivation (IMD) quintiles.

5.3 Survival models

The survival models developed for ages 60, 65, 70, and 75 are presented in forest plots. The results on AMI and related treatments across the age cohorts can be found in this Section Figures 5.4 and 5.5, respectively. The results of the survival models by age cohort can be found in Appendix B Figures B.3-B.6.

The adjusted hazard of all-cause mortality for AMI patients was constant during follow-up of 24 years; it did not matter how many years the cases had already survived, they were still at a higher risk of dying than the controls. This relative hazard was the greatest in the youngest cohort while the absolute hazard was the greatest in the oldest cohort, see Figure 5.4 and Appendix B Figure B.7.

Due to the way AMI and angina interacted, a factor representing IHD was generated as described above. Compared with no history of IHD by age 60, 65, 70, or 75, having had one AMI was associated with a hazard of mortality of 1.80 (95% confidence interval 1.60-2.02), 1.71 (1.59-1.84), 1.50 (1.42-1.59), or 1.45 (1.38-1.53), respectively. This translates to an increase in effective age of 5.9 (4.7-7.0), 5.4 (4.6-6.1), 4.1 (3.5-4.6), or 3.7 (3.2-4.3) years, respectively. Compared with no history of IHD by age 60, 65, 70, or 75, having had multiple AMIs was associated with a hazard of mortality of 1.92 (1.60-2.29), 1.87 (1.68-2.07), 1.66 (1.53-1.80), or 1.63 (1.51-1.76), respectively. This translates to an increase in effective age of 6.5 (4.7-8.3), 6.2 (5.2-7.3), 5.1 (4.3-5.9), or 4.9 (4.1-5.6) years, respectively. The hazard of mortality did not differ between cases with or without a history of angina. Cases and controls differed in survival benefits of prescriptions of beta blockers and calcium-channel blockers, which are described below. There were no other interactions with a history of AMI, meaning that the effect of AMI on the hazard of mortality was the same for different subgroups of patients, such as for men and women.



Cohort	Ischaemic heart disease	Unadjusted HR (95%CI)	Adjusted HR (95%CI)
Age 60	No		
	Angina	2.05 (1.75-2.41)	1.50 (1.25-1.80)
	Single AMI	2.05 (1.90-2.22)	1.80 (1.60-2.02)
	Multiple AMIs	2.38 (2.07-2.73)	1.92 (1.60-2.29)
Age 65	No		
	Angina	1.57 (1.44-1.70)	1.21 (1.10-1.34)
	Single AMI	1.77 (1.68-1.85)	1.71 (1.59-1.84)
	Multiple AMIs	2.18 (2.00-2.36)	1.87 (1.68-2.07)
Age 70	No		
	Angina	1.37 (1.29-1.45)	1.15 (1.08-1.23)
	Single AMI	1.64 (1.58-1.70)	1.50 (1.42-1.59)
	Multiple AMIs	1.98 (1.86-2.11)	1.66 (1.53-1.80)
Age 75	No		
	Angina	1.36 (1.30-1.43)	1.16 (1.10-1.22)
	Single AMI	1.55 (1.50-1.61)	1.45 (1.38-1.53)
	Multiple AMIs	1.91 (1.80-2.03)	1.63 (1.51-1.76)

Figure 5.4: Unadjusted and adjusted hazards of all-cause mortality associated with ischaemic heart disease

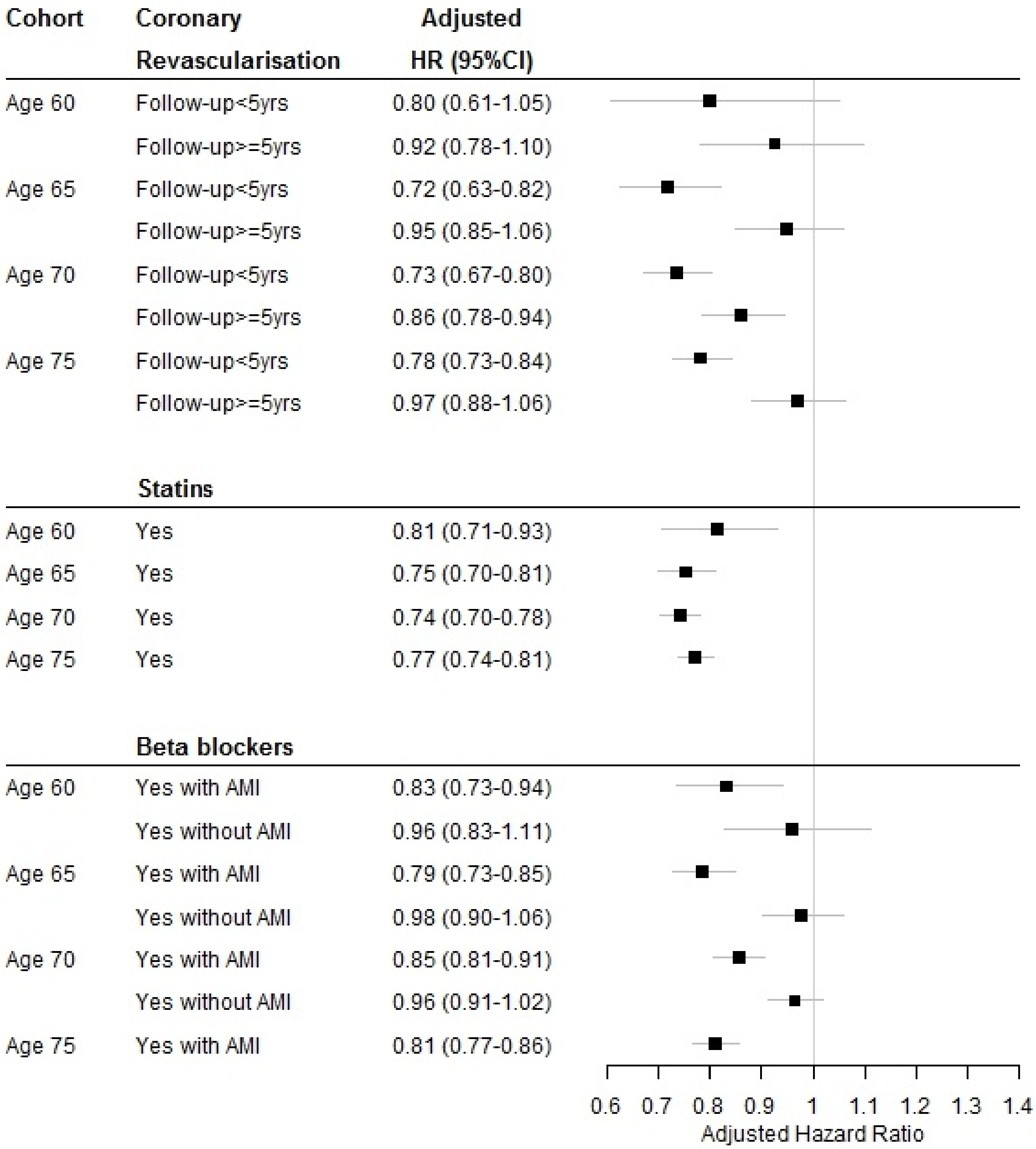
The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Due to the way AMI and angina interacted, a factor representing ischaemic heart disease was generated that had the following levels: no history, angina only, single AMI with possibly angina, or multiple AMIs with possibly angina. The hazard ratios (95% confidence interval) were adjusted for sex, year of birth, socioeconomic status, heart failure, cardiovascular system conditions, chronic kidney disease (only at ages 70 and 75), diabetes, hypertension, hypercholesterolaemia, coronary revascularisation, ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, statins, alcohol consumption status, body mass index, smoking status, and general practice.

The comorbidities that had the greatest impact on survival were cardiovascular system conditions and heart failure, see Appendix B Figures B.3-B.6. The hazard of mortality associated with a history of medical conditions was constant over time; it did not matter how many years patients have had the history, they were still at a higher risk of dying than patients without the history. The exception to this was cardiovascular system conditions, where the relative hazard decreased over time and thus the survival prospects improved over time. In general, the relative hazard associated with the comorbidities was greatest in the youngest cohort. On average the comorbidities led to an additional increase in effective age of 3.5 to 7.0 years at all ages.

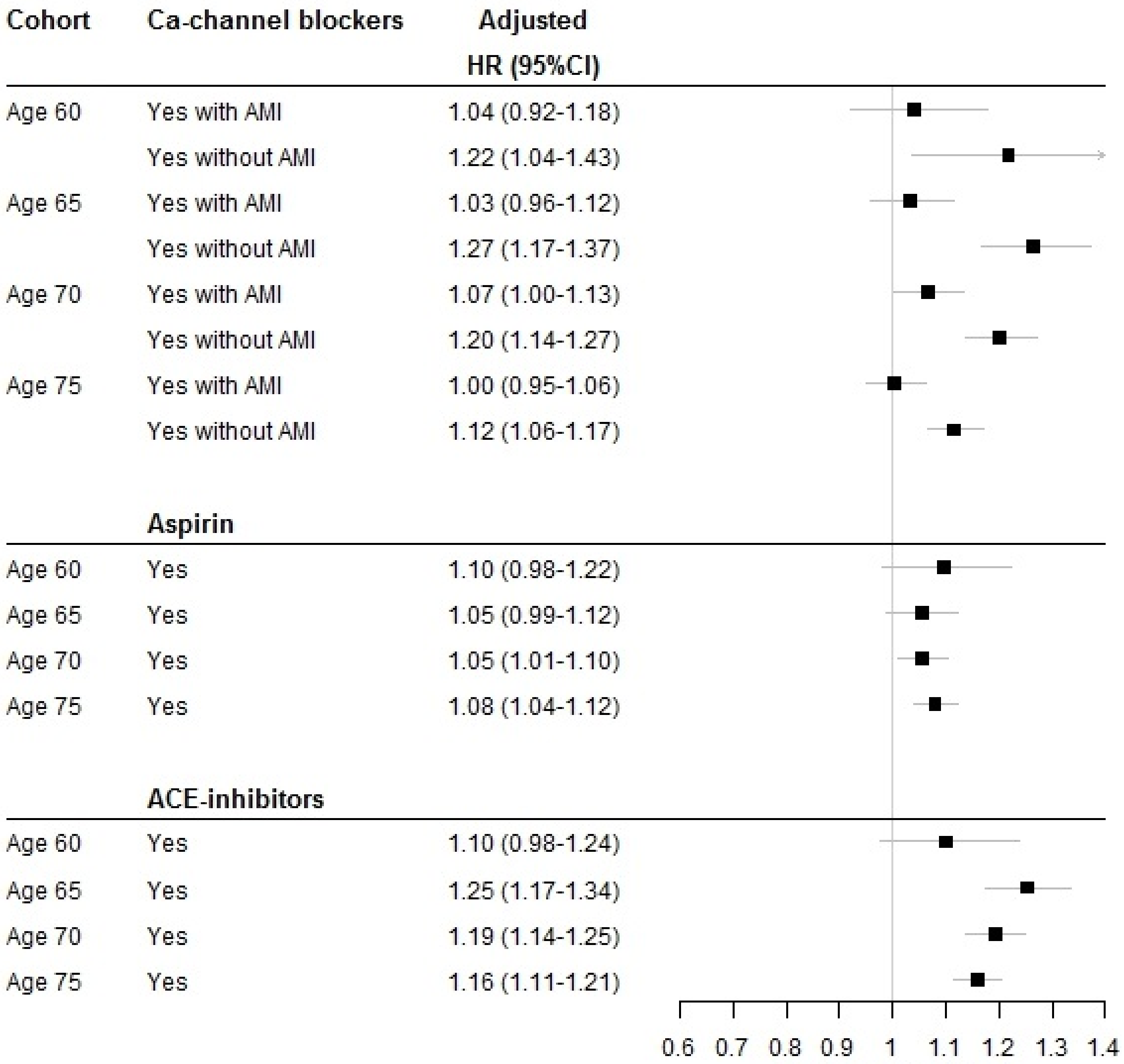
The lifestyle factor that had the greatest impact on survival was smoking, see Appendix B Figures B.3-B.6. The hazard of mortality associated with smoking compared with non-smoking was greater in normal weight patients than in overweight or obese patients. This translates to an increase in effective age across the four age cohorts of 7.5 to 10.0 and 5.5 to 8.0 years, respectively. The relative hazard was greatest in the youngest cohort. In non-smokers, the survival prospects of overweight patients were not significantly different from normal weight patients. Alcohol consumption was associated with improved survival benefits, translating to a decrease in effective age of 0.5 to 1.5 years at all ages.

Coronary revascularisation was associated with a significant improvement in the survival prospects in the short-term, see Figure 5.5. Compared with no history of coronary revascularisation by age 60, 65, 70, or 75, having had revascularisation was associated with a hazard of mortality of 0.80 (0.61-1.05), 0.72 (0.63-0.82), 0.73 (0.67-0.80), or 0.78 (0.73-0.84), respectively, in the first five years of follow-up. This translates to a decrease in effective age of 2.3 (-0.5-5.0), 3.3 (2.0-4.7), 3.1 (2.2-4.0), or 2.5 (1.7-3.2) years, respectively. After five years of follow-up, a history of coronary revascularisation was no longer associated with a significant improvement in the survival prospects. These prospects were the same for different subgroups of patients, such as for men and women.

Drug therapy was associated with mixed survival prospects and could differ by subgroups of patients, see Figure 5.5. The drug therapy that was associated with the greatest improved survival prospects was prescription of statins; the prescription translated to an average decrease in effective age of 2.5 years at all ages. The hazard of mortality associated with statin prescription did not differ between patients with or without a history of hypercholesterolaemia. Prescription of beta blockers was associated with mixed survival prospects; prescription translated to an average decrease in effective age of 2.0 years at all ages in AMI patients versus no decrease in patients without AMI. Prescription of calcium-channel blockers was also associated with mixed survival prospects; prescription translated to no decrease in effective age in AMI patients versus an average increase in effective age of 2.0 years in patients without AMI. Prescription of aspirin or ACE inhibitors was associated with worsened survival prospects; the prescription translated to an average increase in effective age of 1.0 or 1.5 years, respectively, at all ages. There were no significant differences in the effects of the treatments by other subgroups of patients than described above, such as for men and women.



Cohort	Coronary Revascularisation	Adjusted HR (95%CI)
Age 60	Follow-up<5yrs	0.80 (0.61-1.05)
	Follow-up>=5yrs	0.92 (0.78-1.10)
Age 65	Follow-up<5yrs	0.72 (0.63-0.82)
	Follow-up>=5yrs	0.95 (0.85-1.06)
Age 70	Follow-up<5yrs	0.73 (0.67-0.80)
	Follow-up>=5yrs	0.86 (0.78-0.94)
Age 75	Follow-up<5yrs	0.78 (0.73-0.84)
	Follow-up>=5yrs	0.97 (0.88-1.06)
	Statins	
Age 60	Yes	0.81 (0.71-0.93)
Age 65	Yes	0.75 (0.70-0.81)
Age 70	Yes	0.74 (0.70-0.78)
Age 75	Yes	0.77 (0.74-0.81)
	Beta blockers	
Age 60	Yes with AMI	0.83 (0.73-0.94)
	Yes without AMI	0.96 (0.83-1.11)
Age 65	Yes with AMI	0.79 (0.73-0.85)
	Yes without AMI	0.98 (0.90-1.06)
Age 70	Yes with AMI	0.85 (0.81-0.91)
	Yes without AMI	0.96 (0.91-1.02)
Age 75	Yes with AMI	0.81 (0.77-0.86)



Cohort	Ca-channel blockers	Adjusted HR (95%CI)
Age 60	Yes with AMI	1.04 (0.92-1.18)
	Yes without AMI	1.22 (1.04-1.43)
Age 65	Yes with AMI	1.03 (0.96-1.12)
	Yes without AMI	1.27 (1.17-1.37)
Age 70	Yes with AMI	1.07 (1.00-1.13)
	Yes without AMI	1.20 (1.14-1.27)
Age 75	Yes with AMI	1.00 (0.95-1.06)
	Yes without AMI	1.12 (1.06-1.17)
	Aspirin	
Age 60	Yes	1.10 (0.98-1.22)
Age 65	Yes	1.05 (0.99-1.12)
Age 70	Yes	1.05 (1.01-1.10)
Age 75	Yes	1.08 (1.04-1.12)
	ACE-inhibitors	
Age 60	Yes	1.10 (0.98-1.24)
Age 65	Yes	1.25 (1.17-1.34)
Age 70	Yes	1.19 (1.14-1.25)
Age 75	Yes	1.16 (1.11-1.21)

Figure 5.5: Adjusted hazards of all-cause mortality associated with treatments for ischaemic heart disease

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. The hazard ratios (95% confidence interval) were adjusted for sex, year of birth, socioeconomic status, ischaemic heart disease, heart failure, cardiovascular system conditions, chronic kidney disease (only at ages 70 and 75), diabetes, hypertension, hypercholesterolaemia, alcohol consumption status, body mass index, smoking status, general practice, and the listed treatments. The hazard associated with coronary revascularisation was split at five years of follow-up after the cohort's age. The hazard associated with statin was the same in patients with and without hypercholestolaemia.

Survival prospects also differed by socioeconomic status, in which the difference was greater at a younger age. The Mosaic category 5 ('neighbourhood with mainly young couples') was associated with the worst survival prospects for patients aged 60 and older, this ranged from a hazard of mortality of 1.73 (1.42-2.10) at age 60 to 1.33 (1.22-1.44) at age 75, see Appendix B Figures B.3-B.6. This translates to an increase in effective age of 5.5 years at age 60 and 3.0 years at age 75. In addition, survival prospects varied considerably between general practices. The 95% tolerance interval of the hazard of mortality associated with general practice was (0.79 to 1.21) at age 60 and (0.55 to 1.45) at an older age. This translates to a maximum of 4.5 and 10 years difference in effective age between general practices, respectively. A general practice could serve a range of patients with regards to health status, ethnic background, deprivation, urbanisation, and pollution. These factors, however, were not correlated with the hazard of mortality associated with general practice, see Appendix B Table B.8. Moreover, adjusting for these factors in the final survival models did not attenuate the hazard of mortality associated with general practice.

The variables contributing the most to the survival models in explaining survival variations in the age cohorts were IHD, cardiovascular system conditions, interaction between BMI and smoking status, and general practice. The variables contributing the least to the survival models were hypertension, prescription of aspirin and calcium-channel blockers, and alcohol consumption status.

5.4 Evaluation

5.4.1 Performance statistics

The final survival models explained 20 to 29% of survival differentials. There was 68 to 70% concordance between the estimated hazard of mortality and survival time. The shrinkage slopes indicated that the adjusted effects were overestimated by less than 3%. These performance statistics are typical for survival analysis and the small shrinkage slopes suggest that the results are robust.

The survival models that only included history of IHD estimated 1.08 and 1.19 times higher hazards of mortality associated with single and multiple AMIs, respectively, see Figure 5.4. These models explained less than 1% of survival differentials and had between 56-59% concordance between the estimated hazard of mortality and survival time. The difference between the unadjusted and adjusted estimates and the respective model performances demonstrate the importance of adjusting for confounders when estimating the effects of medical conditions and treatments associated with the hazard of mortality.

5.4.2 Internal validation

The complete case analysis, estimated on only complete medical records, provided similar results as the final survival models that were estimated on both complete and incomplete records, see Appendix B Figures B.3-B.6. Both types of analyses adjusted for the same variables. The complete case analysis did not find a significant difference in the hazard of mortality associated with calcium-channel blockers in AMI patients compared with other patients, and also did not find a significant difference in the hazards of hypertension over time in the oldest two cohorts.

Both analyses estimated similar hazard ratios with overlapping confidence intervals. In the youngest cohort, the difference in estimates was the greatest in socioeconomic status, BMI, and smoking status. The complete case analysis estimated the hazard ratio of socioeconomic status on average 0.17 lower and the hazard ratios of BMI and smoking status on average 0.15 higher. The difference in these estimates decreased with increasing age, where the HR difference in the oldest cohort was only 0.03 and 0.01, respectively. In the oldest cohort, the difference in estimates was the greatest in AMI and cardiovascular system conditions after five years of follow-up, where the complete case analysis estimated a HR of 0.10 to 0.23 greater.

The overall performance, discrimination, and validation statistics of the two types of analyses were also alike, see Appendix B Table B.9. The survival models developed in the complete case analysis explained 23 to 30% of the survival variations, determined 69 to 70% concordance between the estimated hazard of mortality and survival time, and overestimated the effects by 1 to 4%.

5.4.3 External validation

This matched cohort study estimated the adjusted hazards of all-cause mortality associated with a history of AMI and respective treatments by age 60, 65, 70, or 75 in UK residents using medical records from primary care between 1987 and 2011. In accordance with the literature, this study found that AMI survivors have a long-term, increased hazard of mortality, in which younger survivors and survivors of multiple events were worse off (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). However, this study estimated lower hazards of mortality than previously estimated. The lower estimated hazards of mortality associated with a history of AMI reported by this study compared with previous studies could be due to the different data source used and the range of confounders adjusted for. This study made use of primary care data, whereas most studies used hospital and register data. Research showed that the 1-year mortality rate of AMI is lower in primary care probably because of a lower proportion of severe cases (Herrett et al., 2013b). Furthermore, this study adjusted for a range of confounders, which attenuated the estimated hazards of mortality associated with a history of AMI. There is a smaller difference between the unadjusted estimates of this study and the sex-standardised and age-standardised mortality ratios estimated in residents of England based on hospital and register data from 2004 to 2010 by Smolina et al. (2012b). That study concluded that after 7 years, people with a first or recurrent AMI had double or triple the risk of mortality, respectively, compared with the general population of equivalent sex and age (Smolina et al., 2012b). It is unlikely that the lower estimated hazards of mortality reported by this study are due to the shifting epidemiological trends in CVD because there were no interactions between a history of AMI and year of birth category or other risk factors with the exception of angina, beta blockers, and calcium-channel blockers. The medical advances and shifting prevalence of risk factors over time were adjusted for in the analysis and had no different survival effects in AMI patients compared with patients without AMI.

This study tested all second-order interaction effects between AMI and the confounders and found only a significant interaction with angina, beta blockers, and calcium-channel blockers at all initial ages. Two studies that estimated the long-term survival prospects after angina or AMI also found that angina was less hazardous than AMI, however the studies did not estimate the survival prospects when both types of ischaemic heart diseases were present (Capewell et al., 2006; Chang et al., 2003). This study found a borderline insignificant interaction between AMI and heart failure. The insignificance could be due to the low prevalence of heart failure and therefore not having enough power to test the interaction. This study did not find a sex difference in survival prospects after AMI. This is supported by some studies (Chang et al., 2003; Gottlieb et al., 2000; Koek et al., 2007; Rosengren et al., 2001) but contradicted by another (Smolina et al., 2012b). The difference could be explained by (the lack of) adjustment for comorbidities and treatments.

This study found that the lower uptake of coronary revascularisation by women could not be explained by age, diabetes, or deprivation, as suggested by a previous study (Chang et al., 2003). A study with data from the UK from 2003 to 2008, showed that coronary revascularisation was more prevalent in non-STEMIs than in STEMIs (Herrett et al., 2013a). As non-STEMIs are more common among women than among men (Herrett et al., 2013a), it seems that type of AMI could not explain the sex difference in uptake of surgery present in this study. In 2012, the European Society for Cardiology reviewed the sex differences in treatment after AMI, taking into account sex differences in risk profiles, and concluded that sex differences exist (Chieffo et al., 2012). Patients from deprived areas had also lower uptake of coronary revascularisation.

This study also found that a history of coronary revascularisation was no longer associated with a significantly improved survival prospects after 5 years of follow-up. This is in accordance with another study that reported a protective effect in the 1-year mortality rate but an insignificant effect in the 5-year mortality rate of AMI (HR=0.76 (0.67-0.85) and HR=0.91 (0.78-1.08), respectively) (Chang et al., 2003). The findings suggest that coronary revascularisation might mainly be beneficial in reducing early mortality. No sex difference in survival after coronary revascularisation was found in this study, which is supported by some studies (Chang et al., 2003; Smolina et al., 2012b) but contradicted by another (Lagerqvist et al., 2001).

This study found no difference in drugs prescriptions by sex by 2010. A previous study that looked at treatment after AMI in the UK from 1991 to 2002 did report differences in prescription (Hardoon et al., 2011). It seems that the sex difference in drugs prescription converged over time. The current study found that survival was better in those who were prescribed statins or beta blockers, but worse in those prescribed aspirin or ACE inhibitors, and unchanged in those prescribed calcium-channel blockers. The estimated hazards of mortality associated with these treatments were almost the same at each age, implying that the effectiveness of treatments does not differ by age. The findings of this study agree with the clinical evidence reviewed by NICE (2013a) on the effectiveness of statins and calcium-channel blockers, but disagree with the effectiveness of ACE inhibitors, aspirin, and beta blockers.

This study found that prescription of ACE inhibitors was associated with increased hazard of mortality, whereas the NICE review estimated a protective effect in AMI patients with left ventricular systolic dysfunction (LVSD; relative risk (RR) of 0.84 (0.78-0.91)) and an inconclusive effect in AMI patients with unselected LVSD (RR=1.02 (0.57-1.84)) (NICE, 2013a). The studies included used data from 1986-1993 (AIRE Study, 1993; Borghi et al., 1998; Køber et al., 1995; Pfeffer et al., 1992; SOLVD Investigators, 1992). A study using data from the same period found inconclusive effects of ACE inhibitors in AMI patients with diabetes (HR=1.15 (0.79-1.66)) and hazardous effects in AMI patients with no diabetes (HR=1.52 (1.15-2.01)), which the authors explained by confounding in respect to heart failure and the use of old data from 1985-1992 (Löwel et al., 2000). Two studies that made use of longer follow-up also found hazardous effects of ACE inhibitors in AMI patients: a hazard of 1.54 (1.07-2.23) was found using data from 1988-2001 of the United States (Nigam et al., 2006), and a hazard of 1.91 (1.64-2.27) or 2.23 (1.89-2.62) for ACE inhibitors initiated in hospital or at discharge, respectively, were found using data from 1984-2005 of Australia as part of the MONICA study (Briffa et al., 2009). The authors of MONICA study suggested that these findings on the hazardous effects of ACE inhibitors might be due to confounding by indication. The current study controlled for heart failure, which lowered the HR of ACE inhibitors by ∼0.05, and made use of more recent data from 1987 to 2011, thereby suggesting that ACE inhibitors might in fact be harmful to survival.

This study found that prescription of aspirin was associated with increased hazard of mortality, whereas the NICE review only included one study that estimated an inconclusive protective effect of the drug versus placebo on all-cause mortality (NICE, 2013a). That study included men with a recent AMI aged 30-64 in 1972-1974 (CDP, 1976). The current study made use of more recent data with longer follow-up of older patients of both sexes. Aspirin is associated with an increased risk of bleeding, where the risk increases with age (NICE, 2013a). Since the elderly are excluded from most clinical trials, it could be that aspirin might actually be harmful in the elderly as the findings of the current study suggest. Since 2013, DAPT is recommend as a first line treatment to AMI patients (NICE, 2013b). Studies included in the NICE review reported that DAPT is more effective than aspirin on its own (NICE, 2013a). For example, a study in Australia estimated a hazard of 0.77 (0.65-0.90) or 0.74 (0.64-0.85) for antiplatelet drugs initiated in AMI patients in the hospital or at discharge, respectively (Briffa et al., 2009).

This study found that prescription of beta blockers was associated with significant survival benefits, whereas the NICE review reported uncertain survival benefits in AMI patients who received the drugs in the first 72 hours (RR=0.87 (0.67-1.20)) or after the first 72 hours to a year (RR=0.76 (0.49-1.16)) (NICE, 2013a). The studies included in the review used data from 1976 to 1985 (BHAT Research Group, 1982; LIT Research Group, 1987; Olsson et al., 1985; Pedersen, 1983; Roberts et al., 1984; Roque et al., 1987). Studies using more recent data estimated significant survival benefits associated with beta blockers: a study in Germany using data from 1985-1992 estimated a hazard of 0.62 (0.45-0.85-0.85) or 0.64 (0.52-0.78) in AMI patients with diabetes or no diabetes, respectively (Löwel et al., 2000); a study in the United States using data from 1988-2001 estimated a hazard of 0.60 (0.45-0.80) in AMI patients; a study in Australia using data from 1984-2005 estimated a hazard of 0.55 (0.48-0.63) or 0.56 (0.50-0.64) for beta blockers initiated in AMI patients in the hospital or at discharge, respectively (Briffa et al., 2009); and a study in England using data from

2003-2008 estimated a hazard of 0.59 (0.44-0.79) or 0.50 (0.36-0.69) for beta blockers initiated in patients with chronic obstructive pulmonary disease (COPD) before or after AMI, respectively (Quint et al., 2013).

Finally, this study found that survival prospects varied greatly across general practices, which was independent from health status, ethnic background, deprivation, urbanisation, and air pollution. A study by Gerber et al. (2010) estimated the effect of neighbourhood and individual socioeconomic status on survival after AMI and suggested that higher level measured socioeconomic status might capture residual confounding of unequal hospital resources and social characteristics of an area such as social cohesion and attitudes towards health. Other studies on survival prospects after AMI assumed that there were no survival variations by hospitals (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; Quint et al., 2013; Smolina et al., 2012b), although the study by Chang et al. (2003) did adjust for cardiovascular specialist, hospitals capable of performing coronary artery bypass graft, and health region. Failing to adjust for survival variations by care provider could result in false precision of survival prospects (Therneau and Grambsch, 2000).

5.4.4 Strengths and limitations

This study used routinely collected primary care data that were representative of the UK (Blak et al., 2011; Hall, 2009). The advantage of using primary care data was that there was more information on socio-demographic and lifestyle factors available and there was a higher coverage of AMI cases (Herrett et al., 2013b). Another strength could be found in the study design; the matched cohort study design allowed to estimate the effect of a history of AMI on mortality compared with no history of AMI while adjusting for a wide range of confounders. The confounders included comorbidities, treatments, lifestyle factors, and demographics, and interactions between these factors. This has not been done before; previous studies were either population-based which has a tendency to overestimate the hazardous effect of AMI on survival, or previous studies only included AMI cases which meant that only survival variations among AMI survivors could be estimated. An additional strength of the study was that by estimating the survival prospects given a possible history of AMI at different ages, the results could be used for planning ongoing medical management and planning resources allocation in the UK population. Finally, the study had a long follow-up of almost 25 years.

Data on the type of AMI were not available in THIN. This meant that the study could not distinguish between STEMI and non-STEMI and thus could not provide specific survival prospects for them. Although the major confounders of AMI were adjusted for, there could potentially be some residual confounding by a number of other factors such as family history of AMI or CVD, psychosocial factors, fruit and vegetable intake, and physical activity. These factors were not adjusted for in the survival models due to the unsystematic or no recording in the medical records. AMI severity indicators, such as left ventricular function, were also not included in the survival models because this information was only available for the cases and not the controls. Another limitation of the study was incomplete medical records with respect to lifestyle factors. Missing data were dealt with by multiple imputations, which is a widely accepted method to deal with bias and imprecision when missing data are present (van Buuren, 2012). An additional limitation of the study was that adherence to drug therapy was unknown and therefore the survival prospects associated with prescription of drug therapy might not accurately reflect the effect of the drugs themselves on the hazard of mortality. Furthermore, no dose-response effect could be estimated as the prescribed doses were not included in the survival models. Finally, there might be bias by indication in which patients receiving treatment were somehow sicker than those not receiving the treatment, despite the adjustment for important confounders.

5.5 Conclusions

The findings of this study suggest that surviving an AMI was associated with a permanent increased hazard of mortality and that coronary revascularisation, statins prescription, and beta blockers prescription could reduce this hazard. This is of clinical importance, because not every AMI survivor received these treatments. In 2010, beta blockers were not widely prescribed to AMI survivors; the survival prospects of 35% of the AMI survivors might be improved by such a prescription.

This study also found that the prescription of aspirin and/or ACE inhibitors was associated with an increased hazard of mortality. This might be of potential concern as the previous explanations for similar findings on the hazardous effects associated with ACE inhibitors on survival, such as confounding by heart failure and use of old data, were addressed by this study. By 2010, 94% and 85% of AMI survivors were prescribed aspirin and ACE inhibitors, respectively. Further research is required to assess the effectiveness of aspirin and ACE inhibitors in the light of these findings that such commonly used medications may be of little benefit, or even cause harm.

Furthermore, the findings of this research suggest that there were sex and deprivation inequalities in uptake of coronary revascularisation while all subgroups benefit equally from it. Further research is needed to assess whether there might be a barrier in access to surgery or whether the difference in uptake is due to confounding by indication such as by prognosis of CVD.

Finally, this research found there were up to ten years difference in effective age between general practices and these differences could not be explained by the socio-demographic factors, health status, and environmental characteristics of the districts their patients resided. Further research is needed to explore the reasons for the unexplained survival variations between general practices and how the differences can be minimised.

Chapter 6

Survival models for statin prescription

This Chapter presents the development of the survival models for statin prescription. The survival models estimated the adjusted hazard of all-cause mortality associated with statins prescribed as primary prevention of cardiovascular disease (CVD) by key ages in residents of the United Kingdom (UK). The findings were published in PLoS One (Gitsels et al., 2016).

This Chapter starts with explaining the analysis procedure including the study design, selected medical history, and model development. This is followed by a description of the age cohorts. Next, the survival models are presented. Then, the survival models are evaluated with respect to model performance, internal validation, and external validation. Finally, recommendations based on the findings are provided.

6.1 Analysis procedure

6.1.1 Study design

The research objective of investigating how prescription of statins affects longevity in the general population with no previous history of CVD, was addressed by a retrospective cohort design. Four cohorts of patients aged 60, 65, 70, and 75 were followed up. These are key ages at which individuals and general practitioners would make a decision about statin uptake. Before the age 60, the initiation rate of statin prescription for primary prevention of CVD in the UK is less than 3% (O'Keeffe et al., 2015). The age cohorts were split by the estimated cardiovascular risk at baseline. The risk groups consisted of patients with a QRISK2 of <10%, 10-19%, or ≥20% risk of a first cardiovascular event in the next ten years. These risk groups were chosen because the UK National Institute for Health and Clinical Excellence (NICE) recently lowered the QRISK2 estimated risk threshold at which to prescribe statins from 20% to 10% (NICE, 2015). As discussed in Chapter 2 Section 1, the effectiveness of the drug in the recently eligible patients is uncertain.

The data restrictions and the identification of eligible general practices and patients from The Health Improvement Network (THIN) primary care database, were discussed in Chapter 3 Section 4. For this particular research objective, patients with a history of CVD (acute myocardial infarction, angina, cerebrovascular disease, heart failure, peripheral vascular disease, valvular heart disease, or other cardiovascular system disorders) were excluded from the analysis. The prevalence of CVD at ages 60, 65, 70, and 75, was 8.7, 13.5, 18.6, and 24.5%, respectively. Patients with a missing Townsend deprivation score were also excluded from the analysis because this information was needed to calculate the baseline risk of CVD using QRISK2. The Townsend scores in THIN were derived from the 2001 census data of England and Wales (UKDS, 2006). This means that patients living in Scotland and Northern Ireland were excluded from the analysis. Only a small percentage of patients from England and Wales had a missing Townsend score. The prevalence of missing Townsend score at ages 60, 65, 70, and 75, was 7.8, 11.6, 12.3, and 12.4%, respectively, with less than 0.2% contributed by patients from England and Wales. Patients with and without a Townsend score were similar to each other with respect to the mortality rate and prevalence of statin prescription, see Appendix C Table C.1.

Patients were part of multiple age cohorts if they remained at the same general practice for more than five years and did not develop CVD during that time, see Figure 6.1. Patients were followed up for maximum 24 years and the oldest patients were 91 years old by the end of study.

6.1.2 Selected medical history

At the cohort's age, a snapshot of the patient's medical history was obtained. Additionally, the patient's survival was recorded during the follow-up. Chapter 3 Section 5 described the raw information selected from the medical records for the entire research. This Section describes the edited information selected for the survival models aimed at estimating the hazard of mortality associated with statin prescription prescribed as primary prevention of CVD.

To obtain cardiovascular risk group, the baseline risk of the patient needed to be calculated. The NICE guideline on lipid modification recommends using QRISK2 to calculate the 10-year risk of a first cardiovascular event (NICE, 2015). The QRISK2 algorithm is available online (`http://svn.clinrisk.co.uk/opensource/qrisk2/`). It incorporates information on multiple demographic, medical, and lifestyle factors to estimate the risk (Hippisley-Cox et al., 2008), see Chapter 2 Figure 2.1. Not all this information was available in THIN or in its subset used in this research, leading to a number of substitutions or exclusions in the algorithm, see Appendix C Table C.2. The QRISK2 was calculated for all patients with complete information in JAVA version 8.

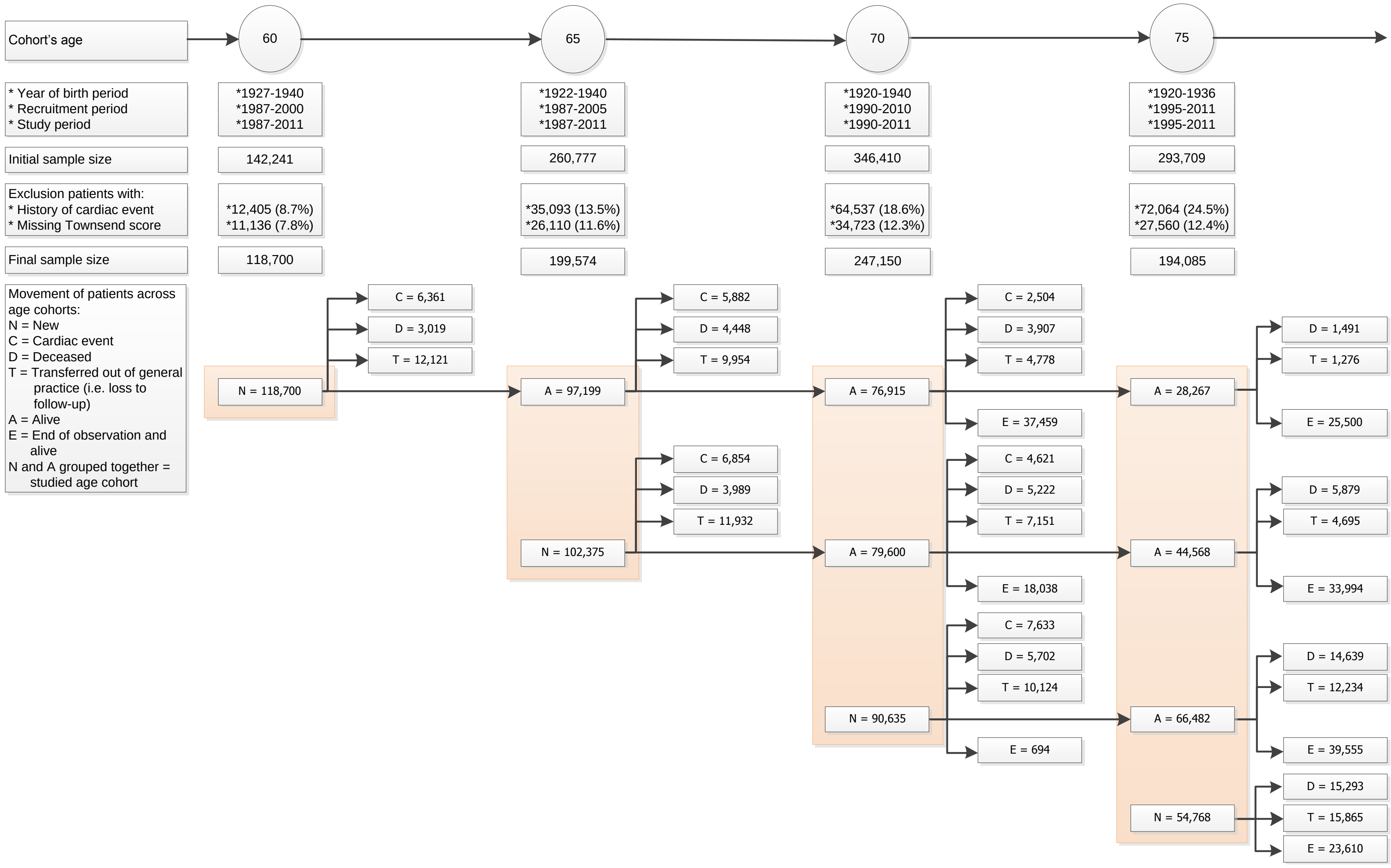


Figure 6.1: Selected age cohorts without cardiovascular disease

This Figure is an altered version of Figure 3.1, where only patients with no history of cardiovascular disease by an initial age were followed up.

Ethnicity was not recorded in THIN and could therefore not be included in the QRISK2 calculation. The QRISK2 calculator has white ethnicity as the reference category. 93% of the UK population is white and the QRISK2 score risk would be underestimated for people with an Indian, Pakistani, or Bangladeshi background and overestimated for people with a black Caribbean background and for black African and Chinese men (Hippisley-Cox et al., 2008). The QRISK2 calculator uses Townsend deprivation score measured on a continuous scale. The data cut from THIN, however, provided the Townsend score in quintiles. Based on the 2001 census data, the associated median values for each quintile were used for calculating the cardiovascular risk (UKDS, 2006). The selected medical history for this research did not include atrial fibrillation and rheumatoid arthritis. It was assumed that the patients in the studied cohorts did not have these medical conditions. This would be true for the majority of the patients, as the prevalence of atrial fibrillation and rheumatoid arthritis is less than one percent in the United Kingdom (Collins and Altman, 2012; Hippisley-Cox et al., 2008). Family history of CVD substituted family history of ischaemic heart disease in the QRISK2 calculation. Family history of CVD is also used by the JBS3, another accepted risk calculator developed by the Joint British Societies for the prevention of cardiovascular disease (Joint Formulary Committee, 2016a). With the studied cohorts, it was assumed that diabetes was type two, which would be true for 90% of the cases (NHS, 2014a). It was also assumed that current smokers smoked moderately (10 to 19 cigarettes a day), which would be true for the majority (ONS, 2014). The diagnosis of hypercholesterolaemia was used instead of a specific cholesterol level, because it was only recorded in the minority of primary care records (27 to 50% recordings) (Hippisley-Cox et al., 2008). When a patient did not have a diagnosis of hypercholesterolaemia, a ratio of total cholesterol to high density lipids of four was ascribed. This is the default value of the online QRISK2 calculator when no ratio is given. When a patient had a diagnosis of hypercholesterolaemia, a ratio value of five was ascribed.

For the survival analysis, the outcome of interest was time to death and the primary exposure was statin prescription prior to the cohort's age. The selected confounders of the potential survival benefit associated with statin prescription were based on the literature review, and consisted of: sex, year of birth, socioeconomic status, chronic kidney disease, diabetes, hypertension, hypercholesterolaemia, blood pressure regulating drugs, body mass index, and smoking status, see Appendix C Table C.3. Socioeconomic status was measured by Mosaic, which is based on demographics, lifestyles, and behaviour of people at a postcode level (Experian Ltd., 2009).

There were missing values in systolic blood pressure (15-28%), smoking status (14-30%), and body mass index (22-37%). The fraction of incomplete medical records decreased with age; 46% of the youngest cohort and 29% of the oldest cohort had incomplete records. The proportion of complete records was greater among participants born at a later year, with a medical condition, or who were prescribed drugs, see Appendix C Table C.4. This is in agreement with previous research that found that recording is likely be related to ill health and that recording improved since the introduction of the Quality and Outcomes Framework (QOF) in 2004 (Marston et al., 2010; Shephard et al., 2011; Taggar et al., 2012).

Incomplete medical records were dealt with by multiple imputation. The joint imputation model consisted of all factors included in the QRISK2 calculation and the outcome variable of the analysis model which was time to death. The imputation model was multilevel to adjust for the correlation between patients from the same general practice. QRISK2 score was imputed instead of being calculated based on the imputed values to ensure that the imputed values were consistent with the analysis model (van Buuren, 2012). Imputations were done in REALCOM-Imputation software. The Monte Carlo Markov Chain (MCMC) estimation had a burn-in length of 100 iterations and was in total 1,000 iterations long. The imputed values of every 100^{th} iteration were used. This resulted in ten imputed datasets. The distribution of recorded and imputed values for systolic blood pressure, body mass index, smoking status, and QRISK2 scores were similar, see Appendix C Table C.5.

6.1.3 Model development

A Cox's proportional hazards regression was fitted to estimate the effect of statin prescription on the hazard of all-cause mortality for different risk groups at various ages, with the outcome variable being time to death in days. All-cause mortality was used as the primary outcome because multiple previous studies showed the protective effect of statins on the risk of cardiovascular events and related deaths, but reported uncertain overall survival benefit (CTTC, 2012). The exposure was any statin prescription at any time before the participant reached the baseline age. The analysis on statins excluded patients who were prescribed other lipid-lowering therapy to ensure that the effect of statins was not mediated by those drugs. The number of patients excluded at age 60, 65, 70, and 75 were 876 (0.7%), 1,718 (0.9%), 2,840 (1.1%), and 2,475 (1.3%), respectively. A separate analysis was performed to estimate the effect of lipid-lowering therapy prescription including and excluding statins. The analysis was on an intention to treat basis, as it was based on the prescription of drugs and not intake of them. The models were fitted to each age and risk group. The ages were 60, 65, 70, and 75 years old. The risk groups consisted of patients with a QRISK2 of <10%, 10-19%, or ≥20% risk of a first cardiovascular event in the next ten years. This would mean that twelve age-risk groups would be studied. However, as only a few patients were at low risk by age 70 (n=325) and none by age 75, these two groups were not studied.

Starting from a model with second-order interaction effects of all factors with the main exposure statin prescription, sex, and age, backward elimination was performed to obtain the most parsimonious model possible. The final model consisted of the factors that were found significant with the alpha level set to 5% for fixed effects and 1% for interaction effects. The survival models were multilevel on general practice and patient level, to adjust for the interdependence of patients from the same general practice.

The Cox's regression assumes no time-tied deaths, uninformative censoring, and proportional hazards. Time-tied deaths were handled by Efron's approximation. Censoring due to loss of follow-up was examined by comparing patients who did and did not transfer out of their general practice with respect to health status and lifestyle factors. For the youngest two age cohorts, distinction was made between patients observed until age 70 or for longer. This was because previous studies indicated that there are two groups of people who move at an older age, where the younger group comes from more affluent areas and move in good health and the older group comes from more deprived areas and move in worse health (Pennington, 2013; Uren and Goldring, 2007). These trends were not observed in the age cohorts, see Appendix C Table C.6. Patients who transferred were similar to patients who did not transfer with respect to medical conditions and lifestyle factors. However, patients who transferred after age 70 were approximately half as likely to be prescribed lipid-lowering therapy as patients who transferred before age 70 or who did not transfer at all, even though the QRISK2 scores between the groups were not significantly different. Uninformative censoring was assumed because transfers did not differ in health status and lifestyle factors other than prescription of lipid-lowering therapy. The assumption of proportional hazards was checked by Grambsch and Therneau's test. When the assumption was violated for a variable, being significant at alpha level set to 1%, the variable's effect on survival time was made time-variant. Follow-up time was split in intervals in which the assumption was no longer violated, where the time intervals of 10, 5, or 1 year were tested.

A unified model for all risk groups and ages was attempted in order to have the same interpretation of the hazards. However, some factors could not be adjusted for because only one subgroup of patients specified by the factor was present in the particular risk group at the specific age. All men were either at moderate or high risk of CVD by age 65 and 99% were at high risk by age 75. There were no patients with diabetes in the low risk group at ages 60 and 65, and in the moderate risk group at ages 65, 70, and 75. Furthermore there were no patients with hypercholesterolaemia in the moderate risk group at age 75. Where possible, the final models adjusted for sex, year of birth, socioeconomic status, diabetes, hypercholesterolaemia, blood pressure regulating drugs, body mass index, smoking status, and general practice. There were no interactions with statin prescription, sex, or age, and no time-varying hazards of variables. The contribution of each variable to the model in explaining survival variations was assessed by decomposing the Cox's regression by ANOVA (analysis of variance).

The number of years lost or gained in effective age associated with statin prescription were calculated. The models were assessed on overall performance, discrimination, and external validation, using Royston's R^2, Harrell's concordance, and the shrinkage slope, respectively. Furthermore, the results were compared internally with the complete case analysis and externally with previous studies. The analyses were performed in R version 3.1.1, using the packages 'hmisc', 'rms', and 'survival'.

6.2 Description of cohorts

Four cohorts of patients who were either 60, 65, 70, or 75 years old and who had no history of CVD at baseline were studied. The profile of the patients with respect to medical conditions, treatments, and lifestyle factors differed by age and changed over time, see Table 6.1, Figure 6.2 and Appendix C Figure C.1.

The older the cohort, the more prevalent the outcome death, the exposure statin prescription, and the medical conditions were. The prevalences of these factors were lower than in the general population because the cohorts excluded patients with a history of CVD. Family history of CVD, however, was exceptionally low in the age cohorts, with prevalence less than one percent.

From the age of 60 to 75, the annual death rate per 1,000 individuals for men increased from 12.8 to 41.0, and for women from 8.7 to 28.4. Likewise, the 10-year risk of a first cardiovascular event in men increased from 13.3 to 33.2%, and in women from 7.8 to 25.8%. Of the medical conditions, diabetes was the most common; from the age of 60 to 75, its prevalence in men increased from 4.1 to 10.6% and in women from 3.0 to 8.0%. From 1995 to 2011, the prevalence of diabetes increased by 6-9% across age cohorts, see Appendix C Figure C.1.

From the age of 60 to 75, prescription of lipid-lowering therapy in men increased from 2.0 to 20.0%, and in women from 2.6 to 19.2%. Statins were the most common prescribed type of lipid-lowering therapy; 70% of lipid-lowering therapy prescription was a statin at age 60 and this increased to 94% at age 75. These prevalences were affected by calendar year as statins became more popular over time, see Figure 6.2. Since 2006, all patients with a cardiovascular risk of 20% or greater are eligible for statin therapy (NICE, 2016). In 2010, only 45% of this group were prescribed statins. Given risk group and calendar year, statins were prescribed less in older patients and in men, see Table 6.2 and Figure 6.2. Given age-risk group, 50% of the therapy durations started between one and four years prior to the cohort's age, 75% under six years, and only 10% more than eight years, see Appendix C Figure C.2. The majority of the prescriptions (87-93%) were in the year prior to the cohort's age.Adherence of treatment arm was ascertained for patients who were observed in multiple cohorts. Assuming that patients who were lost to follow-up stayed in the initial treatment arm, 88 to 98% of the cases and 78 to 92% of the controls never switched treatment arm, see Appendix C Table C.7. Five percent of the patients changed treatment arms multiple times.

Given calendar year, the prevalence of the clinical and lifestyle factors hypertension and ex-smoking increased with age, while smoking decreased, and overweight and obesity remained approximately the same, see Appendix C Figure C.1. The prevalence of hypercholesterolaemia was more common in younger patients prior to 2000, and was approximately the same across age after 2000. From 1995 to 2011, hypertension, obesity, and ex-smoking became more prevalent, while smoking became less prevalent, and overweight remained approximately at the same level. Interestingly, the prevalence of hypercholesterolaemia increased from 1995 to 2005, after which it levelled off. This trend was the most pronounced in the oldest cohort, in which the prevalence increased from 6% in 1995 to 40% in 2005, after which it levelled off at 42% in 2011. The prevalence order of the clinical and lifestyle factors in patients with no history of CVD changed over time and this varied by age. In 2000, overweight was with a prevalence of approximately 42% the most common while diabetes with a prevalence of approximately 7% was the least common at all ages. In 2010, the prevalence of hypertension was between 40-60%; hypercholesterolaemia and overweight between 40-45%; ex-smoking and obesity between 20-40%, and smoking and diabetes between 10-20%.

Table 6.1: Characteristics of men and women in age cohorts without cardiovascular disease

[1] Missing values in smoking status, body mass index, and consequently QRISK2 score, were dealt with by multiple imputation. The reported prevalences of these variables are the means across ten imputed datasets. The prevalences of comorbidities and lifestyle factors at the cohort's age were affected by calendar year, see Appendix C Figure C.1.

	Age 60		Age 65		Age 70		Age 75	
	Women	Men	Women	Men	Women	Men	Women	Men
Number of patients	61,715	56,985	106,633	92,941	138,355	108,794	114,434	79,651
Total person-years of follow-up data (mean)	759,967 (12.3)	676,579 (11.9)	1,045,437 (9.8)	857,273 (9.2)	1,033,903 (7.5)	754,810 (6.9)	681,791 (6.0)	438,204 (5.5)
Deaths (%)	6,628 (10.7%)	8,668 (15.2%)	12,975 (12.2%)	15,873 (17.1%)	19,515 (14.1%)	21,184 (19.5%)	19,379 (16.9%)	17,977 (22.6%)
Loss to follow-up (%)	14,992 (24.3%)	14,844 (26%)	21,415 (20.1%)	20,620 (22.2%)	23,747 (17.2%)	18,541 (17.0%)	20,707 (18.1%)	13,363 (16.8%)
Family history of cardiovascular disease (%)	86 (0.1%)	51 (0.1%)	363 (0.3%)	198 (0.2%)	820 (0.6%)	398 (0.4%)	664 (0.6%)	308 (0.4%)
Chronic kidney disease (%)	4 (0.0%)	1 (0.0%)	60 (0.1%)	18 (0.0%)	3,324 (2.4%)	2,099 (1.9%)	5,528 (4.8%)	3,151 (4.0%)
Diabetes (%)	1,825 (3.0%)	2,363 (4.1%)	5,050 (4.7%)	6,024 (6.5%)	9,174 (6.6%)	9,864 (9.1%)	9,148 (8.0%)	8,462 (10.6%)
Hypercholesterolaemia (%)	8,760 (14.2%)	7,638 (13.4%)	25,656 (24.1%)	19,129 (20.6%)	42,483 (30.7%)	26,649 (24.5%)	36,792 (32.2%)	18,914 (23.7%)
Treated hypertension (%)	11,617 (18.8%)	8,144 (14.3%)	29,873 (28.0%)	21,311 (22.9%)	50,742 (36.7%)	34,854 (32.0%)	51,864 (45.3%)	30,612 (38.4%)
Reported systolic blood pressure (%)	47,535 (77.0%)	37,802 (66.3%)	86,916 (81.5%)	68,466 (73.7%)	115,548 (83.5%)	88,023 (80.9%)	98,494 (86.1%)	67,259 (84.4%)
Reported smoking status (%)	44,689 (72.4%)	37,844 (66.4%)	85,080 (79.8%)	70,020 (75.3%)	115,189 (83.3%)	89,115 (81.9%)	98,553 (86.1%)	68,394 (85.9%)
Ex-smoker[1] (%)	14,097 (22.8%)	19,990 (35.1%)	24,493 (23.0%)	34,030 (36.6%)	32,850 (23.7%)	42,624 (39.2%)	26,247 (22.9%)	31,703 (39.8%)

Continued on next page

Table 6.1 – *Continued from previous page*

	Age 60		Age 65		Age 70		Age 75	
	Women	Men	Women	Men	Women	Men	Women	Men
Smoker[1] (%)	10,469 (17.0%)	12,403 (21.8%)	16,128 (15.1%)	18,288 (19.7%)	16,989 (12.3%)	17,034 (15.7%)	11,923 (10.4%)	10,599 (13.3%)
Reported BMI (%)	41,759 (67.7%)	33,110 (58.1%)	79,055 (74.1%)	62,168 (66.9%)	105,789 (76.5%)	80,251 (73.8%)	89,174 (77.9%)	61,381 (77.1%)
BMI[1] mean (sd)	26.3 (4.1)	26.3 (3.0)	26.6 (4.4)	26.4 (3.3)	26.7 (4.5)	26.5 (3.5)	26.5 (4.9)	26.2 (3.5)
Calculated QRISK2 score (%)	36,325 (58.9%)	27,586 (48.4%)	70,023 (65.7%)	52,799 (56.8%)	96,015 (69.4%)	72,252 (66.4%)	81,980 (71.6%)	56,273 (70.6%)
QRISK2[1] mean score (sd)	7.8 (3.3)	13.3 (4.0)	12.1 (4.4)	18.9 (5.2)	18.0 (5.5)	25.5 (6.3)	25.8 (6.4)	33.2 (6.9)

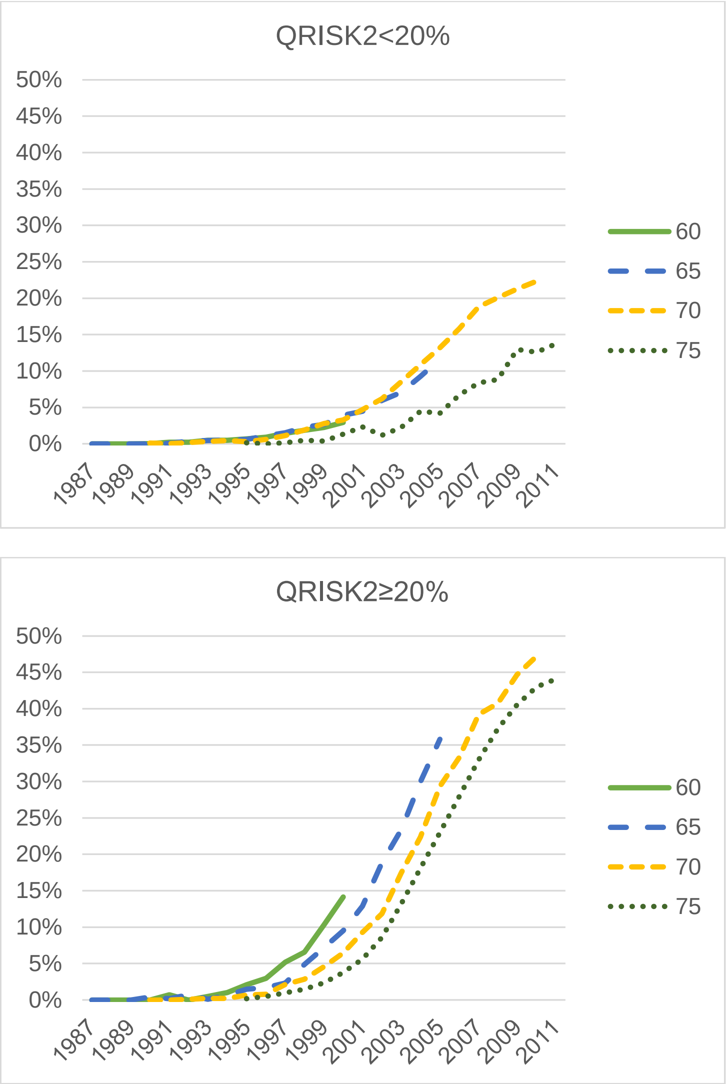


Figure 6.2: Prevalence of statin prescription by cohort's age in patients without cardiovascular disease

The prevalence in a given year was the percentage of patients who turned the cohort's age in that year and were prescribed statins prior to that age. The reported QRISK2 risk group is the mean 10-year risk of a first cardiovascular event across ten imputed datasets.

Table 6.2: Prevalence of statin prescription by cohort's age, cardiovascular risk group, and sex

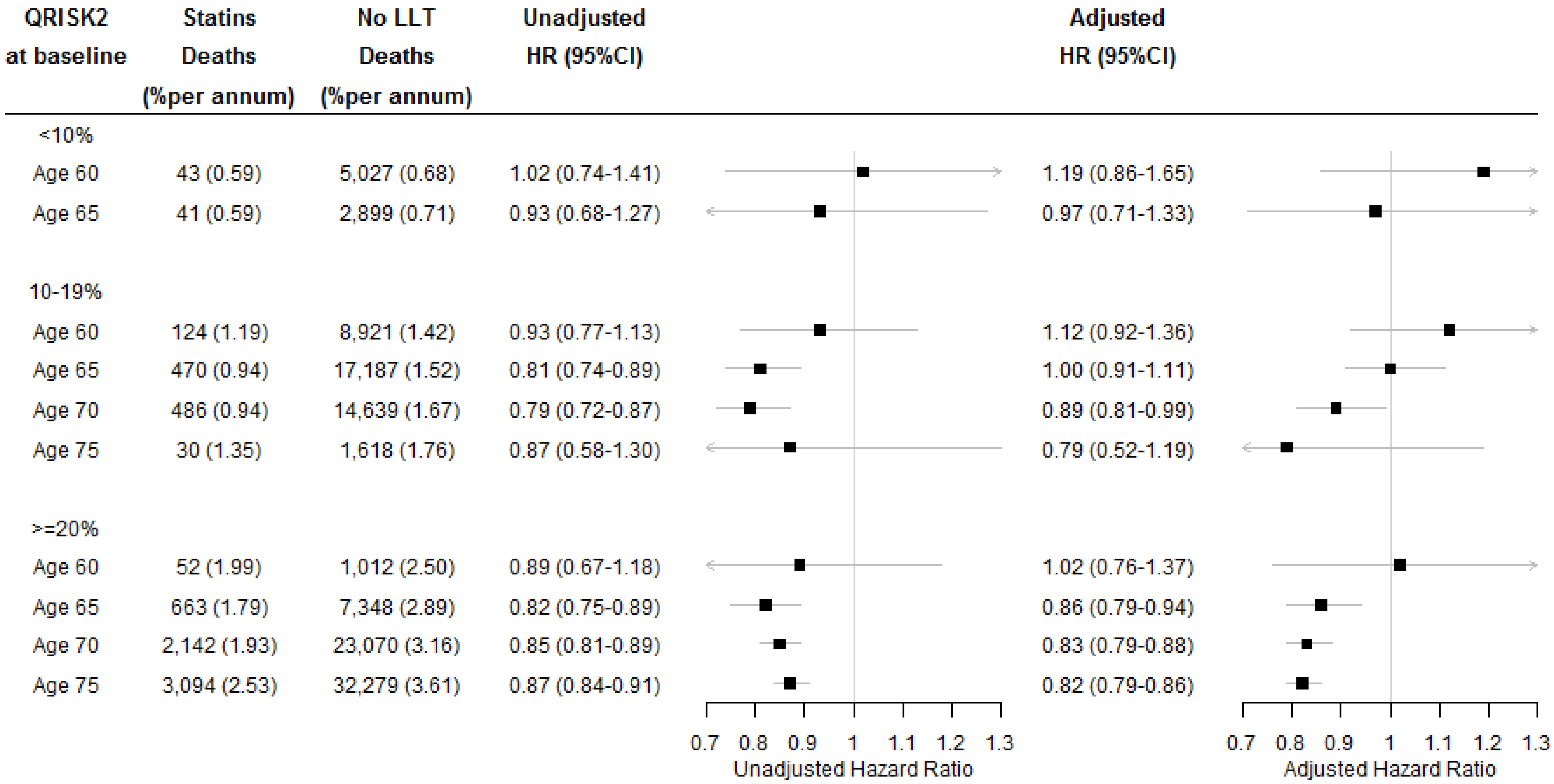
The age cohorts included patients with no history of cardiovascular disease. The prevalence of statin prescription by the cohort's age was affected by calendar year, see Figure 6.2. The number of patients per QRISK2 group is the mean 10-year risk of a first cardiovascular event across ten imputed datasets.

	QRISK2 score	Women (% Statins)	Men (% Statins)	Total (% Statins)
Age 60	<10%	50,668 (1.2%)	9,234 (0.4%)	59,902 (1.1%)
	10-19%	9,834 (3.8%)	44,037 (1.3%)	53,871 (1.8%)
	≥20%	692 (12.3%)	3,359 (5.3%)	4,051 (6.5%)
	Total	61,194 (1.7%)	56,630 (1.4%)	117,824 (1.6%)
Age 65	<10%	40,749 (2.2%)	0 (0.0%)	40,749 (2.2%)
	10-19%	58,475 (7.5%)	64,203 (3.2%)	122,678 (5.2%)
	≥20%	6,274 (27.8%)	28,155 (12.5%)	34,429 (15.3%)
	Total	105,498 (6.6%)	92,358 (6.0%)	197,856 (6.4%)
Age 70	<10%	325 (0.9%)	0 (0.0%)	325 (0.9%)
	10-19%	102,938 (9.6%)	16,587 (5.4%)	119,525 (9.1%)
	≥20%	33,196 (28.8%)	91,263 (17.5%)	124,459 (20.5%)
	Total	136,459 (14.3%)	107,850 (15.7%)	244,309 (14.9%)
Age 75	10-19%	14,345 (4.6%)	1 (0.0%)	14,346 (4.6%)
	≥20%	98,365 (19.9%)	78,899 (19.2%)	177,264 (19.6%)
	Total	112,710 (17.9%)	78,900 (19.2%)	191,610 (18.5%)

6.3 Survival models

Survival models were developed for the ten different age, and cardiovascular risk groups and the results on statin prescription are presented in Figure 6.3. The survival models estimated no survival benefit from statin prescription by any age in patients with a cardiovascular risk lower than 10%. There was also no survival benefit associated with statin prescription in patients younger than 60, no matter how high their cardiovascular risk was. In patients with a risk of 10 to 19% and aged over 60, statin prescription had uncertain survival benefit. In these patients, statin prescription by age 65, 70, or 75 was associated with a hazard of mortality of 1.00 (95% confidence interval 0.91-1.11), 0.86 (0.81-0.99), or 0.79 (0.52-1.19), respectively. Patients with a risk of 20% or greater and aged over 60, had a significant survival benefit from statin prescription. In these patients, statin prescription by age 65, 70, or 75 was associated with a hazard of mortality of 0.86 (0.79-0.94), 0.83 (0.79-0.88), or 0.82 (0.79-0.86), respectively. This translates to a decrease in effective age of 1.5 (0.6-2.4), 1.9 (1.3-2.4), or 2.0 (1.5-2.4) years, respectively, compared with patients without statin prescription by those ages.

There were no interactions found with statin prescription. This means that the potential survival benefit associated with statin prescription was the same for different subgroups of patients, such as for men and women. As no interaction between statin prescription and year of birth category was found, the results suggest that the effectiveness of statin prescription does not vary by year of prescription category. Lastly, there was no survival benefit from lipid-lowering therapy excluding statins by any age for any risk group. This resulted in smaller estimated survival benefit of lipid-lowering therapy as a whole compared with its subclass of statins, see Appendix C Figure C.3.



QRISK2 at baseline	Statins Deaths (%per annum)	No LLT Deaths (%per annum)	Unadjusted HR (95%CI)	Adjusted HR (95%CI)
<10%				
Age 60	43 (0.59)	5,027 (0.68)	1.02 (0.74-1.41)	1.19 (0.86-1.65)
Age 65	41 (0.59)	2,899 (0.71)	0.93 (0.68-1.27)	0.97 (0.71-1.33)
10-19%				
Age 60	124 (1.19)	8,921 (1.42)	0.93 (0.77-1.13)	1.12 (0.92-1.36)
Age 65	470 (0.94)	17,187 (1.52)	0.81 (0.74-0.89)	1.00 (0.91-1.11)
Age 70	486 (0.94)	14,639 (1.67)	0.79 (0.72-0.87)	0.89 (0.81-0.99)
Age 75	30 (1.35)	1,618 (1.76)	0.87 (0.58-1.30)	0.79 (0.52-1.19)
>=20%				
Age 60	52 (1.99)	1,012 (2.50)	0.89 (0.67-1.18)	1.02 (0.76-1.37)
Age 65	663 (1.79)	7,348 (2.89)	0.82 (0.75-0.89)	0.86 (0.79-0.94)
Age 70	2,142 (1.93)	23,070 (3.16)	0.85 (0.81-0.89)	0.83 (0.79-0.88)
Age 75	3,094 (2.53)	32,279 (3.61)	0.87 (0.84-0.91)	0.82 (0.79-0.86)

Figure 6.3: Unadjusted and adjusted hazards of all-cause mortality associated with statin prescription

The age cohorts included patients with no history of cardiovascular disease. The hazard ratios (95% confidence interval) were adjusted for sex, year of birth, socioeconomic status, diabetes, hypercholesterolaemia, blood pressure regulating drugs, body mass index, smoking status, and general practice. QRISK2=10-year risk of first cardiovascular event. LLT=lipid-lowering therapy.

The variables contributing the most to the survival models in explaining survival variations in the age cohorts were smoking status and general practice. The variables contributing the least to the survival models were statins, sex, and year of birth category.

6.4 Evaluation

6.4.1 Performance statistics

The final survival models for each age and cardiovascular risk group performed similarly with the exception of the model estimated on patients aged 60 at high risk, see Appendix C Table C.8. The survival model estimated on this group of patients performed worse than the others. This might be due to only 0.3% of patients at that age had a cardiovascular risk of 20% or greater (n=4,051). The survival model explained 5% of the survival variations, had 59% concordance between the estimated hazard of mortality and survival time, and overestimated the effects by 22%. The high percentage of overestimation, i.e. the large shrinkage slope, suggests that the results are not robust and may not be trusted. It may be the case that with a larger sample size, results would show that statins prescription is associated with a survival benefit for this age-risk group. The survival models for the other age-risk groups explained 11 to 13% of the survival variations, had 62 to 63% concordance between the estimated hazard of mortality and survival time, and overestimated the effects by less than 2%. These performance statistics are typical for survival analysis and the small shrinkage slope suggest that the results are robust.

The unadjusted hazards of mortality indicate that prescription of statins is beneficial for survival by age 65 in patients with a cardiovascular risk of 10% or greater, see Figure 6.3. The difference in unadjusted and adjusted survival benefits is greater in patients with a risk of 10 to 19% than in patients with a risk of 20% or greater, the difference was on average 12% and 4%, respectively, in no particular direction. The unadjusted model explained 0% of the survival variations, had only 51% concordance between the estimated hazard of mortality and survival time, and had an unstable shrinkage slope. These performance statistics support the importance of controlling for confounders when estimating the effect of a drug on the hazard of mortality. The stratification on cardiovascular risk groups alone was not sufficient, even though the risk calculation was based on multiple demographic, medical, and lifestyle factors. The risk groups were based on the risk of cardiovascular event whereas the study was concerned with the risk of premature death.

6.4.2 Internal validation

The complete case analysis, estimated on only complete medical records, provided similar results and performance statistics as the final survival models that were estimated on both complete and incomplete records, see Appendix C Table C.8 and Figure C.4. The complete case analyses estimated 3 to 5% greater survival benefits associated with statin prescription. This could maybe be explained by the fact that completeness of medical records was associated with ill health, indicating that sicker patients might benefit more from the drugs. The complete case models explained 9 to 14% of survival variations, had 61 to 65% concordance between the estimated hazard of mortality and survival time, and overestimated the effects by 5%. Again, the survival model estimated on patients aged 60 at high risk performed worse; the model explained 6% of the survival variations, had 60% concordance between the estimated hazard of mortality and survival time, and overestimated the effects by 26%.

6.4.3 External validation

This large population-based cohort study estimated the hazard of mortality associated with statin prescription for groups with <10%, 10-19%, or ≥20% risk of a first cardiovascular event in the next ten years, using QRISK2, over almost 25 years. It shows that there was no mortality reduction associated with statin prescription in the 60-year old cohort, and in participants at less than 10% risk aged 65, 70, or 75 at baseline. Participants aged 70 or 75 at baseline and who were at moderate risk (QRISK2 score 10-19%) showed uncertain mortality reduction associated with statin prescription. Participants aged 65, 70, or 75 at baseline and who were at high risk (QRISK2 score $\geq$20%) showed significant mortality reduction associated with statin prescription. The hazard ratios reduce with increasing age, with the greatest benefit seen in the oldest cohort. In keeping with previous research, there was no difference between men and women in mortality associated with statin prescription (Kostis et al., 2012).

These findings reinforce the considerable benefits of statin treatment in high-risk groups (where this study found substantial and important undertreatment) reported by the extremely influential Cholesterol Treatment Trialists' Collaboration (CTTC) meta-analysis on 27 randomised control trials (RCTs) (CTTC, 2012). However, the findings of this cohort study clearly differ from the CTTC results for those at low risk of CVD. Where this cohort study found no benefit, despite widespread treatment of patients (particularly women) at low risk, the CTTC reported an overall reduction in all-cause mortality (rate ratio 0.91) in patients without a history of cardiovascular disease, and concluded that statins 'are effective for people with a 5-year risk of major vascular events lower than 10%' (CTTC, 2012). The use of statins in primary prevention remains controversial and contested, perhaps at least in part due to the limitations of RCTs discussed in the Chapter 2, including lack of generalisability due to strict inclusion criteria, lack of comparability with clinical risk scores such as QRISK2 due to trials use of observed events as a comparator (the CTTC trial 5-year risk of major vascular events lower than 10% is hard to compare to a general population QRISK2 score of 10% and is not equivalent), the small number of older patients in trials, and the relatively short follow-up time of 4-5 years in trial (compared with 6-12 years in the age cohorts studied here), and perhaps most importantly, the concerns that anonymised individual patient data from statins trials have still not been made available for independent scrutiny, and remain under the control of a single group of respected researchers, whilst statins are among the most widely prescribed drugs globally (Krumholz, 2016; Parish et al., 2015). Large scale observational studies of the effects of statins in everyday diverse clinical practice over many years are an under-explored source of information on the effects of statins on mortality. No data source is perfect and there are well rehearsed uncertainties and unanswered questions arising from both observational and trial data, as set out in previous chapters. This analysis of cohort data fills in some of the gaps and provides an important new source of information on the possible effects of statins in routine practice.

6.4.4 Strengths and limitations

One of the study's strengths was that it used routinely collected primary care data that were representative of the UK population and widely available. The large sample size included a great number of patients aged over 80 and many patients at low risk of a cardiovascular event, with almost 25 years of follow-up data. The 10-year risk of a first cardiovascular event was calculated by QRISK2, which is recommended by NICE and widely used in routine practice, and broadly comparable to other widely used risk assessment tools such as SCORE and PCE (Mortensen and Falk, 2014; Perk et al., 2012; Stone et al., 2014). The use of all-cause mortality, meant that the overall effect of statin prescription on mortality could be assessed as opposed to estimating the possible shift of the hazard of mortality from one medical condition to another. Estimating the effect of statin prescription by age group meant that age-group specific recommendations could be given.

The analysis was performed on an intention to treat basis to more accurately assess the effect of routine current practice in the general population. THIN had information on prescription of drugs, and not on dispense and intake of them. This means that the actual statin uptake could be lower than THIN records indicated. Therefore, the benefits of statin intake might be greater than the estimated benefits of statin prescription. The analysis did not include the duration of therapy as a possible predictor of survival. As the majority of the patients had started statin therapy within the six years prior to each key age for each risk group, it seems more likely that the benefit of statins is age-dependant rather than duration-dependant. In other words, it is unlikely that the younger patients did not experience a survival benefit from statins due to a shorter duration of therapy. Limitations of a prevalence-user (rather than new-user) study design are that bias might be introduced when the exposure's effect on the outcome varies by time and alters other health indicators (Ray, 2003). As the hazard of mortality associated with statin prescription did not differ by time, it is unlikely that underascertainment of deaths between the time statin therapy is initiated and the target age is reached would have biased the results. Due to limitations of the data, the QRISK2 score could only be approximated. Although there were missing data, sensitivity analyses showed that it was unlikely they influenced the results. Finally, a limitation of using routinely collected observational data to estimate the effects of interventions is that the results might be affected by unexplained confounding. This was minimised by stratifying by cardiovascular risk group and age, and by controlling in the regression models for a wide range of potential confounders.

6.5 Conclusions

This large population-based cohort study estimated the adjusted hazard of all-cause mortality associated with statin prescription by age and CVD risk groups, using QRISK2. As expected, patients at high risk (QRISK2 score of 20% or greater) had reduced mortality associated with statin prescription. This is of clinical importance, because not every patient receives the drug. In 2010, statins were not widely prescribed to patients at high risk of CVD; the survival prospects of 55% of these patients might be improved by such prescriptions.

In addition, the newly eligible patients who are at moderate risk (QRISK2 score of 10 to 19%) showed uncertain mortality reduction associated with statin prescription. Furthermore, the study found no mortality reduction associated with statin prescription in patients younger than 60 years, and in patients at less than 10% risk. Further research is needed on the effects of statins over the long-term for younger patients at low risk of a first cardiovascular event. The recent revision of guidelines to extend treatment to younger and lower risk groups may need to be reconsidered. Clinicians may want to use this new information when discussing the risks and benefits of statins initiation with their patients.

Chapter 7

Discussion

This thesis concerns the development of survival models using primary care data to estimate all-cause mortality hazard indices of cardiovascular disease (CVD) and to evaluate related treatments in routine clinical practice in the United Kingdom (UK). This Chapter discusses these newly developed models, focussing on their validity and utility in medicine and retirement planning. First, the main findings are summarised and the contributions to the existing clinical evidence are provided. Second, the strengths and limitations of this research are reviewed. Third, this research' aims are addressed and the implications in medical management and retirement planning are discussed. Finally, the overall conclusions are presented.

7.1 Main findings

For this research, medical records from 1987 to 2011 from general practices contributing to The Health Improvement Network (THIN) database were used to develop two survival models specified at ages 60, 65, 70, and 75. The first model was developed to estimate the hazards of all-cause mortality associated with a history of acute myocardial infarction (AMI) and related treatments while adjusting for other risk factors. As the prevalence of AMI was relatively rare, especially in the youngest age cohort, patients with a history of AMI were selected and each matched to three controls without this history on sex, year of birth category, and general practice. The second survival model was developed to estimate the hazard of all-cause mortality associated with statins prescribed as primary prevention of CVD while adjusting for other risk factors. The age cohorts excluded patients with a history of CVD.

7.1.1 Survival models for acute myocardial infarction

This research found that AMI survivors had a long-term, increased hazard of all-cause mortality, in which younger survivors and survivors of multiple events were worse off. These hazards were lower than estimated by previous studies (Briffa et al., 2009; Capewell et al., 2000; Chang et al., 2003; Gerber et al., 2010, 2009; Herzog et al., 1998; Kirchberger et al., 2014; Koek et al., 2007; Nigam et al., 2006; Quint et al., 2013; Smolina et al., 2012b). The difference could be due to this study's sample included a wider range of AMI patients and this research adjusted not only for sex and age but also for comorbidities, treatments, lifestyle choices, and socio-demographic factors, resulting in more accurate estimates. Thus, a new finding is that the hazards of all-cause mortality associated with AMI in the general population are most likely less severe than previously estimated.

This research found that coronary revascularisation was mainly beneficial in reducing early mortality, up to five years of follow-up. This finding is in accordance with a previous study by Chang et al. (2003). There was lower uptake of coronary revascularisation by women and by patients from the most deprived areas, even though the survival benefits of the procedure did not differ by sex or deprivation. The lower uptake of coronary revascularisation by women could not be explained by age, diabetes, or deprivation as suggested by a previous study (Chang et al., 2003). The European Society for Cardiology recognises the difference in treatment after ischaemic heart disease (IHD) and advocates equality in treatment (Chieffo et al., 2012).

This research found mixed survival prospects associated with prescription of statins, beta blockers, calcium-channel blockers, aspirin, and ACE inhibitors, which could differ by subgroups of patients. The findings only partly agree with the clinical evidence of drug therapy in AMI patients reviewed by the UK National Institute of Health and Care Excellence (NICE) (NICE, 2013a). In accordance with the existing clinical evidence, this research found that survival prospects were improved in patients prescribed statins. This research also found that beta blockers improved survival prospects in patients with a history of AMI but did not change survival prospects in patients without this history. NICE's guideline, however, reported uncertain survival benefits associated with beta blockers in AMI patients due to the wide confidence intervals. This research found that survival prospects were not improved by calcium-channel blockers in patients with a history of AMI and were worsened in patients without this history. This is in accordance with the existing clinical evidence and that is why the calcium-channel blockers changed from a first line to a second line prescription in 2007 (Joint Formulary Committee, 2016b; NICE, 2013b). This research found that patients prescribed aspirin had worse survival prospects. NICE's guideline included limited evidence on the effectiveness of aspirin compared to placebo on long-term survival; it included only one study, which reported inconclusive survival benefits (CDP, 1976). Finally, this research found that survival prospects were worsened in patients prescribed ACE inhibitors. NICE's guideline included relatively old studies on ACE inhibitors, which reported a survival benefit in AMI patients with left ventricular systolic dysfunction (LVSD) and an inconclusive effect in AMI patients with unselected LVSD (AIRE Study, 1993; Borghi et al., 1998; Køber et al., 1995; Pfeffer et al., 1992; SOLVD Investigators, 1992). More recent studies that were not included in NICE's guideline estimated, like this research, significantly increased hazard of mortality associated with ACE inhibitors (Briffa et al., 2009; Nigam et al., 2006).

7.1.2 Survival models for statin prescription

This research found that the survival benefit associated with statin prescription increased by age and by risk of a first cardiovascular event in the next ten years. Patients aged younger than 60 or with <10% cardiac risk had no survival benefit from statin prescription. Patients aged over 60 with 10-19% cardiac risk had uncertain survival benefit from statin prescription. Patients aged over 60 with ≥20% cardiac risk had significant survival benefit from statin prescription. These findings are in contrast with the findings of the Cholesterol Treatment Trialists' Collaboration (CTTC) meta-analysis and the lipid modification guideline by NICE (CTTC, 2012; NICE, 2015). The CTTC recommended the prescription of statins to people with more than 10% risk of a first major vascular event in the next five years, even though it estimated uncertain survival benefits by statins for the individual risk groups due to the small number of deaths observed during the study period (CTTC, 2012). Based on the CTTC findings, NICE lowered the risk threshold at which statins should be prescribed from 20% to 10% cardiac risk NICE (2015). With the change in guideline, NICE recommended further research into the effectiveness of statins in older patients because the CTTC did not differentiate by age and included only a small number of older patients. Thus, a contribution by this research to the existing clinical evidence is that the effectiveness of statins most likely differs by age in which older patients benefit the most. In addition, statins might not be effective in the newly eligible patients with 10-19% cardiac risk.

7.2 Strengths

The age cohorts studied in this research were drawn from general practices that were representative of the UK population given adjustment for sex, age, and deprivation (Blak et al., 2011; Hippisley-Cox and Coupland, 2010b; MacDonald and Morant, 2008; Massó González et al., 2009). The medical histories of the age cohorts provided insights in the clinical practice of diagnosing medical conditions and offering treatments in the general population. Primary care data has a higher coverage of AMI patients compared to hospital data and disease registers (Herrett et al., 2013b), therefore the matched age cohorts for AMI were more representative of AMI patients in the UK than previous studies that selected patients through hospitals admissions. Moreover, AMI cases and controls were selected from the same source population therefore valid comparisons could be made as there was no selection bias (Hennekens et al., 1987).

The recruitment period of the age cohorts was up to 21 years and the study period was up to 24 years. The long recruitment period meant that changes in the prevalences of medical conditions, treatments, and lifestyle choices could be observed. This provided insights in the past and current well-being of the study population and clinical practice. By testing whether the effects of these risk factors on the outcome changed over time, potential longevity risks could be identified. The long study period meant that more deaths could be observed and that life expectancy could be more accurately estimated. With increasing life expectancy, accurate life expectancy becomes even more relevant for medical resource allocation and retirement planning.

Survival was estimated at four key ages, namely 60, 65, 70, and 75. These are key ages in primary and secondary prevention of CVD and when people would typically retire from work. CVD is mainly prevalent from the age of 60 and the effect of CVD on mortality rate differs by age (Townsend et al., 2014). In addition, existing CVD research and NICE recommended further research to examine the effectiveness of age alone to identify people at high risk of developing CVD and the effectiveness of treatments at older ages, 70 or older, as they are often excluded from clinical trials (Godlee, 2014; NICE, 2015; Zoungas et al., 2014).

The survival models for AMI included both cases and controls. This allowed an estimate of the effect of a history of AMI on mortality compared to no history of AMI. The estimates were adjusted for a wide range of risk factors that are known to explain survival variations, resulting in more precise estimates of the hazard of mortality associated with AMI. All interactions between risk factors were tested, instead of restricting the interactions to the main exposure, sex, and age as most epidemiology studies would do (Hennekens et al., 1987). This meant that survival variations could be explored in greater detail, resulting in recommendations for more fine-tuned patient tailored care and retirement planning. In addition, the interdependence between patients from the same general practice was taken into account by introducing a random effect to the survival model. This meant that the inferences of the findings were not restricted to the practices included in the research but could be generalised to the whole of the UK (Brown and Prescott, 2006).

The way the survival models were developed, as explained above, meant that the findings could be used in medical management, and retirement planning. The survival models included risk factors that are routinely recorded by general practitioners and the records are accessible to patients, therefore survival prospects could simply be calculated for any interested individual. With respect to medical management, the findings of this research are informative for administration of preventative health measures, ongoing therapy, and strategic resources allocation. With respect to retirement planning, the findings of this research are informative for financial planning of retirement for individuals, pricing of annuities for actuaries, and shaping of the pension system for the government.

7.3 Limitations

The survival models assumed uninformative censoring, i.e. that patients who transferred from general practice had the same mortality rate as patients who stayed at their practice. This assumption can only be proven when national death records are linked to THIN database. However, by examining the profile of patients who transferred compared to who stayed, the assumption may be supported or opposed. National trends indicate that there are two groups of people who move at an older age, where the younger group aged 60 to 69 comes from more affluent areas and move in good health and the older group aged 70 or older comes from more deprived areas and move in worse health (Pennington, 2013; Uren and Goldring, 2007). These trends were not observed in the studied cohorts, thereby supporting the assumption of uninformative censoring.

This research made use of data from 1987 to 2011. Over time, the incentives and methods for data recording by general practitioners have changed (Marston et al., 2010). For example, in 1990 clinical audits became a contractual requirement, in 1999 national standards for treatment for coronary heart disease were introduced, and in 2004 financial incentives by means of the Quality and Outcomes Framework (QOF) were introduced (Campbell et al., 2007). Especially as a result of QOF, the recording and managing of common chronic diseases, preventative measures, and lifestyle choices improved (Campbell et al., 2007; Langley et al., 2011; NICE, 2014a; Szatkowski et al., 2012). After the introduction of QOF, patients might have been more likely to receive preventative measures, be followed-up more consistently, thereby obtain better control of risk factors and in turn have better survival prospects after diagnosis. This research found no interactions of the risk factors with year of birth category, suggesting that the hazards of mortality associated with the risk factors do not vary by year of treatment category.

The medical records were incomplete with respect to lifestyle factors. In the cohorts without a history of CVD, 46% of the youngest and 29% of the oldest cohort had incomplete records, while in the matched cohorts for AMI this was 45% and 23%, respectively. The presence of missing data led to additional analysis and model assumptions, and loss of precision in the estimates. Missing data were dealt with by multiple imputation, which is a widely accepted method to deal with bias and imprecision when missing data are present (van Buuren, 2012). The distributions of recorded and imputed values were similar. Furthermore, the survival models estimated on only complete medical records provided similar hazard ratios and performance statistics as the models estimated on both complete and incomplete records.

The survival models were as complete as possible. There were, however, some risk factors that could not be included due to the unsystematic or lack of recording in the medical records during the study period. This meant that there could potentially be some residual confounding by indication such as by family history of CVD, family history of AMI, ratio of total cholesterol to high density lipids, psychosocial factors, and ethnicity. The recording of these risk factors are part of QOF and has improved over time (Mathur et al., 2014; Tucker, 2014). Thus, with future updates of the survival models using more recent data, these factors could potentially be included.

The analysis was performed on an intention to treat basis to more accurately assess the effect of routine current practice in the general population. With intention to treat, both sick-user bias and healthy-user bias could not be excluded. Sick-user bias would arise where at a given health status, general practitioners recognised people who were at greater risk and consequently provided treatment. This bias would have led to an underestimation of the survival benefit associated with the treatment. Healthy-user bias would arise where at a given health status, patients who were more proactive about their health, were more likely to be treated (Dormuth et al., 2009; Ray, 2003;

Vonesh et al., 2000). This bias would have led to an overestimation of the survival benefit associated with the treatment. Sick- and healthy-user biases were minimised by inclusion of comorbidities, lifestyle choices, and socio-demographic factors in the survival models.

7.4 Implications

The implications of the current research for the medical management and retirement planning are discussed by addressing the aims of the research.

The first aim was to investigate how the presence and duration of comorbidities and treatments affect the hazard of mortality at each age and whether they can be related to age-specific medical management. The hazards of mortality associated with the comorbidities were constant during follow-up; it did not matter how many years patients had already lived with the comorbidity, they were still at a higher risk of dying than patients without the comorbidity. This shows the importance of follow-up of care. Interestingly, the hazard associated with hypertension at ages 70 and 75 was protective in the first five years and hazardous after five years of follow-up. People at those ages were likely to have high blood pressure but not necessarily diagnosed with hypertension or treated with blood pressure regulating drugs. The results suggest that diagnosing improves survival prospects even though the condition is hazardous for survival. General practitioners might want to screen for hypertension in people aged 70 and above as part of their medical management by measuring their blood pressure on a regular basis.

The relative hazards of mortality associated with treatments were approximately the same at each age, implying that the effectiveness of treatments did not differ by age. Therefore, cardiovascular treatments should not be age-specific. The relative hazards of mortality associated with drugs prescriptions were constant during follow-up. This supports follow-up of AMI patients with regards to their cardiac rehabilitation and adherence to drugs.

The second aim was to investigate the survival benefits of statins prescribed as primary prevention of CVD for various cardiovascular risk groups at each age and whether this can inform risk thresholds for action. The survival benefits of statins therapy increased with the 10-year risk of a first cardiac event and with age. Only in people with $\geq$20% cardiac risk and aged 65 and older did statins therapy significantly prolong life. The current recommended thresholds for statins therapy for primary prevention of CVD in routine practice may be too low and lead to overtreatment, particularly in people with $<$10% cardiac risk or younger than 60 years old. Revision of the guidelines on lipid modification by statins therapy should consider not only having a risk threshold for action but also an age barrier.

The third aim was to investigate how modifiable risk factors such as cholesterol level, blood pressure, body mass index, alcohol consumption, and smoking affect the hazard of mortality at each age and whether they can inform public health measures. The research' findings are in line with NICE's current guidelines on cardiac rehabilitation and prevention of further AMI, with the exception of weight management (NICE, 2013b). As part of the cardiac rehabilitation, overweight and obese patients are recommended to maintain a healthy weight. This research, however, found that in non-smokers, the survival prospects of overweight patients were not significantly different from healthy weight patients. Also, survival prospects of smokers were less poor when they were overweight or obese than when they were healthy weight. This is in line with well-known obesity paradoxes in CVD patients as well as in the general population, of which one demonstrates that when physical activity is taken into account, the survival prospects of obese and normal weight are no longer significantly different

(McAuley and Blair, 2011). Based on these findings, the recommended weight management as part of cardiac rehabilitation might be too strict, and a revision of the guideline might be considered by focussing on obese patients and emphasising cardio fitness.

The fourth aim was to investigate the effect of general practice on the hazard of mortality at each age and whether this is a factor additional to the socio-demographic factors of a district to consider in resource allocation. With both survival models developed for this research, general practice was one of the factors that contributed the most in explaining survival variations. The adjusted survival prospects differed by maximum ten years in effective age between general practices. This suggests that the average period expectation of life at the cohort's ages of 60, 65, 70, and 75 between general practices for males differed up to 3.3, 7.9, 7.2, and 6.5 years, respectively, and for females differed up to 3.5, 8.4, 7.9, and 7.3 years, respectively.

With the post-hoc analyses, the survival variations by general practice were not found to be associated with health status, ethnic background, deprivation, urbanisation, or air pollution. A study by Gerber et al. (2010), who developed a survival model for AMI with individual socioeconomic status as covariate and neighbourhood socioeconomic status as random effect, suggested that neighbourhood socioeconomic status might capture residual confounding of unequal hospital resources and social characteristics of an area such as social cohesion and attitudes towards health. This explanation for unexplained survival variations might also be the case for general practices. It could be that general practices differ in their availability, quality, and follow-up of care, such as providing support in cardiac rehabilitation. Survival variations by general practices might reduce when their performance is considered in medical resource allocation.

The fifth aim was to estimate the years lost or gained in effective age for each of the medical conditions, treatments, lifestyle choices, and socio-demographic factors at each age, and investigate how this could inform individuals about financial planning for retirement. With the reforms of the UK pension system in 2015, individuals have now greater freedom in what they can do with their pension pots during retirement. This freedom, however, comes with greater responsibility and more complex decision making in planning their finances for retirement. A key factor in retirement planning is knowing the average period expectation of life for a certain medical history at a certain age. This research provided estimates of years lost or gained in effective age for different scenarios compared to the average period expectation of life at four key ages of retirement. These estimates are informative when setting up and reviewing financial plans for retirement.

This research found that the highest increase in effective age were associated with a single or multiple AMIs (range across age cohorts 4-5 or 5-6 years, respectively), cardiovascular system conditions (6-10 years), diabetes (4-6 years), and smoking (7-10 years). In contrast, the highest decrease in effective age were associated with prescription of beta blockers (1 year) or statins (2-3 years), and with coronary revascularisation (2 years). The prevalence of multimorbidity increases with age, where the most common morbidities in descending order are hypertension, lipid metabolism disorder, chronic low back pain, diabetes, joint arthritis, and ischaemic heart disease (Schäfer et al., 2010). Based on the presence of these morbidities in the survival models for AMI and the most prevalent drugs prescribed in these age cohorts, the three most likely scenarios are (1) a single AMI with prescription of statins and aspirin, (2) additionally with hypertension, and (3) additionally with diabetes, see Table 7.1. Having a medical history of scenario 1 by age 60, 65, 70, or 75 would be associated with a decrease in the average longevity of 3.9, 2.5, 1.2, and 1.2 years, respectively, for men, and of 4.2, 2.6, 1.3, and 1.4 years, respectively, for women. Having a medical

Table 7.1: Average period expectation of life for various scenarios based on the survival models for acute myocardial infarction (AMI)

Using the UK life tables of 2010-2012 (ONS, 2016), the number of years lost or gained in effective age associated with a single AMI, statins prescription, aspirin prescription, hypertension, and diabetes were translated into the average period expectation of life at different ages.

		60	65	70	75
Average	Men	22.2	18.2	14.5	11.1
	Women	25.0	20.7	16.7	12.9
Single AMI + statins + aspirin (1)	Men	18.3	15.7	13.3	9.9
	Women	20.8	18.1	15.4	11.5
Scenario 1 + hypertension (2)	Men	17.1	15.7	13.1	9.5
	Women	19.5	18.1	15.1	11.1
Scenario 2 + diabetes (3)	Men	15.9	12.0	10.3	7.4
	Women	18.3	13.9	12.0	8.7

history of scenario 2 by age 60, 65, 70, or 75 would be associated with a decrease in the average longevity of 5.1, 2.5, 1.4, and 1.6 years, respectively, for men, and of 5.5, 2.6, 1.6, and 1.8 years, respectively, for women. Having a medical history of scenario 2 by age 60, 65, 70, or 75 would be associated with a decrease in the average longevity of 6.3, 6.2, 4.2, and 3.7 years, respectively, for men, and of 6.7, 6.8, 4.7, and 4.2 years, respectively, for women.

The sixth aim was to investigate which medical conditions, treatments, lifestyle factors, socio-demographic factors, and interactions of risk factors at each age do and do not contribute in explaining survival variations and therefore to minimise the basis risk of estimating life expectancy for the pricing of annuities. This research found that the risk factors contributing the most in explaining survival variations were history of different types of CVD, the interaction between body mass index and smoking status, and general practice. Actuaries might want to focus on these risk factors to improve longevity estimation. The following risk factors might not be of importance for longevity estimation as these contributed the least in explaining survival variations: hypertension, alcohol consumption status, and prescription of aspirin or calcium-channel blockers.

The last aim was to investigate whether the effects of treatments, lifestyle choices, or other risk factors on longevity change over time and might form longevity risks that should be taken into account with pricing of annuities. Over the last five years of the study period, from 2006 to 2011, 10% more AMI patients were prescribed beta blockers or had coronary revascularisation. Both treatments could prolong life by 1 to 2 years. Actuaries should be aware that the effect of coronary revascularisation on the hazard of mortality changes over time, where it is mainly effective in the reducing early mortality. Over the last five years of the study period, 15% more people without a history of CVD were prescribed statins. Depending on the age and cardiac risk at prescription, statins therapy could prolong life up to 2 years. These upward trends in treatments might form a longevity risk as a considerable group of patients could life longer than expected.

7.5 Conclusions

The primary objectives of this research were to investigate how a history of CVD affects longevity and which treatments improve longevity in the general population based on a secondary data analysis derived from UK general practice patient records of 1987 to 2011.

This research reported on the survival prospects associated with a history of a single or multiple AMIs and how the survival prospects could be modified by coronary revascularisation and drug therapy for secondary prevention. Based on previous studies and evidence, it was unclear by how much AMI survivors in routine clinical practice are worse off and to what the extent of survival benefits by various treatments are. This is because previous studies excluded controls or limited adjustment for confounders, and made use of hospital and register data. This research fills in some of the gaps in evidence, and provides an important new source of information on the survival prospects of AMI survivors in routine clinical practice. The findings suggest that AMI survivors are to a lesser extent worse off than previously estimated and that the current recommended guidelines of drug therapy for secondary prevention of AMI may not necessarily be associated with improved longevity.

This research also reported on the survival benefits of statins prescribed as primary prevention of CVD. The main evidence base to date came from clinical trials and this is the first large observational population-based cohort study on the matter. Clinical trials have a high internal reliability but may be poorly generalizable due to excluding many patients who will be treated in routine practice. This research fills in some of the gaps in evidence, and provides an important new source of information on the possible effects of statins in routine clinical practice. The findings suggest that the current internationally recommended thresholds for statins therapy for primary prevention of CVD in routine practice may be too low and lead to overtreatment.

Bibliography

Abildstrom, S., Rasmussen, S., and Madsen, M. (2005). Changes in hospitalization rate and mortality after acute myocardial infarction in denmark after diagnostic criteria and methods changed. *European Heart Journal*, 26(10):990–995.

Acute Infarction Ramipril Efficacy Study (AIRE Study) (1993). Effect of ramipril on mortality and morbidity of survivors of acute myocardial infarction with clinical evidence of heart failure. *The Lancet*, 342(8875):821–828.

Ali, S. and Rouse, A. (2002). Practice audits: reliability of sphygmomanometers and blood pressure recording bias. *Journal of Human Hypertension*, 16(5):359–361.

Allison, P. D. (2001). *Missing data*. Number 136 in Quantitative Applications in the Social Sciences. SAGE Publications, Inc.

Antman, E., Bassand, J.-P., Klein, W., Ohman, M., Sendon, J., Rydén, L., Simoons, M., and Tendera, M. (2000). Myocardial infarction redefineda consensus document of the Joint European Society of Cardiology/American College of Cardiology committee for the redefinition of myocardial infarction: the Joint European Society of Cardiology/American College of Cardiology Committee. *Journal of the American College of Cardiology*, 36(3):959–969.

Ashman, K., Cawood, A., and Stratton, R. (2012). OC-038 Healthcare use according to body mass index (BMI) category in individuals registered to GP practices. *Gut*, 61(Suppl 2):A16–A17.

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.

Bartley, M. (2004). *Health inequality: An introduction to theories, concepts and methods*. Polity Press.

Bata, I., Gregor, R., Wolf, H., and Brownell, B. (2006). Trends in five-year survival of patients discharged after acute myocardial infarction. *Canadian Journal of Cardiology*, 22(5):399–404.

Baxter, L. (2015a). *Pensions freedoms: the first three months - access to the freedoms, take-up, advice issues, and the FCA review.* The Chartered Insurance Institute: London.

Baxter, L. (2015b). *Pensions freedoms: the unfolding picture access to the freedoms, take-up, and 'advice' issues.* The Chartered Insurance Institute: London.

Baxter, L. (2016). *What consumers want: pensions freedoms and the "new normal".* The Chartered Insurance Institute: London.

Beta-Blocker Heart Attack Trial Research Group (BHAT Research Group) (1982). A randomized trial of propranolol in patients with acute myocardial infarction. *JAMA*, 247(12):1707–1714.

Biondi-Zoccai, G., Lotrionte, M., Agostoni, P., Abbate, A., Fusaro, M., Burzotta, F., Testa, L., Sheiban, I., and Sangiorgi, G. (2006). A systematic review and meta-analysis on the hazards of discontinuing or not adhering to aspirin among 50 279 patients at risk for coronary artery disease. *European Heart Journal*, 27(22):2667–2674.

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. *Informatics in Primary Care*, 19(4):251–255.

Borghi, C., Marino, P., Zardini, P., Magnani, B., Collatina, S., Ambrosioni, E., et al. (1998). Short-and long-term effects of early fosinopril administration in patients with acute anterior myocardial infarction undergoing intravenous thrombolysis: results from the Fosinopril in Acute Myocardial Infarction Study. *American Heart Journal*, 136(2):213–225.

Brenner, H., Gefeller, O., and Greenland, S. (1993). Risk and rate advancement periods as measures of exposure impact on the occurrence of chronic diseases. *Epidemiology*, 4(3):229–236.

Breslow, N. (1972). Discussion of professor cox's paper. *Journal of the Royal Statistical Society, Series B*, 34:216–217.

Briffa, T., Hickling, S., Knuiman, M., Hobbs, M., Hung, J., Sanfilippo, F., Jamrozik, K., and Thompson, P. (2009). Long term survival after evidence based treatment of acute myocardial infarction and revascularisation: follow-up of population based perth monica cohort, 1984-2005. *BMJ*, 338:b36.

Brown, H. and Prescott, R. (2006). *Applied mixed models in medicine.* Wiley.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Campbell, S., Reeves, D., Kontopantelis, E., Middleton, E., Sibbald, B., and Roland, M. (2007). Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine*, 357(2):181–190.

Capewell, S., Livingston, B., MacIntyre, K., Chalmers, J., Boyd, J., Finlayson, A., Redpath, A., and et al. (2000). Trends in case-fatality in 117 718 patients admitted with acute myocardial infarction in scotland. *European Heart Journal*, 21(22):1833–1840.

Capewell, S., Murphy, N. F., MacIntyre, K., Frame, S., Stewart, S., Chalmers, J., Boyd, J., Finlayson, A., Redpath, A., and McMurray, J. J. (2006). Short-term and long-term outcomes in 133 429 emergency patients admitted with angina or myocardial infarction in Scotland, 1990–2000: population-based cohort study. *Heart*, 92(11):1563–1570.

Carlin, J. B., Sterne, A. C., White, I. R., Royston, P., Kenward, M. G., Wood, A. M., et al. (2007). Multiple imputation needs to be used with care and reported in detail. *BMJ*, 335:136.

Centre for Multilevel Modelling (2016). REALCOM: Developing multilevel models for REAListically COMplex social science data. `http://www.bris.ac.uk/cmm/software/realcom/`. Accessed: 2017-01-19.

Cepeda, M., Boston, R., Farrar, J., and Strom, B. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3):280–287.

Chang, W.-C., Kaul, P., Westerhout, C. M., Graham, M. M., Fu, Y., Chowdhury, T., and Armstrong, P. W. (2003). Impact of sex on long-term mortality from acute myocardial infarction vs unstable angina. *Archives of Internal Medicine*, 163(20):2476–2484.

Chapman, I. (2010). *Obesity paradox during aging*. Karger Publishers.

Charlson, M., Pompei, P., Ales, K., and MacKenzie, C. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40(5):373–383.

Chieffo, A., Buchanan, G. L., Mauri, F., Mehilli, J., Vaquerizo, B., Moynagh, A., Mehran, R., and Morice, M.-C. (2012). ACS and STEMI treatment: gender-related issues. *EuroIntervention*, 8:P27e35.

Cholesterol Treatment Trialists Collaborators (CTTC) (2012). The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet*, 380(9841):581–590.

ClinicalCodes.org (2016). All publications with clinical code lists. `https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/articles/`. Accessed: 2017-01-19.

ClinRisk Ltd (2015). Welcome to the QRISK®2-2015 risk calculator. `http://www.qrisk.org`. Accessed: 2017-01-19.

Collins, G. S. and Altman, D. G. (2009). An independent external validation and evaluation of qrisk cardiovascular risk prediction: a prospective open cohort study. *BMJ*, 339:b2584.

Collins, G. S. and Altman, D. G. (2012). Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ*, 344.

Cologne, J., Sharp, G., Neriishi, K., Verkasalo, P., Land, C., and Nakachi, K. (2004). Improving the efficiency of nested case-control studies of interaction by selecting controls using counter matching on exposure. *International Journal of Epidemiology*, 33(3):485–492.

Coronary Drug Project Research Group (CDP) (1976). Aspirin in coronary heart disease. *Journal of Chronic Diseases*, 29(10):625–642.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34(2):187–220.

Cox, D. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC Press.

CPRD (2016). The Clinical Practice Research Datalink. `https://www.cprd.com`. Accessed: 2017-01-19.

Dalal, H. M., Doherty, P., and Taylor, R. S. (2015). Cardiac rehabilitation. *BMJ*, 351:h5000.

Department of Health (2011). The Good Practice Guidelines for GP electronic patient records. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/215680/dh_125350.pdf`. Accessed: 2017-01-19.

Dormuth, C. R., Patrick, A. R., Shrank, W. H., Wright, J. M., Glynn, R. J., Sutherland, J., and Brookhart, M. A. (2009). Statin adherence and risk of accidents a cautionary tale. *Circulation*, 119(15):2051–2057.

Douglas, L. and Szatkowski, L. (2013). Socioeconomic variations in access to smoking cessation interventions in UK primary care: insights using the Mosaic classification in a large dataset of primary care records. *BMC Public Health*, 13(1):546.

Downs, J. R., Clearfield, M., Weis, S., Whitney, E., Shapiro, D. R., Beere, P. A., Langendorfer, A., Stein, E. A., Kruyer, W., and Jr, G. A. M. (1998). Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of AFCAPS/TEXCAPS. *JAMA*, 279(20):1615–1622.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565.

Experian Ltd. (2009). Mosaic United Kingdom: The consumer classification of the United Kingdom. `http://www.experian.co.uk/assets/business-strategies/brochures/Mosaic_UK_2009_brochure.pdf`. Accessed: 2017-01-19.

García Rodríguez, L., Cea-Soriano, L., Martín-Merino, E., and Johansson, S. (2011). Discontinuation of low dose aspirin and risk of myocardial infarction: case-control study in uk primary care. *BMJ*, 343:d4094.

Gerber, Y., Benyamini, Y., Goldbourt, U., and Drory, Y. (2010). Neighborhood socioeconomic context and long-term survival after myocardial infarction. *Circulation*, 121(3):375–383.

Gerber, Y., Rosen, L., Goldbourt, U., Benyamini, Y., and Drory, Y. (2009). Smoking Status and Long-Term Survival After First Acute Myocardial InfarctionA Population-Based Cohort Study. *Journal of the American College of Cardiology*, 54(25):2382–2387.

Gitsels, L. A., Kulinskaya, E., and Steel, N. (2016). Survival benefits of statins for primary prevention: a cohort study. *PloS One*, 11(11):e0166847.

Gitsels, L. A., Kulinskaya, E., and Steel, N. (2017). Survival prospects after acute myocardial infarction in the UK: a matched cohort study 1987-2011. *BMJ Open*, 6:e013570.

Godlee, F. (2014). Adverse effects of statins. *BMJ*, 348.

Gottlieb, S., Harpaz, D., Shotan, A., Boyko, V., Leor, J., Cohen, M., Mandelzweig, L., Mazouz, B., Stern, S., Behar, S., et al. (2000). Sex Differences in Management and Outcome After Acute Myocardial Infarction in the 1990s: A Prospective Observational Community-Based Study. *Circulation*, 102(20):2484–2490.

Gourlay, S., Rundle, A., and Barron, H. (2002). Smoking and mortality following acute myocardial infarction: results from the National Registry of Myocardial Infarction 2 (NRMI 2). *Nicotine & Tobacco Research*, 4(1):101–107.

Graham, I., Atar, D., Borch-Johnsen, K., Boysen, G., Burell, G., Cifkova, R., Dallongeville, J., de Backer, G., Ebrahim, S., Gjelsvik, B., and et al (2007). European guidelines on cardiovascular disease prevention in clinical practice: executive summary Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts). *European Heart Journal*, 28(19):2375–2414.

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

Greenland, S. and Morgenstern, H. (1990). Matching and efficiency in cohort studies. *American Journal of Epidemiology*, 131(1):151–159.

Greenwood, M. (1926). *A Report on the Natural Duration of Cancer*. Reports on public health and medical subjects. H.M. Stationery Office.

Hall, G. (2009). Validation of death and suicide recording on the thin uk primary care database. *Pharmacoepidemiology and Drug Safety*, 18(2):120–131.

Hardoon, S. L., Whincup, P. H., Petersen, I., Capewell, S., and Morris, R. W. (2011). Trends in longer-term survival following an acute myocardial infarction and prescribing of evidenced-based medications in primary care in the UK from 1991: a longitudinal population-based study. *Journal of Epidemiology and Community Health*, 65(9):770–774.

Harrell, F. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Harrison, W., Lancashire, R., and Marshall, T. (2007). Variation in recorded blood pressure terminal digit bias in general practice. *Journal of Human Hypertension*, 22(3):163–167.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

Health & Social Care Information Centre (HSCIC) (2006). QOF 2014/15 results. `http://qof.hscic.gov.uk/`. Accessed: 2017-01-19.

Health & Social Care Information Centre (HSCIC) (2013). Quality and outcomes framework - 2012-13. `http://www.hscic.gov.uk/catalogue/PUB12262`. Accessed: 2017-01-19.

Health & Social Care Information Centre (HSCIC) (2015). Health Survey for England, 2014 [NS]. `http://www.hscic.gov.uk/catalogue/PUB19295`. Accessed: 2017-01-19.

Health & Social Care Information Centre (HSCIC) (2016a). Deaths. `http://systems.hscic.gov.uk/demographics/pds/contents/deaths`. Accessed: 2017-01-19.

Health & Social Care Information Centre (HSCIC) (2016b). Health Survey for England; Health, social care and lifestyles. `http://www.hscic.gov.uk/healthsurveyengland`. Accessed: 2017-01-19.

Hemingway, H. and Marmot, M. (1999). Psychosocial factors in the aetiology and prognosis of coronary heart disease: systematic review of prospective cohort studies. *BMJ*, 318(7196):1460–1467.

Hennekens, C., Buring, J., and Mayrent, S. (1987). *Epidemiology in medicine*. Boston: Little Brown and Company.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., and Smeeth, L. (2015). Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*, 44(3):827–836.

Herrett, E., George, J., Denaxas, S., Bhaskaran, K., Timmis, A., Hemingway, H., and Smeeth, L. (2013a). Type and timing of heralding in ST-elevation and non-ST-elevation myocardial infarction: an analysis of prospectively collected electronic healthcare records linked to the national registry of acute coronary syndromes. *European Heart Journal: Acute Cardiovascular Care*, 2(3):235–245.

Herrett, E., Shah, A. D., Boggon, R., Denaxas, S., Smeeth, L., van Staa, T., Timmis, A., and Hemingway, H. (2013b). Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ*, 346:f2350.

Herrett, E., Thomas, S. L., Schoonen, W. M., Smeeth, L., and Hall, A. J. (2010). Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *British Journal of Clinical Pharmacology*, 69(1):4–14.

Herzog, C., Ma, J., and Collins, A. (1998). Poor long-term survival after acute myocardial infarction among patients on long-term dialysis. *New England Journal of Medicine*, 339(12):799–805.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G. M., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., and Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ*, 346:e5793.

Hippisley-Cox, J. and Coupland, C. (2010a). Individualising the risks of statins in men and women in England and Wales: population-based cohort study. *Heart*, 96(12):939–947.

Hippisley-Cox, J. and Coupland, C. (2010b). 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.

Hippisley-Cox, J. and Coupland, C. (2010c). Unintended effects of statins in men and women in England and Wales: population based cohort study using the QResearch database. *BMJ*, 340:c2197.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., and Brindle, P. (2008). Predicting cardiovascular risk in england and wales: prospective derivation and validation of qrisk2. *BMJ*, 336(7659):1475–1482.

HM Treasury (2014). Freedom and choice in pensions. `http://www.gov.uk/government/consultations/freedom-and-choice-in-pensions`. Accessed: 2017-01-19.

Hosmer, D., Lemeshow, S., and May, S. (2008). *Model development*. Wiley.

IMS Health Incorporated (2015a). THIN Access to Data. `http://csdmruk.cegedim.com/our-data/accessing-the-data.shtml`. Accessed: 2017-01-19.

IMS Health Incorporated (2015b). THIN Data Collection. http://www.csdmruk.imshealth.com/our-data/data-collection.shtml. Accessed: 2017-01-19.

IMS Health Incorporated (2015c). THIN Data Content. http://csdmruk.cegedim.com/our-data/data-content.shtml. Accessed: 2017-01-19.

IMS Health Incorporated (2015d). THIN Data Statistics. http://www.csdmruk.imshealth.com/our-data/statistics.shtml. Accessed: 2017-01-19.

Joint Formulary Committee (2016a). 2.12 lipid-regulating drugs. http://www.evidence.nhs.uk/formulary/bnf/current/2-cardiovascular-system/212-lipid-regulating-drugs. Accessed: 2017-01-19.

Joint Formulary Committee (2016b). *British National Formulary*. London: BMJ Group and Pharmaceutical Press.

Jordan, K., Porcheret, M., and Croft, P. (2004). Quality of morbidity coding in general practice computerized medical records: a systematic review. *Family Practice*, 21(4):396–412.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Khan, N. F., Harrison, S. E., and Rose, P. W. (2010). Validity of diagnostic coding within the general practice research database: a systematic review. *British Journal of General Practice*, 60(572):e128–e136.

Kirchberger, I., Meisinger, C., Golüke, H., Heier, M., Kuch, B., Peters, A., Quinones, P., von Scheidt, W., and Mielck, A. (2014). Long-term survival among older patients with myocardial infarction differs by educational level: results from the monica/kora myocardial infarction registry. *International Journal for Equity in health*, 13(1):19.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer.

Kleinbaum, D. and Klein, M. (2011). *Survival Analysis: A self-learning text*. New York, Springer-Verlag.

Køber, L., Torp-Pedersen, C., Carlsen, J. E., Bagger, H., Eliasen, P., Lyngborg, K., Videbæk, J., Cole, D. S., Auclert, L., Pauly, N. C., et al. (1995). A clinical trial of the angiotensin-converting–enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine*, 333(25):1670–1676.

Koek, H., Soedamah-Muthu, S., Kardaun, J., Gevers, E., de Bruin, A., Reitsma, J., Bots, M., and Grobbee, D. (2007). Short-and long-term mortality after acute myocardial infarction: comparison of patients with and without diabetes mellitus. *European Journal of Epidemiology*, 22(12):883–888.

Kostis, W. J., Cheng, J. Q., Dobrzynski, J. M., Cabrera, J., and Kostis, J. B. (2012). Meta-analysis of statin effects in women versus men. *Journal of the American College of Cardiology*, 59(6):572–582.

Krumholz, H. M. (2016). Statins evidence: when answers also raise questions. *BMJ*, 354:i4963.

Kulinskaya, E. and Gitsels, L. A. (2016). Use of big health and actuarial data for understanding longevity and morbidity risk. *Longevity Bulletin*, (9):15–18.

Lagerqvist, B., Säfström, K., Ståhle, E., Wallentin, L., and Swahn, E. (2001). Is early invasive treatment of unstable coronary artery disease equally effective for both women and men? *Journal of the American College of Cardiology*, 38(1):41–48.

Langholz, B. and Clayton, D. (1994). Sampling strategies in nested case-control studies. *Environmental Health perspectives*, 102(Suppl 8):47.

Langley, T., Szatkowski, L., Wythe, S., and Lewis, S. (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.

Lopressor Intervention Trial Research Group (LIT Research Group) (1987). The Lopressor Intervention Trial: multicentre study of metoprolol in survivors of acute myocardial infarction. *European Heart Journal*, 8(10):1056–1064.

Löwel, H., Koenig, W., Engel, S., Hörmann, A., and Keil, U. (2000). The impact of diabetes mellitus on survival after myocardial infarction: can it be modified by drug treatment? *Diabetologia*, 43(2):218–226.

Ltd, C. (2015). QRISK®2015 Annual Update Information. `http://www.qrisk.org/QRISK2-2015-Annual-Update-Information.pdf`. Accessed: 2017-01-19.

Luepker, R. V. (2011). WHO MONICA Project: what have we learned and where to go from here. *Public Health Reviews*, 33(2).

MacDonald, T. and Morant, S. (2008). Prevalence and treatment of isolated and concurrent hypertension and hypercholesterolaemia in the United Kingdom. *British Journal of Clinical Pharmacology*, 65(5):775–786.

Maguire, A., Blak, B. T., and Thompson, M. (2009). The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiology and Drug Safety*, 18(1):76–83.

Marshall, A., Altman, D. G., and Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology*, 10(1):112.

Marston, L., Carpenter, J., Walters, K., Morris, R., Nazareth, I., and Petersen, I. (2010). Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiology and Drug Safety*, 19(6):618–626.

Massó González, E., Johansson, S., Wallander, M.-A., and García Rodríguez, L. (2009). Trends in the prevalence and incidence of diabetes in the uk: 1996–2005. *Journal of Epidemiology and Community Health*, 63(4):332–336.

Mathur, R., Bhaskaran, K., Chaturvedi, N., Leon, D. A., Grundy, E., Smeeth, L., et al. (2014). Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *Journal of Public Health*, 36(4):684–692.

McAuley, P. A. and Blair, S. N. (2011). Obesity paradoxes. *Journal of Sports Sciences*, 29(8):773–782.

Mortensen, M. B. and Falk, E. (2014). Real-life evaluation of European and American high-risk strategies for primary prevention of cardiovascular disease in patients with first myocardial infarction. *BMJ Open*, 4(10).

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Naghavi, M., Wang, H., Lozano, R., Davis, A., Liang, X., Zhou, M., Vollset, S. E., Ozgoren, A. A., Abdalla, S., Abd-Allah, F., et al. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease study 2013. *The Lancet*, 385(9963):117–171.

Nakamura, H., Arakawa, K., Itakura, H., Kitabatake, A., Goto, Y., Toyota, T., Nakaya, N., Nishimoto, S., Muranaka, M., Yamamoto, A., et al. (2006). Primary prevention of cardiovascular disease with pravastatin in japan (mega study): a prospective randomised controlled trial. *The Lancet*, 368(9542):1155–1163.

National Health Service (NHS) (2013). Improving general practice - a call to action. `https://www.england.nhs.uk/wp-content/uploads/2013/08/igp-cta-slide.pdf`. Accessed: 2017-01-19.

National Health Service (NHS) (2014a). Type 2 diabetes. http://www.nhs.uk/conditions/Diabetes-type2/Pages/Introduction.aspx. Accessed: 2017-01-19.

National Health Service (NHS) (2014b). What is blood pressure? http://www.nhs.uk/chq/Pages/what-is-blood-pressure.aspx. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2011). Hypertension in adults: diagnosis and management. clinical guideline 127. https://www.nice.org.uk/guidance/cg127. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2013a). MI - secondary prevention: Secondary prevention in primary and secondary care for patients following a myocardial infarction. Partial update of Clinical Guideline CG48. https://www.nice.org.uk/guidance/cg172/evidence/myocardial-infarction-secondary-prevention-full-guideline-248682925. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2013b). Myocardial infarction: cardiac rehabilitation and prevention of further ention of further MI. Clinical Guideline 172. https://www.nice.org.uk/guidance/cg172. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2014a). About the Quality and Outcomes Framework (QOF). http://www.nice.org.uk/aboutnice/qof/qof.jsp. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2014b). Wider use of statins could cut deaths from heart disease. https://www.nice.org.uk/news/article/wider-use-of-statins-could-cut-deaths-from-heart-disease. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2015). Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. clinical guideline 181. https://www.nice.org.uk/guidance/cg181. Accessed: 2017-01-19.

National Institute for Health and Care Excellence (NICE) (2016). Statins for the prevention of cardiovascular events in patients at increased risk of developing cardiovascular disease or those with established cardiovascular disease guidance. clinical guideline 181. https://www.nice.org.uk/guidance/ta94. Accessed: 2017-01-19.

Newton, J. N., Briggs, A. D., Murray, C. J., Dicker, D., Foreman, K. J., Wang, H., Naghavi, M., Forouzanfar, M. H., Ohno, S. L., Barber, R. M., et al. (2015). Changes in health in England, with analysis by English regions and areas of deprivation, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 386(10010):2257–2274.

Nicholson, A., Tate, A. R., Koeling, R., and Cassell, J. A. (2011). What does validation of cases in electronic record databases mean? the potential contribution of free text. *Pharmacoepidemiology and Drug Safety*, 20(3):321–324.

Nietert, P., Wessell, A., Feifer, C., and Ornstein, S. (2006). Effect of terminal digit preference on blood pressure measurement and treatment in primary care. *American Journal of Hypertension*, 19(2):147–152.

Nigam, A., Wright, R., Allison, T., Williams, B., Kopecky, S., Reeder, G., Murphy, J., and Jaffe, A. (2006). Excess weight at time of presentation of myocardial infarction is associated with lower initial mortality risks but higher long-term risks including recurrent re-infarction and cardiac death. *International Journal of Cardiology*, 110(2):153–159.

O'Connell, R. and Lim, L. (2000). Utility of the charlson comorbidity index computed from routinely collected hospital discharge diagnosis codes. *Methods of Information in Medicine*, 39(1):7–11.

Office for National Statistics (2013). Pension Trends, Chapter 6: Private Pensions, 2013 edition. `http://www.ons.gov.uk/ons/dcp171766_313466`. Accessed: 2017-01-19.

Office for National Statistics (ONS) (2014). Opinions and lifestyle survey, adult smoking habits in great britain, 2013. `http://www.ons.gov.uk/ons/dcp171778_386291.pdf`. Accessed: 2017-01-19.

Office for National Statistics (ONS) (2016). National Life Tables: United Kingdom. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables`. Accessed: 2017-01-19.

O'Keeffe, A. G., Petersen, I., and Nazareth, I. (2015). Initiation rates of statin therapy for the primary prevention of cardiovascular disease: an assessment of differences between countries of the uk and between regions within england. *BMJ Open*, 5(3):e007207.

Olsson, G., Rehnqvist, N., Sjögren, A., Erhardt, L., and Lundman, T. (1985). Long-term treatment with metoprolol after myocardial infarction: effect on 3 year mortality and morbidity. *Journal of the American College of Cardiology*, 5(6):1428–1437.

O'Quigley, J., Xu, R., and Stare, J. (2005). Explained randomness in proportional hazards models. *Statistics in Medicine*, 24(3):479–489.

Parish, E., Bloom, T., and Godlee, F. (2015). Statins for people at low risk. *BMJ*, 351:h3908.

Pedersen, T. R. (1983). A multicentre study on timolol in secondary prevention after myocardial infarction. *Acta Medica Scandinavica*, 674:1–129.

Pennington, J. (2013). *Moving on: migration trends in later life.* Institute for Public Policy Research, London.

Perk, J., de Backer, G., Gohlke, H., Graham, I., Reiner, Ž., Verschuren, M., Albus, C., , et al. (2012). European guidelines on cardiovascular disease prevention in clinical practice (version 2012). *European Heart Journal*, 33(13):1635–1701.

Persson, I. (2002). *Essays on the assumption of proportional hazards in Cox regression.* Acta Universitatis Upsaliensis.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer*, 35(1):1–39.

Pfeffer, M. A., Braunwald, E., Moyé, L. A., Basta, L., Brown Jr, E. J., Cuddy, T. E., Davis, B. R., Geltman, E. M., Goldman, S., Flaker, G. C., et al. (1992). Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction: results of the Survival and Ventricular Enlargement Trial. *New England Journal of Medicine*, 327(10):669–677.

Platt, R., Madre, L., Reynolds, R., and Tilson, H. (2008). Active drug safety surveillance: a tool to improve public health. *Pharmacoepidemiology and Drug Safety*, 17(12):1175–1182.

QResearch (2016). QResearch. `http://www.qresearch.org`. Accessed: 2017-01-19.

Quint, J., Herrett, E., Bhaskaran, K., Timmis, A., Hemingway, H., Wedzicha, J., and Smeeth, L. (2013). Effect of β blockers on mortality after myocardial infarction in adults with COPD: population based cohort study of UK electronic healthcare records. *BMJ*, 347:f6650.

Raboud, J. and Breslow, N. (1989). Efficiency gains from the addition of controls to matched sets in cohort studies. *Statistics in Medicine*, 8(8):977–985.

Rasbash, J., Steele, F., Browne, W., and Goldstein, H. (2012). A user's guide to MLwiN: Version 2.26.

Ray, W. A. (2003). Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*, 158(9):915–920.

Reeves, D., Springate, D. A., Ashcroft, D. M., Ryan, R., Doran, T., Morris, R., Olier, I., and Kontopantelis, E. (2014). Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case-control analysis. *BMJ Open*, 4:e004952.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-validation.* Springer.

Research, C. M. (2011). CSD EPIC Research Format THIN Data, version 2.2.

Richards, S. (2008). Applying survival models to pensioner mortality data. *British Actuarial Journal*, 14(02):257–303.

Ridker, P. M., Danielson, E., Fonseca, F., Genest, J., Gotto Jr, A. M., Kastelein, J., Koenig, W., Libby, P., Lorenzatti, A. J., MacFadyen, J. G., et al. (2008). Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine*, 359(21):2195.

Ridsdale, B. and Gallop, A. (2010). Mortality by cause of death and by socio-economic and demographic stratification 2010. *International Congress of Actuaries*, page 183.

Roberts, R., Croft, C., Gold, H. K., Hartwell, T. D., Jaffe, A. S., Muller, J. E., Mullin, S. M., Parker, C., Passamani, E. R., Poole, W. K., et al. (1984). Effect of propranolol on myocardial-infarct size in a randomized blinded multicenter trial. *New England Journal of Medicine*, 311(4):218–225.

Roger, V., Weston, S., Gerber, Y., Killian, J., Dunlay, S., Jaffe, A., Bell, M., Kors, J., Yawn, B., and Jacobsen, S. (2010). Trends in incidence, severity, and outcome of hospitalized myocardial infarction. *Circulation*, 121(7):863–869.

Roque, F., Amuchastegui, L., Morillos, M. L., Mon, G., Girotti, A., Drajer, S., Fortunato, M., Moreyra, E., Tuero, P., and Solchaga, J. (1987). Beneficial effects of timolol on infarct size and late ventricular tachycardia in patients with acute myocardial infarction. *Circulation*, 76(3):610–617.

Rosengren, A., Spetz, C., Köster, M., Hammar, N., Alfredsson, L., and Rosen, M. (2001). Sex differences in survival after myocardial infarction in Sweden. Data from the Swedish National Acute Myocardial Infarction register. *European Heart Journal*, 22(4):314–322.

Rothman, K., Greenland, S., and Lash, T. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.

Royston, P. (2006). Explained variation for survival models. *Stata Journal*, 6(1):83–96.

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

Salomaa, V., Ketonen, M., Koukkunen, H., Immonen-Räihä, P., Lehtonen, A., Torppa, J., Kuulasmaa, K., Kesäniemi, Y., and Pyörälä, K. (2006). The effect of correcting for troponins on trends in coronary heart disease events in finland during 1993–2002: the finami study. *European Heart Journal*, 27(20):2394–2399.

Sanfilippo, F., Hobbs, M., Knuiman, M., and Hung, J. (2008). Impact of new biomarkers of myocardial damage on trends in myocardial infarction hospital admission rates from population-based administrative data. *American Journal of Epidemiology*, 168(2):225–233.

Scandinavian Simvastatin Survival Study (4S) (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *The Lancet*, 344(8934):1383–1389.

Schäfer, I., von Leitner, E.-C., Schön, G., Koller, D., Hansen, H., Kolonko, T., Kaduszkiewicz, H., Wegscheider, K., Glaeske, G., and van den Bussche, H. (2010). Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS One*, 5(12):e15941.

Schafer, J. and Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147.

Schmidt, M., Jacobsen, J., Lash, T., Bøtker, H., and Sørensen, H. (2012). 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. *BMJ*, 344:e356.

Sharma, A., Lewis, S., and Szatkowski, L. (2010). Insights into social disparities in smoking prevalence using mosaic, a novel measure of socioeconomic status: an analysis using a large primary care dataset. *BMC Public Health*, 10(1):755.

Shephard, E., Stapley, S., and Hamilton, W. (2011). The use of electronic databases in primary care research. *Family Practice*, 28(4):352–354.

Shepherd, J., Cobbe, S. M., Ford, I., Isles, C. G., Lorimer, A. R., Macfarlane, P. W., McKillop, J. H., and Packard, C. J. (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *New England Journal of Medicine*, 333(20):1301–1308.

Smolina, K., Wright, F., Rayner, M., and Goldacre, M. (2012a). Determinants of the decline in mortality from acute myocardial infarction in england between 2002 and 2010: linked national database study. *BMJ*, 344:d8059.

Smolina, K., Wright, F., Rayner, M., and Goldacre, M. (2012b). Long-term survival and recurrence after acute myocardial infarction in england, 2004 to 2010. *Circulation: Cardiovascular Quality and Outcomes*, 5(4):532–540.

Spiegelhalter, D. (2016). How old are you, really? Communicating chronic risk through 'effective age' of your body and organs. *BMC Medical Informatics and Decision Making*, 16:104.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Bairey Merz, C. N., Blum, C. B., Eckel, R. H., Goldberg, A. C., et al. (2014). 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63:2889–2934.

Studies of Left Ventricular Dysfunction Investigators (SOLVD Investigators) (1992). Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. *New England Journal of Medicine*, 327(10):685–91.

Swanton, R. and Banerjee, S. (2009). *Swanton's Cardiology.* Wiley.

Szatkowski, L., Lewis, S., McNeill, A., Huang, Y., and Coleman, T. (2012). Can data from primary care medical records be used to monitor national smoking prevalence? *Journal of Epidemiology and Community Health*, 66(9):791–795.

Taggar, J., Coleman, T., Lewis, S., and Szatkowski, L. (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.

Thavarajah, S., White, W., and Mansoor, G. (2003). Terminal digit bias in a specialty hypertension faculty practice. *Journal of Human Hypertension*, 17(12):819–822.

Therneau, T. (2014). Package 'survival' version 2.37-7. `http://r-forge.r-project.org`. Accessed: 2017-01-19.

Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer.

Thompson, R., O'Regan, C., Morant, S., Phillips, B., and Ong, S. (2008). Measurement of baseline total cholesterol: new data from the health improvement network (thin) database. *Primary Care Cardiovascular Journal*, 1(2):107–111.

Thurley, D. (2015). Pensions: annuities. House of Commons Library.

Townsend, N., Bhatnagar, P., Wilkins, E., Wickramasinghe, K., and Rayner, M. (2015). Cardiovascular disease statistics, 2015. British Heart Foundation: London.

Townsend, N., Williams, J., Bhatnagar, P., Wickramasinghe, K., and Rayner, M. (2014). Cardiovascular disease statistics, 2014. British Heart Foundation: London.

Tucker, N. (2014). The keep it simple guide to qof 2014/2015. `http://www.nbmedical.com/pdf/keep_simple_qof_2014.pdf`. Accessed: 2017-01-19.

UK Parliament (2015). Political challenges relating to an aging population: Key issues for the 2015 parliament. `http://www.parliament.uk/business/publications/research/key-issues-parliament-2015/social-change/ageing-population/`. Accessed: 2017-01-19.

UK Data Service (UKDS) (2006). Deprivation data. `https://census.ukdataservice.ac.uk/get-data/related/deprivation`. Accessed: 2017-01-19.

Uren, Z. and Goldring, S. (2007). *Migration trends at older ages in England and Wales*. Office for National Statistics, Ageing Unit.

Vaccarino, V., Parsons, L., Every, N., Barron, H., and Krumholz, H. (1999). Sex-based differences in early mortality after myocardial infarction. *New England Journal of Medicine*, 341(4):217–225.

van Baal, P. H., Engelfriet, P. M., Boshuizen, H. C., van de Kassteele, J., Schellevis, F. G., and Hoogenveen, R. T. (2011). Co-occurrence of diabetes, myocardial infarction, stroke, and cancer: quantifying age patterns in the dutch population using health survey data. *Popullation Health Metrics*, 9:51.

van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3).

Vaupel, J. W. (2010). Biodemography of human ageing. *Nature*, 464(7288):536–542.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291.

Vonesh, E., Schaubel, D., Hao, W., and Collins, A. (2000). Statistical methods for comparing mortality among ESRD patients: examples of regional/international variations. *Kidney International*, 57:S19–S27.

White, C. and Butt, A. (2015). *Inequality in Health and Life Expectancies within Upper Tier Local Authorities: 2009 to 2013*. Office for National Statistics.

Wijlaars, L. (2013a). Thin data: Strengths & limitations. `http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database/pros-cons`. Accessed: 2017-01-19.

Wijlaars, L. (2013b). Thin database. `http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database`. Accessed: 2017-01-19.

Wingfield, D., Cooke, J., Thijs, L., Staessen, J., Fletcher, A., Fagard, R., Bulpitt, C., and et al. (2002a). Terminal digit preference and single-number preference in the syst-eur trial: influence of quality control. *Blood Pressure Monitoring*, 7(3):169–177.

Wingfield, D., Freeman, G., and Bulpitt, C. (2002b). Selective recording in blood pressure readings may increase subsequent mortality. *QJM*, 95(9):571–577.

World Health Organization (WHO) (2006). BMI classification. `www.who.int/bmi`. Accessed: 2017-01-19.

World Health Organization (WHO) (2010). International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). `http://apps.who.int/classifications/icd10/browse/2016/en#/IX`. Accessed: 2017-01-19.

World Health Organization (WHO) (2015a). Cardiovascular diseases (CVDs). `http://www.who.int/mediacentre/factsheets/fs317`. Accessed: 2017-01-19.

World Health Organization (WHO) (2015b). Obesity and overweight. `http://www.who.int/mediacentre/factsheets/fs311`. Accessed: 2017-01-19.

World Health Organization (WHO) (2015c). World report on ageing and health. `http://www.who.int/ageing/events/world-report-2015-launch`. Accessed: 2017-01-19.

Wright, C. and Dent, T. (2014). *Quality standards in risk prediction*. PHG Foundation: Cambridge. Accessed: 2017-01-19.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., and Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*, 364(9438):937–952.

Zoungas, S., Curtis, A., Tonkin, A., and McNeil, J. (2014). Statins in the elderly: an answered question? *Current Opinion in Cardiology*, 29(4):372–380.

Appendix A

Appendix statistical methods

Table A.1: Prevalence missing observations in 60-year old cohort

Prevalence of missing observations in blood pressure, body mass index, alcohol status, and smoking status by medical history is listed in percentages. Associations of missingness with the medical history were significant (χ^2(df), p<.001).

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
Sex	Female	71,615	22.9	32.0	35.4	27.0
	Male	70,626	32.0	40.6	40.2	32.5
Year of birth	1930-35	51,082	31.9	51.9	56.6	44.5
	1936-40	91,159	24.9	27.5	27.2	21.4
Socioeconomic status	1	16,872	27.7	35.8	38.5	29.5
	2	9,181	26.7	33.1	34.7	28.3
	3	32,659	27.3	36.1	37.8	29.9
	4	20,551	27.9	37.6	38.1	30.1
	5	8,072	28.3	37.7	38.4	29.8
	6	16,443	28.2	37.4	39.4	30.0
	7	6,527	27.0	39.3	40.4	30.9
	8	13,853	23.8	33.0	34.6	27.0
	9	8,484	27.2	33.9	34.9	29.7
	10	9,599	30.1	39.9	40.6	32.4
Acute myocardial infarction	No	138,055	27.8	36.5	38.0	29.9
	Yes	4,186	14.2	29.0	31.3	22.5
Angina	No	136,281	28.0	36.7	38.1	30.1
	Yes	5,960	13.2	27.2	30.1	20.0
Heart failure	No	141,497	26.1	36.2	37.8	29.8
	Yes	744	12.0	29.4	31.3	21.0
Cardiovascular system conditions	No	135,353	28.0	36.7	38.2	30.2
	Yes	6,888	15.3	29.0	30.0	19.5
Diabetes	No	136,956	28.1	37.0	38.1	30.2

Continued on next page

Table A.1 – *Continued from previous page*

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
	Yes	5,285	8.9	17.2	28.5	17.6
Blood pressure	Normal	26,178	0.0	18.8	20.8	12.5
	Pre	49,948	0.0	20.3	22.4	13.9
	Hyper	27,150	0.0	23.5	26.0	17.3
Hypercholesterolaemia	No	118,126	31.4	40.7	41.9	33.4
	Yes	24,115	7.5	14.7	17.3	11.5
Coronary revascularisation	No	140,540	27.6	36.4	37.9	29.8
	Yes	1,701	14.2	26.0	26.5	19.2
Blood pressure lowering drugs	No	105,040	33.2	40.7	41.8	33.8
	Yes	37,201	11.0	24.0	26.3	18.2
Lipid-lowering therapy	No	135,945	28.2	37.2	38.6	30.4
	Yes	6,296	11.0	17.9	20.3	14.5
Body mass index	Normal	36,321	11.2	0.0	10.5	5.0
	Overweight	37,364	9.4	0.0	10.5	5.8
	Obese	16,918	6.9	0.0	13.0	7.2
Alcohol status	No	17,033	10.2	9.4	0.0	1.7
	Yes	71,495	10.1	8.7	0.0	2.8
Smoking status	No	59,157	10.7	13.2	12.9	0.0
	Ex	14,217	8.9	12.8	12.2	0.0
	Yes	26,594	15.0	18.7	16.4	0.0
Total		142,241	27.4	36.3	37.8	29.7

Table A.2: Prevalence missing observations in 65-year old cohort

Prevalence of missing observations in blood pressure, body mass index, alcohol status, and smoking status by medical history is listed in percentages. Associations of missingness with the medical history were significant (χ^2(df), p<.001).

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
Sex	Female	133,130	18.8	25.5	27.4	20.0
	Male	127,647	24.1	31.2	30.2	23.2
Year of birth	1925-30	48,160	31.4	52.8	56.9	44.6
	1931-35	82,920	22.7	26.9	26.5	20.7
	1936-40	129,697	16.9	20.0	19.7	13.5
Socioeconomic status	1	29,725	21.3	27.8	28.8	21.2
	2	17,101	19.6	25.3	25.8	20.0
	3	55,987	21.2	27.8	28.3	21.4
	4	35,731	21.9	29.0	28.6	21.6

Continued on next page

Table A.2 – *Continued from previous page*

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
	5	14,512	22.2	30.2	30.2	21.7
	6	30,032	22.5	29.5	30.6	22.1
	7	13,554	21.7	30.6	30.9	23.0
	8	28,332	18.7	25.2	26.3	19.8
	9	15,573	22.0	27.7	27.8	22.5
	10	20,230	23.7	31.6	31.3	23.5
Acute myocardial infarction	No	249,894	21.9	28.7	29.1	21.9
	Yes	10,883	9.8	19.2	20.4	12.8
Angina	No	243,786	22.3	28.9	29.3	22.2
	Yes	16,991	9.1	18.9	20.4	12.9
Heart failure	No	257,932	20.1	28.4	28.8	21.7
	Yes	2,845	6.3	18.7	19.3	10.4
Cardiovascular system conditions	No	239,188	22.4	28.9	29.4	22.3
	Yes	21,589	10.6	21.0	21.2	12.8
Diabetes	No	244,904	22.5	29.5	29.5	22.4
	Yes	15,873	4.8	9.7	17.5	8.7
Blood pressure	Normal	60,606	0.0	15.0	15.8	8.9
	Pre	99,642	0.0	17.1	17.8	10.6
	Hyper	44,694	0.0	21.3	22.3	14.4
Hypercholesterolaemia	No	183,692	28.4	35.2	35.3	27.4
	Yes	77,085	4.8	11.9	13.0	7.6
Coronary revascularisation	No	255,190	21.7	28.6	29.0	21.8
	Yes	5,587	7.5	15.2	16.2	9.8
Blood pressure lowering drugs	No	161,179	30.0	34.8	34.8	27.2
	Yes	99,598	7.5	17.7	19.0	12.4
Lipid-lowering therapy	No	227,414	23.8	30.8	31.1	23.6
	Yes	33,363	5.1	11.2	12.7	7.3
Body mass index	Normal	69,853	11.6	0.0	8.0	4.0
	Overweight	78,135	9.1	0.0	8.0	4.4
	Obese	39,020	6.5	0.0	9.3	5.0
Alcohol status	No	37,858	10.5	8.3	0.0	1.5
	Yes	147,982	9.6	7.5	0.0	2.2
Smoking status	No	116,320	10.6	11.9	10.4	0.0
	Ex	41,773	7.0	10.4	9.6	0.0
	Yes	46,474	14.5	16.4	14.0	0.0
Total		260,777	21.4	28.3	28.7	21.6

Table A.3: Prevalence missing observations in 70-year old cohort

Prevalence of missing observations in blood pressure, body mass index, alcohol status, and smoking status by medical history is listed in percentages. Associations of missingness with the medical history were significant (χ^2(df), p<.001), except for prevalence of missing smoking status across sexes ($\chi^2(1)<.01$, p=.97) and body mass index categories ($\chi^2(2)=1.55$, p=.46).

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
Sex	Female	182,873	16.3	22.5	23.9	16.0
	Male	163,537	16.9	23.5	23.2	16.0
Year of birth	1920-30	120,730	27.2	39.4	40.2	31.6
	1931-35	113,904	15.3	19.3	19.4	13.0
	1936-40	111,776	6.4	8.9	9.9	2.1
Socioeconomic status	1	38,143	15.6	21.8	22.7	14.5
	2	21,612	14.1	19.0	19.5	13.1
	3	71,863	15.7	21.6	22.4	15.2
	4	46,142	17.3	23.9	24.0	16.4
	5	19,082	18.5	25.7	26.4	17.7
	6	39,966	17.6	24.5	25.4	16.8
	7	20,291	18.1	26.4	27.1	18.7
	8	41,917	15.9	22.2	22.7	16.0
	9	19,899	16.1	22.0	22.9	15.6
	10	27,495	18.4	25.2	25.2	17.4
Acute myocardial infarction	No	327,973	17.1	23.5	24.1	16.4
	Yes	18,437	8.2	13.9	14.6	8.8
Angina	No	315,850	17.4	23.9	24.4	16.6
	Yes	30,560	7.7	13.7	15.0	9.0
Heart failure	No	339,411	15.0	23.2	23.8	16.1
	Yes	6,999	4.6	15.0	15.7	8.2
Cardiovascular system conditions	No	301,149	17.8	24.0	24.7	17.0
	Yes	45,261	8.5	16.1	16.5	9.1
Chronic kidney disease	No	337,002	16.9	23.5	24.1	16.4
	Yes	9,408	3.7	2.8	4.2	0.2
Diabetes	No	316,017	17.7	24.6	24.8	17.0
	Yes	30,393	4.8	5.8	11.4	4.8
Blood pressure	Normal	109,043	0.0	10.8	11.8	5.0
	Pre	131,434	0.0	14.9	15.8	7.8
	Hyper	48,487	0.0	21.6	22.6	13.7
Hypercholesterolaemia	No	200,545	26.1	33.8	34.0	24.9
	Yes	145,865	3.5	8.1	9.3	3.7
Coronary revascularisation	No	335,522	16.9	23.5	24.0	16.3
	Yes	10,888	6.0	8.4	9.4	5.0
Blood pressure	No	174,281	27.1	32.3	32.3	23.5

Continued on next page

Table A.3 – *Continued from previous page*

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
lowering drugs	Yes	172,129	6.0	13.6	14.8	8.3
Lipid-lowering	No	261,389	20.7	28.3	28.6	20.2
therapy	Yes	85,021	3.8	6.7	8.1	3.0
Body mass index	Normal	95,681	9.5	0.0	7.3	2.8
	Overweight	112,167	6.8	0.0	6.7	2.9
	Obese	58,936	5.0	0.0	7.4	2.9
Alcohol status	No	59,774	8.4	7.2	0.0	1.0
	Yes	204,919	7.4	6.1	0.0	1.4
Smoking status	No	163,771	8.5	10.9	10.1	0.0
	Ex	76,459	5.4	8.2	8.5	0.0
	Yes	50,881	12.6	15.2	13.4	0.0
Total		346,410	16.6	23.0	23.6	16.0

Table A.4: Prevalence missing observations in 75-year old cohort

Prevalence of missing observations in blood pressure, body mass index, alcohol status, and smoking status by medical history is listed in percentages. Associations of missingness with the medical history were significant (χ^2(df), p<.001), except for prevalence of missing blood pressure across sexes (χ^2(1)=1.76, p=.18).

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
Sex	Female	161,907	13.9	20.8	21.1	13.1
	Male	131,802	14.0	19.8	18.9	12.1
Year of birth	1920-25	79,616	24.1	36.6	34.8	26.6
	1926-30	99,425	14.8	21.7	21.4	14.2
	1931-36	114,668	6.2	8.1	8.9	1.6
Socioeconomic	1	31,068	13.1	19.7	19.7	11.5
status	2	16,677	11.7	17.3	17.0	10.0
	3	58,079	13.0	18.6	18.6	11.6
	4	37,518	14.7	21.3	20.8	13.2
	5	15,757	15.9	22.8	22.4	14.2
	6	33,450	14.6	21.3	21.6	13.2
	7	20,037	14.4	22.4	21.7	14.2
	8	40,845	12.9	19.7	19.1	12.2
	9	15,572	14.1	19.9	19.8	12.6
	10	24,706	16.6	23.3	22.8	15.5
Acute myocardial	No	274,583	14.4	21.0	20.7	13.1
infarction	Yes	19,126	8.1	11.4	11.5	6.4

Continued on next page

Table A.4 – *Continued from previous page*

Covariate	Category	Size	Blood pressure	Body mass index	Alcohol status	Smoking status
Angina	No	260,026	14.8	21.5	21.2	13.4
	Yes	33,683	7.3	11.5	11.8	6.5
Heart failure	No	283,830	11.8	20.6	20.4	12.8
	Yes	9,879	3.6	14.5	14.0	7.0
Cardiovascular system conditions	No	237,061	15.3	21.8	21.6	13.9
	Yes	56,648	8.1	14.7	14.1	7.2
Chronic kidney disease	No	277,120	14.5	21.4	21.2	13.4
	Yes	16,589	4.9	3.0	3.8	0.1
Diabetes	No	263,224	15.0	22.2	21.4	13.7
	Yes	30,485	5.1	4.4	9.3	3.2
Blood pressure	Normal	109,265	0.0	10.3	10.7	4.2
	Pre	107,002	0.0	15.1	14.9	7.1
	Hyper	36,485	0.0	21.7	20.8	12.0
Hypercholesterolaemia	No	156,153	22.8	31.6	30.4	21.1
	Yes	137,556	3.9	7.7	8.5	3.0
Coronary revascularisation	No	282,452	14.2	20.9	20.7	13.0
	Yes	11,257	6.8	6.8	7.4	3.6
Blood pressure lowering drugs	No	120,274	25.4	31.2	30.1	20.9
	Yes	173,435	6.0	12.9	13.2	6.9
Lipid-lowering therapy	No	204,433	18.1	26.6	25.8	17.1
	Yes	89,276	4.3	6.2	7.2	2.4
Body mass index	Normal	87,919	8.9	0.0	6.4	2.3
	Overweight	97,187	6.3	0.0	6.0	2.2
	Obese	48,711	4.9	0.0	6.6	2.1
Alcohol status	No	59,629	8.0	7.3	0.0	0.8
	Yes	174,901	7.0	6.3	0.0	1.1
Smoking status	No	147,447	8.0	11.1	9.6	0.0
	Ex	72,806	5.4	7.9	7.7	0.0
	Yes	36,301	12.6	15.9	13.1	0.0
Total		293,709	13.9	20.4	20.1	12.7

Appendix B

Appendix survival models for acute myocardial infarction

Table B.1: Description and coding of variables in matched age cohorts

Values were the latest reading before entering the study, which was at the 1$^{\text{st}}$ of January of the year the patient turned the cohort's age. The first category functioned as the baseline. For information on the raw data, see Table 3.2.

	Category	Coding
Medical condition	Acute myocardial infarction (AMI) diagnosis, multiple AMIs had $\geq$30 days between events	No/single/multiple
	Angina pectoris diagnosis	No/yes
	Cardiovascular system conditions, which include diagnosis of: valvular heart disease, peripheral vascular disease, cerebrovascular disease, and other cardiovascular system disorders	No/yes
	Chronic kidney disease at end stage (GFR<15)	No/yes
	Diabetes mellitus diagnosis	No/yes
	Heart failure diagnosis	No/yes
	Hypercholesterolaemia diagnosis or a total cholesterol reading >5mmol/L	No/yes
	Hypertension diagnosis	No/yes
	Ischaemic heart disease, which include diagnosis of: angina pectoris, acute myocardial infarction, and subsequent events or complications of these conditions	No/angina/single AMI and possible angina/multiple AMIs and possible angina
Treatment	ACE inhibitor prescription, which include: angiotensin-converting enzyme inhibitors and angiotensin-II receptor antagonists	No/yes
	Aspirin prescription	No/yes

Continued on next page

Table B.1 – *Continued from previous page*

	Category	Coding
	Beta-adrenoceptor blocking drugs prescription	No/yes
	Calcium-channel blocker prescription	No/yes
	Coronary revascularisation, which include coronary artery bypass graft (CABG) and percutaneous coronary intervention (PCI)	No/yes
	Statin prescription, which include: atorvastatin, cerivastatin, fluvastatin, pravastatin, rosuvastatin, and simvastatin	No/yes
Lifestyle	Alcohol consumption status	No/yes
	Body mass index (weight in kg)/(height in m)^2	Normal weight ($<$25)/overweight (25-30)/obese ($\geq$30)
	Smoking status	No/ex/yes
Demography	Sex	Female/male
	Socioeconomic status measured by Mosaic	10 categories, see Table 3.3
	Year of birth category	1920-25, 1926-29, 1930-35, 1936-40
District	Air pollution, which includes separate variables for 2001 estimated level of nitrogen dioxide, nitrogen oxide, sulphur dioxide, and particulate matter	Quintiles
	Ethnicity, which includes separate variables for proportion of district population defining themselves as white, mixed, Asian or Asian British, black or black British, and other	Quintiles
	Index of multiple deprivation (IMD)	Quintiles
	Limiting long-term illness	Quintiles
	Urbanisation	Urban/town and fringe/village, hamlet and isolated dwelling

Table B.2: Prevalence antiplatelet therapy in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. The prevalence of the treatment by the cohort's age was affected by calendar year, see Figure 5.2.

Cohort	Dual antiplatelet therapy	Aspirin only	Other antiplatelet agent only
Age 60	122 (1%)	2,213 (13%)	119 (1%)
Age 65	1,079 (3%)	10,152 (23%)	387 (1%)
Age 70	4,097 (6%)	22,639 (31%)	802 (1%)
Age 75	5,565 (7%)	28,451 (37%)	1,009 (1%)

Table B.3: Characteristics of patients with complete and incomplete medical records in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Patients with incomplete medical records had a missing observation for alcohol consumption status, body mass index, or smoking status.

	Records	Year of birth	Size	Annual death rate (/1,000)	AMI (%)
Age 60	Complete	1936-40	6,901	14.9	1,869 (27.1%)
		1930-35	2,374	19.1	674 (28.4%)
	Incomplete	1936-40	3,691	14.3	779 (21.1%)
		1930-35	3,778	18.1	864 (22.9%)
	Total		16,744	16.4	4,186 (25.0%)
Age 65	Complete	1936-40	16,357	17.9	4,607 (28.2%)
		1931-35	9,568	23.0	2,595 (27.1%)
		1925-30	2,946	29.7	810 (27.5%)
	Incomplete	1936-40	5,339	15.8	817 (15.3%)
		1931-35	4,608	22.8	949 (20.6%)
		1925-30	4,710	30.2	1,104 (23.4%)
	Total		43,528	22.5	10,882 (25.0%)
Age 70	Complete	1936-40	20,790	21.6	5,631 (27.1%)
		1931-35	19,645	27.1	5,486 (27.9%)
		1920-30	13,718	38.7	3,760 (27.4%)
	Incomplete	1936-40	2,774	19.4	260 (9.4%)
		1931-35	6,143	27.1	961 (15.6%)
		1920-30	10,658	41.2	2,334 (21.9%)
	Total		73,728	32.5	18,432 (25.0%)
Age 75	Complete	1931-36	27,648	32.7	7,393 (26.7%)
		1926-30	20,266	44.8	5,624 (27.8%)
		1920-25	11,035	59.0	3,016 (27.3%)
	Incomplete	1931-36	3,240	32.1	329 (10.2%)
		1926-30	6,854	46.5	1,156 (16.9%)
		1920-25	7,349	58.0	1,580 (21.5%)
	Total		76,392	47.6	19,098 (25.0%)

Table B.4: Distribution of recorded and imputed values of variables with missing data in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Patients with incomplete medical records had a missing observation for alcohol consumption status, body mass index, or smoking status. The distribution of imputed values is the mean across ten imputed datasets.

	Values	Alcohol	BMI (sd)	Ex-smoker	Smoker
Age 60	Recorded	83.8%	26.8 (4.2)	19.7%	29.8%
	Imputed	84.0%	26.3 (4.3)	19.4%	30.0%
Age 65	Recorded	81.8%	27.0 (4.3)	26.9%	24.6%
	Imputed	79.8%	26.4 (4.3)	24.8%	24.2%
Age 70	Recorded	79.9%	27.1 (4.2)	33.0%	18.8%
	Imputed	78.7%	26.2 (4.3)	32.3%	17.4%
Age 75	Recorded	76.7%	26.8 (4.4)	35.1%	14.7%
	Imputed	75.0%	25.9 (4.3)	34.1%	13.2%

Table B.5: Characteristics of patients lost to follow-up in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Patients who transferred out of their practice during the study were lost to follow-up. The prevalence of smoker is the mean across ten imputed datasets. Affluent area was measured as the first quintile of the Index of Multiple Deprivation (IMD), where the prevalence excluded missing observations.

	Transferred, observed age	Size	AMI	Diabetes	Smoker	Affluent area
Age 60	No, <70	1,882	40.0%	13.4%	43.7%	13.5%
	No, ≥70	10,927	23.2%	5.2%	27.1%	21.7%
	Yes, <70	2,967	23.0%	6.5%	31.6%	20.8%
	Yes, ≥70	968	22.6%	5.8%	28.7%	19.0%
Age 65	No, <70	3,077	37.6%	17.0%	38.9%	14.4%
	No, ≥70	31,868	24.3%	8.4%	22.6%	22.1%
	Yes, <70	4,511	23.1%	9.5%	27.0%	20.7%
	Yes, ≥70	4,072	23.2%	7.4%	25.7%	21.5%
Age 70	No	62,254	25.3%	12.4%	18.0%	22.3%
	Yes	11,474	23.5%	10.8%	22.2%	20.3%
Age 75	No	64,688	25.3%	14.3%	14.1%	22.7%
	Yes	11,704	23.4%	11.7%	18.0%	20.8%

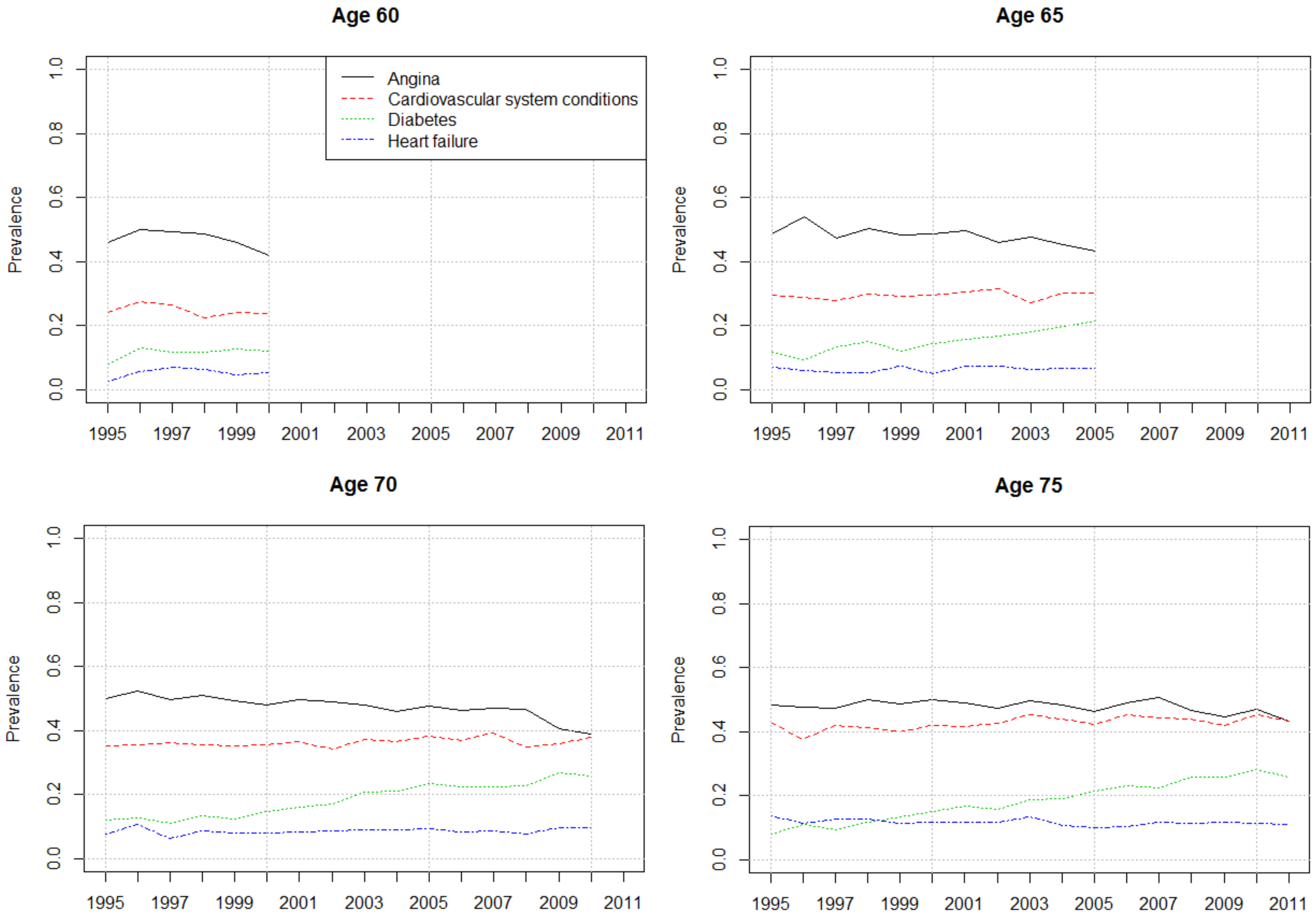


Figure B.1: Prevalence of comorbidites by cohort's age in patients with acute myocardial infarction

The age cohorts differed in recruitment period. The prevalence prior to 1995 is not presented due to the small numbers of medical records available.

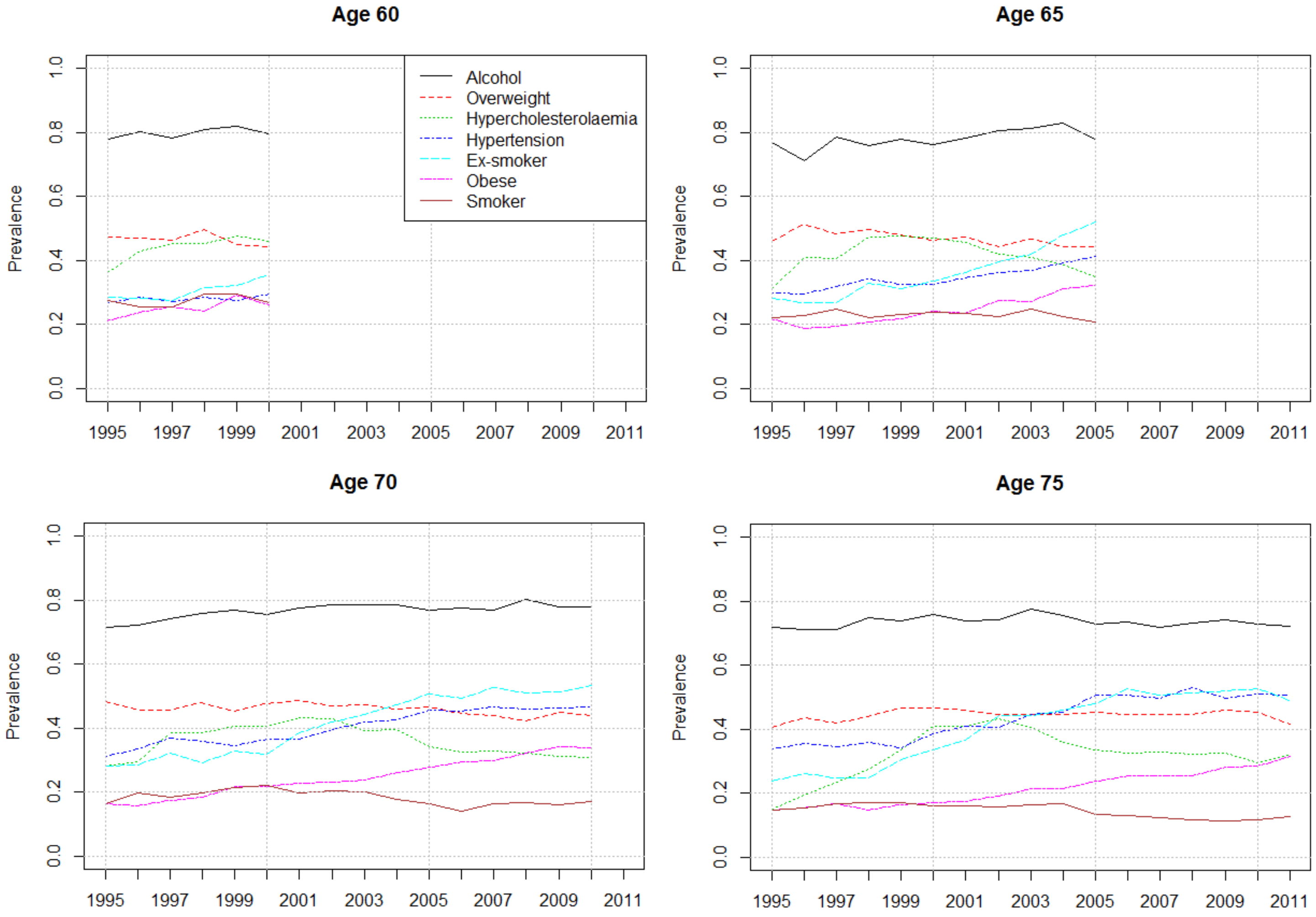


Figure B.2: Prevalence of lifestyle factors by cohort's age in patients with acute myocardial infarction

The age cohorts differed in recruitment period. The prevalence prior to 1995 is not presented due to the small numbers of medical records available.

Table B.6: Prevalence coronary revascularisation given ischaemic heart disease (IHD)

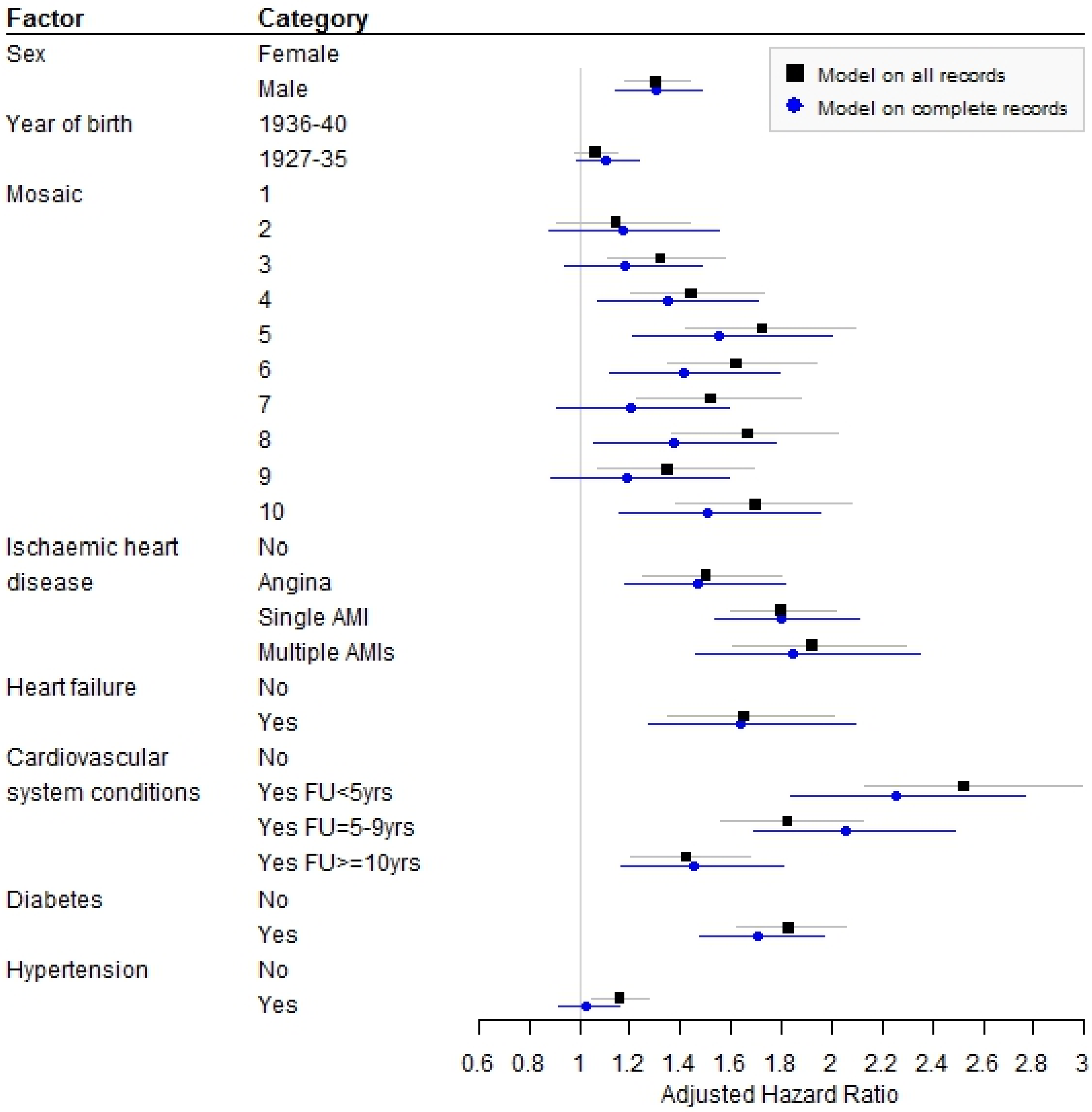
The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. History of IHD consisted of AMI and angina. Patients of both subtypes of IHD could be eligible for coronary revascularisation, which consisted of coronary artery bypass graft (CABG) and percutaneous coronary intervention (PCI). Some IHD patients had both CABG and PCI. The prevalence of the treatment by the cohort's age was affected by calendar year, see Figure 5.2.

Cohort	Coronary revascularisation	Number of patients (%)
Age 60	CABG	751 (16%)
	PCI	167 (3%)
	Total	881 (18%)
Age 65	CABG	2,479 (19%)
	PCI	750 (6%)
	Total	3,069 (23%)
Age 70	CABG	4,606 (19%)
	PCI	1,869 (8%)
	Total	6,113 (26%)
Age 75	CABG	5,036 (19%)
	PCI	1,958 (7%)
	Total	6,601 (25%)

Table B.7: Prevalence of diabetes in men and women with ischaemic heart disease (IHD)

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. History of IHD consisted of AMI and angina. Patients of both subtypes of IHD could be eligible for coronary revascularisation.

Coronary revascularisation	Sex	Age 60	Age 65	Age 70	Age 75
No	Men	315 (10%)	1,093 (14%)	2,209 (17%)	2,518 (19%)
	Women	98 (12%)	375 (15%)	864 (18%)	1,171 (18%)
Yes	Men	86 (11%)	430 (16%)	1,081 (21%)	1,219 (23%)
	Women	24 (23%)	80 (19%)	217 (21%)	306 (24%)


Factor
Category
Sex
Female
Male
Year of birth
1936-40
1927-35
Mosaic
1
2
3
4
5
6
7
8
9
10
Ischaemic heart disease
No
Angina
Single AMI
Multiple AMIs
Heart failure
No
Yes
Cardiovascular system conditions
No
Yes FU<5yrs
Yes FU=5-9yrs
Yes FU>=10yrs
Diabetes
No
Yes
Hypertension
No
Yes
■ Model on all records
● Model on complete records
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
Adjusted Hazard Ratio

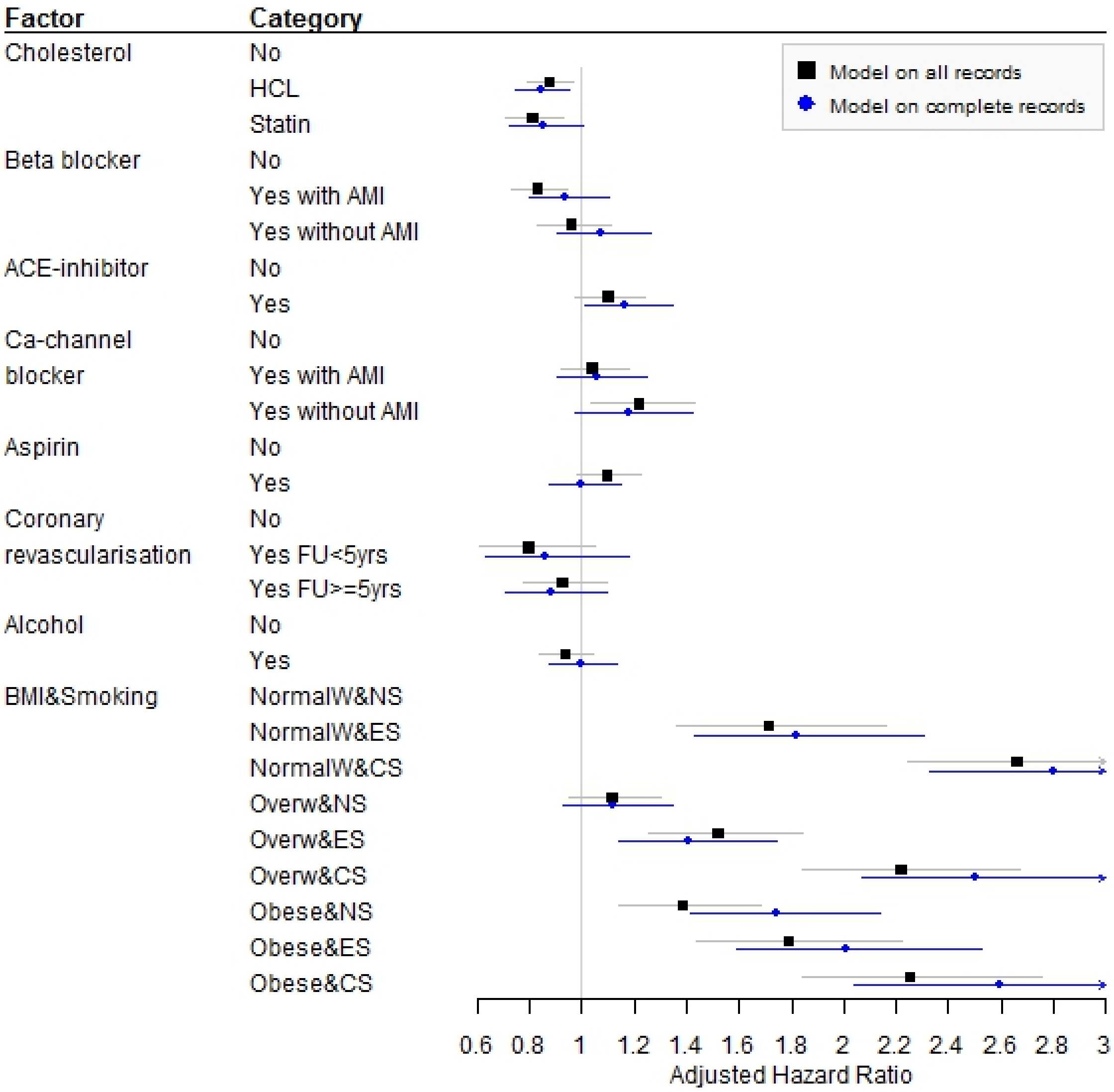
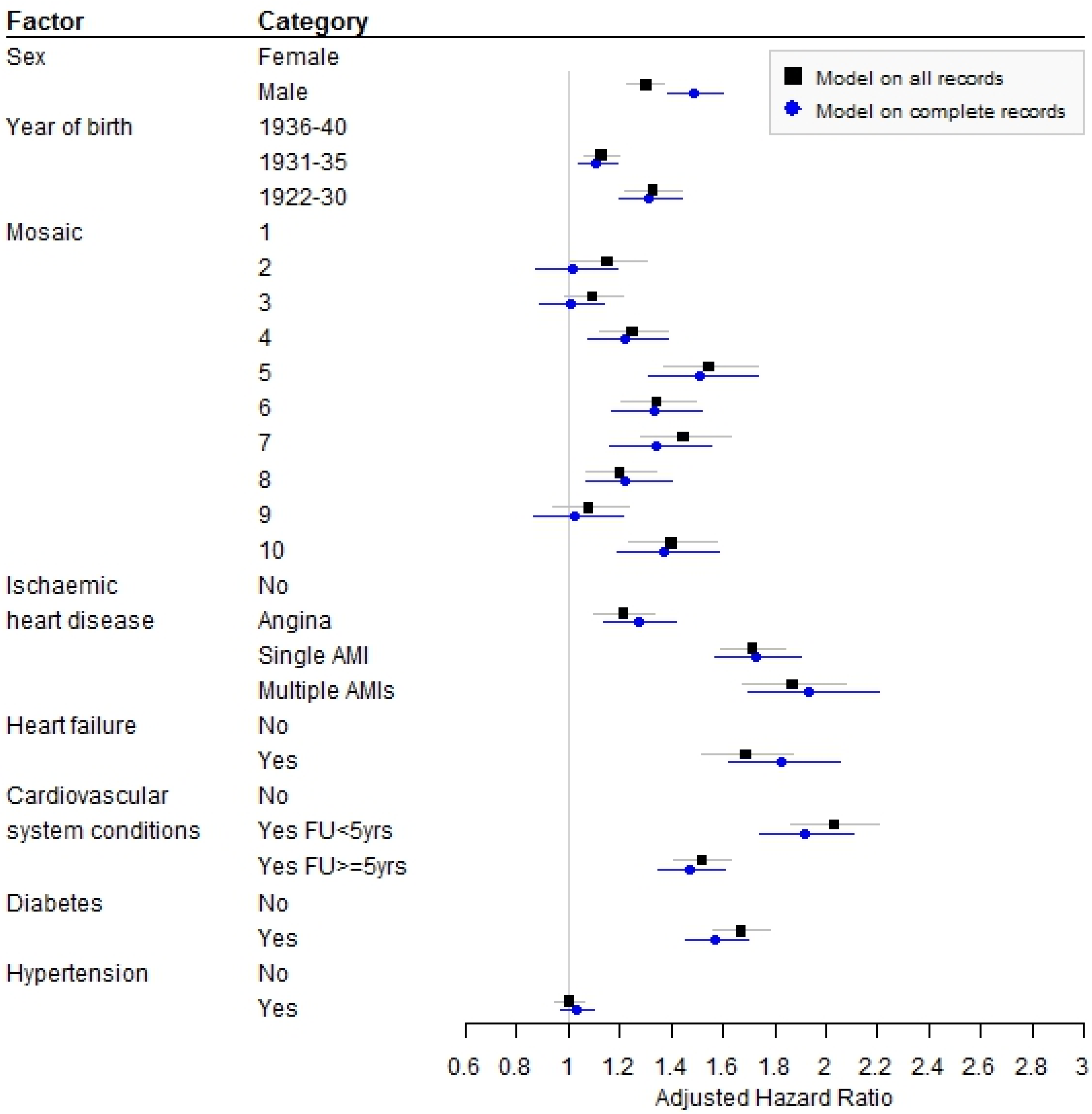


Figure B.3: Survival model for 60-year old matched cohort

The hazard of all-cause mortality were adjusted for the listed risk factors and general practice. The hazard associated with single/multiple acute myocardial infarction (AMI) was the same in patients with and without angina, and the hazard associated with statin prescription was the same in patients with and without hypercholesterolaemia (HCL). Time-varying hazards were split at five or ten years of follow-up (FU) after the cohort's age. Abbreviations: NormalW=normal weight, Overw=overweight, NS=non-smoker, ES=ex-smoker, and CS=current-smoker.


Factor
Category
Sex
Female
Male
Year of birth
1936-40
1931-35
1922-30
Mosaic
1
2
3
4
5
6
7
8
9
10
Ischaemic heart disease
No
Angina
Single AMI
Multiple AMIs
Heart failure
No
Yes
Cardiovascular system conditions
No
Yes FU<5yrs
Yes FU>=5yrs
Diabetes
No
Yes
Hypertension
No
Yes
■ Model on all records
● Model on complete records
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
Adjusted Hazard Ratio

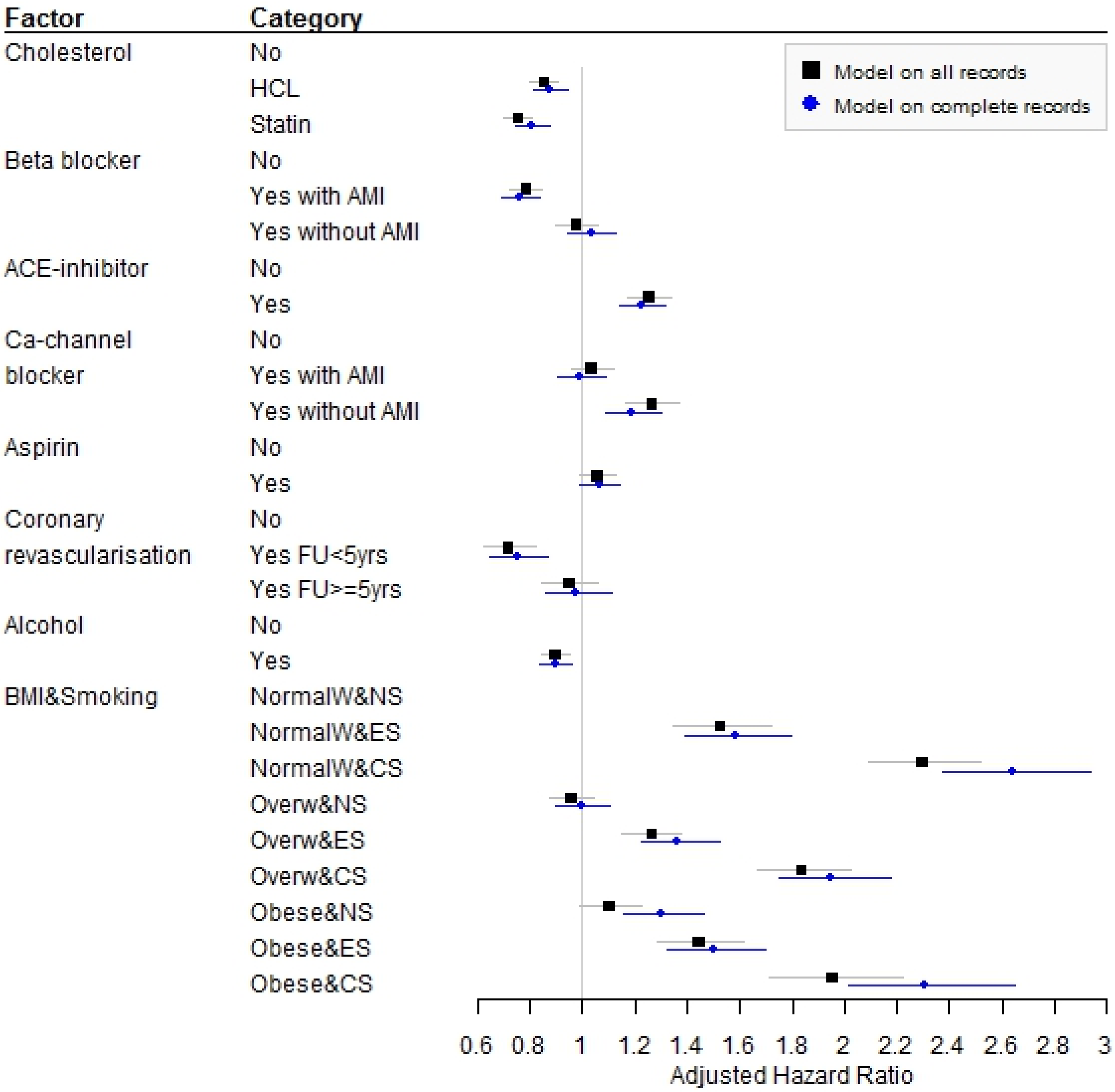
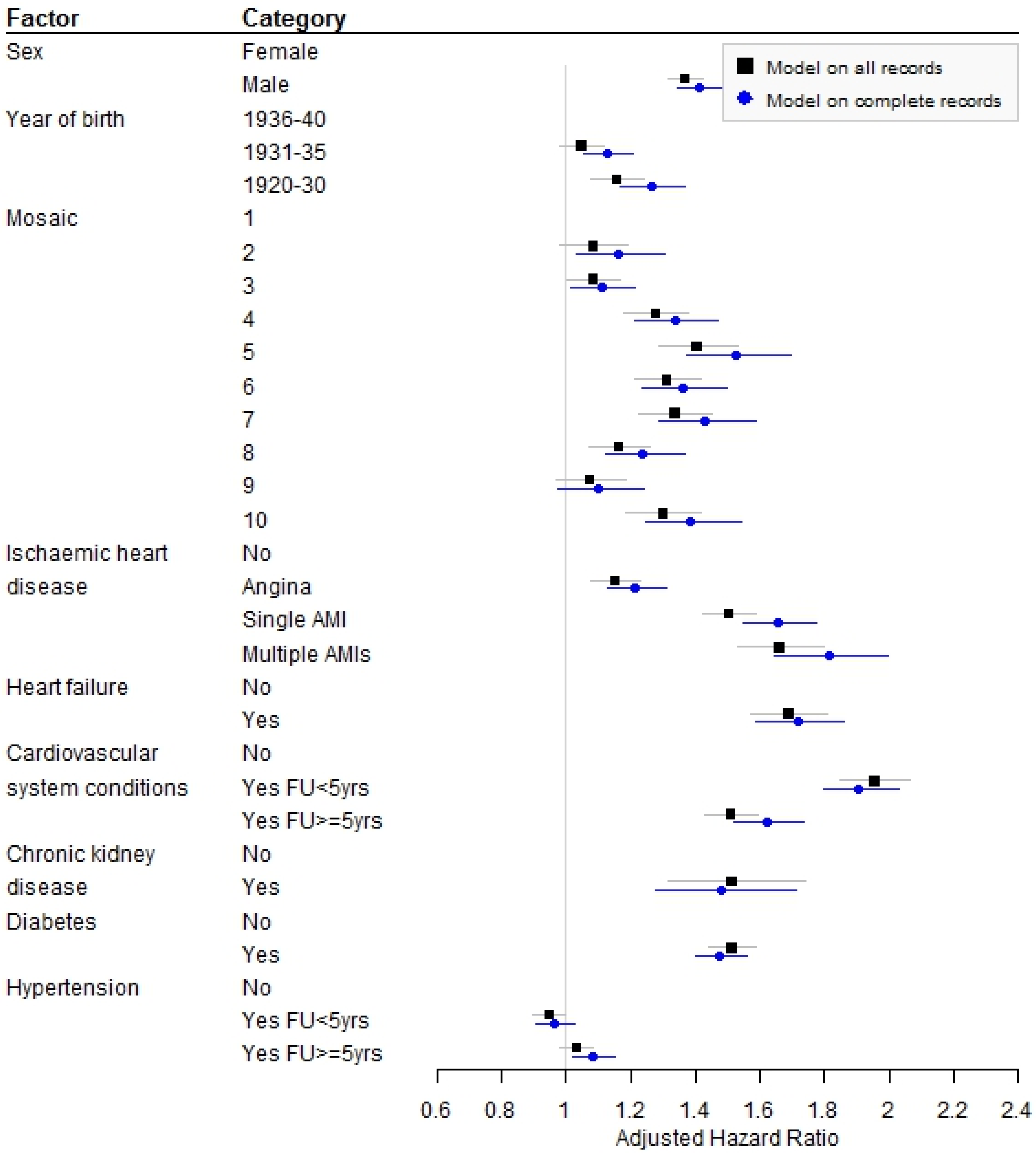


Figure B.4: Survival model for 65-year old matched cohort

The hazard of all-cause mortality were adjusted for the listed risk factors and general practice. The hazard associated with single/multiple acute myocardial infarction (AMI) was the same in patients with and without angina, and the hazard associated with statin prescription was the same in patients with and without hypercholesterolaemia (HCL). Time-varying hazards were split at five years of follow-up (FU) after the cohort's age. Abbreviations: NormalW=normal weight, Overw=overweight, NS=non-smoker, ES=ex-smoker, and CS=current-smoker.


Factor
Category
Sex
Female
Male
Year of birth
1936-40
1931-35
1920-30
Mosaic
1
2
3
4
5
6
7
8
9
10
Ischaemic heart disease
No
Angina
Single AMI
Multiple AMIs
Heart failure
No
Yes
Cardiovascular system conditions
No
Yes FU<5yrs
Yes FU>=5yrs
Chronic kidney disease
No
Yes
Diabetes
No
Yes
Hypertension
No
Yes FU<5yrs
Yes FU>=5yrs
■ Model on all records
● Model on complete records
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
Adjusted Hazard Ratio

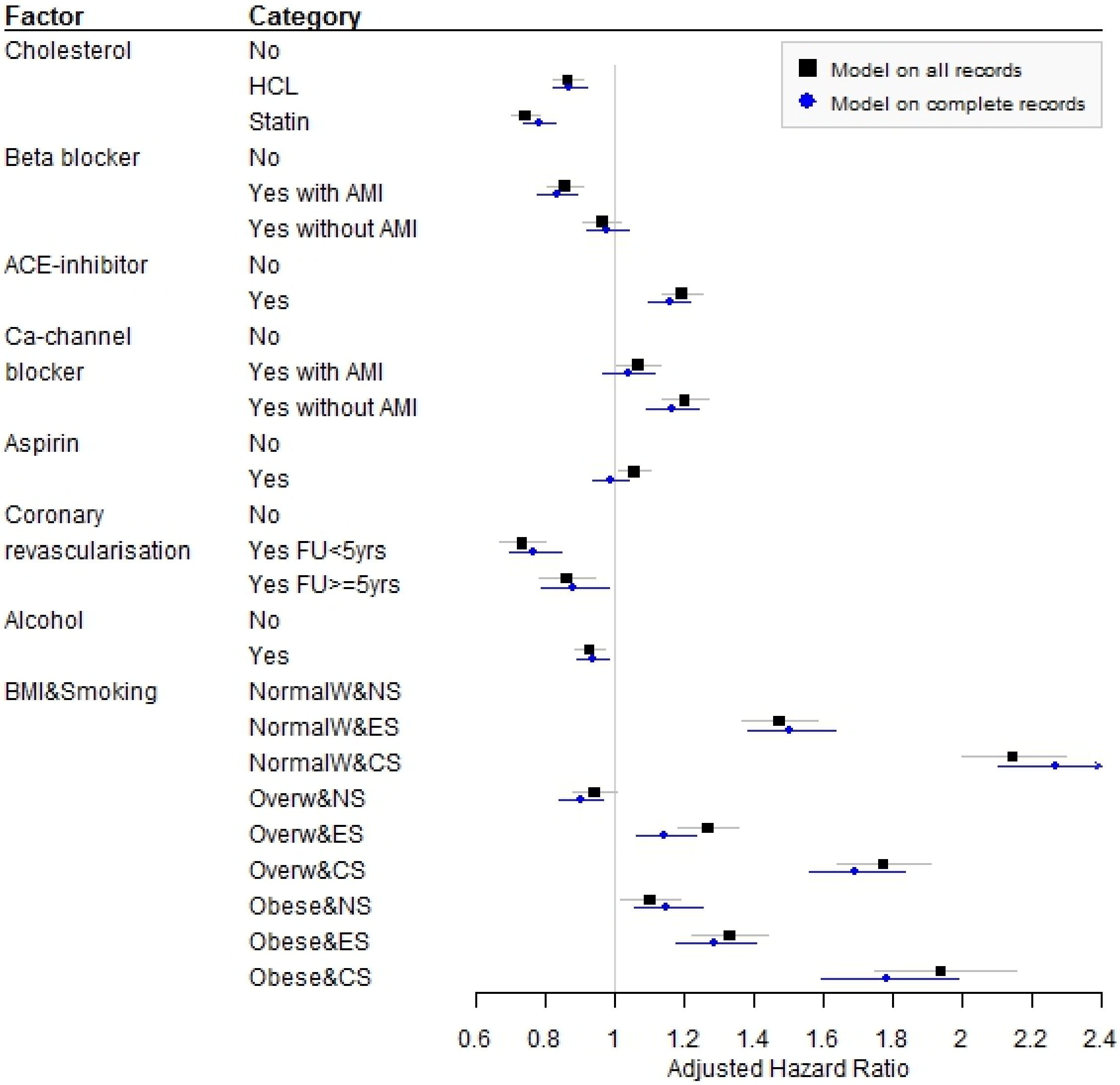
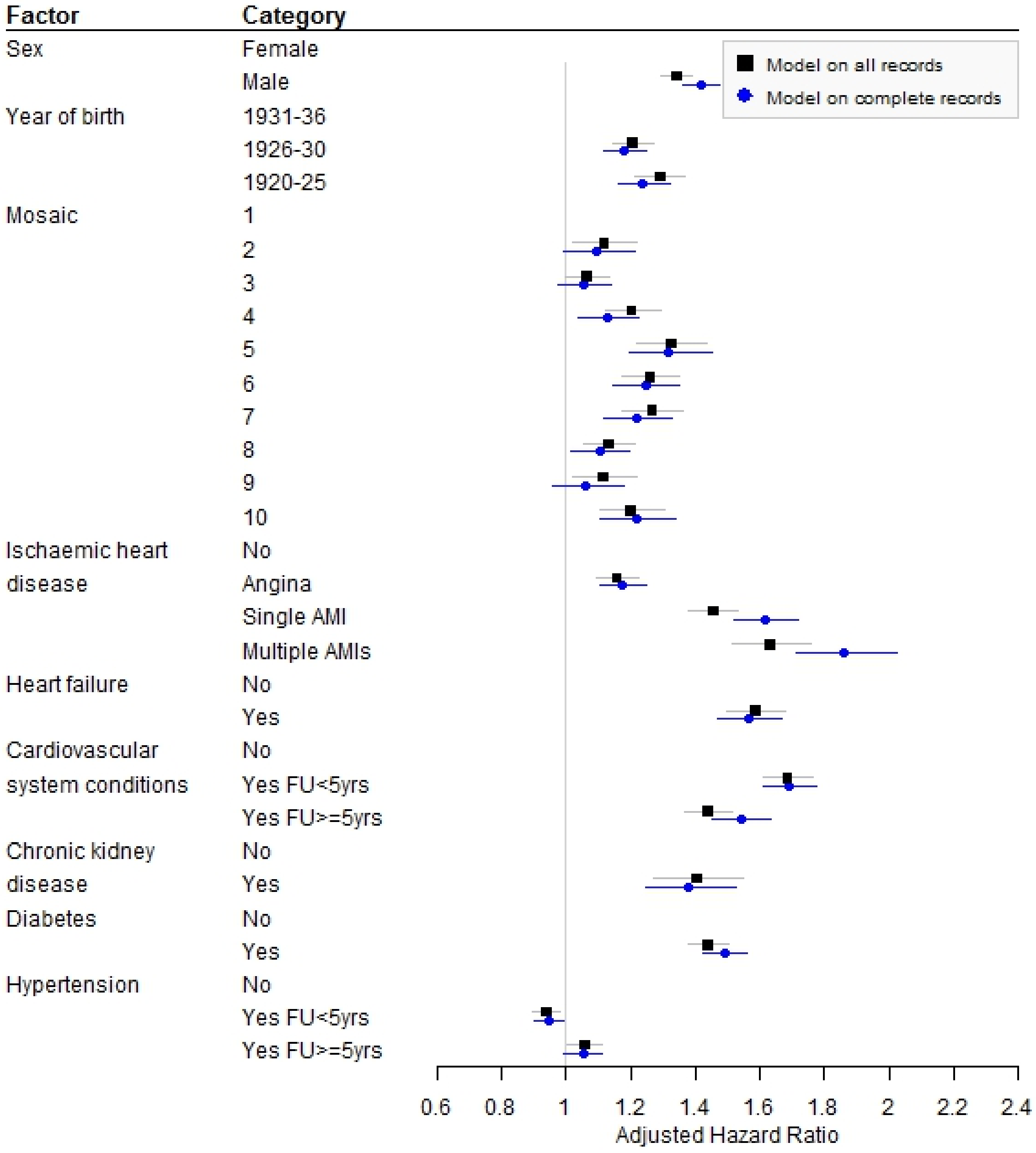


Figure B.5: Survival model for 70-year old matched cohort

The hazard of all-cause mortality were adjusted for the listed risk factors and general practice. The hazard associated with single/multiple acute myocardial infarction (AMI) was the same in patients with and without angina, and the hazard associated with statin prescription was the same in patients with and without hypercholesterolaemia (HCL). Time-varying hazards were split at five years of follow-up (FU) after the cohort's age. Abbreviations: NormalW=normal weight, Overw=overweight, NS=non-smoker, ES=ex-smoker, and CS=current-smoker.


Factor
Category
Sex
Female
Male
Year of birth
1931-36
1926-30
1920-25
Mosaic
1
2
3
4
5
6
7
8
9
10
Ischaemic heart disease
No
Angina
Single AMI
Multiple AMIs
Heart failure
No
Yes
Cardiovascular system conditions
No
Yes FU<5yrs
Yes FU>=5yrs
Chronic kidney disease
No
Yes
Diabetes
No
Yes
Hypertension
No
Yes FU<5yrs
Yes FU>=5yrs
■ Model on all records
● Model on complete records
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
Adjusted Hazard Ratio

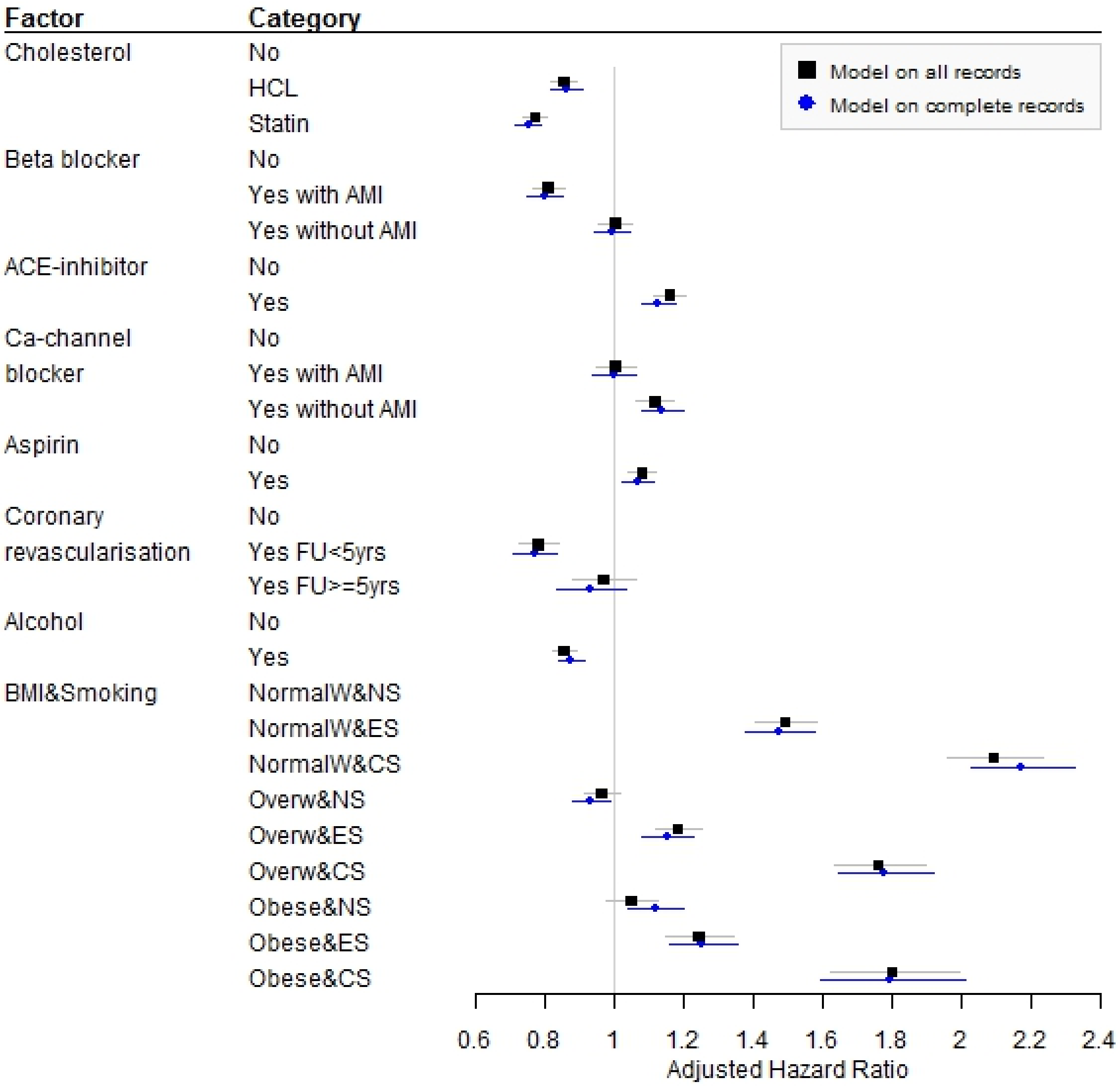


Figure B.6: Survival model for 75-year old matched cohort

The hazard of all-cause mortality were adjusted for the listed risk factors and general practice. The hazard associated with single/multiple acute myocardial infarction (AMI) was the same in patients with and without angina, and the hazard associated with statin prescription was the same in patients with and without hypercholesterolaemia (HCL). Time-varying hazards were split at five years of follow-up (FU) after the cohort's age. Abbreviations: NormalW=normal weight, Overw=overweight, NS=non-smoker, ES=ex-smoker, and CS=current-smoker.

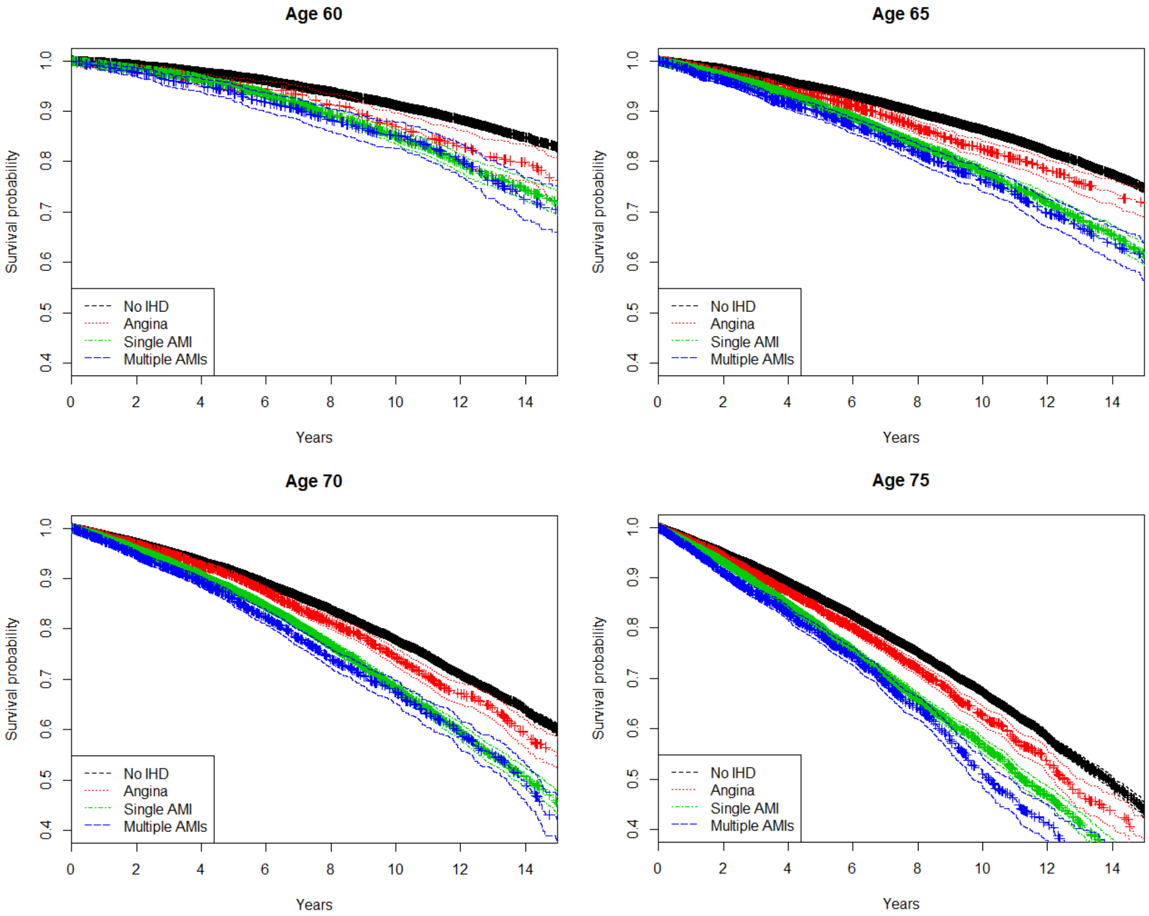


Figure B.7: Adjusted survival curves associated with ischaemic heart disease (IHD)

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Due to the way AMI and angina interacted, a factor representing IHD was generated that had the following levels: no history, angina only, single AMI with possibly angina, or multiple AMIs with possibly angina. The survival curves were adjusted for sex, year of birth, socioeconomic status, heart failure, cardiovascular system conditions, chronic kidney disease (only at ages 70 and 75), diabetes, hypertension, hypercholesterolaemia, coronary revascularisation, ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, statins, alcohol consumption status, body mass index, smoking status, and general practice.

Table B.8: Correlations of district's characteristics and the adjusted hazards of all-cause mortality associated with general practices

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. The hazards were adjusted for sex, year of birth, socioeconomic status, ischaemic heart disease, heart failure, cardiovascular system conditions, chronic kidney disease (only at ages 70 and 75), diabetes, hypertension, hypercholesterolaemia, coronary revascularisation, ACE inhibitors, aspirin, beta blockers, calcium-channel blockers, statins, alcohol consumption status, body mass index, and smoking status. The table reports the Spearman's rank correlations r, where r=0 stands for no correlation and $|r|$=1 for perfect correspondence between two variables (Harrell et al., 1996).

Characteristic	Category	Age 60	Age 65	Age 70	Age 75
Deprivation (IMD)		0.03	0.02	0.10	-0.05
Urbanisation		-0.08	0.03	0.04	0.05
Limiting long-term illness		0.03	0.03	0.10	-0.08
Ethnicity	White	-0.03	0.01	0.01	0.05
	Mixed	0.00	0.00	0.01	-0.03
	Asian	0.01	-0.01	0.00	-0.05
	Black	0.03	0.01	0.01	-0.07
	Other	0.02	0.02	0.00	-0.04
Air pollution	Nitrogen dioxide	0.00	0.06	-0.01	-0.05
	Nitrogen oxide	0.03	0.07	0.00	-0.08
	Sulphur dioxide	0.08	0.08	-0.01	-0.03
	Particulate matter	0.00	0.06	0.00	-0.05

Table B.9: Performance statistics of survival models based on complete medical records and irrespective of completeness in matched age cohorts

The age cohorts included cases with history of acute myocardial infarction (AMI) who were matched to three controls on sex, year of birth, and general practice. Matching was carried out once on complete medical records and once irrespective of completeness of records. Patients with incomplete medical records had a missing observation for alcohol consumption status, body mass index, or smoking status.

Model based on	Statistic	Age 60	Age 65	Age 70	Age 75
Complete records	R^2	0.300	0.290	0.254	0.231
	C	0.703	0.702	0.694	0.685
	Shrinkage	0.958	0.986	0.989	0.987
All records	R^2	0.288	0.264	0.222	0.199
	C	0.702	0.698	0.683	0.675
	Shrinkage	0.974	0.988	0.992	0.993

Appendix C

Appendix survival models for statin prescription

Table C.1: Characteristics of patients with and without Townsend deprivation score in age cohorts without cardiovascular disease

Townsend scores were based on 2001 Census Data of England and Wales (UKDS, 2006). Consequently, patients living in Scotland or Northern Ireland had a missing record for this deprivation measure. Less than 0.2% of patients living in England or Wales had a missing score.

	Townsend score	Year of birth	Size	Death/1,000 person-years	Men (%)	Lipid-lowering therapy (%)
Age 60	Recorded	1936-40	75,379	9.2	36,217 (48.0%)	2,229 (3.0%)
		1930-35	43,321	12.5	20,768 (47.9%)	520 (1.2%)
	Missing	1936-40	7,727	10.2	3,566 (46.1%)	306 (4.0%)
		1930-35	3,409	14.8	1,607 (47.1%)	26 (0.8%)
	Total		129,836	10.8	62,158 (47.9%)	3,081 (2.4%)
Age 65	Recorded	1936-40	95,322	11.0	44,846 (47.0%)	11,229 (11.8%)
		1931-35	64,932	14.8	30,154 (46.4%)	2,569 (4.0%)
		1925-30	39,320	21.0	17,941 (45.6%)	500 (1.3%)
	Missing	1936-40	16,603	12.2	7,543 (45.4%)	1,959 (11.8%)
		1931-35	6,660	16.4	2,950 (44.3%)	326 (4.9%)
		1925-30	2,847	23.8	1,230 (43.2%)	30 (1.1%)
	Total		225,684	15.2	104,664 (46.4%)	16,613 (7.4%)
Age 70	Recorded	1936-40	77,025	12.8	34,514 (44.8%)	25,326 (32.9%)
		1931-35	78,278	17.7	34,782 (44.4%)	11,373 (14.5%)
		1920-30	91,846	28.1	39,498 (43.0%)	2,525 (2.7%)
	Missing	1936-40	13,513	13.5	5,801 (42.9%)	4,615 (34.2%)
		1931-35	13,361	19.7	5,669 (42.4%)	1,879 (14.1%)
		1920-30	7,850	29.2	3,221 (41.0%)	251 (3.2%)
	Total		281,873	22.7	123,485 (43.8%)	45,969 (16.3%)
Age 75	Recorded	1931-36	73,154	20.7	30,787 (42.1%)	27,689 (37.9%)
		1926-30	63,971	30.1	25,997 (40.6%)	8,910 (13.9%)
		1920-25	56,960	41.0	22,867 (40.1%)	1,280 (2.2%)
	Missing	1931-36	12,333	21.1	4,920 (39.9%)	4,883 (39.6%)
		1926-30	10,342	32.8	4,052 (39.2%)	1,391 (13.5%)
		1920-25	4,885	42.7	1,816 (37.2%)	114 (2.3%)
	Total		221,645	33.3	90,439 (40.8%)	44,267 (20.0%)

Table C.2: Original and modified QRISK2 algorithm

The QRISK2 estimates the risk of developing a first cardiovascular event in the next ten years based on a person's information on multiple demographic, medical, and lifestyle factors (Hippisley-Cox et al., 2008)

Variables in original QRISK2	Variables in modified QRISK2
Self-assigned ethnicity (white/not recorded, Indian, Pakistani, Bangladeshi, other Asian, black African, black Caribbean, Chinese, other including mixed)	No: excluded
Age (years)	Yes
Sex (males versus females)	Yes
Smoking status (no/ex/light/moderate/heavy)	No: classified smokers as moderate smoker
Systolic blood pressure (continuous)	Yes
Ratio of total serum cholesterol/high density lipoprotein cholesterol (continuous)	No: substituted based on hypercholesterolaemia (HCL) diagnosis. Patients with no HCL diagnosis, were ascribed a value of 4. Patients with a HCL diagnosis, were ascribed a value of 5.
Body mass index (continuous)	Yes
Family history of ischaemic heart disease in first degree relative under 60 years (no/yes)	No: substituted by family history of cardiovascular disease
Townsend deprivation score (continuous)	No: used corresponding median value for quintiles: -3.15, -2.17, -1.05, 0.84, and 4.51
Treated hypertension (diagnosis of hypertension and at least one current prescription of at least one antihypertensive agent)	Yes
Rheumatoid arthritis (no/yes)	No: excluded
Chronic renal disease (no/yes)	Yes
Diabetes (no/type1/type2)	No: classified diabetes as type two
Atrial fibrillation (no/yes)	No: excluded

Table C.3: Description and coding of variables in age cohorts without cardiovascular disease

Values were the latest reading before entering the study, which was at the 1st of January of the year the patient turned the cohort's age. The first category functioned as the baseline. For information on the raw data, see Table 3.2.

Category	Coding
Body mass index (weight in kg)/(height in m)^2	Normal weight (<25)/ overweight (25-30)/ obese (≥30)
Blood pressure regulating drugs prescription includes: beta-adrenoceptor blocking drugs, thiazides and related diuretics, adrenergic neurone blocking drugs, alpha-adrenoceptor blocking drugs, angiotensin-converting enzyme inhibitors, angiotensin-II receptor antagonists, centrally acting antihypertensive drugs, drugs affecting the renin-angiotensin system, drugs related to hypertension and heart failure, renin inhibitors, vasodilator antihypertensive drugs, and calcium-channel blockers	No/yes
Chronic kidney disease (CKD) at end stage (GFR<15)	No/yes
Diabetes mellitus diagnosis	No/yes
Hypercholesterolaemia (HCL) diagnosis or a total cholesterol reading >5mmol/L	No/yes
Hypertension diagnosis	No/yes
Lipid-lowering therapy prescription includes a type of statin or one of the following: colesevelam, colestipol, colestyramine, ezetimibe, bezafibrate, ciprofibrate, clofibrate, fenofibrate, gemfibrozil, acipimox, nicotinic acid, and omega-3-triglycerides including other esters and acids	No/yes
Sex	Female/male
Smoking status	No/ex/yes
Socioeconomic status measured by Mosaic	10 categories, see Table 3.3
Statin prescription, which include: atorvastatin, cerivastatin, fluvastatin, pravastatin, rosuvastatin, and simvastatin	No/yes
Year of birth category	1920-25/1926-29/ 1930-35/1936-40

Table C.4: Characteristics of patients with complete and incomplete medical records in age cohorts without cardiovascular disease

Patients with incomplete medical records had a missing observation for systolic blood pressure, body mass index, or smoking status, and consequently a missing QRISK2 score. The reported QRISK2 score is the mean (standard deviation) 10-year risk of a first cardiovascular event across ten imputed datasets.

	Records	Year of	Size	Death/1,000	Lipid lowering	QRISK2
		birth		person-years	therapy	(sd)
Age 60	Complete	1936-40	45,807	8.9	1,734 (3.8%)	10.6 (5.3)
		1930-35	18,104	11.7	324 (1.8%)	10.4 (5.0)
	Incomplete	1936-40	29,572	9.8	495 (1.7%)	10.4 (3.7)
		1930-35	25,217	13	196 (0.8%)	10.2 (3.8)
	Total		118,700	10.6	2,749 (2.3%)	10.4 (4.6)
Age 65	Complete	1936-40	66,764	10.7	9,884 (14.8%)	15.9 (6.7)
		1931-35	40,060	14.1	1,986 (5.0%)	15.2 (6.3)
		1925-30	15,998	20	299 (1.9%)	15.1 (6.1)
	Incomplete	1936-40	28,558	11.6	1,345 (4.7%)	14.9 (4.7)
		1931-35	24,872	16	583 (2.3%)	14.6 (4.7)
		1925-30	23,322	21.6	201 (0.9%)	14.5 (4.7)
	Total		199,574	15.2	14,298 (7.2%)	15.2 (5.9)
Age 70	Complete	1936-40	66,733	12.7	24,235 (36.3%)	22.5 (7.7)
		1931-35	55,423	17	9,970 (18.0%)	22.0 (7.4)
		1920-30	46,111	26.3	1,849 (4.0%)	21.2 (7.1)
	Incomplete	1936-40	10,292	13.6	1,091 (10.6%)	19.9 (5.4)
		1931-35	22,855	19.3	1,403 (6.1%)	20.1 (5.5)
		1920-30	45,735	29.8	676 (1.5%)	19.7 (5.3)
	Total		247,149	22.8	39,224 (15.9%)	21.3 (7.0)
Age 75	Complete	1931-36	64,358	20.4	26,585 (41.3%)	30.4 (8.2)
		1926-30	43,947	29.1	7,761 (17.7%)	29.7 (7.9)
		1920-25	29,948	40.3	921 (3.1%)	28.6 (7.6)
	Incomplete	1931-36	8,796	22.7	1,104 (12.6%)	26.5 (5.8)
		1926-30	20,024	32.5	1,149 (5.7%)	26.8 (5.7)
		1920-25	27,012	41.8	359 (1.3%)	26.6 (5.5)
	Total		194,085	33.4	37,879 (19.5%)	28.9 (7.5)

Table C.5: Distribution of recorded and imputed values of variables with missing data in age cohorts without cardiovascular disease

Patients with incomplete medical records had a missing observation for systolic blood pressure, body mass index, or smoking status, and consequently a missing QRISK2 score. The reported distributions of imputed values are the mean across ten imputed datasets.

		Ex-smoker	Smoker	BMI (sd)	SBP (sd)	QRISK2 (sd)
Age 60	Recorded	12.9%	25.5%	26.4 (4.4)	139.1 (18.4)	10.5 (5.2)
	Imputed	13.0%	28.1%	26.2 (4.4)	136.4 (17.7)	10.3 (4.9)
Age 65	Recorded	18.3%	21.5%	26.6 (4.5)	141.6 (17.9)	15.6 (6.5)
	Imputed	19.9%	23.3%	26.2 (4.4)	138.8 (17.6)	14.7 (6.3)
Age 70	Recorded	23.4%	16.5%	26.7 (4.6)	142.4 (17.5)	22.0 (7.5)
	Imputed	24.4%	18.0%	26.0 (4.5)	142.0 (17.5)	19.8 (7.2)
Age 75	Recorded	24.8%	13.4%	26.5 (4.6)	143.7 (17.7)	29.8 (8.0)
	Imputed	26.5%	15.0%	25.9 (4.5)	143.1 (17.7)	26.7 (7.9)

Table C.6: Characteristics of patients lost to follow-up in age cohorts without cardiovascular disease

Patients who transferred out of their practice during the study were lost to follow-up. The reported mean (standard deviation) of QRISK2 and the prevalence of smokers are the mean across ten imputed datasets. Affluent area was measured as the first quintile of the Index of Multiple Deprivation (IMD), where the prevalence excluded missing observations.

	Transferred, observed age	Size	QRISK2 (sd)	Lipid lowering therapy	Diabetes	Smoker	Affluent area
Age 60	No, <70	8,108	12.8 (5.4)	2.8%	6.8%	45.5%	21.0%
	No, ≥70	80,756	10.2 (4.5)	2.3%	3.2%	24.1%	29.0%
	Yes, <70	22,732	10.5 (4.6)	2.3%	3.6%	27.6%	27.6%
	Yes, ≥70	7,104	10.3 (4.3)	1.5%	3.0%	24.6%	26.4%
Age 65	No, <70	8,437	18.0 (6.7)	8.2%	9.9%	39.4%	22.4%
	No, ≥70	149,102	15.1 (5.8)	7.5%	5.4%	20.7%	30.6%
	Yes, <70	21,886	15.6 (6.0)	7.3%	6.0%	24.4%	27.7%
	Yes, ≥70	20,149	14.9 (5.5)	4.1%	4.6%	22.3%	26.5%
Age 70	No	204,861	21.4 (7.0)	17.4%	7.9%	16.1%	30.9%
	Yes	42,288	20.9 (6.7)	8.3%	6.6%	19.6%	26.0%
Age 75	No	160,015	29.0 (7.6)	21.8%	9.4%	13.0%	30.6%
	Yes	34,070	28.2 (7.2)	9.0%	7.5%	16.7%	26.4%

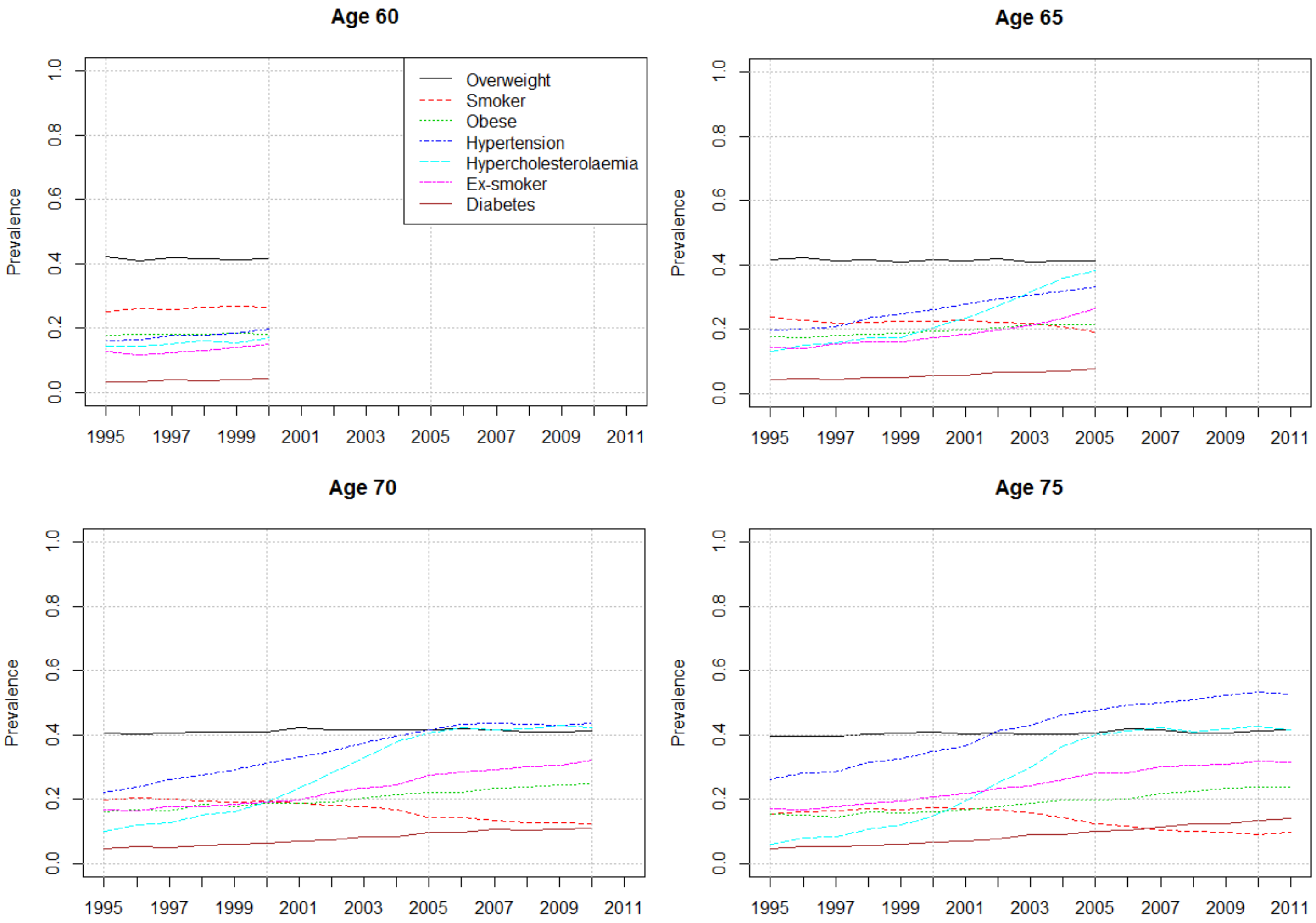


Figure C.1: Prevalence of comorbidites and lifestyle factors by cohort's age in patients without cardiovascular disease

The age cohorts differed in recruitment period. The prevalence prior to 1995 is not presented due to the small numbers of medical records available.

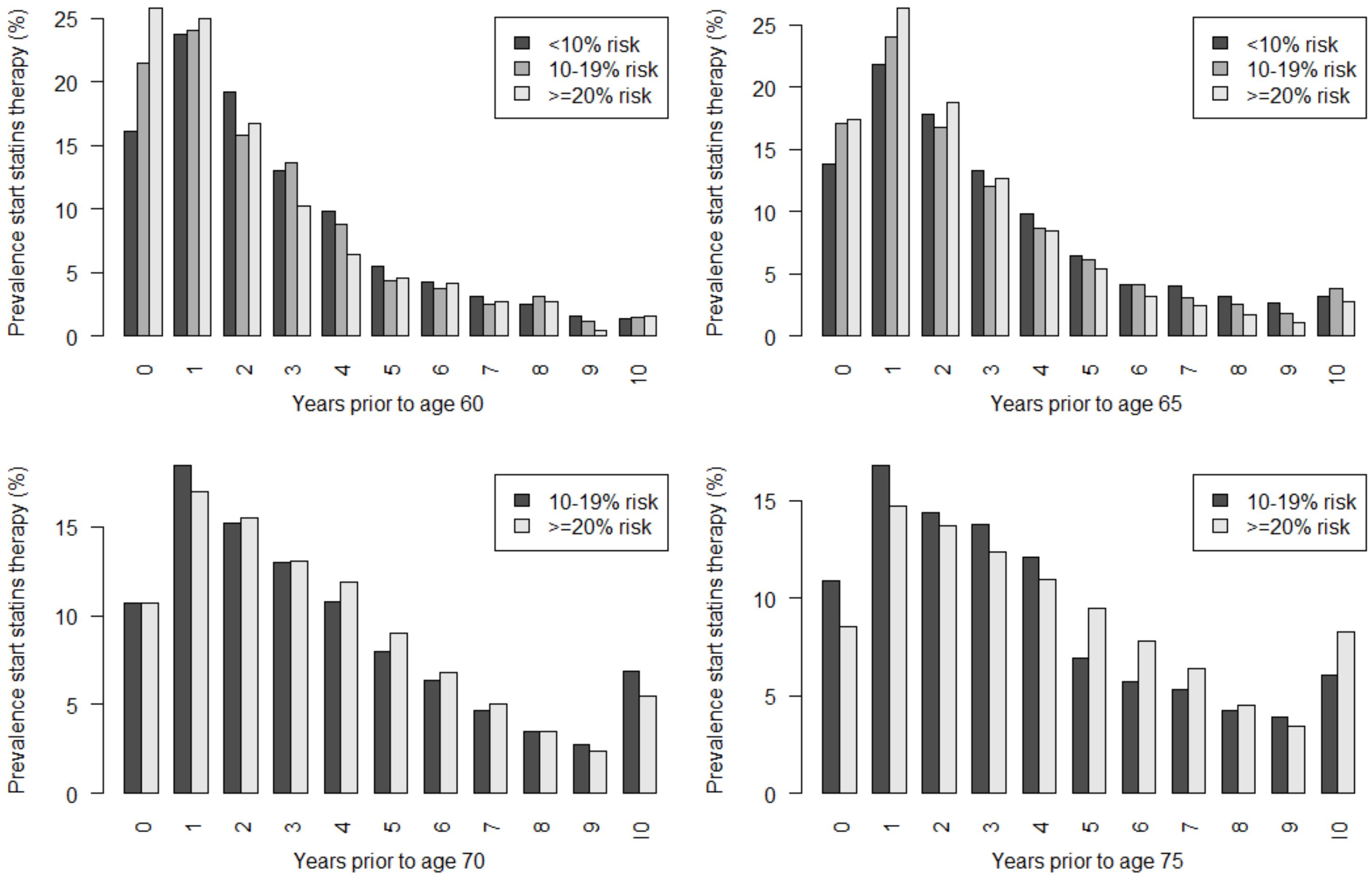


Figure C.2: Prevalence start of statin therapy given prescription prior to cohort's age by QRISK2 group

The number of patients in each cardiovascular risk group is the mean across ten imputed datasets.

Table C.7: Cases and controls staying in initial treatment arm of statin prescription during follow-up

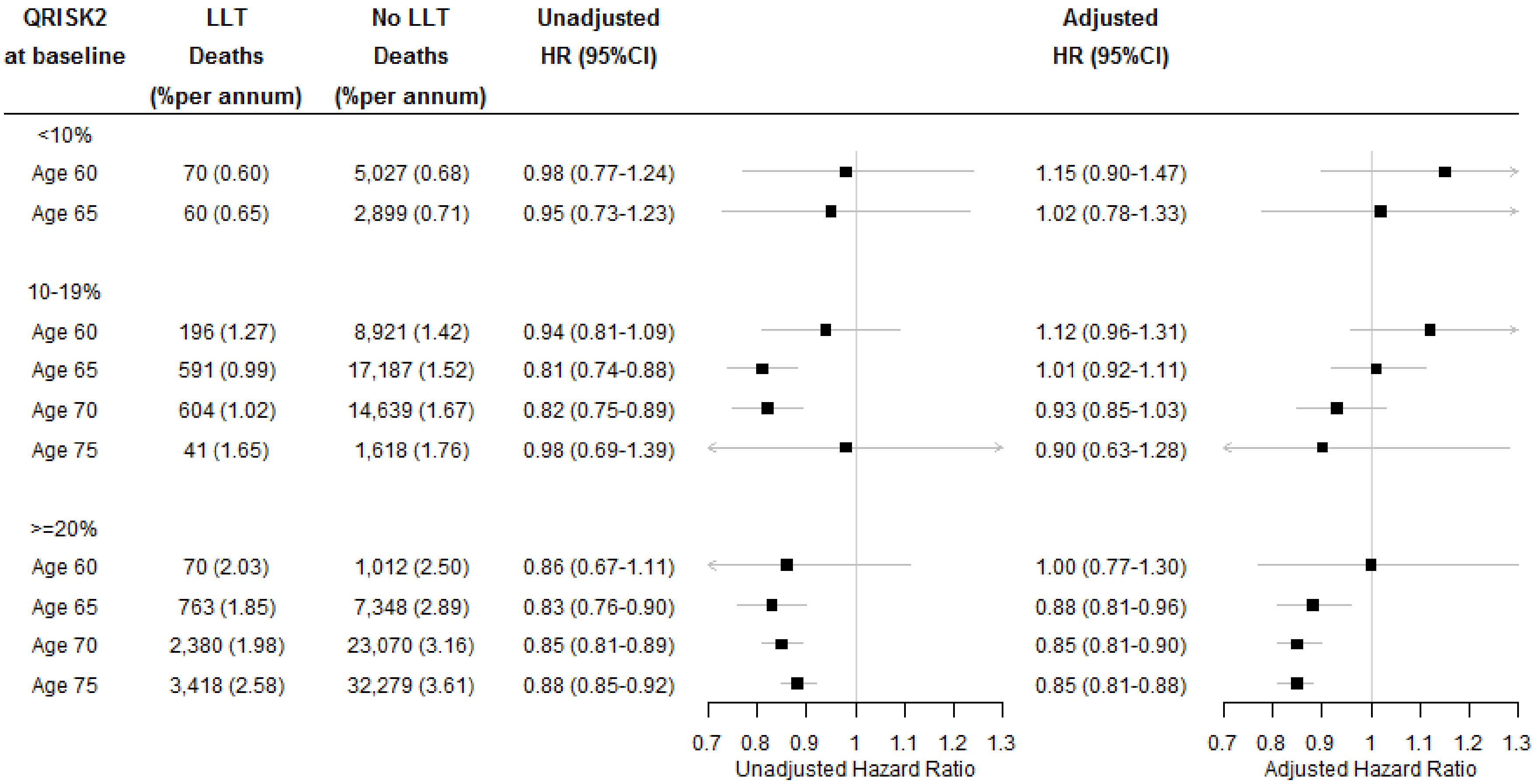
The age cohorts included patients with no history of cardiovascular disease. Adherence of treatment arm of statin prescription was ascertained in patients who were observed in multiple cohorts. Patients were not observed in an older age cohort when one of the following events happened: cardiovascular event, death, transfer out from general practice, or end of study. It was assumed that patients lost to follow-up stayed in the initial treatment arm.

	Cases	Controls
Age 60	1,469/1,664 (88%)	91,691/117,036 (78%)
Age 65	3,935/4,278 (92%)	82,399/98,097 (84%)
Age 70	4,560/4,673 (98%)	78,565/85,961 (92%)

Table C.8: Performance statistics of survival models based on complete medical records and irrespective of completeness in age cohorts without cardiovascular disease

Patients with incomplete medical records had a missing observation for systolic blood pressure, body mass index, or smoking status, and consequently a missing QRISK2 score.

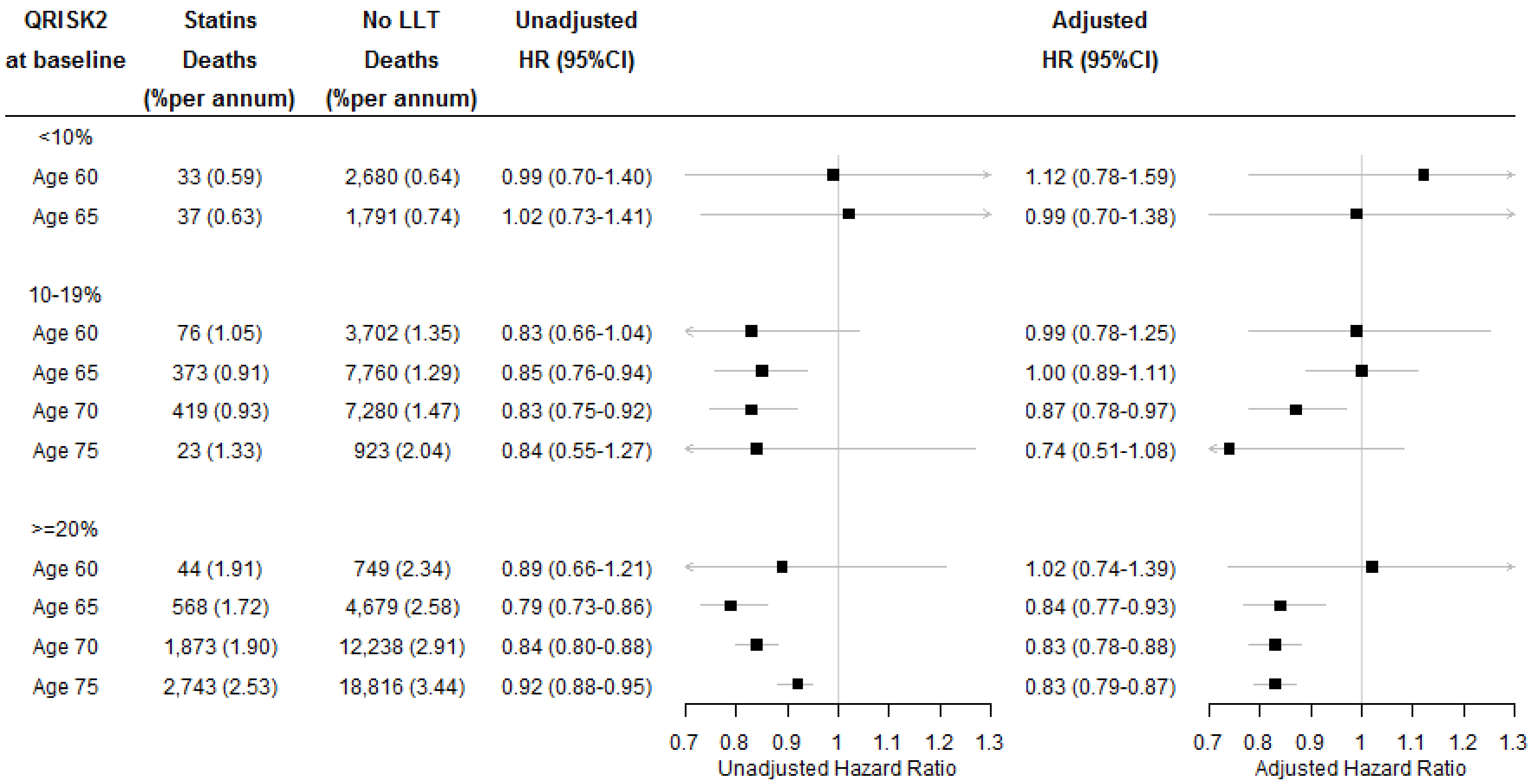
	QRISK2	Complete records			All records		
		R^2	C	Shrinkage	R^2	C	Shrinkage
Age 60	<10%	0.110	0.615	0.952	0.122	0.624	0.976
	10-19%	0.111	0.629	0.956	0.123	0.634	0.987
	≥20%	0.055	0.597	0.733	0.047	0.592	0.775
Age 65	<10%	0.050	0.578	0.872	0.127	0.633	0.992
	10-19%	0.118	0.629	0.950	0.127	0.633	0.975
	≥20%	0.090	0.614	0.956	0.091	0.619	0.975
Age 70	10-19%	0.104	0.625	0.978	0.106	0.626	0.990
	≥20%	0.120	0.635	0.992	0.107	0.629	0.993
Age 75	10-19%	0.132	0.636	0.969	0.103	0.634	0.984
	≥20%	0.138	0.647	0.995	0.134	0.644	0.997



QRISK2 at baseline	LLT Deaths (%per annum)	No LLT Deaths (%per annum)	Unadjusted HR (95%CI)	Adjusted HR (95%CI)
<10%				
Age 60	70 (0.60)	5,027 (0.68)	0.98 (0.77-1.24)	1.15 (0.90-1.47)
Age 65	60 (0.65)	2,899 (0.71)	0.95 (0.73-1.23)	1.02 (0.78-1.33)
10-19%				
Age 60	196 (1.27)	8,921 (1.42)	0.94 (0.81-1.09)	1.12 (0.96-1.31)
Age 65	591 (0.99)	17,187 (1.52)	0.81 (0.74-0.88)	1.01 (0.92-1.11)
Age 70	604 (1.02)	14,639 (1.67)	0.82 (0.75-0.89)	0.93 (0.85-1.03)
Age 75	41 (1.65)	1,618 (1.76)	0.98 (0.69-1.39)	0.90 (0.63-1.28)
>=20%				
Age 60	70 (2.03)	1,012 (2.50)	0.86 (0.67-1.11)	1.00 (0.77-1.30)
Age 65	763 (1.85)	7,348 (2.89)	0.83 (0.76-0.90)	0.88 (0.81-0.96)
Age 70	2,380 (1.98)	23,070 (3.16)	0.85 (0.81-0.89)	0.85 (0.81-0.90)
Age 75	3,418 (2.58)	32,279 (3.61)	0.88 (0.85-0.92)	0.85 (0.81-0.88)

Figure C.3: Unadjusted and adjusted hazards of all-cause mortality associated with lipid-lowering therapy prescription

The age cohorts included patients with no history of cardiovascular disease. The hazard ratios (95% confidence interval) were adjusted for sex, year of birth, socioeconomic status, diabetes, hypercholesterolaemia, blood pressure regulating drugs, body mass index, smoking status, and general practice. QRISK2=10-year risk of first cardiovascular event. LLT=lipid-lowering therapy.



QRISK2 at baseline	Statins Deaths (%per annum)	No LLT Deaths (%per annum)	Unadjusted HR (95%CI)	Adjusted HR (95%CI)
<10%				
Age 60	33 (0.59)	2,680 (0.64)	0.99 (0.70-1.40)	1.12 (0.78-1.59)
Age 65	37 (0.63)	1,791 (0.74)	1.02 (0.73-1.41)	0.99 (0.70-1.38)
10-19%				
Age 60	76 (1.05)	3,702 (1.35)	0.83 (0.66-1.04)	0.99 (0.78-1.25)
Age 65	373 (0.91)	7,760 (1.29)	0.85 (0.76-0.94)	1.00 (0.89-1.11)
Age 70	419 (0.93)	7,280 (1.47)	0.83 (0.75-0.92)	0.87 (0.78-0.97)
Age 75	23 (1.33)	923 (2.04)	0.84 (0.55-1.27)	0.74 (0.51-1.08)
>=20%				
Age 60	44 (1.91)	749 (2.34)	0.89 (0.66-1.21)	1.02 (0.74-1.39)
Age 65	568 (1.72)	4,679 (2.58)	0.79 (0.73-0.86)	0.84 (0.77-0.93)
Age 70	1,873 (1.90)	12,238 (2.91)	0.84 (0.80-0.88)	0.83 (0.78-0.88)
Age 75	2,743 (2.53)	18,816 (3.44)	0.92 (0.88-0.95)	0.83 (0.79-0.87)

Figure C.4: Unadjusted and adjusted hazards of all-cause mortality associated with statin prescription in complete case analysis

The age cohorts included patients with no history of cardiovascular disease. Patients with a missing observation for systolic blood pressure, body mass index, or smoking status, and consequently a missing QRISK2 score, were excluded from this analysis. The hazard ratios (95% confidence interval) were adjusted for sex, year of birth, socioeconomic status, diabetes, hypercholesterolaemia, blood pressure regulating drugs, body mass index, smoking status, and general practice. QRISK2=10-year risk of first cardiovascular event. LLT=lipid-lowering therapy.